



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

Yiannis Papargyris, Language Assessment and Deputy Quality Director

Cathy Jones, Assessment Development Specialist

Leda Lampropoulou, Head of Research

LanguageCert Higher Education Summit
Athens, June 2026



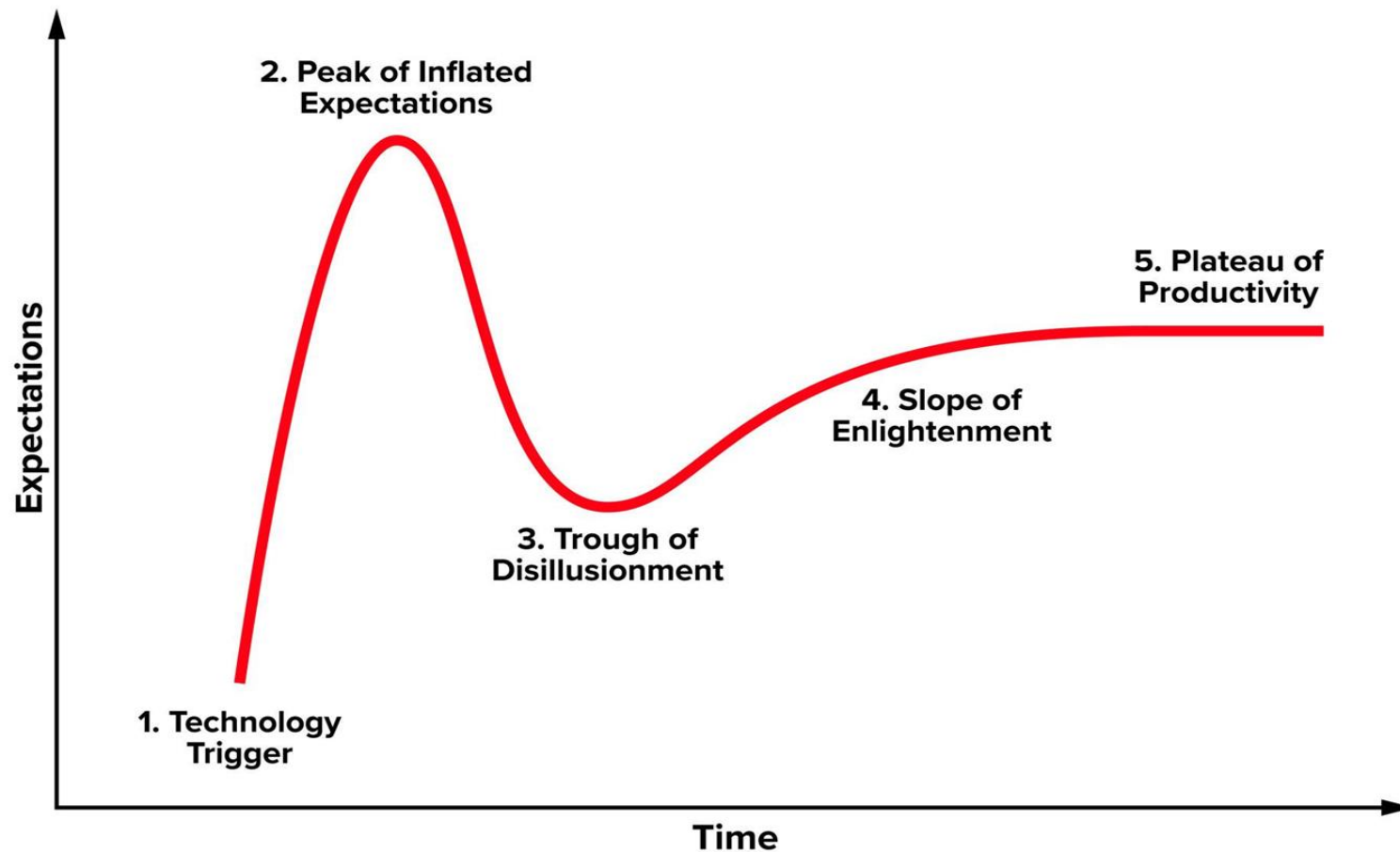
The Transformative Role of Technology in Assessment

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)

Gartner Hype Cycle



1: New Technology emerges; early proof-of-concept stories and media interest trigger significant publicity.

