From Assessment to Human-AI Co-Creation in Language Learning

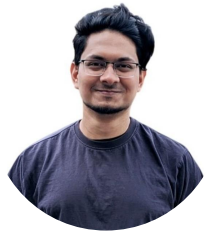
Adaptive, Inclusive, and Game-Based Design in the Generative AI Era

Zheng Yuan
University of Sheffield
UK

Okan Bulut
University of Alberta
Canada

Thierry Geoffre
University of Luxembourg
Luxembourg

Qiao Wang
Hosei University
Japan



Adnan Labib
King's College London
UK

June 28, 2026

AIED SEOUL 2026

Who are we?

Okan Bulut

- Professor of Measurement, Evaluation, and Data Science
- Faculty of Education, University of Alberta, Canada
- Research Interests:
  - Computational psychometrics
  - Digital assessments
  - Learning analytics
  - AI-powered personalized learning
- Email: bulut@ualberta.ca
- Website: www.okanbulut.com

Who are we?

Thierry Geoffre

- Associate Professor of Teaching and Learning French in Pre- and Primary School
- Department of Education and Social Work, University of Luxembourg
- Research Interests:
  - AI for Inclusive Education
  - Language Education
  - Adaptive Learning
  - Praxeological approach to language education
- Email: thierry.geoffre@uni.lu

Who are we?

Qiao (Judy) Wang

- Associate Professor
- Faculty of Economics, Hosei University
- Research Interests:
  - Computer Assisted Language Learning
  - Educational Technology
  - Second Language Acquisition
- Email: judy.wang@hosei.ac.jp

Who are we?

Adnan Labib

- PhD Candidate
- Department of Informatics, King’s College London
- Research Interests:
  - Educational NLP
- Email: adnan.1.labib@kcl.ac.uk

Who are we?

Zheng Yuan

- Associate Professor in Natural Language Processing
- School of Computer Science, The University of Sheffield
- Research Interests:
  - AI4Education
  - Educational NLP
  - EdTech
  - Computer Assisted Language Learning
  - Human-AI Collaboration
- Email: zheng.yuan1@sheffield.ac.uk

Agenda (Part I)

- 13:30 - 13:40: **1. Welcome & Introduction**
- 13:40 - 14:25: **2. Automated Assessment Pipeline**
- 14:25 - 14:30: Q&A + Short Break
- 14:30 - 15:15: **3. Human-AI Collaboration for Inclusive Education**
- 15:15 - 15:45: Q&A + ☕ Coffee Break

Agenda (Part II)

- 15:45 - 16:30: **4. Game-Based Language Learning**
- 16:30 - 16:40: Q&A + Short Break
- 16:40 - 17:10: **5. Demos**
- 17:10 - 17:30: **6. Challenges and Future Directions**

Information & Resources

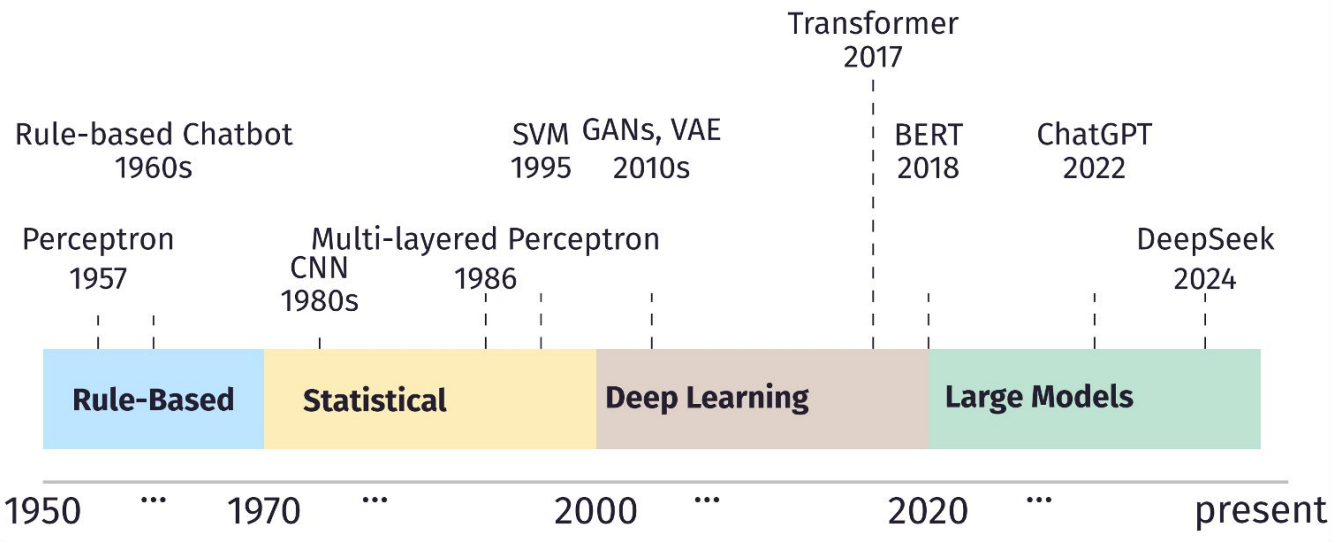
- Tutorial website:

https://aied2026-hacc-tutorial.github.io/

- Slides, materials, and further reading will be available online
- Please reach out with questions or feedback

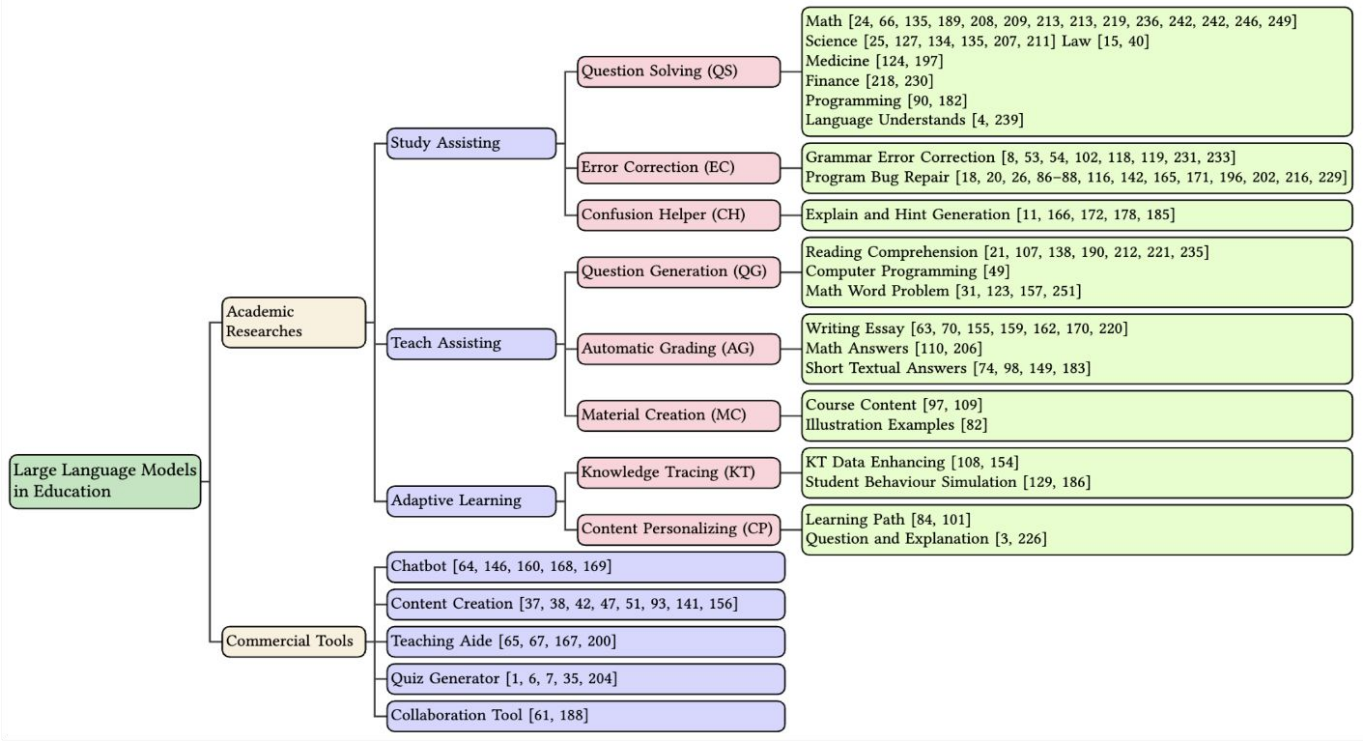
1. Introduction

Brief History



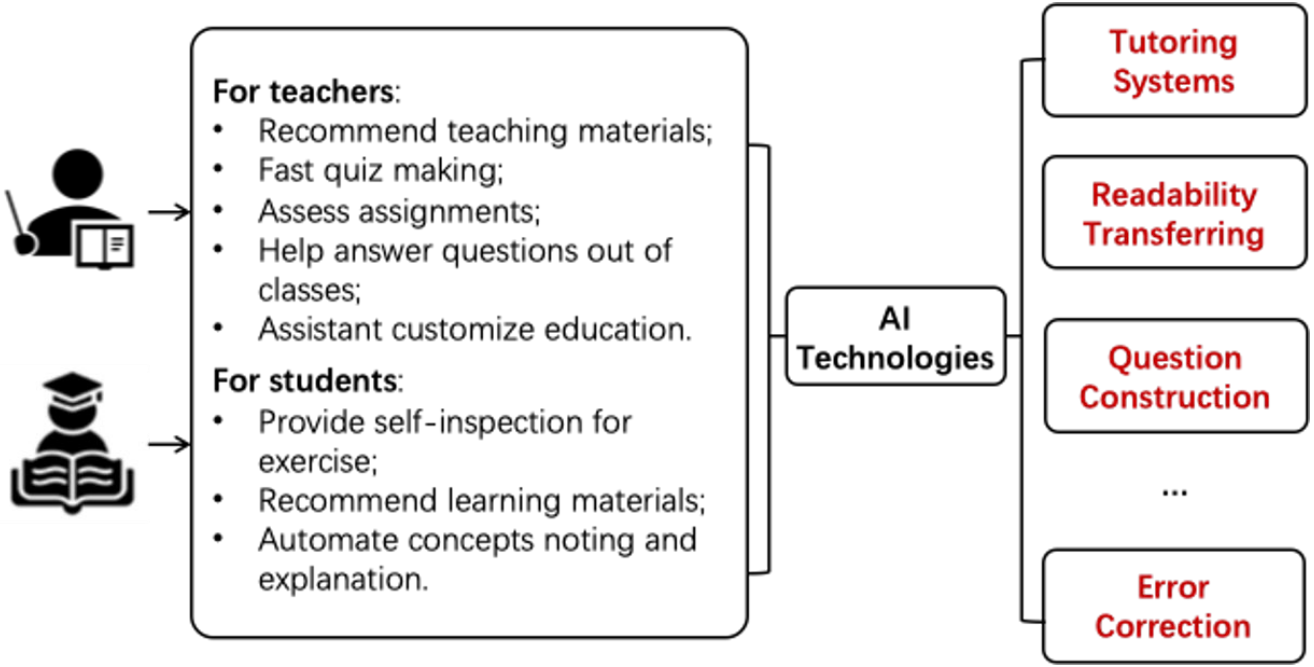
Yuan et al. (2025). NLP and Generative AI for Language Learning and Assessment

