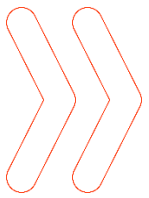


# AI in Language Assessment:

- **AI-assisted Item Generation**
- **Automated Scoring of Writing Performance**

**Yiannis Papargyris (PhD, CEA), Assessment Director**  
**Corina Dourda, Head of Assessment Design & Innovation**



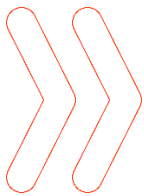


# The Transformative Role of Technology in Assessment

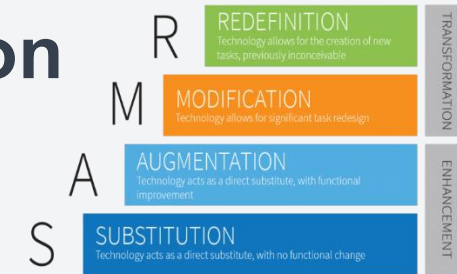
New technologies [will] permit a transformation in assessment by:

- › allowing us to create tests that are more firmly grounded in conceptualizations of what one needs to know and be able to do to succeed in a domain;
- › making performance assessment practical and routine through the use of computer-based simulation, automatic item generation, and automated essay scoring;
- › changing the ways in which we deliver, and the purposes for which we use, large-scale tests.

(Bennett, 1999a, p. 11)



# The SAMR Model for Technology Integration



Puentedura (2009) identified four levels through which we progress in our use of technology:

1. **Substitution**, in which the technology is a direct substitute and there is no functional change, through
2. **Augmentation**, in which the technology is still a direct substitute but now with some functional improvement, to
3. **Modification**, in which the technology allows or even catalyses significant redesign of the tasks, and finally,
4. **Redefinition**, where the technology enables us to create new tasks that were previously inconceivable.

# AI-assisted Item Generation





# Overview of the Project

1



**Purpose:** Explore AI's potential to enhance our item development process while keeping quality standards high.

2



**Who's involved:**

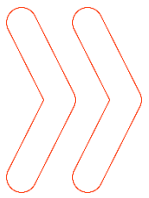
- Assessment Development: → Provide training data, design prompts, coordinate trials.
- Software Development: → Build & refine the system (database, prompting, interfaces).
- Expert consultants: → Review AI-generated items and evaluate quality during trials.

