



AI in High-Stakes Language Assessment: An Evidence-Led Approach

Corina Dourda, Head of Assessment Design & Innovation

Leda Lampropoulou, Head of Research

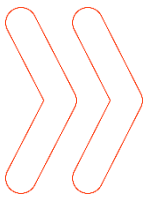
LANGUAGECERT Higher Education Summit
Athens, 16 October 2026



by PeopleCert

AI-assisted item generation





AI in language assessment

“ We’re trying to create AI that can do things like humans, but also do things that humans can’t do.

Demis Hassabis

“ AI’s success must be measured in economic impact, not benchmarks.

Satya Nadella

In language assessment, AI is already being used for:

› **Test development**

Item and asset generation

› **Candidate response scoring**

Automated scoring of Writing and Speaking

› **Test delivery**

Automated delivery of Speaking tests

› **Test security**

Detecting suspicious behaviour

The key question: What can AI contribute to language assessment, and what value does it add?



Overview of the project



Purpose

Explore how AI can support item development while maintaining assessment quality and expert oversight.



Who's involved

Assessment Development

Specifications, prompt design, trial coordination and quality assurance.

Software Development

System development, integration in the item banking system.

Expert consultants

Item review, editing, vetting and feedback.



Trialling

Focus

LANGUAGECER Academic
High-stakes, multilevel English proficiency test (B1-C2) for academic study.

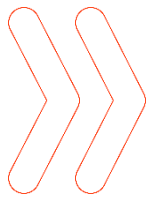
Task types

Reading: word substitution

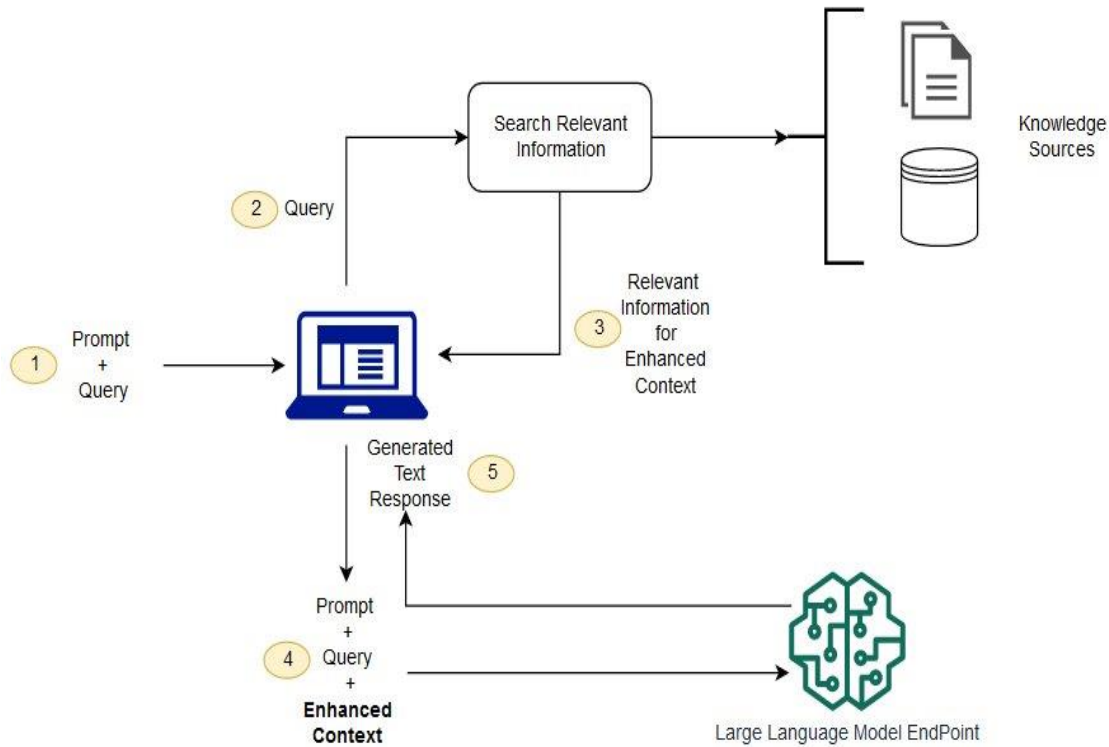
Writing: essay prompts

Speaking: discrete questions, role-play scenarios and read aloud items

Goal Evaluate AI-generated content and iteratively refine prompts, resources, workflows and system functionality.