LLMs in Education Taxonomy



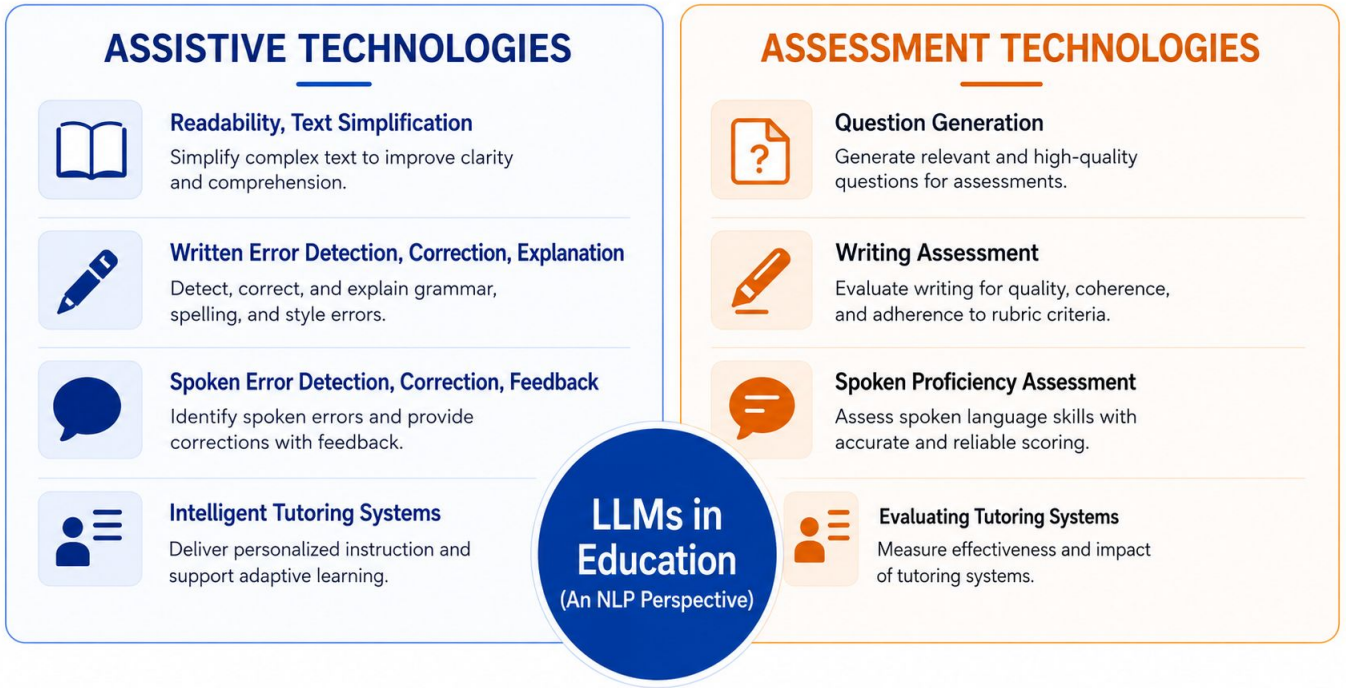
Wang et al. (2024). [Large Language Models for Education: A Survey and Outlook]

Teacher-side vs. Student-side

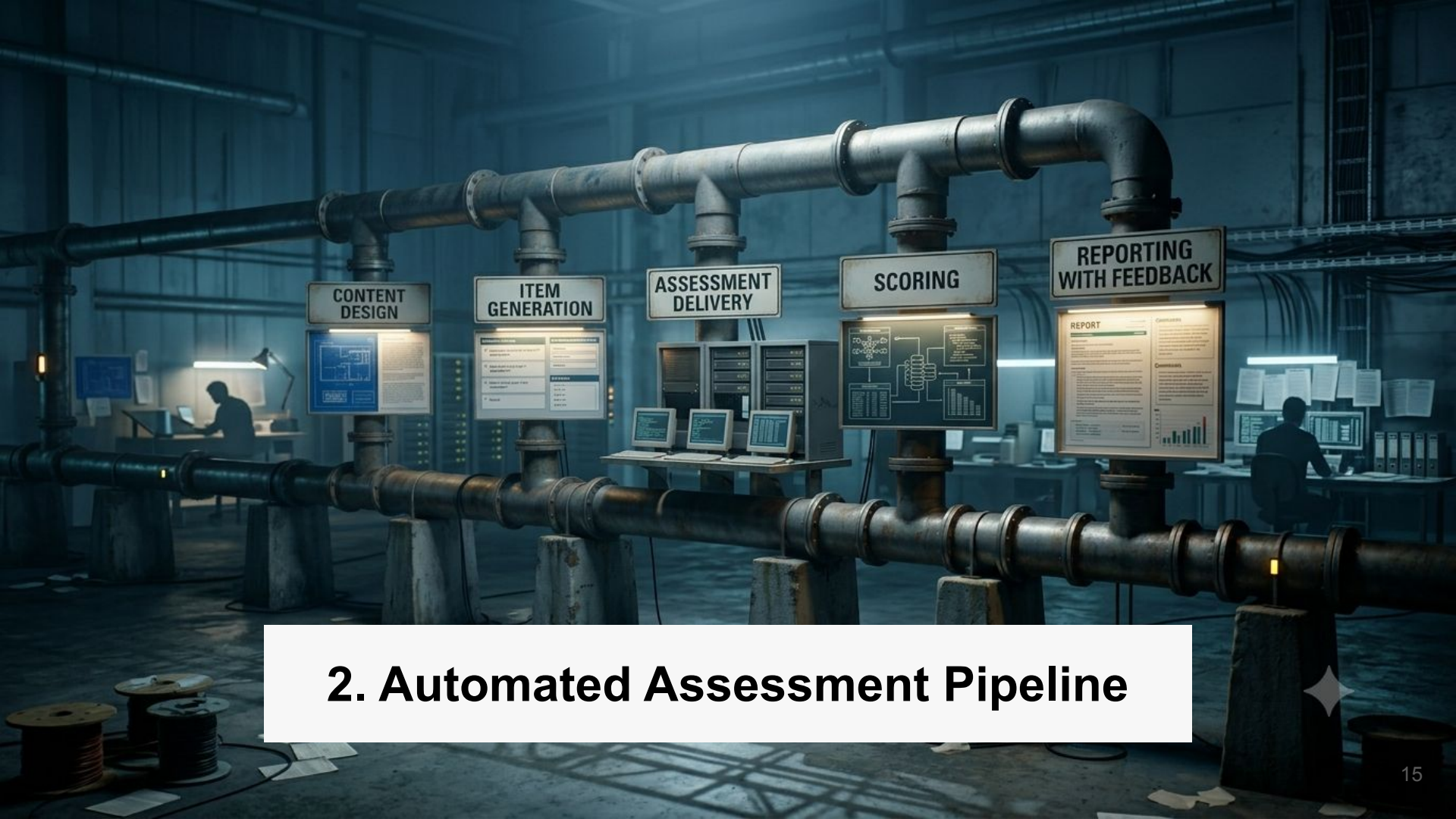


Yuan et al. (2025). NLP and Generative AI for Language Learning and Assessment

Assistive vs. Assessment



Vajjala et al. (2025). Opportunities and Challenges of LLMs in Education: An NLP Perspective



2. Automated Assessment Pipeline

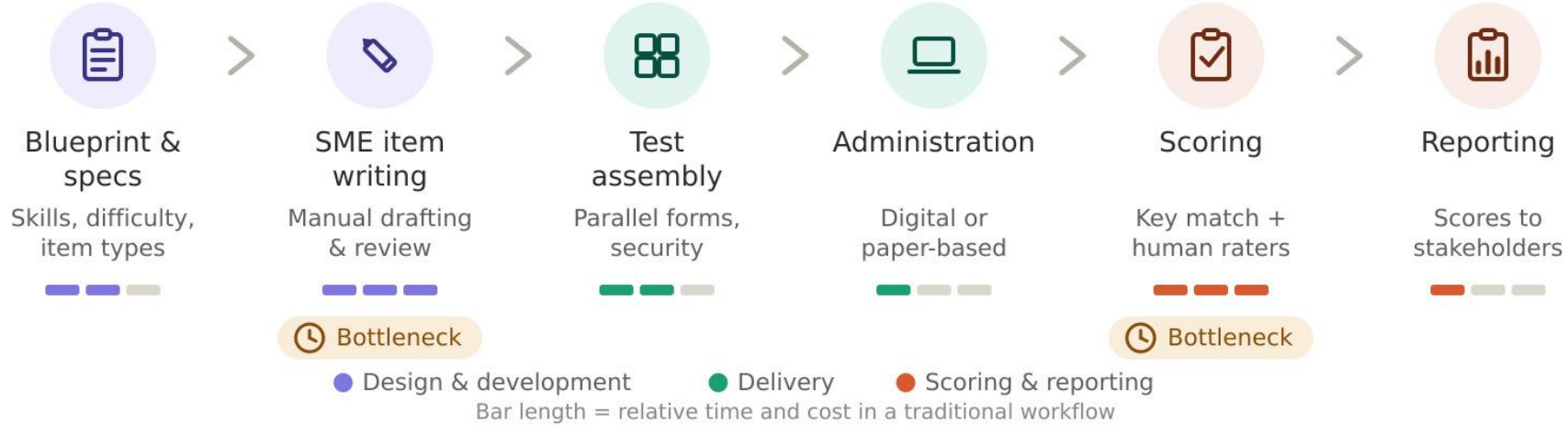
Automation (Broadly Speaking)

"*the use of technology to execute tasks with minimal human intervention by following predefined rules, algorithms, or learned patterns*"

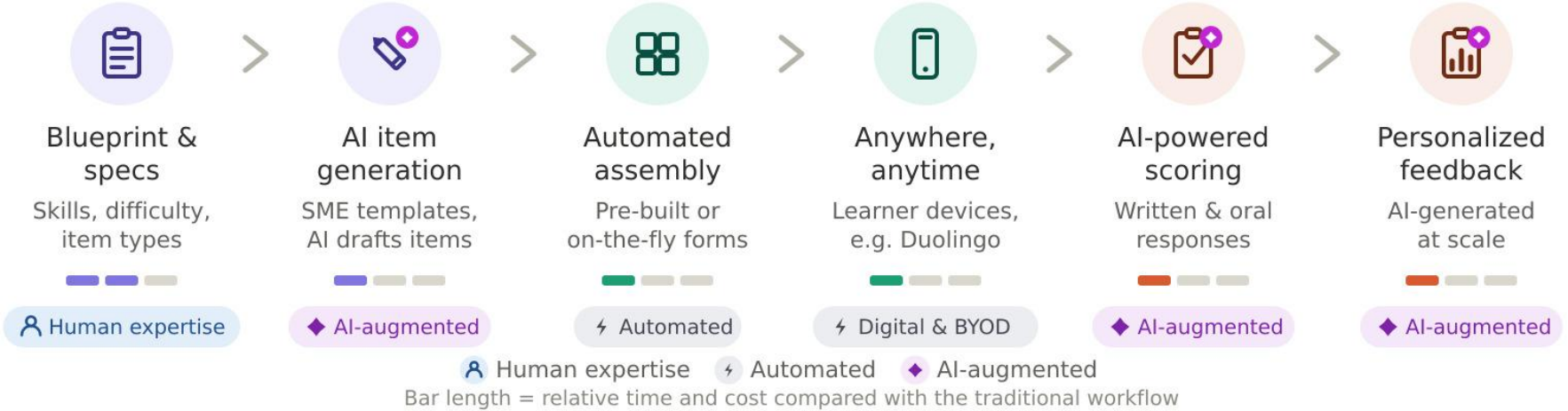
Automation in Educational Assessment

"*the application of computational methods to streamline or replace manual processes across the assessment cycle*"