2: Early publicity produces a number of success stories.

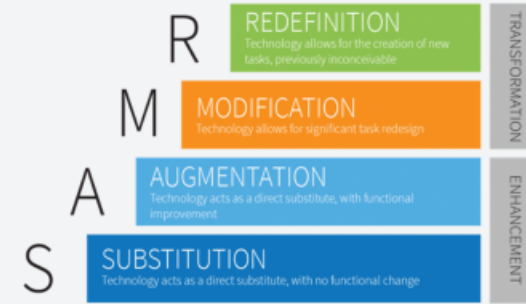
3: Interest wanes as experiments and implementations fail to deliver.

4: More instances of how the technology can benefit the enterprise start to crystallize and become more widely understood. Second- and third-generation products appear from technology providers.

5: Mainstream adoption starts to take off. Criteria for assessing viability are more clearly defined.



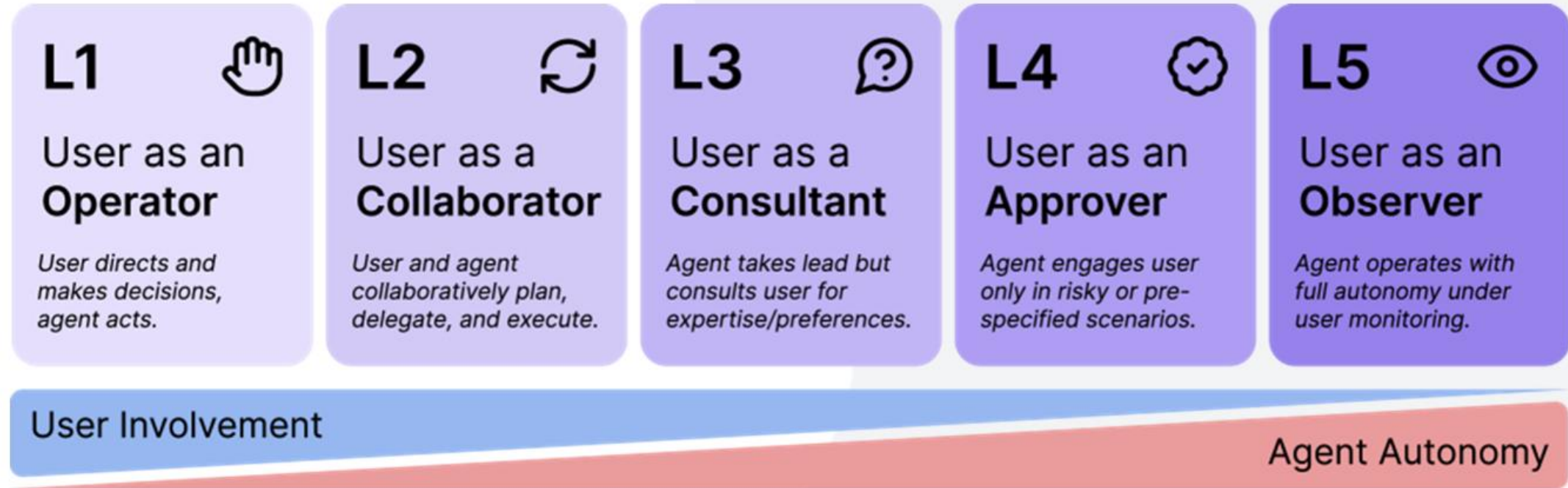
The SAMR Model for Technology Integration



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.