How the system works



The conceptual flow of using RAG with LLMs

What is Retrieval-Augmented Generation (RAG)?

Retrieves relevant information from curated resources to support more controlled, assessment-specific generation.

Two essential components

01 Structured prompt (blueprint)

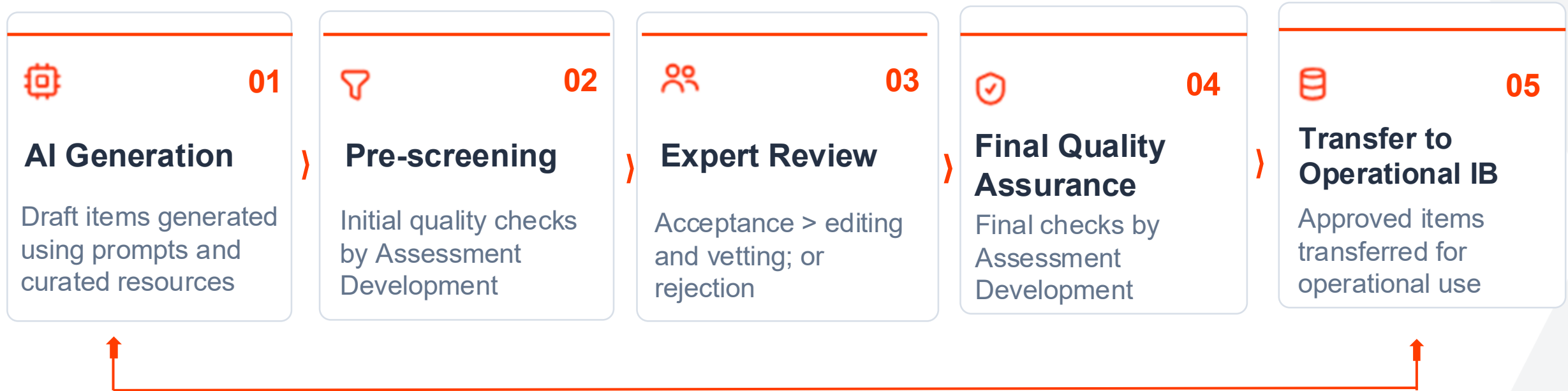
Defines task requirements, structure and generation constraints.

02 Curated knowledge base

Provides reliable assessment resources to support level-appropriate, construct-relevant generation.

The human-AI item development workflow

Integrated within LANGUAGECERT's item-banking environment (IB).

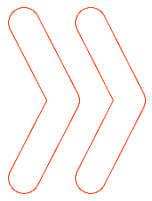


Feedback and iterative refinement

Review decisions and rejection reasons inform prompts, knowledge resources and system improvements.



Human oversight and established quality standards throughout the process



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 CEFR levels (B1-C2).
- **Key challenge:** accessible to lower-level candidates (B1) while allowing stronger candidates (C1+) to demonstrate their abilities.

Focus Task: Role-Play

- Six scenarios in socio-academic contexts.
 - Group A: Respond to an Interlocutor's opening statement.
 - Group B: Initiate a conversation.
- Assesses:
 - Fluency & interactional competence
 - Appropriate language use
 - Listening & interpretation

Illustrative role-play scenarios

Group A: Respond

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?”

Group B: Initiate

I'm your course tutor at university. You want to arrange a time to discuss your essay grade. 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:

- › Insufficient guidance on testing focus (functional language).
- › Broad reference to CEFR levels B1-C2, without explicit guidance on language level and complexity.

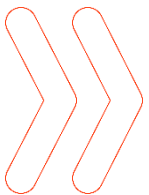
Knowledge Base

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

BUT:

- › IW guidelines rely on implicit professional knowledge, particularly regarding CEFR expectations; not explicit AI-facing instructions.

Next step: Pre-trial review to evaluate initial output and identify areas for refinement.



Pre-trial: review outcomes

› Key issues identified:

- Language frequently too advanced (C1+).
- Use of academic jargon and specialised terms.
- Weak alignment with testing focus – tasks assumed prior/subject knowledge rather than eliciting functional language.

› Some early positives:

- Correct task structure (two sets of three scenarios).
- Some fresh topics and new takes on familiar ones.
- Reasonable 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.

Examples of AI-generated scenarios

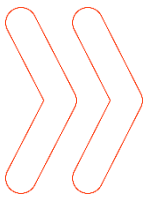
Group A: Respond

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?”

Group B: Initiate

We're participants in an academic conference.
We're contemplating attending the workshop on AI applications. You start.