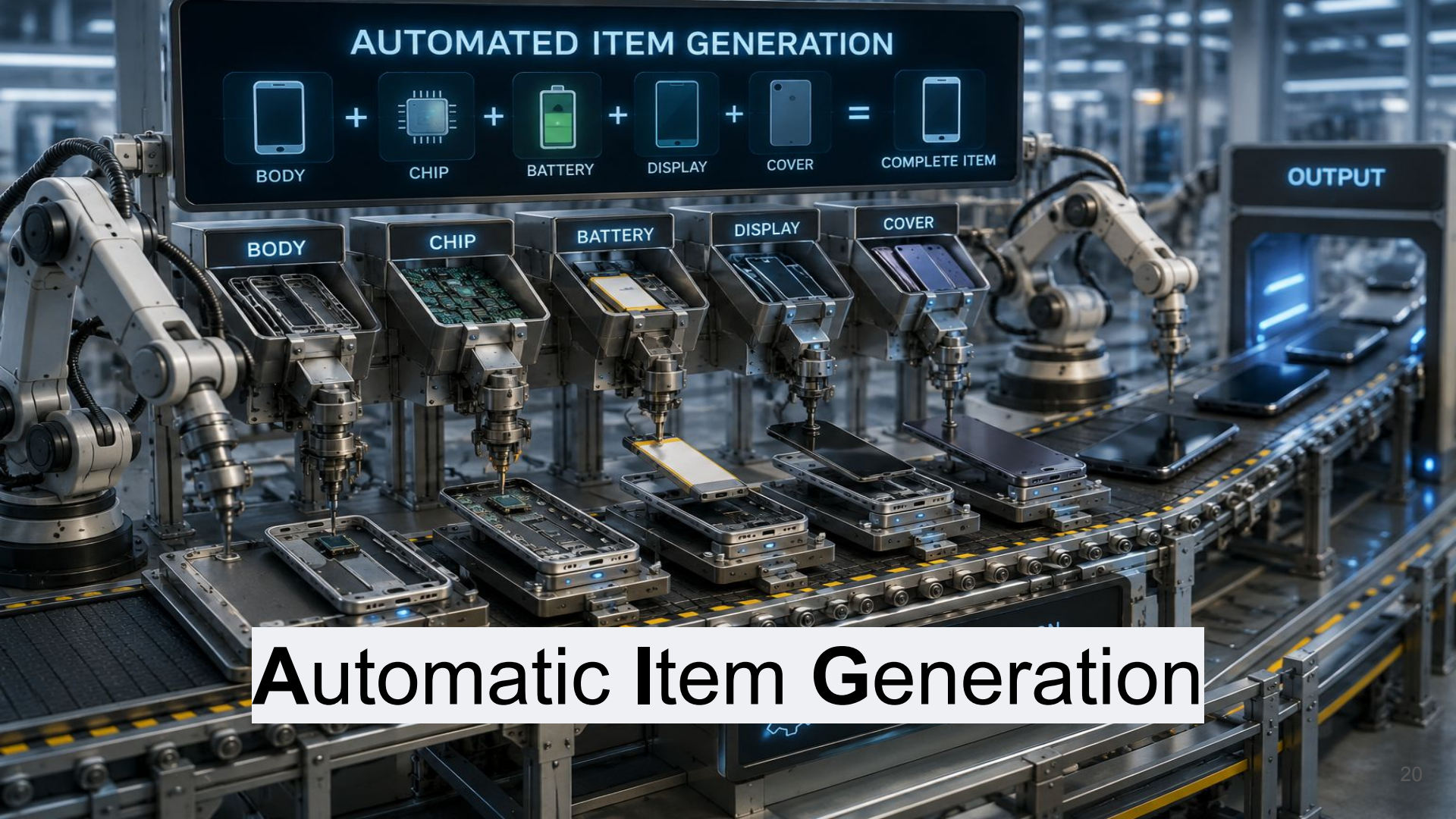
Traditional Assessment Pipelines



Automated Assessment Pipelines



Some Automation Applications

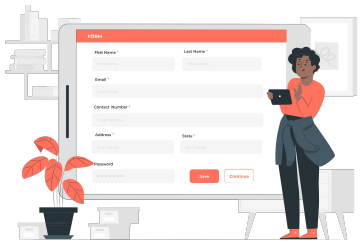


Automatic **I**tem **G**eneration

Automatic **I**tem **G**eneration (AIG)

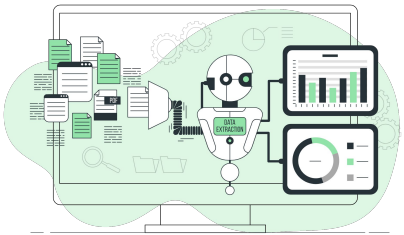
Template-Based AIG

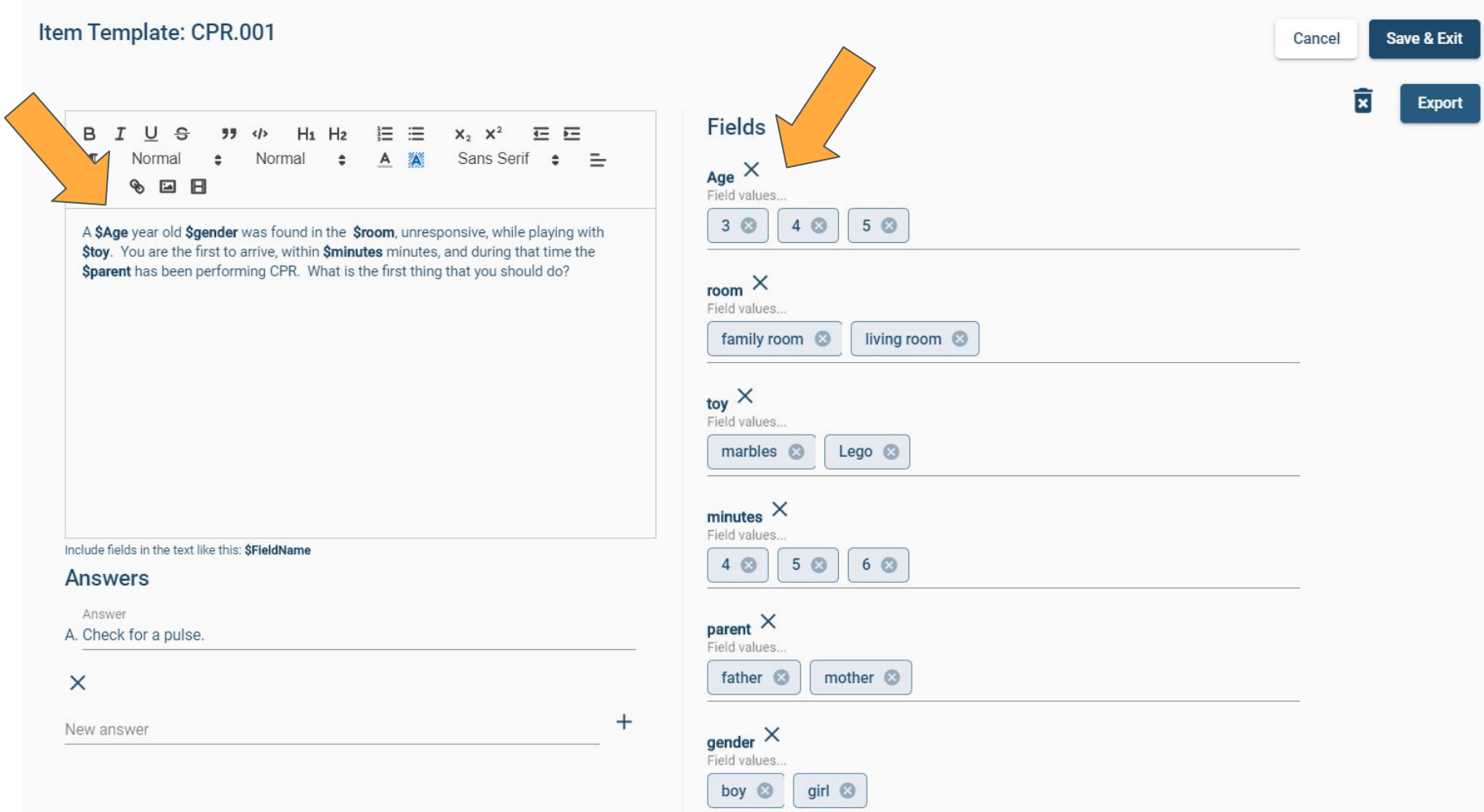
- Still the most common form of AIG
- Mainly driven by subject matter experts
- Works in 3 steps:
  - Cognitive modeling of the target skills
  - Template creation
  - Algorithmic execution



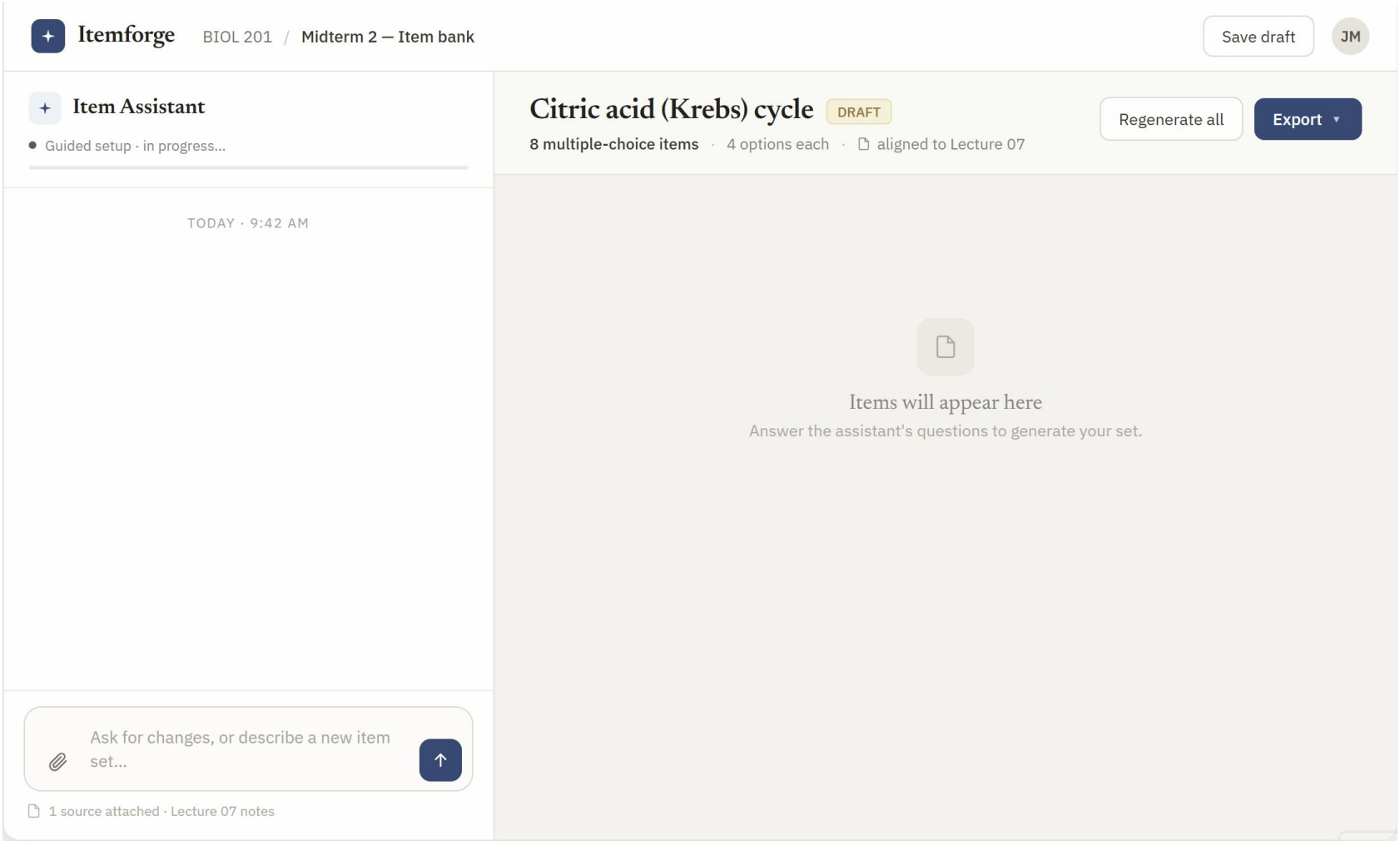
NLP/Generative AI-Driven AIG

- Requires NLP techniques, neural networks and Large Language Models (LLMs)
- Unstructured inputs → text generation
- Less upfront human involvement





Source: https://assess.com/what-is-automated-item-generation/


Itemforge
BIOL 201 / Midterm 2 — Item bank
Save draft
JM
Item Assistant
Guided setup · in progress...
TODAY · 9:42 AM
Ask for changes, or describe a new item set...
1 source attached · Lecture 07 notes
Citric acid (Krebs) cycle
DRAFT
8 multiple-choice items · 4 options each · aligned to Lecture 07
Regenerate all
Export
Items will appear here
Answer the assistant's questions to generate your set.