3

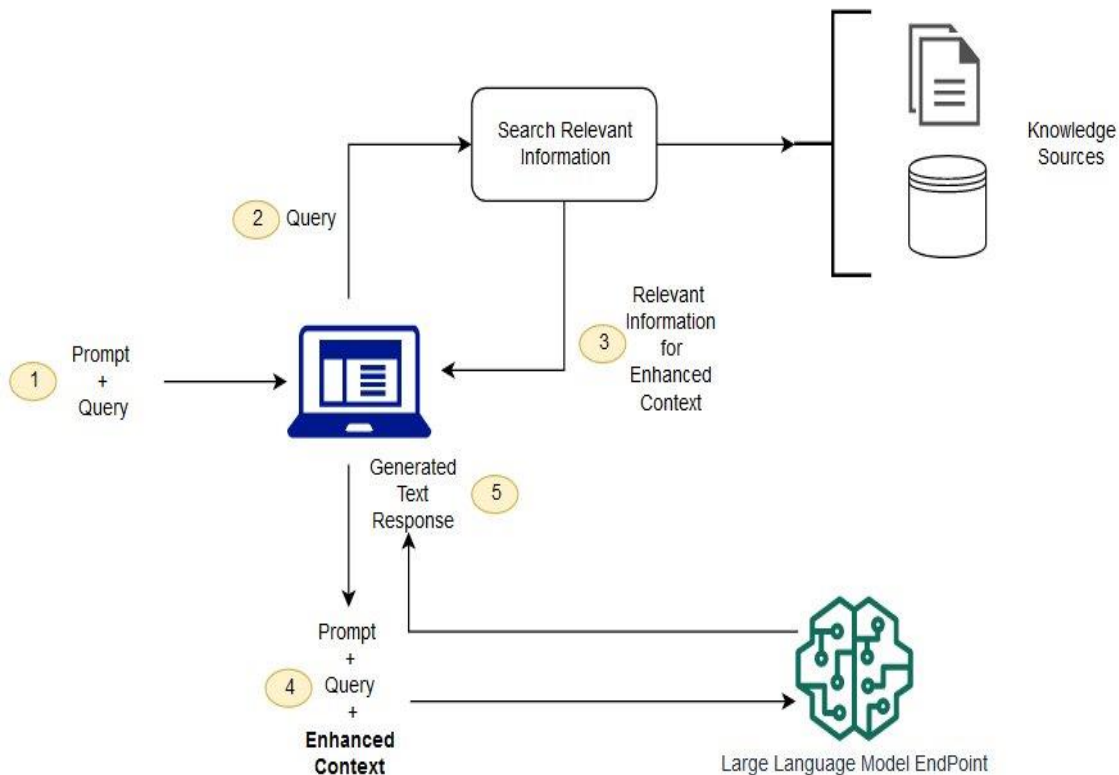


**Trialling:**

- **Focus** → **LANGUAGECERT Academic Test:** a high-stakes, four-skill, multi-level test (B1-C2) for academic study.
- **Progress** → **Trials** on Reading (word substitution), Writing (essay prompts), Speaking (discrete questions, role-play scenarios, read aloud items).
- **Goal** → **Evaluate** AI output quality and **refine** the system (interfaces, workflows, prompts, training data).



# How the System Works



## › What is Retrieval-Augmented Generation (RAG)?

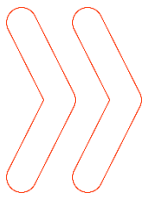
- Enhances the output of Large Language Models (LLMs).
- Retrieves relevant information first → then generates more accurate, targeted responses.

## › Key elements for quality AI output:

- 1 **A well-structured prompt** → Guides the AI on what to create, how to structure it, and what specific requirements to follow.
- 2 **A strong knowledge base** → Provides reliable reference materials to ensure accuracy and quality in responses.

The conceptual flow of using RAG with LLMs.

Retrieved from: <https://aws.amazon.com/what-is/retrieval-augmented-generation/>



# Trialling Speaking Tasks

## › Why Speaking tasks?

- More complex than sentence-level tasks but don't require large AI-generated output.
- Opportunity to test AI's ability to adjust output across proficiency levels (B1-C2).
- **Key challenge:** ensuring tasks are accessible for lower-level test takers (B1) while still challenging for stronger test takers (C1+).

## › Focus Task: **Role-Play**

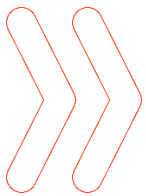
- Real-world interactions in a socio-academic context.
  - Group A: Respond to a prompt.
  - Group B: Initiate a conversation.
- Assesses:
  - Fluency & interactional competence
  - Appropriate language use
  - Listening & interpretation

### Group A

- We're friends studying together in the library. Other students near us are talking loudly. I start.  
*It's so loud in here, and I can't concentrate! What should we do?*
- I'm your tutor on a college course. I start.  
*I notice you haven't submitted your coursework yet. Is there anything I can help with?*
- We're classmates on a language course. I start.  
*My laptop's not working, and I really need to finish my essay. Any ideas where I can use a computer?*

### Group B

- I'm your course tutor at university. You want to arrange a time to discuss your essay grade. You start.
- We're classmates. You're going on a college trip to an art gallery and want me to come too. You start.
- I'm a student accommodation officer. You're having trouble finding suitable accommodation near campus. You start.



# Initial Approach to Prompting and Knowledge Base

## Prompt

- › Simple structure with essential elements:
  - **Test overview**
  - **Target audience & context**
  - **Task focus**
  - **Task structure**

### BUT:

- › Minimal guidance on testing focus (functional language).
- › Broad reference to B1-C2 levels, with no clear direction on level adjustment.  
*e.g. “test takers form B1-C2 should all be able to attempt the tasks.”*

## Knowledge Base

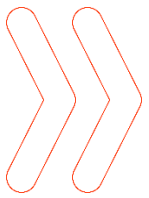
- › Included:
  - **Item Writer guidelines** with **task specifications** and **sample tasks**.

### BUT:

- › IW guidelines rely on human inference of CEFR expectations; not explicit instructions for the AI.

**Next step:** Run a pre-trial review to check how the system handled the task and identify what needed adjustment.





# Pre-Trial: Review Outcomes

## › Key issues identified:

- Language level too high (B2/C1+).
- Use of academic jargon and specialised terms.
- Wrong testing focus – tasks assumed prior/subject knowledge beyond test takers' experience.

## › Some early positives:

- Correct task structure (two sets of three scenarios).
- Some fresh topics and new takes on familiar ones.
- Decent attempts at context-setting and defining speaker roles.
- Variety of interactions (but not the right mix of formality).