» Levels of Autonomy



» Levels of Autonomy for AI Agents. K. Feng, D. W. McDonald, A. Zhang, University of Washington

AI-assisted item generation



Item writing is difficult

- › Assessment items must measure specific constructs reliably
- › Language, level and task demands must align precisely
- › Items must be accessible across diverse candidate populations
- › Small wording changes can alter difficulty or validity
- › Only one correct answer: distractors and prompts must function consistently
- › Tasks must elicit authentic but controlled language use
- › Content must avoid cultural bias and unintended assumptions



Assessment items are carefully engineered measurement tools

Why explore AI-assisted generation?

Opportunities

- › Faster drafting
- › Greater scale and variety
- › Sustainable item development
- › Structured knowledge capture
- › Iterative development cycles

Impact

- › Wider variety of preparation and practice content
- › Increased item bank coverage
- › Reduced production burden
- › Focusing expertise where it matters most



But efficiency gains must never compromise validity, fairness and quality.

Project overview

Purpose

Explore how AI can support item development while maintaining quality, validity and expert human oversight.

Focus

LANGUAGECER Academic: high-stakes, four-skill, multi-level test (B1-C2)

Trials

Reading, Writing and Speaking tasks (discrete word-substitution questions, essay prompts, role-play scenarios)

Goal

Evaluate AI output quality and refine workflows, prompts and guidance

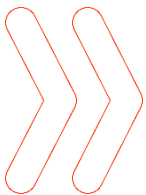
»» Operational progress to date

6 operational item types across **Reading, Writing** and **Speaking**.