International Journal of Assessment Tools in Education
2025, Vol. 12, No. 2, 317–340
https://doi.org/10.21449/ijate.1602294
journal homepage: https://dergipark.org.tr/en/pub/ijate Research Article

A review of automatic item generation techniques leveraging large language models

Bin Tan[1*], Nour Armoush[1], Elisabetta Mazzullo[1], Okan Bulut[2], Mark J. Gierl[2]

[1]University of Alberta, Faculty of Education, Measurement, Evaluation, and Data Science, Edmonton, Canada
[2]University of Alberta, Faculty of Education, Centre for Research in Applied Measurement and Evaluation, Edmonton, Canada

https://dergipark.org.tr/en/pub/ijate/article/1602294

International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 72–87
https://doi.org/10.21449/ijate.1124382
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate Research Article

Automatic story and item generation for reading comprehension assessments with transformers

Okan Bulut[1,*], Seyma Nur Yildirim-Erbasli[2]

[1]University of Alberta, Centre for Research in Applied Measurement and Evaluation, Edmonton, AB Canada
[2]Concordia University of Edmonton, Faculty of Arts, Department of Psychology, Edmonton, AB Canada

https://dergipark.org.tr/en/pub/ijate/article/1124382

Chapter 12
Applications of Large Language Models in Automatic Item and Feedback Generation for Early Elementary Education

Okan Bulut, Bin Tan, and Elisabetta Mazzullo

https://link.springer.com/content/pdf/10.1007/978-3-032-17190-0_12.pdf

Educational Measurement: Issues and Practice / Volume 44, Issue 1 / pp. 96-107
Original Article | Open Access

Instruction-Tuned Large-Language Models for Quality Control in Automatic Item Generation: A Feasibility Study

Guher Gorgun, Okan Bulut

First published: 19 December 2024
https://doi.org/10.1111/emip.12663

https://onlinelibrary.wiley.com/doi/10.1111/emip.12663

CHAPTER 1

THE ITEM FACTORY

Intelligent Automation in Support of Test Development at Scale

Alina A. von Davier[1]
Duolingo

Yigal Attali
Duolingo

Andrew Runge
Duolingo

Jacqueline Church
Duolingo

Yena Park
Duolingo

Geoff LaFlair
Duolingo

ABSTRACT

Test development for a high-volume, high-stakes test is a highly complex process. In this chapter, we are proposing a new framework for managing test development at scale, *the item factory*, based on the concept of intelligent automation used in manufacturing. This framework leverages automation for item generation and quality review processes, human expert review processes (including training and quality assurance), crowdsourcing (including pilot research studies, A/B testing, and large-sample pilots), and statistical quality assurance.

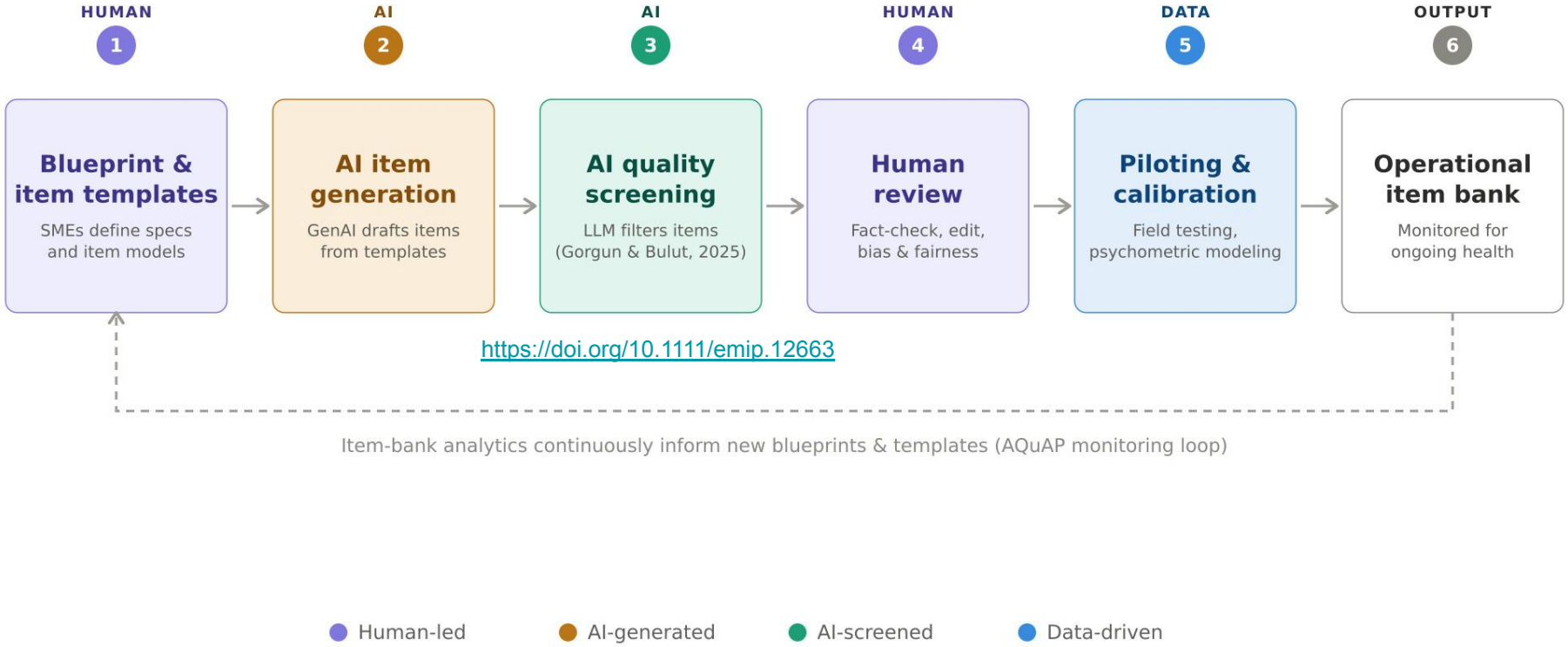
Machine Learning, Natural Language Processing, and Psychometrics, pages 1–25

duolingo english test

https://doi.org/10.1108/979-8-88730-606-320251002

From item writing to a validated item bank

A simplified item-generation and quality-assurance loop, adapted from von Davier et al.'s AQuAP framework (https://arxiv.org/abs/2606.18536)



Adaptive Assessment Delivery

COMPUTERIZED ADAPTIVE TESTING (CAT)

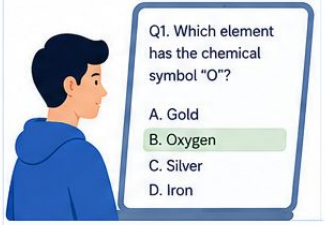
The right question at the right time for every learner

CAT uses psychometric principles to adapt to each learner in real time—selecting the most informative item next to measure ability accurately and efficiently.

HOW IT WORKS

1 START

The test begins with an item of medium difficulty.



2 RESPONSE

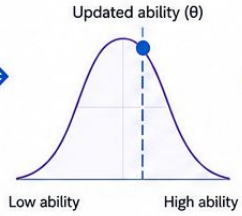
The learner responds to the item.

Correct!

The system records the response.

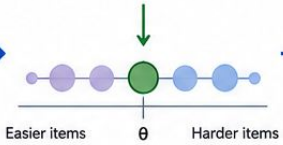
3 ESTIMATE ABILITY

The system updates the learner's ability estimate (θ) using psychometric models (e.g., IRT).



4 SELECT NEXT ITEM

The system selects the item that is most informative for the current ability level (e.g., **medium difficulty** near θ).



Most informative item for current ability

5 REPEAT

Steps 2–4 repeat. The test adapts—getting easier if needed, harder if possible.

6 STOP & SCORE

The test stops when sufficient precision is reached. The final ability estimate is reported.

Estimated ability (θ)

0.85

Standard error

0.23

High precision with fewer items

THE PSYCHOMETRIC FOUNDATION

Item Response Theory (IRT)

Models the relationship between a learner's ability (θ) and the probability of answering an item correctly.

Item Information

Each item provides the most information about ability at a certain point. CAT chooses items to maximize information (and precision).

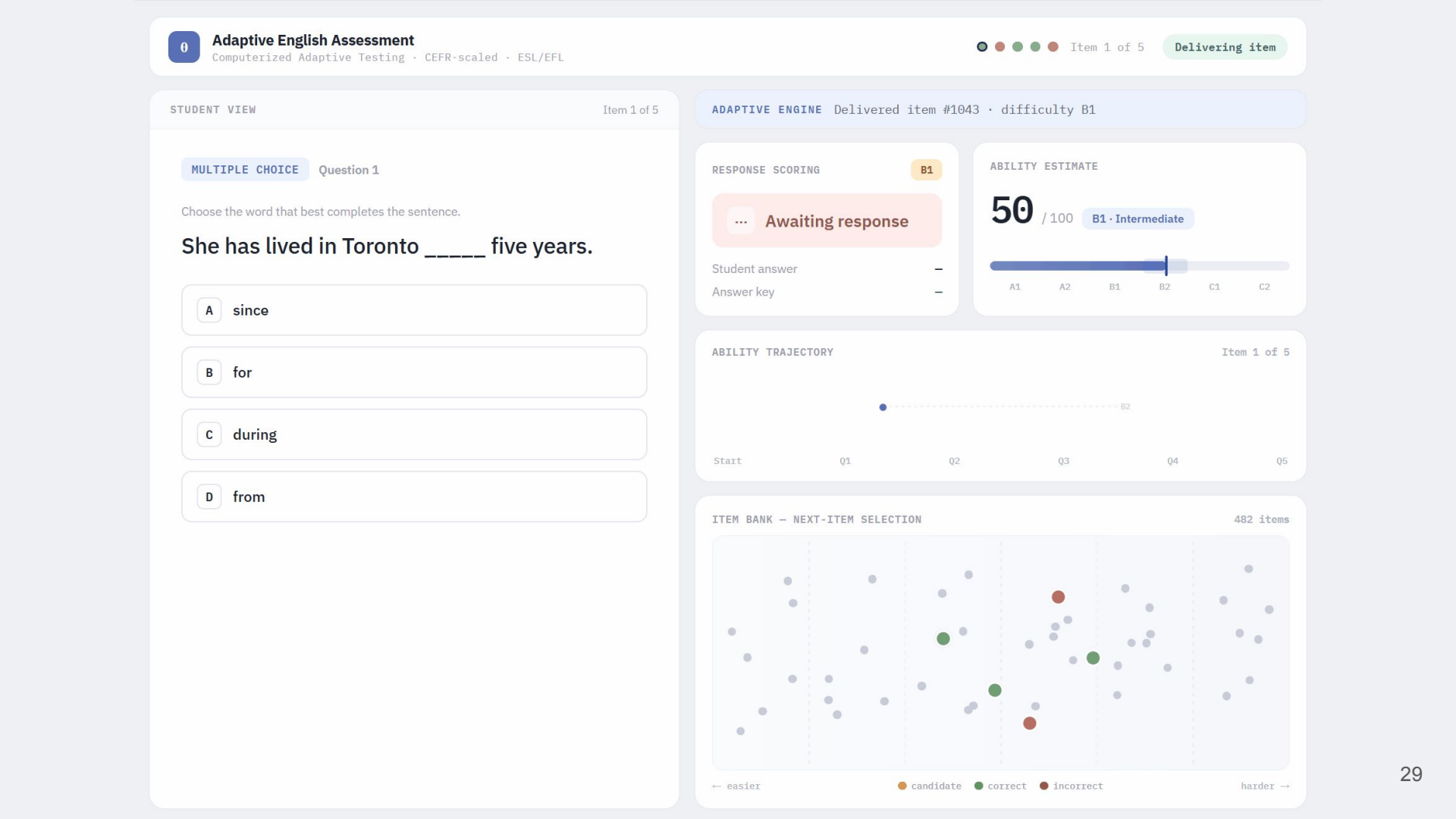
Measurement Precision

More precise estimates come from asking the right questions—not more questions.