## › Acceptance/rejection rate:

✓ 13% accepted (with extensive editing).

✗ 87% rejected.

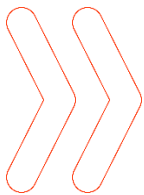
### AI-generated task

#### Group A

- You're speaking with your **academic** advisor. I start.  
Your **thesis** is quite **innovative**, but have you considered the **ethical implications** of your research?
- We're discussing your research project. I start.  
I've been **pondering** over the **methodology** for your research project.  
What's your **stance** on **utilizing qualitative** interviews instead of a survey?
- I'm your literature professor. I start.  
In your **perspective**, what is the central theme of the novel we've been **analyzing**, and how does it **reflect contemporary** society?

#### Group B

- We're **participants** in an **academic** conference. We're **contemplating** attending the **workshop** on AI applications. You start.
- I'm your language exchange partner. You mentioned you've been **struggling** with **idiomatic** expressions in English. You start.
- We're friends discussing the main message you're trying to **convey** in your recent article **draft** about student life. You start.



# Pre-Trial: Post-review Fine-tuning

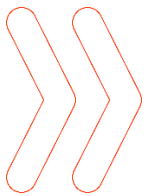
## Refining the Prompt

- › **Added more detail and examples** across core sections to set clearer expectations:
  - **Language style**  
(e.g. ‘conversational’, ‘feels natural when spoken’, ‘accessible’, etc.)
  - **Language level**  
(e.g. “Grammar and vocabulary must be at B1-B2 level”)
  - **Testing focus**  
(e.g. language functions & examples of expected output)
  - **Other constraints**  
(e.g. topics to avoid, length of items, right mix of interaction types)

## Expanding the Knowledge Base

- › **CEFR-linked (receptive) wordlist** to guide level-appropriate vocabulary.
- › **CEFR-linked grammar inventory** to support appropriate structures and complexity.
- › **Topics list** to ensure relevance and variety in tasks.
- › **‘Topics-to-avoid’ list** to steer clear of sensitive or controversial content.

**Next step:** Assess the impact of these changes based on trial results.



# Trial Outcomes

## › Key outcomes:

- Language level closer to B1-B2 (occasionally C1).
- Fewer instances of technical jargon (e.g. 'elasticity in economics').
- Imbalance in interaction types – too many informal (student-to-student) scenarios.

## › Recurring issues:

- Unclear or inappropriate speaker roles (e.g. *candidate acting as a lecturer*).
- Tasks prompting speculation on unfamiliar topics instead of functional language.
- References to UK/US university life not widely familiar. (e.g. *internship, resident advisor, dorm, job fair*)

## › Acceptance/rejection rate:

✓ 31% accepted (with some or extensive editing)

✗ 69% rejected.

### AI-generated task

#### Group A

- You're a university lecturer. I start.  
*Could you give me some **feedback** on my presentation skills?*
- We're in a university study group. I start.  
*I'm **struggling** with the **concept** of elasticity in economics. Can you explain it to me?*
- We're both interested in the same internship. I start.  
*How are you preparing for the internship interview?*

#### Group B

- We're both in the photography club. You want to **propose** a photo exhibition theme. You start.
- I'm your resident advisor, and you have a suggestion for the dorm. You start.
- We're at a job fair on campus. You want to approach a company's booth together. You start.



# Key Takeaways & Next Steps

## What We've Learned

- **AI is a tool** for generating ideas and enhancing efficiency, but **human expertise remains essential** to guide, refine and validate its output.



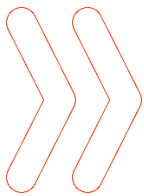
## Where We Go from Here

- **Continued focus:** Refine prompt design, expand knowledge sources, and explore new task types (images, audio)
- **Transition to Agentic AI:** From single-model GenAI to an ecosystem of coordinated AI agents (e.g. *AI Item Writer*, *AI Vetter*, *AI Image Generator*)
  - Gains in efficiency and consistency — but humans remain central for oversight and judgment.




# **Validating Auto-Marking for LCA & LCG Writing Tasks**








# Automated Scoring (Writing): Achievements & Next Steps

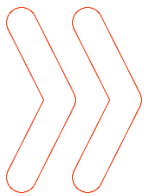
 **Purpose:** Reduce long-term operational costs **by automating one stage of the human marking process**, improve consistency in scoring, accelerate turnaround times for results.

## **Progress so far:**

-  Model **training** and **validation complete**.
-  Reviewed and approved by Advisory Council.
-  The auto-marker **matches or exceeds** human inter-marker agreement (M1-M2-CE) across all criteria.