Pre-trial: post-review ‘fine-tuning’

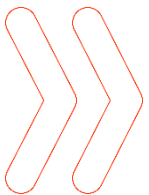
Refining the Prompt

- › **Added clearer instructions & examples covering:**
 - **Language style**
(e.g. ‘conversational’, ‘natural’, ‘accessible’, etc.)
 - **Language level**
(e.g. “Grammar and vocabulary must be at B1-B2 level”)
 - **Testing focus**
(e.g. language functions & examples of expected responses)
 - **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 unsuitable content.

Next step: Evaluate the quality of generated tasks following these refinements.



Trial outcomes

› Key outcomes:

- Language generally closer to B1-B2 (occasionally C1).
- Fewer instances of technical or specialised vocabulary.
- Imbalance in interaction types – too many informal (student-to-student) scenarios.

› Recurring issues:

- Unclear or inappropriate speaker roles.
- Tasks relying on unfamiliar subject knowledge rather than functional language.
- Culturally specific references to UK/US university life.

› Acceptance/rejection rate:

✓ 31% accepted (with some or extensive editing)

✗ 69% rejected.

Examples of AI-generated scenarios

Group A: Respond

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?

You're a university lecturer. I start.

“ Could you give me some feedback on my presentation skills?

Group B: Initiate

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 & what's next

What we've learned

- **AI is a valuable tool** for generating ideas and supporting item development, but **human expertise remains essential** to guide, refine and evaluate its output.
- Rejected items and reviewer feedback are valuable: they **reveal limitations** and **help make expert assessment criteria more explicit**.



Where we go from here

- **Future research:** examine how AI-assisted items perform with candidates, including comparisons with traditionally developed items.
- **Continued development:** refine prompts and knowledge resources and explore new task types involving audio and images.
- **Agentic AI:** explore coordinated AI agents supporting different stages of item development (e.g. *AI Item Writer*, *AI Vetter*), while retaining human oversight and quality assurance.



Validating Auto-Marking for LCA & LCG Writing Tasks





Marking the Writing tasks

Two writing tasks – Four assessment criteria – (at least) Two markers

Part	Task Description	Context
1	Visual and/or text input; describe and summarise/synthesise the information/data presented in a brief report.	Report findings, express opinion and give justification, express trends and certainty, probability and doubt, speculate, describe purpose/cause/result, etc.
2	Discursive piece of writing on a topical 'academic' subject.	Present and justify opinions, compare and contrast opposing points of view, evaluate ideas, analyse and/or describe pros and cons, look at cause-effect relationship, present solutions to issues.

Criteria	Description
Task Achievement	A measure of how far the candidate has achieved/addressed the task and whether or not the candidate has done what was asked.
Accuracy and Range of Grammar	A measure of the range, appropriacy and accuracy of grammar.
Accuracy and Range of Vocabulary	A measure of the range, accuracy and appropriacy of vocabulary as well as spelling accuracy.
Organisation (Coherence)	A measure of how coherently ideas are linked together in the text and how accurate the punctuation is.



Auto-Marking the Writing tasks

Two writing tasks – Four assessment criteria – (at least) Two markers

So why do it?

Auto-marking can be:

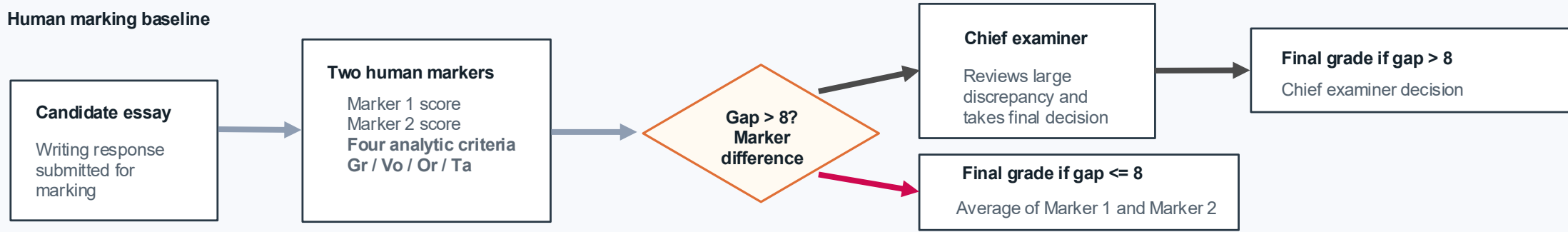
- › Efficient in terms of time and cost
- › Always available
- › Consistent