ADAPTIVE PATH EXAMPLE

KEY OUTCOME

- Accurate ability estimate with fewer items
- Less test fatigue, better experience
- Fairer measurement across learners and populations


Adaptive English Assessment
Computerized Adaptive Testing · CEFR-scaled · ESL/EFL
Item 1 of 5
Delivering item
STUDENT VIEW
Item 1 of 5
MULTIPLE CHOICE
Question 1
Choose the word that best completes the sentence.
She has lived in Toronto _____ five years.
A since
B for
C during
D from
ADAPTIVE ENGINE Delivered item #1043 · difficulty B1
RESPONSE SCORING
B1
Awaiting response
Student answer –
Answer key –
ABILITY ESTIMATE
50 / 100
B1 · Intermediate
A1 A2 B1 B2 C1 C2
ABILITY TRAJECTORY
Item 1 of 5
B2
Start Q1 Q2 Q3 Q4 Q5
ITEM BANK — NEXT-ITEM SELECTION
482 items
← easier
candidate
correct
incorrect
harder →

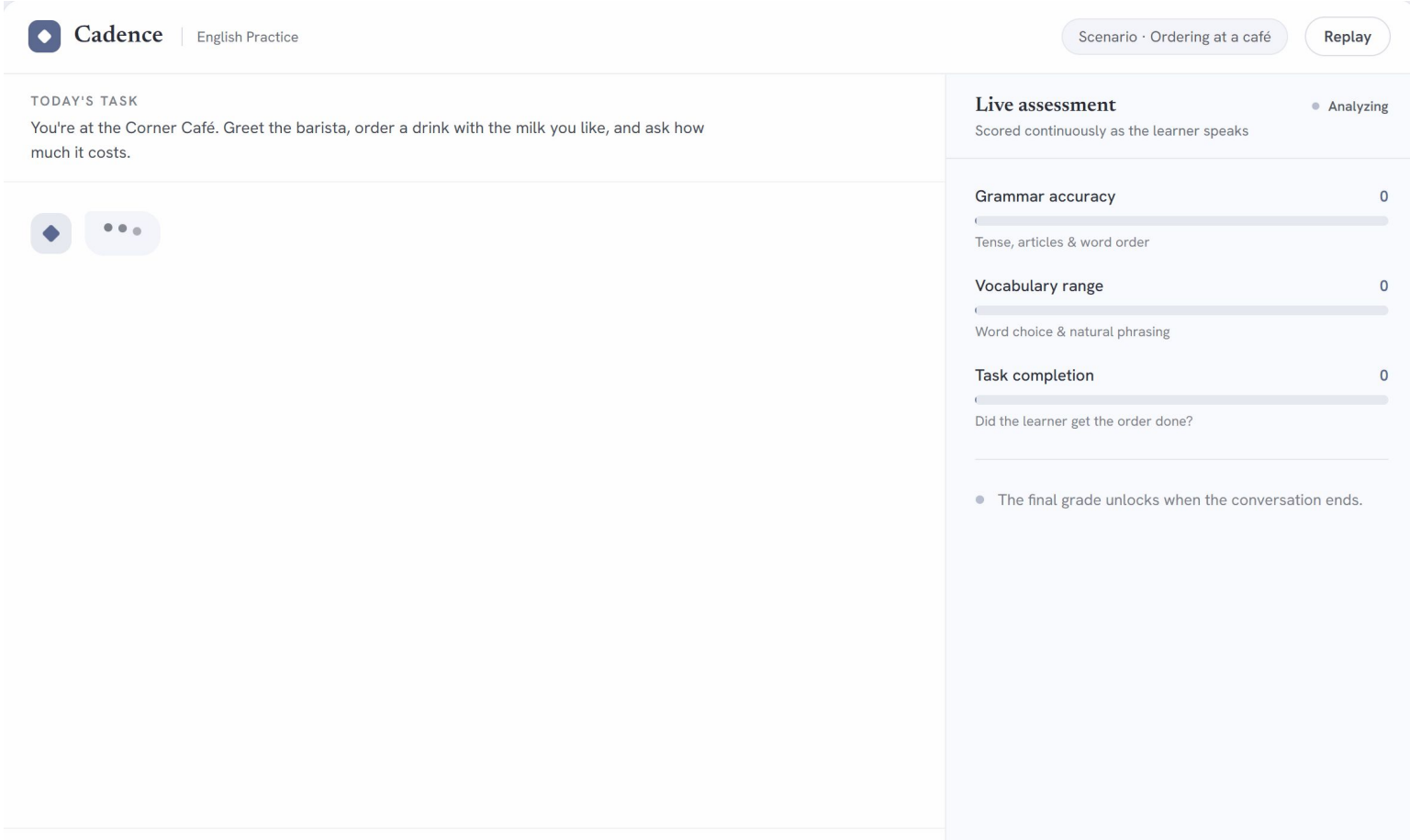
Other Forms of Adaptive Testing

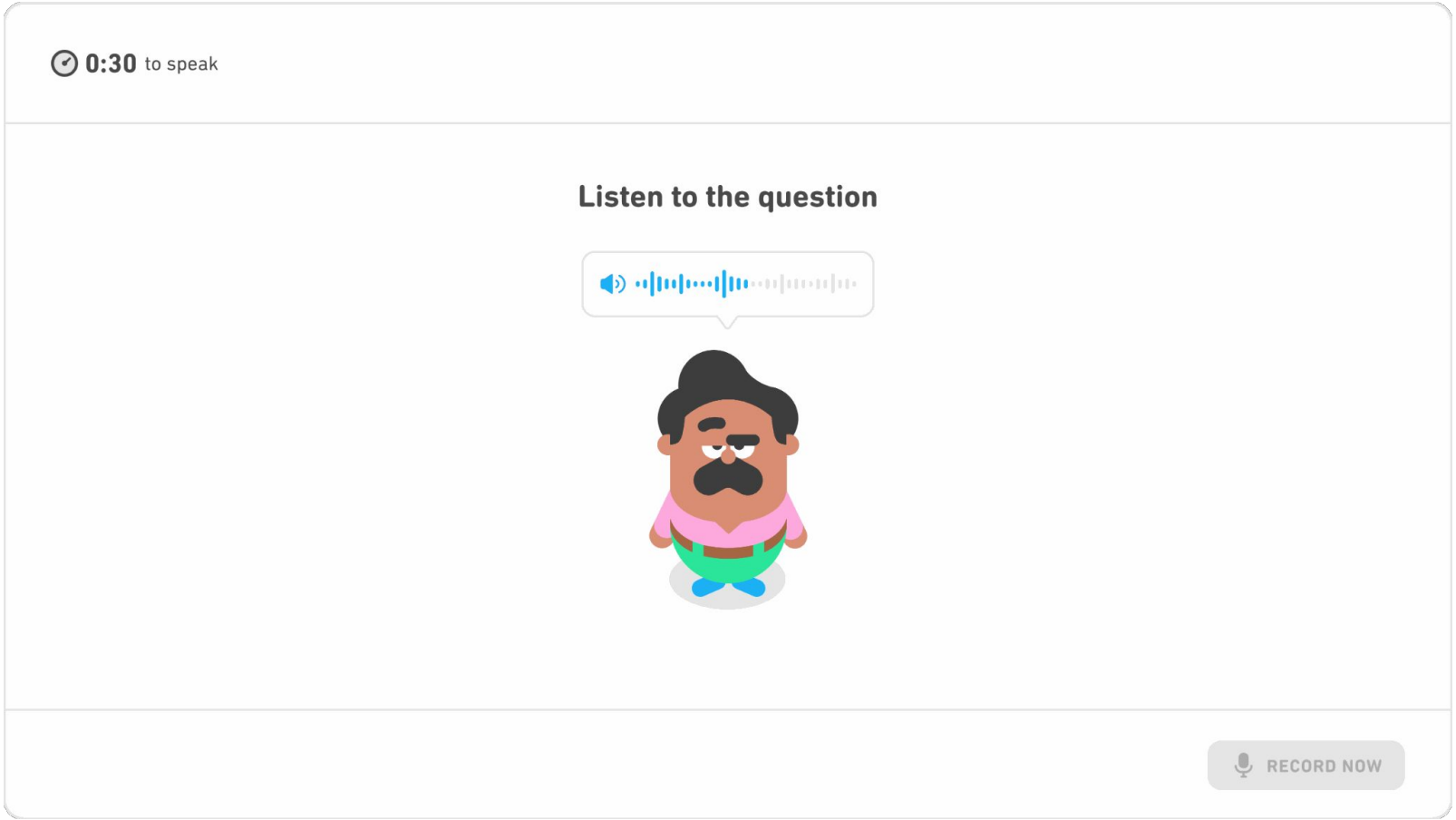
- Computerized adaptive testing (CAT) via **Collaborative Filtering**
  - The underlying idea is similar to recommender systems from commercial applications such as Netflix.
  - The test “recommends” the most suitable items to each test-taker (Xiao & Bulut, 2022).

- **Reinforcement Learning**-based CAT (RL-CAT)
  - A trained RL agent learns how to select items to maximize measurement precision or minimize test length.
  - RL-CAT can consider additional variables such test-taking engagement when looking for the most suitable items (e.g., Bulut et al., 2025).

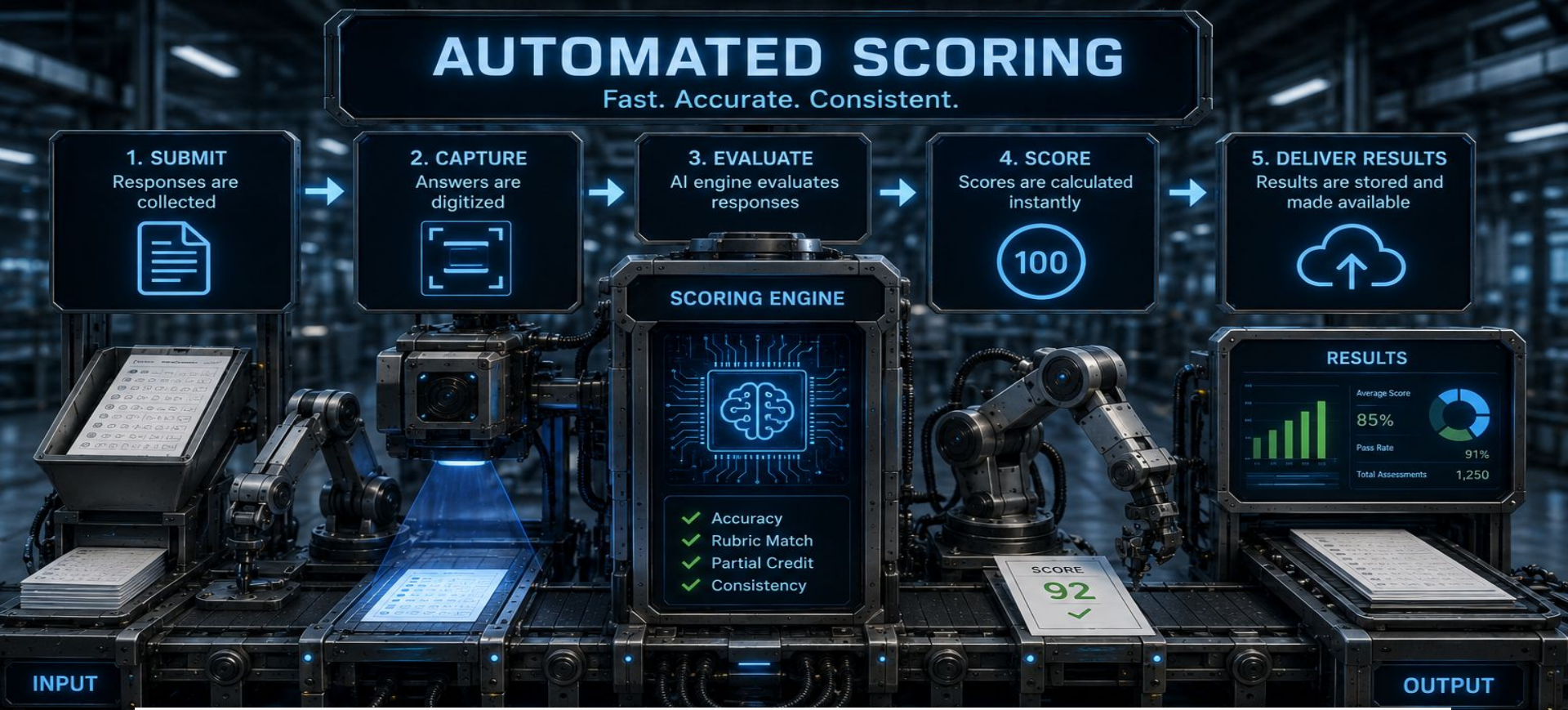
- **Knowledge Graph**-based Adaptive Testing (KGAT)
  - KGAT maps test items and learning concepts into a semantic web of relationships.
  - The model aims to map exactly which branches of the graph learners have mastered and where gaps remain while avoiding too easy and hard items (Qiu & Chen, 2025).