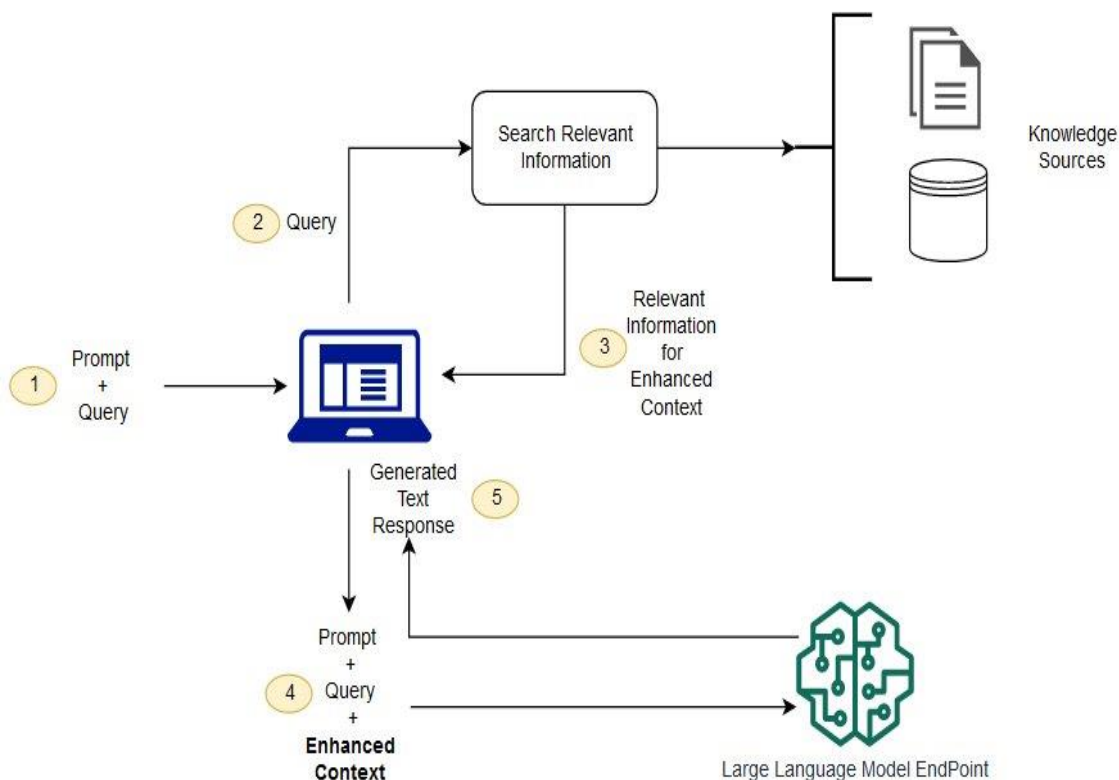
5 commissioned trials completed

2,000+ **AI-generated items** produced

- › Integrated multi-modal workflows (audio and image)
- › Transition from single-model prompting to multi-agent orchestration
- › Developed and integrated **LexiCheck** an AI-powered vocabulary checking assistant built on proprietary wordlist (11,000+ entries).



How the system works (single-model and agentic)



- › **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/>

A shift in the research question

✗ Can AI match human item writers?



✓ What becomes visible when experts must explain their judgement to AI?

AI-assisted development became:

- › a knowledge elicitation process
- › a reflection tool
- › a way of externalising tacit expertise

AI does not have:

- › tacit knowledge
- › shared context
- › professional intuition

AI exposes some assumptions experts can leave unstated.



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

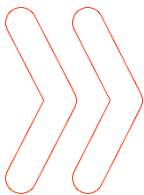
- Real-world interactions in a socio-academic context.
 - Group A: Respond to a prompt.
 - Group B: Initiate a conversation.
- Assesses:
 - Fluency & interactional competence
 - Appropriate language use
 - Listening & interpretation

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 from 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.

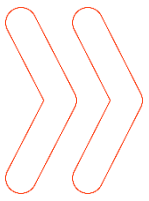
AI-generated task

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’

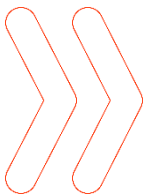
Refining the Prompt

- › **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 performance improved when tacit expertise became operational guidance.

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.

The diagnostic value of failure

AI failure was informative

Rejected items revealed:

- › hidden assumptions
- › unclear specifications
- › construct drift
- › cultural bias
- › under-specified guidance

Through iterative prompting and review, experts translated instinctive judgement into operational guidance.

- › Linguistic principles
- › Construct validity
- › Cultural assumptions
- › Quality criteria
- › CEFR targeting

» What AI could – and could not – do

What was amenable to AI assistance?

More amenable

- › Reading RP1a – lexico-grammatical substitution
- › Writing WP2 – polemical essay question
- › Structured drafting
- › Stable formats



Key finding
AI handled structure better than interaction.

Less amenable

- › Speaking tasks
- › Interactional authenticity
- › Role relationships
- › Discourse progression
- › Fairness-sensitive judgement

» Human-centred approach

AI supported:

- › drafting
- › variation
- › scaling
- › exploratory generation

**Human–AI collaboration was bidirectional;
AI exposed gaps and forced clarification.**

Humans retained :

- › validity judgement
- › fairness
- › construct interpretation
- › ethical judgement
- › final approval

» From automation to articulation

Some expertise could be operationalised:

- › constraints
- › thresholds
- › decision rules

Some resisted codification:

- › fairness
- › contextual sensitivity
- › ethical judgement

AI-assisted item development:

- › supports efficiency
- › improves transparency
- › strengthens documentation
- › surfaces tacit expertise

But the greatest value is reflective.

AI makes professional judgement:
visible

- › discussable
- › inspectable
- › transferable



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.





by PeopleCert

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?