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

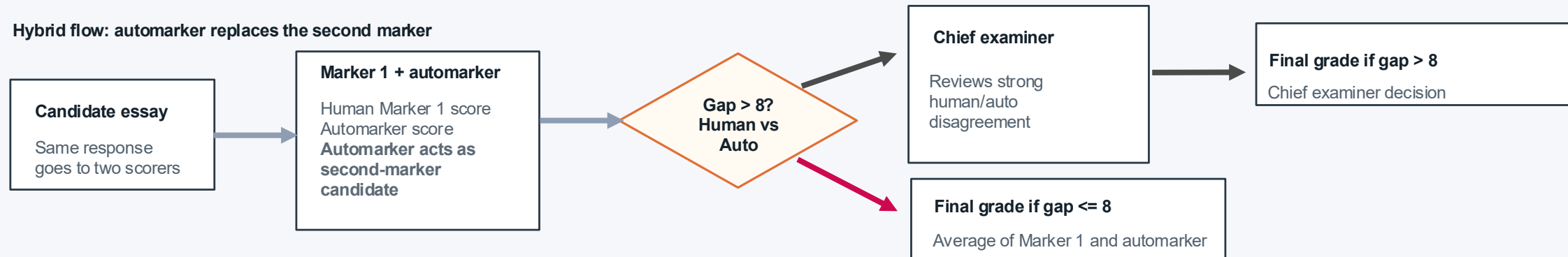
The auto-marker needs to **match or exceed** human inter-marker agreement across all criteria.

Auto-marking mental model

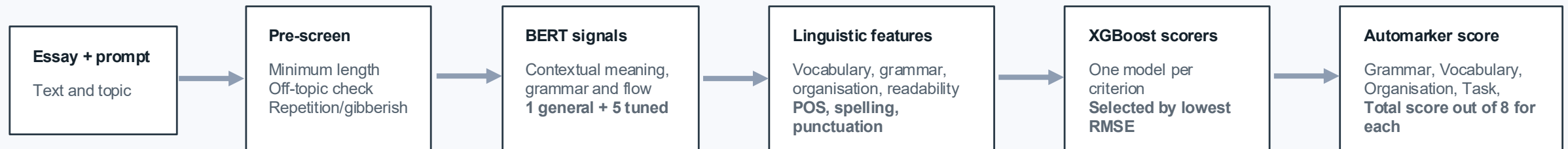
Human marking baseline



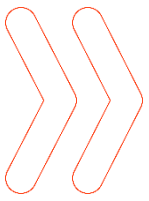
Hybrid flow: automarker replaces the second marker



Automated scoring engine, sequential view



Automarker score feeds the hybrid flow above



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

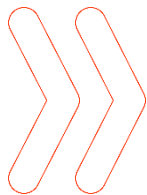
Bert embeddings and handcrafted linguistic features are concatenated into a single input entity.

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 different metrics, including Pearson (r) correlations.

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).

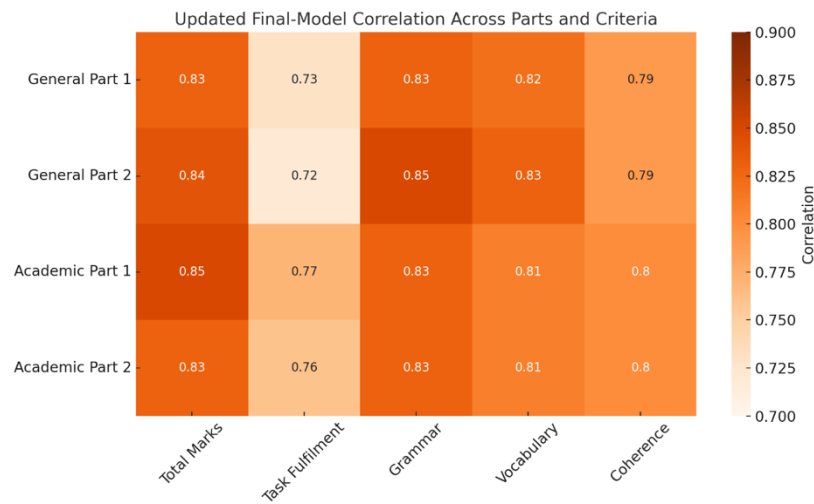
LC Academic_new	Test: 558	Train: 2232
LC General_new	Test: 676	Train: 2702

Training data need to be representative of the testing population and the general population:

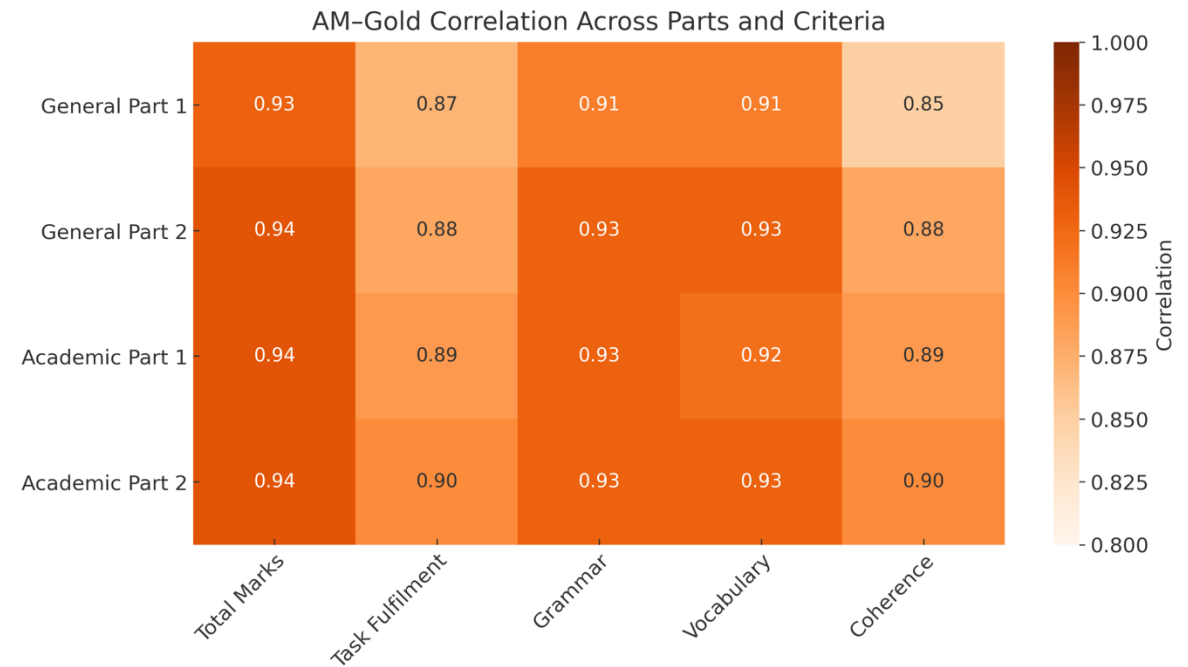
- › Gender
- › Nationalities
- › English proficiency levels

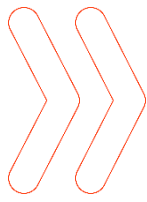
Staged validation

A carefully phased integration of the auto-marking model to replace one of the two first human markers.



✓ The auto-marker **matches or exceeds** human inter-marker agreement across all criteria.





Staged Implementation

CORRELATION

Academic Part 2	Pair	Total Marks	Task Fulfilment	Accuracy & Range of Grammar	Accuracy & Range of Vocabulary	Organisation & Coherence
	Final Mark - Automarker	0.94	0.90	0.93	0.93	0.90
	Automarker - Human Marker	0.81	0.75	0.79	0.78	0.74
	Final Mark - Human Marker	0.89	0.87	0.87	0.87	0.87

- › Model performance exceeds human agreement and correlates more highly with the final score.
- › Model consistently very strong cross all criteria - Consistently in the > 0.9 range across all parts and criteria.
- › AM–Human (M) Agreement at or above HH Baseline.
- › CE Referral Rates lower than Human Double-Marking: Flagged discrepancy rates: 5.4–8.7%, lower than HH baselines of 10%.



In progress and next steps

- › Human review remains in place for quality monitoring.
- › All under-length scripts go to a chief examiner.
- › Escalation protocol retained for scripts with high discrepancies (via Chief Examiners).
- › Ceiling effect: The AM very rarely awards the maximum criterion score of 8/8. This conservative behaviour at the top of the scale is being monitored.
- › A study investigating automarker performance against 5 experienced and highly consistent markers marking the same 600 scripts has just been completed and will be published soon.
- › Must ensure:
 - Accountability → human oversight remains
 - AI supports, but does not replace, expert judgment

A large, stylized red arrow pointing to the right, composed of a solid red arrow and a thin red outline arrow.

Thank you!