Conversation-Based Assessment (Yildirim-Erbasli & Bulut, 2023)





Source: https://blog.englishtest.duolingo.com/interactive-speaking/



Automated (Essay) Scoring

INPUTS

📝 Essay

💬 Short Answer

🎤 Spoken Response

💻 Programming Code

📊 Simulation Log

🎮 Game Activity

AI SCORING ENGINE

✓ Evaluate

✓ Compare with rubric

✓ Assign score

✓ Generate confidence

OUTPUTS

📈 Score

📋 Feedback

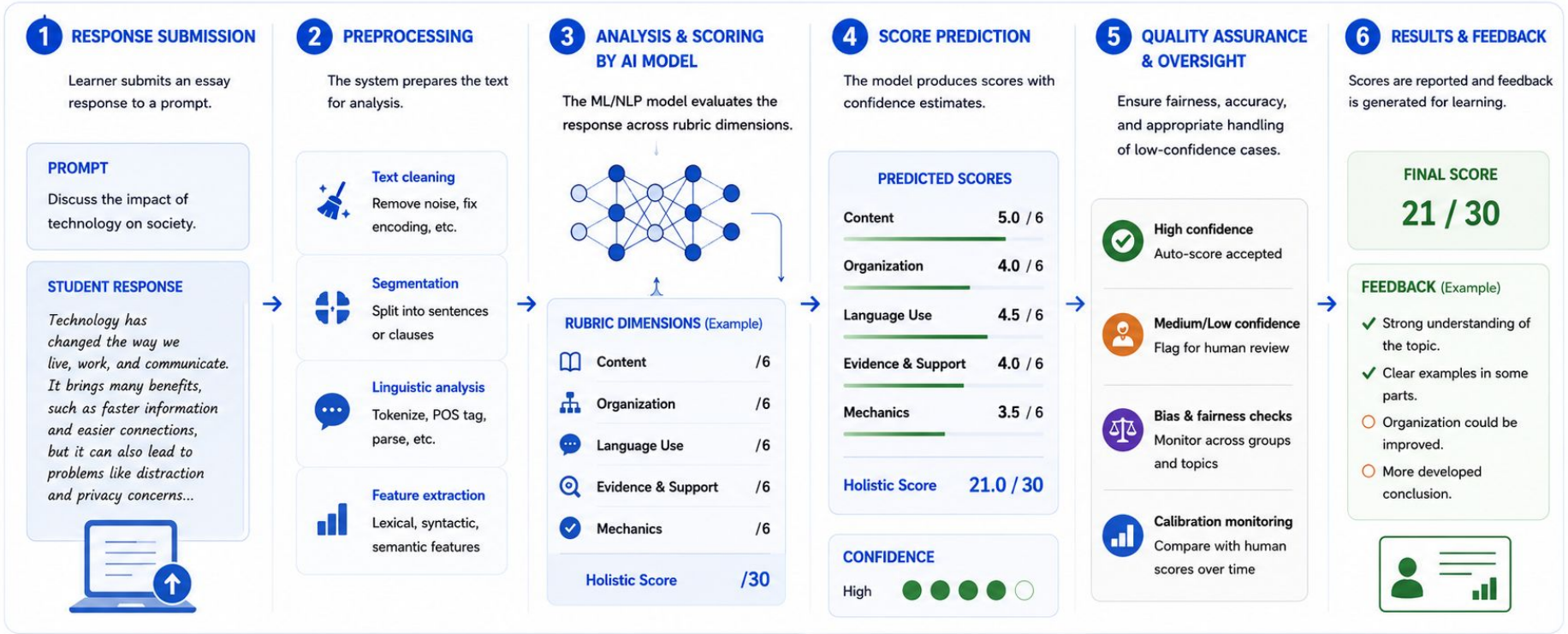
📊 Analytics

🎯 Mastery Level

Real-Life Examples of Automated Scoring

Organization / Platform	Domain	Example
ETS	Essay Writing	GRE Analytical Writing (e-rater®)
Duolingo	Writing & Speaking	Duolingo English Test
Code.org	Programming	Automatic evaluation of coding exercises
Khan Academy	Mathematics	Instant scoring of math problems

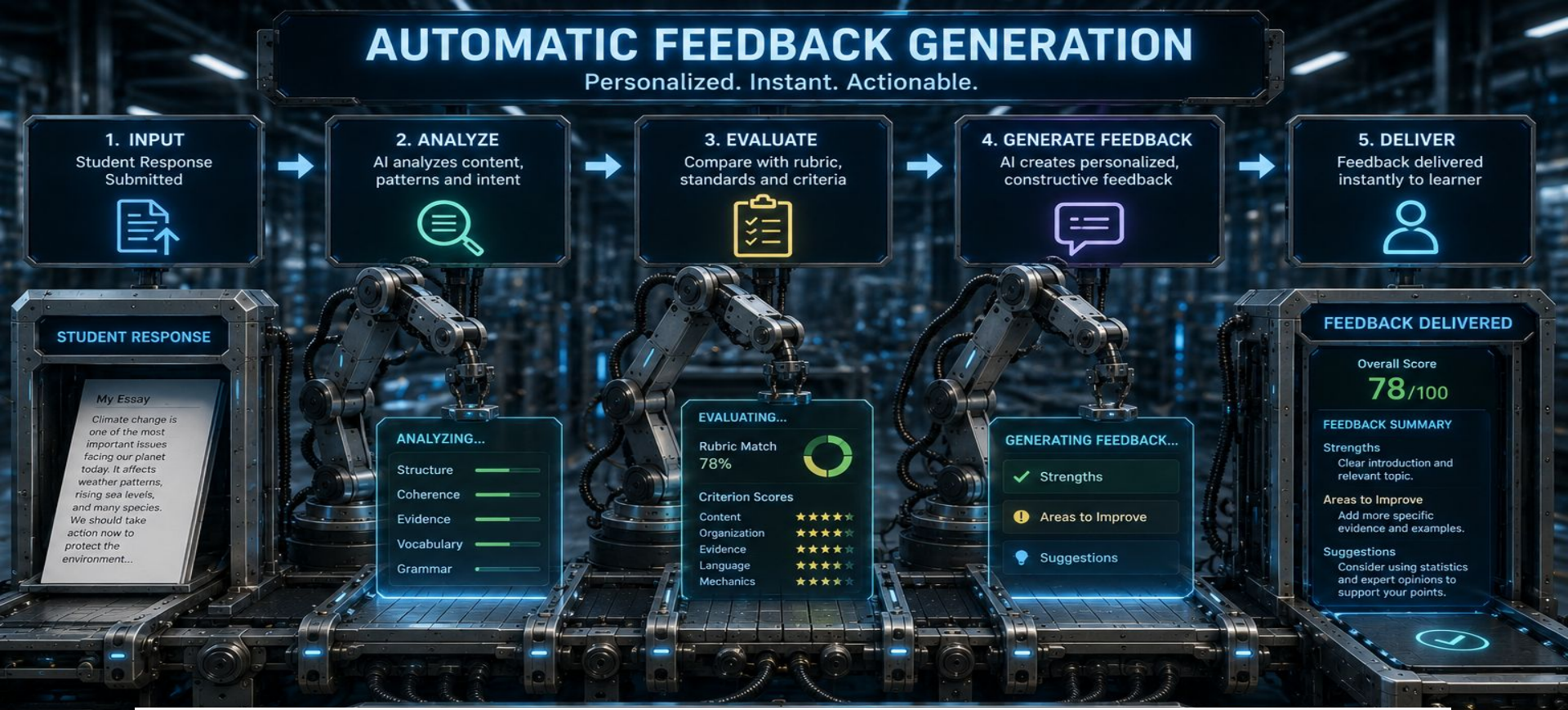
Automated Scoring of Written Responses



CONSTRUCTED-RESPONSE SCORING PIPELINE

How a written answer becomes a score

Pre-process the text, turn it into embeddings, and score four rubric traits with a confidence band on every prediction.



Automatic Feedback Generation

Proceedings of Machine Learning Research 1:1–18, 2024 FM-EduAssess at NeurIPS 2024 Workshop

Automated Feedback Generation for Open-Ended Questions: Insights from Fine-Tuned LLMs

Elisabetta Mazzullo MAZZULLO@UALBERTA.CA
Measurement, Evaluation, and Data Science
University of Alberta, Edmonton, AB T6G 2G5 CANADA

Okan Bulut BULUT@UALBERTA.CA
Centre for Research in Applied Measurement and Evaluation
University of Alberta, Edmonton, AB T6G 2G5 CANADA

https://proceedings.mlr.press/v264/mazzullo25a.html

Education and Information Technologies
https://doi.org/10.1007/s10639-026-13988-0

Check for updates

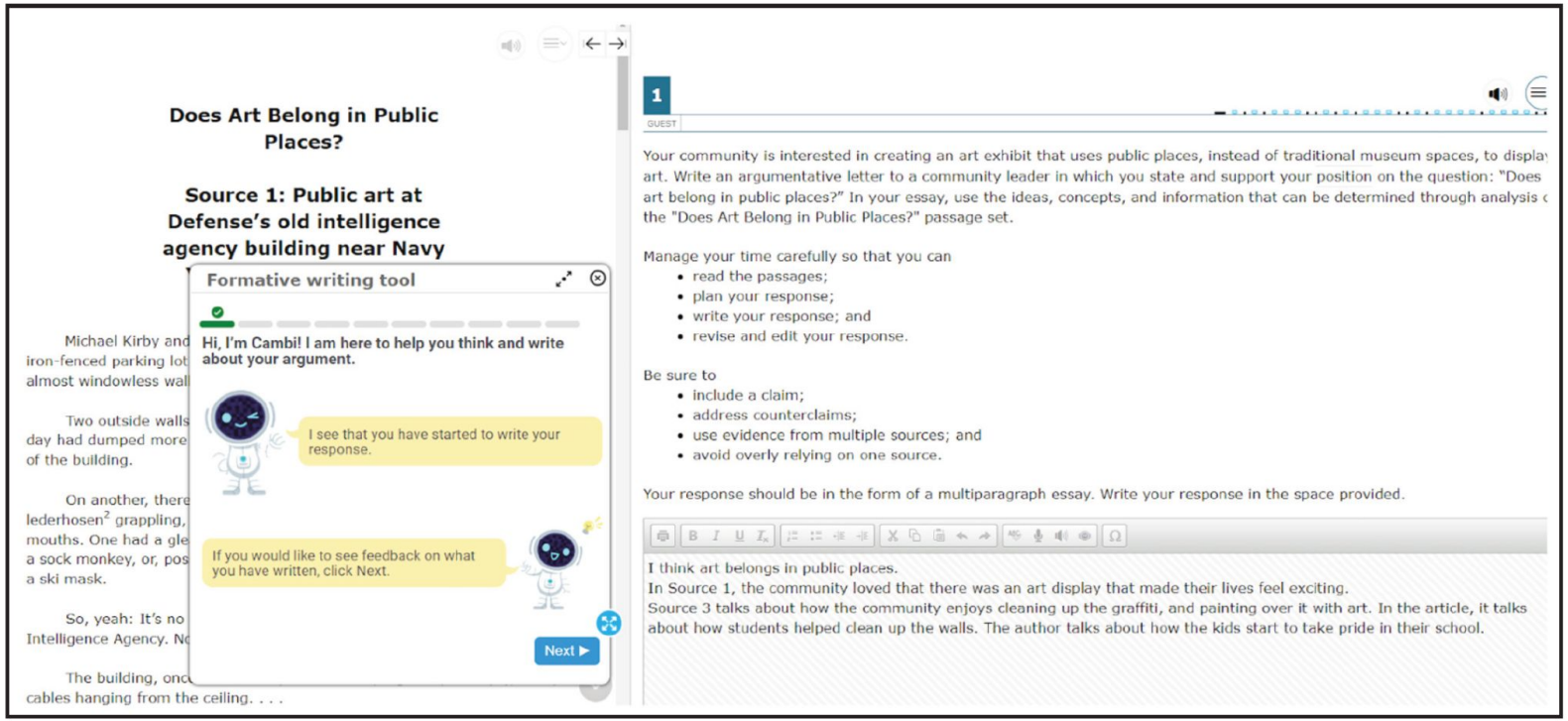
Is it worth the effort? a comparison of automatic educational feedback generated by base and fine-tuned LLMs

Elisabetta Mazzullo[1] · Okan Bulut[2] · Daniel Jerez[1] · Kevin Vo[1] · Cole Walsh[3] · Gill Sitarenios[3] · Alexander MacIntosh[3]

https://link.springer.com/article/10.1007/s10639-026-13988-0

Feedback statements should:

- Be written in the second person to foster a personal and engaging tone
- Be balanced positive and negative aspects
- Maintain a supportive tone
- Offer actionable suggestions
- Address the unique qualities of each response to enhance personalization



"Write On with Cambi" by [Cambium Assessment]

Automation in language assessment: what does it deliver?