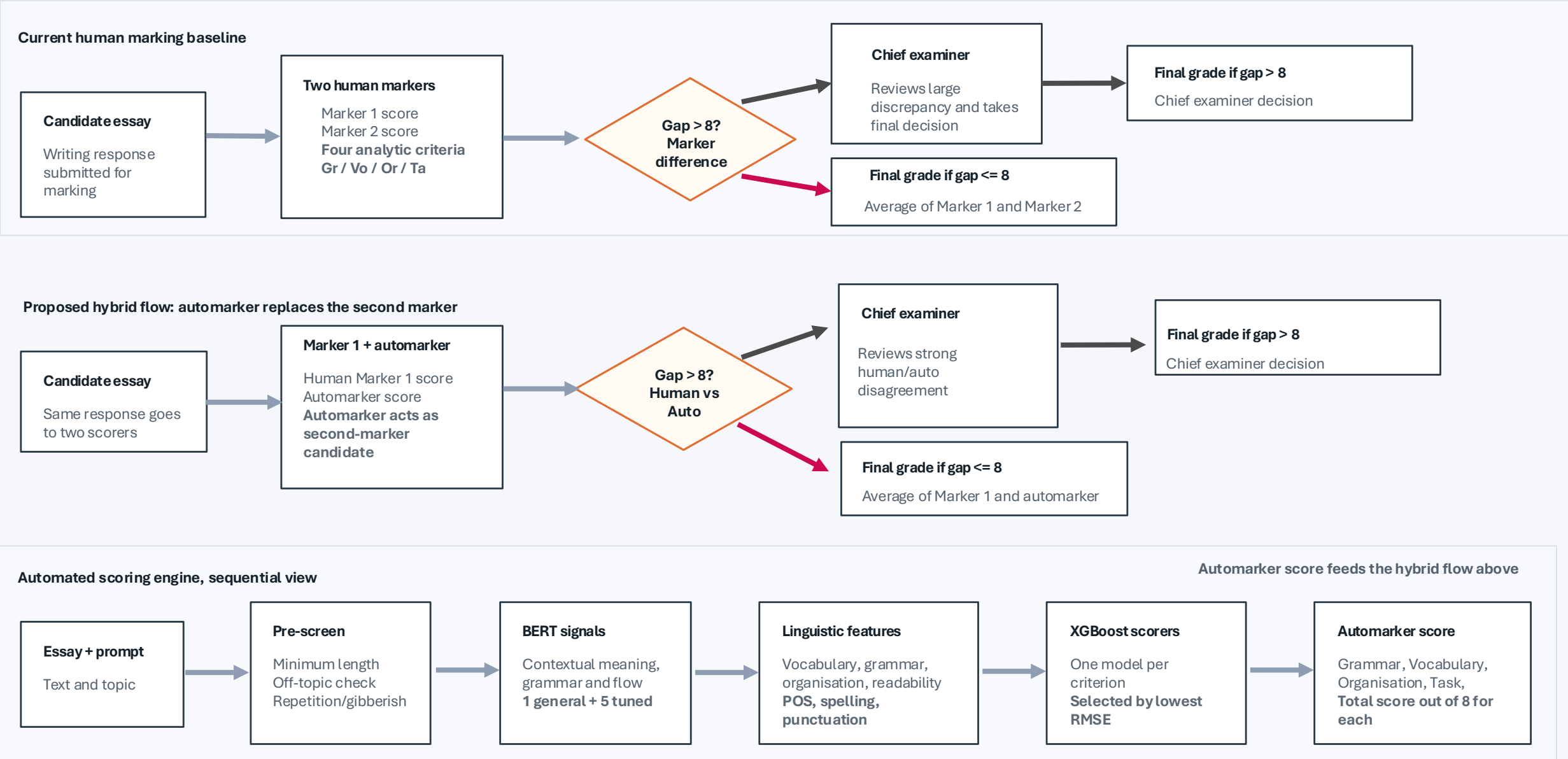
Auto-marking can be:

- › 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



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).