## **2026 Goals:**

- › **Implementation:** Integrate into live systems for **LCA & LCG**.
- › **Human oversight:** → Human review remains in place for quality monitoring.  
→ Escalation protocol retained for scripts with high discrepancies (via Chief Examiners).



# Modelling Framework

The framework integrates linguistic insights with advanced deep learning techniques to form a hybrid scoring engine. Traditional linguistic features - such as readability and syntactic complexity - are combined with contextual representations generated by BERT to capture semantic and stylistic nuances. These elements are fused into a unified model, which is rigorously evaluated for accuracy and reliability prior to deployment in a production environment.

## Data Ingestion & Cleaning

Essays are imported, noisy ones are removed, and datasets are partitioned into training and evaluation sets.

## Readability, Lexical & Syntactic Metrics

Readability indices (e.g., Flesch Reading Ease, SMOG), lexical features (average sentence length, vocabulary complexity) are calculated alongside POS n-gram (uni- to tri-gram) features to represent syntactic patterns.

## Contextual Embedding Extraction via BERT

A pre-trained transformer (BERT) is fine-tuned on scored essays to produce dense embeddings that capture semantic and stylistic nuances.

## Feature Fusion

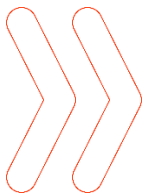
Bert embeddings and handcrafted linguistic features are concatenated into a single input entity.

## Model Training & Validation

A regression model (XGBoost) is trained on the fused features to predict scores across all criteria and performance evaluated using MAE, RMSE and QWK.

## Deployment & Monitoring

The finalized model and feature pipeline are packaged for real-time scoring.



# Model Training and Validation

A regression model (XGBoost) is trained on the fused features to predict scores across all criteria and performance evaluated using MAE, RMSE and QWK.

The initial development of the tool (data till Aug 2024) used the following number of scripts for training and testing:

|                 |           |             |
|-----------------|-----------|-------------|
| LC Academic_old | Test: 460 | Train: 1838 |
| LC General_old  | Test: 563 | Train: 2252 |

The model was then re-trained and re-validated with more data (Sept-Dec 2024).

|                 |           |             |
|-----------------|-----------|-------------|
| LC Academic_new | Test: 558 | Train: 2232 |
| LC General_new  | Test: 676 | Train: 2702 |

For a better understanding of the auto-marker's performance, the QWK values for each exam (LCA & LCG) and each task (Part 1 & Part 2) will be presented, together with Spearman rho correlations.

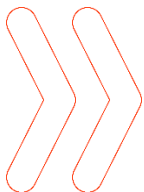
**MAE:** Mean Absolute Error: how far off the predictions are on average.

**RMSE:** Root Mean Squared Error: like MAE but penalises large errors more.

**QWK:** Quadratic Weighted Kappa: measures how well the predictions match human scores.



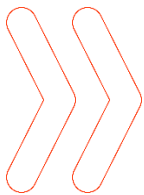




# Key Concepts and Definitions

|         |  |
|---------|--|
| M1 / M2 | One of the two independent markers assigned to the first marking of a Writing task.  |
| Model   | The auto-scoring tool used to generate individual criterion scores.<br>The 'Final mark – Model' correlation is based on the sum of the four criterion scores produced by the model, not on a separately predicted total score. |

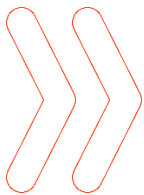
| CORRELATION |               |             |                 |                             |                                |                          | Quadratic Weighted Kappa / QWK |                 |                             |                                |                          |
|-------------|---------------|-------------|-----------------|-----------------------------|--------------------------------|--------------------------|--------------------------------|-----------------|-----------------------------|--------------------------------|--------------------------|
| Task        | Pair          | Total Marks | Task Fulfilment | Accuracy & Range of Grammar | Accuracy & Range of Vocabulary | Organisation & Coherence | Total Marks                    | Task Fulfilment | Accuracy & Range of Grammar | Accuracy & Range of Vocabulary | Organisation & Coherence |
|             | Final - Model |             |                 |                             |                                |                          |                                |                 |                             |                                |                          |
|             | Model - M1    |             |                 |                             |                                |                          |                                |                 |                             |                                |                          |
|             | Model - M2    |             |                 |                             |                                |                          |                                |                 |                             |                                |                          |
|             | M1 - M2       |             |                 |                             |                                |                          |                                |                 |                             |                                |                          |



# Cross tasks observations

- › Model consistently strong in Grammar, Vocabulary, and Organisation (0.79-0.85).
- › Task Fulfilment is the weakest performing criterion across all tests (0.72-0.77).
- › Model performance often exceeds inter-rater agreement (M1-M2).
- › Correlations and QWK align closely, confirming reliability.