APPLICATION	BENEFITS	WHAT IT MEANS IN PRACTICE
Automatic item generation	Efficiency; Cost saving	• Generates large item pools from SME templates instantly, eliminating manual drafting at scale • Reduces item development cost and time-to-deployment; enables rapid bank replenishment
Automated scoring	Efficiency; Cost saving	• Scores constructed responses (essays, speech) instantly without human raters • Scales to high-volume assessments with consistent application of rubric criteria
Adaptive test delivery	Cost saving; Enhanced accuracy	• Fewer items needed per examinee (CAT) — reduces testing time and administration cost • Targets items to ability level, improving measurement precision across the proficiency range
Automated feedback	Efficiency; Pedagogical benefits	• Delivers immediate, personalized feedback at scale — no rater scheduling or delays • Supports learning by providing actionable, targeted guidance aligned to individual learner needs

Automation Myths vs. Reality

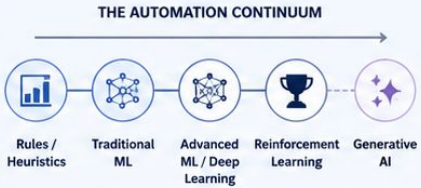
1 AUTOMATION ALWAYS LEADS TO BETTER OUTCOMES

- Not every stage needs automating
- Established frameworks may already be optimal
- Human oversight still required

REALITY

Automate with purpose. Focus on high-impact areas and keep humans in the loop.

2 AUTOMATION MEANS USING GENERATIVE AI



- The pipeline can be automated via rules, ML, deep learning, or RL
- GenAI is not always the right tool
- Choose the right approach for the problem, data, and context

REALITY

Match the right technique to the goal—GenAI is one tool, not the only answer.

3 AUTOMATED SYSTEMS ARE FAIR AND UNBIASED

- Models inherit and scale human biases
- Automated systems can disadvantage underrepresented groups

Non-native speakers · Learners with disabilities · Regional dialect speakers · Low-resource languages

REALITY

Bias doesn't disappear with automation—it can be amplified. Build fairness in by design and test it continuously.

4 ONCE VALIDATED, AN AUTOMATED SYSTEM STAYS VALID

- Language evolves, populations shift, models drift
- "Set and forget" is dangerous
- Continuous monitoring and re-validation are essential

REALITY

Build for change. Monitor, measure, and re-validate to keep systems reliable over time.

THE BOTTOM LINE

Responsible automation is not about replacing humans—it's about combining the right technology with human judgment to create systems that are effective, fair, and resilient.

Be intentional · Use the right tool · Design for fairness · Monitor always

Let’s take a …


BREAK
As a noun, break means when you pause or rest from something. As a verb, you can use it as break or breaking, and it means to shatter or divide something.
grammarist.com

3. Human-AI Collaboration for Inclusive Education

Definition of Inclusive Education

UNESCO's (2005) gives definition of inclusion as

"a process of addressing and responding to the diversity of needs of all learners through increasing participation in learning, cultures and communities, and reducing exclusion within and from education”

Inclusive Education is part of the SDG 4 (Agenda 2030, UN, 2015)

The Epistemological Problem of Inclusion

- **The Global Mandate**: UNESCO (2005) and SDG4 demand we address the diversity of all learners and reduce exclusion.
- **The Translation Gap**: Inclusion is not a monolith; its definition and impact shift drastically depending on the national context (Geoffre, Koonpornpen, Plevková, Yuan)
  - ***Systemic Inclusion*** (e.g., Switzerland, Luxembourg, UK): Extends far beyond disability to embrace linguistic, cultural, and social diversity as integral to the educational project.
  - ***Special-Ed Centric*** (e.g., Slovakia, Thailand): Legal frameworks tether inclusion closely to special education paradigms, often overlooking broader dimensions of diversity.

- **The Computational Challenge**: Educational AI can't be built for a binary paradigm (disabled vs. neurotypical). To achieve systemic inclusion, we must computationally operationalize continuous, multidimensional diversity.

The Linguistic & Cognitive Reality (The PISA Gap)

Focus: Language acquisition as a means of inclusion (language of literacy / instruction)

- **The Data**: OECD PISA results reveal a staggering, systemic lag. Students who speak a different language at home fall 1.5 to 2 years behind their peers on average.
- **The Intersectional Penalty**: For low-SES and multilingual intersections, this gap explodes to 4 years behind by age 15.

  This is not an “ability” gap; it is a massive cognitive friction gap.
- **The Tension**: When language friction meets cognitive overload, the learning process halts entirely. Systemic inclusion requires hyper-personalized, real-time attention. Manual differentiation for 30 diverse profiles is humanly impossible. We must automate the support.

Adaptive Learning systems, promises and limits

The 1980s-90s Intelligent Tutoring Systems

- **Why it was implemented**: To solve “Bloom’s 2 Sigma Problem” by automating 1-on-1 human tutoring using expert systems, cognitive task analysis, and hardcoded misconception networks.

- **Key Limitations**: ITS was incredibly brittle and domain-specific. It could not handle the messy, open-ended reality of primary classrooms, nor could it adapt to diverse inguistic or cultural profiles outside its rigid decision trees.

- **The Paradigm Shift**: The field realized we could not hardcode every possible student misconception. We needed systems that could learn and adapt from data dynamically.

Adaptive Learning systems, promises and limits

The 2010s: Adaptive Learning & The Routing Fallacy

- **Why it was implemented**: The rise of big data and Item Response Theory (IRT) allowed platforms to dynamically route content based on right/wrong answers, adjusting difficulty in real-time.

- **Key Limitations**: It optimizes strictly for outcomes (lagging indicators). It treats diversity as a “content routing” problem rather than a cognitive one. Crucially, **Adaptive Learning is a monologue, not Human-Computer Interaction (HCI)**. It reacts after the failure has occurred, entirely ignoring the real-time cognitive state and the reason for the failure (e.g., language friction vs. lack of math knowledge). (act as “differentiation”)

- **The Paradigm Shift**: The field sought a way to understand the context and nuance of the student’s struggle, leading to the integration of natural language and generative models.

Adaptive Learning systems, promises and limits

The 2020s: The LLM expectations and limitations in Primary Education

- **Why it was implemented:** Generative AI promised infinite, conversational, context-aware tutoring that could explain concepts in any language or style.

- **Key Limitations**:
  - **The Ethical Black-Box**: Even if technical issues like hallucinations or cognitive offloading are mitigated, LLMs remain opaque. Teachers cannot audit why the AI gave a specific explanation or scaffold. In inclusive education, transparency is non-negotiable.
  - **Trust at Scale**: While some early adopters may accept black-box AI, the vast majority of teachers and parents globally will not trust an opaque system with their children's learning. Trust is the foundation of educational adoption.
  - **Computational Feasibility**: LLMs require massive compute resources and low-latency infrastructure. For global applicatio - especially in under-resourced contexts in the Global South - this is prohibitively expensive and technically unfeasible.

➔ We do not just need better content generation; we need a fundamentally **better interaction architecture** that is **transparent**, **trustworthy**, **lightweight**, and respects the student's **cognitive load** and **reflection process**

Documenting Adaptive Learning at school

Swiss National Science Foundation

UNI FR UNIVERSITÉ DE FRIBOURG UNIVERSITÄT FREIBURG — INSTITUT FÜR MEHRSPRACHIGKEIT / INSTITUT DE PLURILINGUISME / ISTITUTO DI PLURILINGUISMO / INSTITUT DA PLURILINGUITAD / INSTITUTE OF MULTILINGUALISM

uni.lu Faculty of Humanities, Education and Social Sciences

Project DOCTALE-FR 2023-2027

- **Adaptive learning to foster reading comprehension of short texts**
- Based on a competency framework: readable for teachers, parents (& students)
- Parameters : objectives, pedagogical intentions

- Differentiation as a retroaction is not enough for inclusion. We must ensure that the activity is **accessible** to all > Universal Design for Learning
- Learner interface can be adapted via additional features (AF) (TTS, Visual lexicon, OpenDyslexic font, …)

Documenting Adaptive Learning at school

- **The Study Design**: We tested adaptive learning in the wild for over 1.5 years, with over 340 students across 12 classes and 6 schools in French-speaking Switzerland.
- **The Conditions Tested**: with or without AF, Adaptive Learning vs. No Adaptive Learning, Adaptive Learning with autonomy, Standardised pathway, Adapted pathway, and ecological teacher use. Game-based Platform.
- **The Data available:**
  - Linguistic complexity (syntactic/textual structure, vocabulary, …)
  - Clickstream and interaction logs: anonymized student identifier, class identifier, session start/end timestamps (enabling duration calculation), validation outcome (correct/incorrect)
  - Temporally ordered sequences of *actions* and *validation*

Data for 107’000 activities

Documenting Adaptive Learning at school

- **The Usage Reality**: Text-to-Speech and Visual Lexicon were used~5% of the time. OpenDyslexic and character spacing were used 0.5% or less
- **The Equity Paradox**: Neurodivergent students used these tools half as much as non-neurodivergent students, despite having a significantly greater objective need
- **The Choice Paradox**: Giving students unassisted choice assumes “Choice Literacy” but, without it, bottom-decile students get trapped in inefficient routines.

The right activity at the right moment (only)

Shift 1: Real-Time Scaffolding via State and Trait

- **Respecting the Reflection Process**: We cannot simply send a blind scaffold message at the 2-minute mark saying “Use the dictionary!” The student might be thinking deeply, or they might need Tool Y, not Tool X.
- **The State (The Immediate Need)**: The real-time cognitive block (e.g., the student is hesitating on a complex sentence, working memory is maxed).
- **The Trait (The Historical Profile)**: The long-term data (e.g., past data shows this specific student responds highly to visual aids or TTS).
- **The Synthesis:** The system uses the State to know exactly when to intervene without interrupting productive struggle, and the Trait to finalize the choice of the right tool to suggest.

Shift 2: From Outcomes to Processes (The “Silly Mistake”)

- **The Platform Paradigm**: Because we are a learning platform, not an exam platform, we care about the process, not just the final outcome.

- **The “Silly Mistake” Reality**: It is entirely normal for a student to make a silly mistake (e.g., saying 4+ 5= 10). This does not mean they lack the underlying competency of addition.

- **The Pedagogical Imperative**: If we only measure outcomes, we penalize silly mistakes and destroy confidence. By measuring the process (how they tried to solve it, where they hesitated), we can scaffold the reasoning, push the students to self-regulate, and help them see their own mistake without penalizing their underlying knowledge

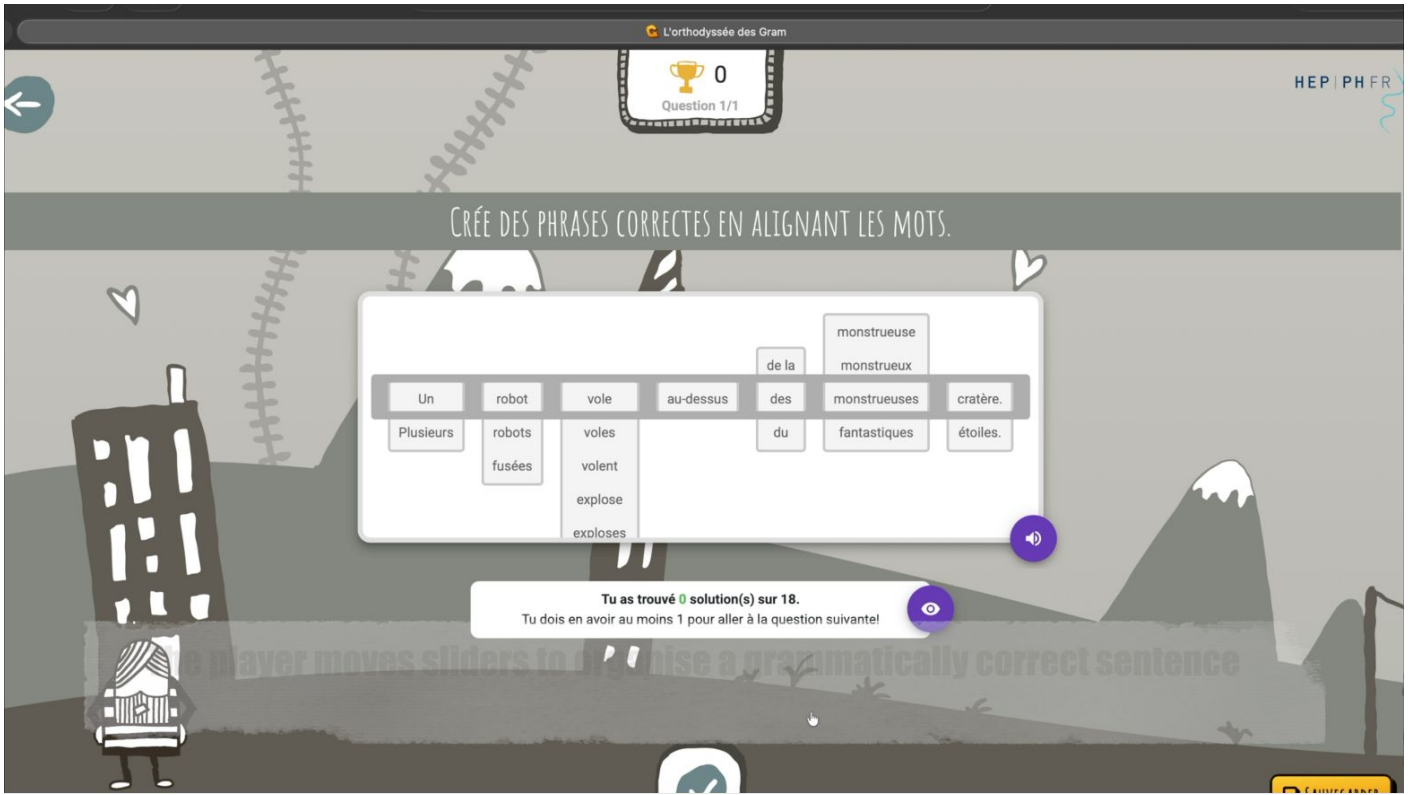
Process-Aware Detection in Action

- **Beyond Scores**: Using action sequences (e.g., tracking convergence, chain interference metrics, navigation paths) to detect cognitive overload before the score drops.