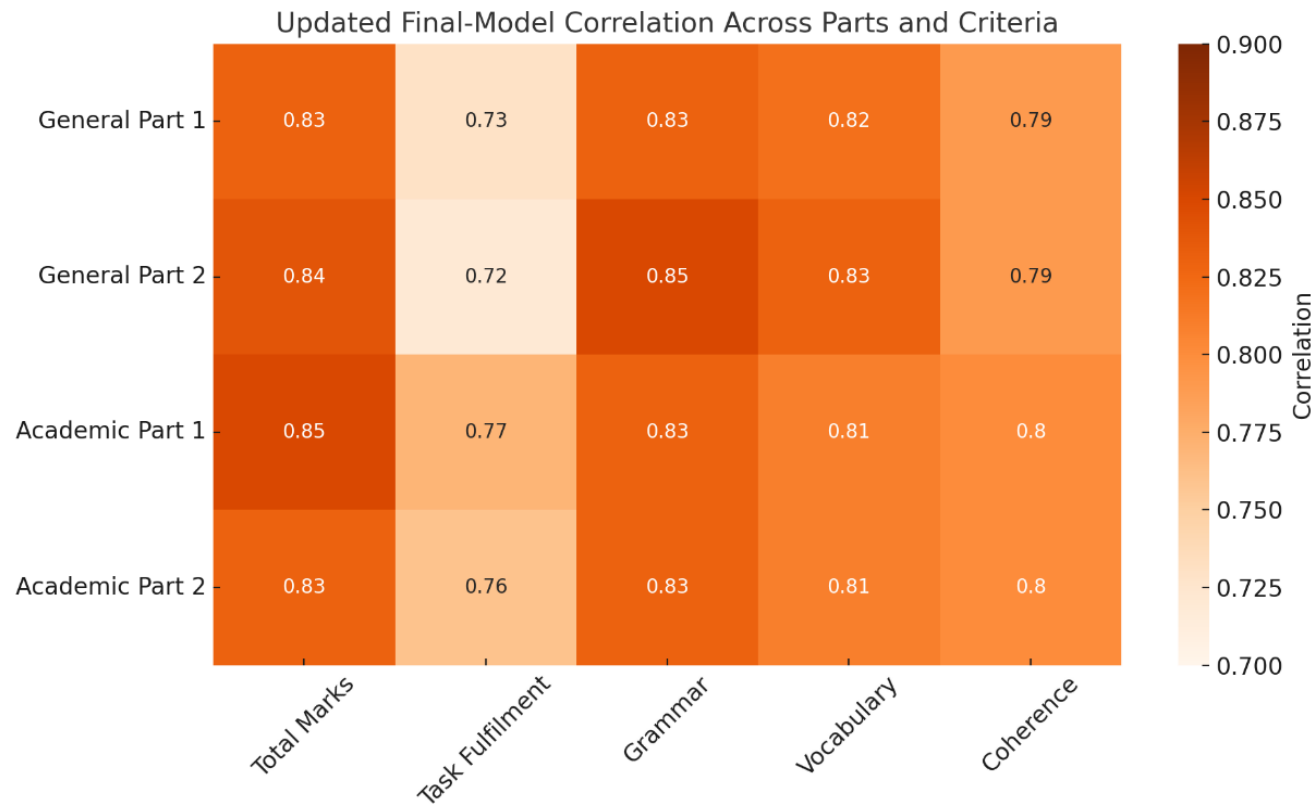
LC Academic_new	Test: 558	Train: 2232
LC General_new	Test: 676	Train: 2702

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

- › Gender
- › Nationalities
- › English proficiency levels

>> First-stage 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 1	Pair	Total Marks	Task Fulfilment	Accuracy & Range of Grammar	Accuracy & Range of Vocabulary	Organisation & Coherence
	Final Mark - Automarker	0.94	0.92	0.93	0.92	0.88
	Automarker – Human Marker	0.78	0.75	0.74	0.73	0.73
	Final Mark - Human Marker	0.87	0.86	0.84	0.84	0.86

- › 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 much lower or comparable to Human Double-Marking:
Scorable-only 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.
- › 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!