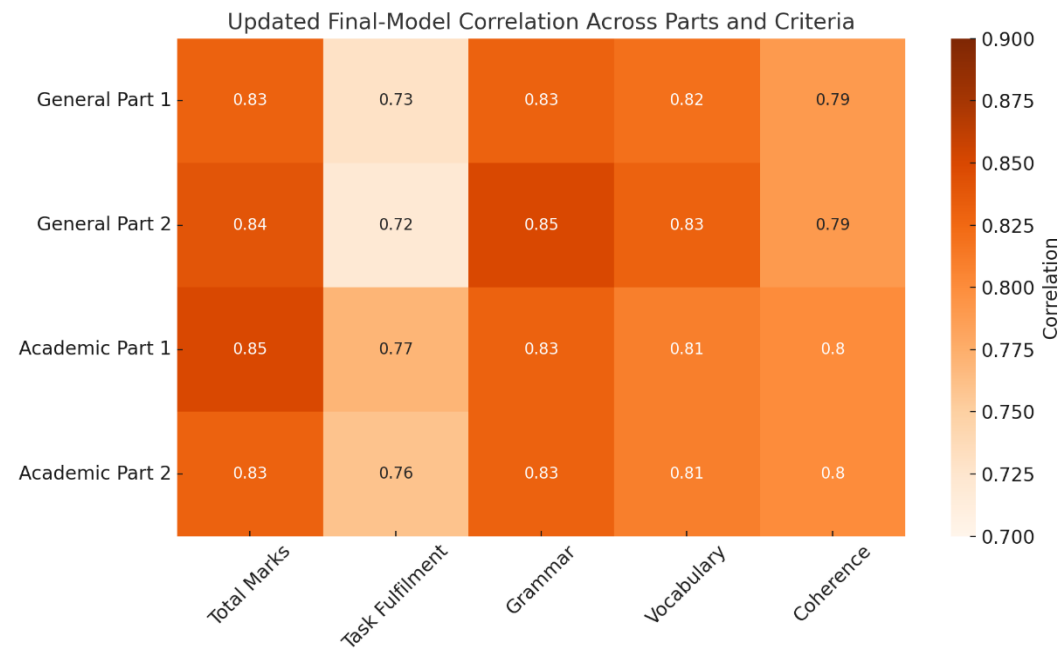
|                 | CORRELATION   |             |                 |                             |                                |                          | Quadratic Weighted Kappa / QWK |                 |                             |                                |                          |
|-----------------|---------------|-------------|-----------------|-----------------------------|--------------------------------|--------------------------|--------------------------------|-----------------|-----------------------------|--------------------------------|--------------------------|
|                 | Pair          | Total Marks | Task Fulfilment | Accuracy & Range of Grammar | Accuracy & Range of Vocabulary | Organisation & Coherence | Total Marks                    | Task Fulfilment | Accuracy & Range of Grammar | Accuracy & Range of Vocabulary | Organisation & Coherence |
| General Part 1  | Final-Model   | 0.83        | 0.73            | 0.83                        | 0.82                           | 0.79                     | 0.82                           | 0.69            | 0.79                        | 0.79                           | 0.75                     |
| General Part 2  | Final-Model   | 0.84        | 0.72            | 0.85                        | 0.83                           | 0.79                     | 0.83                           | 0.67            | 0.82                        | 0.82                           | 0.76                     |
| Academic Part 1 | Final-Model   | 0.85        | 0.77            | 0.83                        | 0.81                           | 0.8                      | 0.84                           | 0.75            | 0.8                         | 0.79                           | 0.75                     |
| Academic Part 2 | Final - Model | 0.83        | 0.76            | 0.83                        | 0.81                           | 0.8                      | 0.83                           | 0.74            | 0.81                        | 0.79                           | 0.77                     |



# Recommendations

For a carefully phased integration of the auto-marking model to supplement one of the two first human markers:

- › Retain escalation mechanism to chief examiner for high discrepancies.
- › Other meaningful steps (partial implementation in one task, monitoring of number of cases sent to CEs, other).
- › What are the next steps?
  - Systems implementation
  - Continuous monitoring



A large, stylized red arrow pointing to the right, composed of a solid red arrow and a thin red outline arrow, positioned on the left side of the slide.

# Thank you!