- **Distinguishing Struggle Types**: Process data allows the system to differentiate between cognitive overload (erratic action patterns requiring tool support) and productive struggle (within the zone of proximal development, requiring the AI to step back).

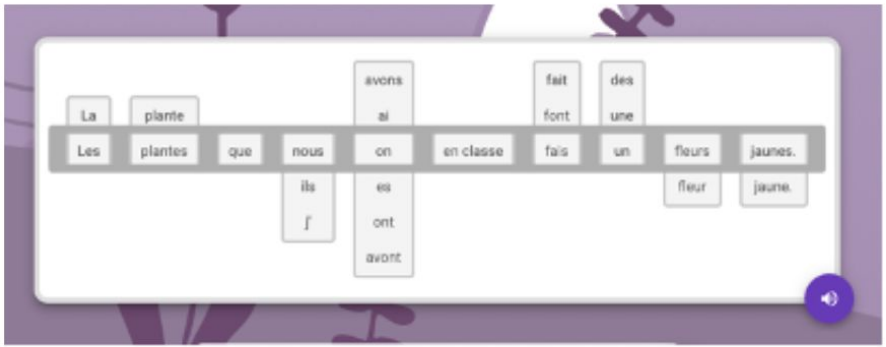
- **Precision Scaffolding**: This prevents the system from interrupting the student with feedback precisely when they should be left alone to think.

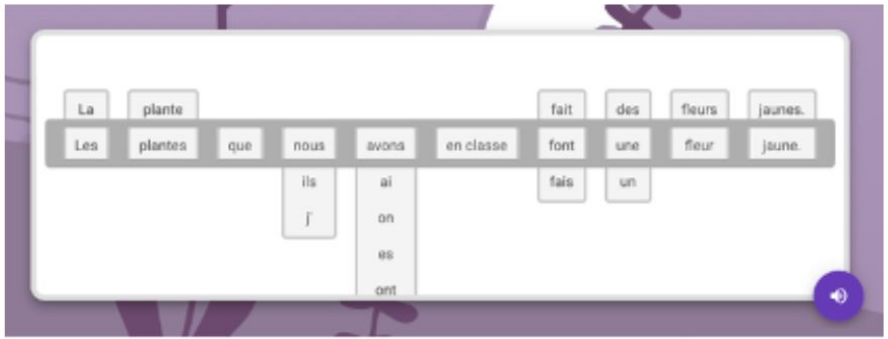
Application on a game : Gram Tirettes



Gram Tirettes: Hamming Distance

- Analyzing **36,234 interactions** from **234 students** across **2,144 sessions** and **42 distinct morphosyntactic exercises**.





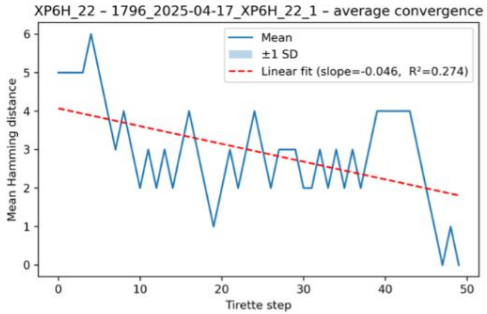
Hamming Distance: Impact of moves

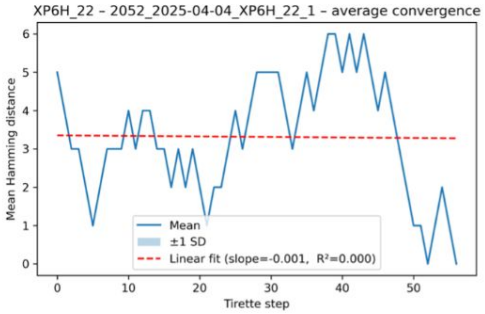
Table 5: Impact of moves on Hamming distance per grammatical category

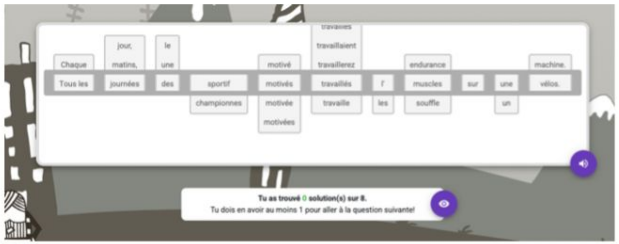
Grammatical category	Improved (%)	Worsened (%)	Unchanged (%)
det	33.0	44.9	22.1
nom	51.2	15.3	33.5
adj	40.8	23.6	35.6
ver	56.6	14.3	29.1

Hamming Distance: Convergence

- **XP6H_22**: 10 sliders; 8 solutions



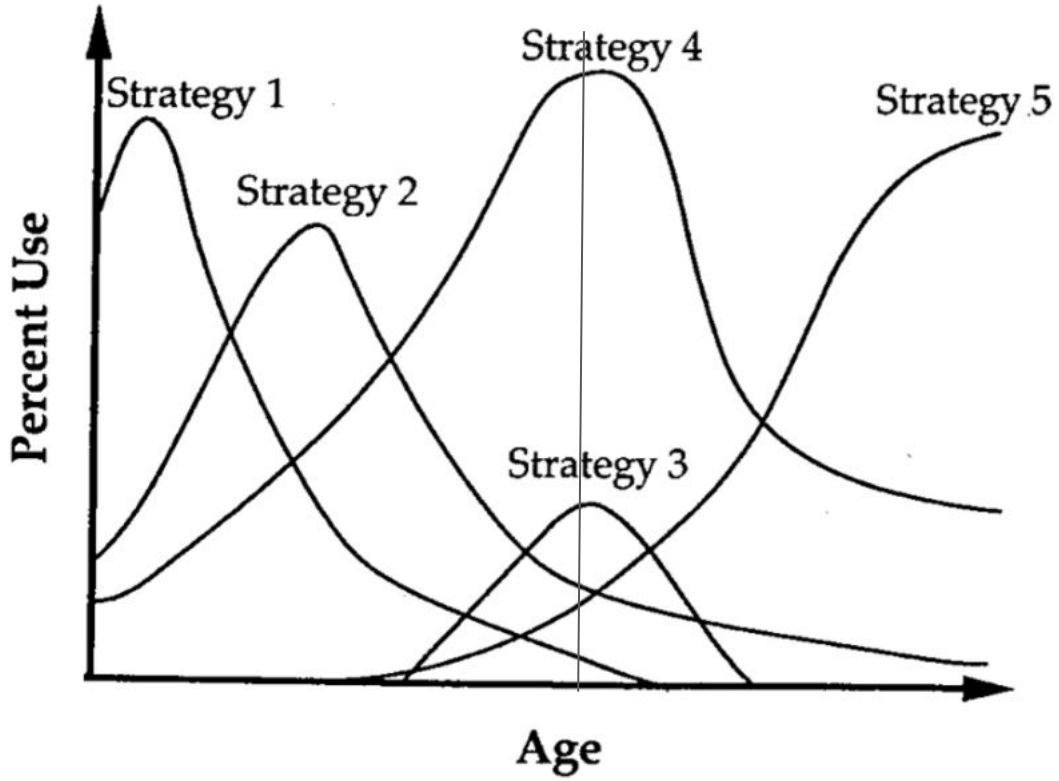




Strategies

- **Verb-First**: Combines early targeting of verb-related slots with sustained focus on the verb agreement chain, rewarding students who prioritize high-impact grammatical dependencies.
- **Chain-Aware**: Measures whether chain switches occur only after making measurable progress on the previous chain, capturing systematic hypothesis testing across morphosyntactic dependencies.
- **Error-Focus**: Computes the proportion of actions that reduce Hamming distance, operationalizing targeted error correction as a metacognitive monitoring strategy.
- **Randomness**: Captures undirected exploration by combining chain entropy with low error focus, distinguishing productive hypothesis generation from unguided guessing.
- **Strategic Commitment**: Identifies sustained pursuit of a non-initial candidate solution until it becomes at least as close as the previously optimal path, reflecting deliberate rule exploration.

Strategies



Overlapping waves Model (Siegler, 2000)

Current work: towards adaptive scaffolding

A theory-driven diagnostic framework that makes hidden grammatical reasoning processes observable and actionable

Analyzing 36,234 interactions from 234 students across 2,144 sessions and 42 distinct morphosyntactic exercises

Strategies > Coordination Demand Index based on strategies and behaviour

Reasoning states > quadrant > transitions

The Teacher-in-the-Loop: Equity Auditing

- **Amplifying Professional Judgment**: AI does not replace the teacher; it handles the real-time micro-scaffolding, freeing the teacher to focus on high-level pedagogical interventions.
-
- **The Equity Dashboard**: Makes the invisible “Access Literacy Gap” and “Tool Gap” visible in real-time.
-
- **No Black Boxes**: The AI explains why it delivered a specific scaffold message (based on the student’s State and Trait), allowing the teacher to exercise professional, calibrated judgment.

Summary of the Epistemological Shift

- **From**: Opaque, outcome-based automation
  →**To**: Transparent, process-aware collaboration.

- **From**: Assuming metacognitive skills
  →**To**: Actively scaffolding access and choice literacy via State/Trait feedback.

- **From**: Content routing (Adaptive Learning)
  →**To**: Cognitive state inference (True HCI).

- **From**: Conflating failure
  →**To**: Disentangling true Competency (Trait) from daily Inefficiency (State).

Human-AI collaboration: adaptive scaffolding

The right activity at the right moment with the right conditions ?

And full agency for the teacher > Integrating the tool into ordinary class preparation can take many forms

Human-AI collaboration: adaptive scaffolding

Thank you!

thierry.geoffre@uni.lu trystan.geoffre@gmail.com


PAWSING FOR A
COFFEE BREAK

4. Game-Based Language Learning

Overview

Section 1

Introduction to GBLL

Definition and Theories

Section 2

Games and Practices

COTS games
Serious games
Empirical examples

Section 3

AI and GBLL

AI in GBLL classes
AI in Game development

- COTS games
- Serious games
- Limitations

1. Introduction

Games-Based Language Learning

Game-Based Language Learning (GBLL): an approach to language teaching and learning where digital games are used as a tool to promote, enhance and sustain language learning (Peterson, 2013).

- 🏆 Immersive exposure to the language learning environment
- ❤️ lowered anxiety and other affective barriers to language learning
- 💬 increased use of the target language for interaction in gaming

🎮 Gaming is already a mainstream activity among today's learners: 85% of U.S. teens play video games, and 41% play them daily (Gottfried & Sidoti, 2024).

Peterson, M. (2013). *Computer games and language learning*. Palgrave Macmillan.
Gottfried, J., & Sidoti, O. (2024, May 9). *Teens and video games today*. Pew Research Center. https://www.pewresearch.org/internet/2024/05/09/teens-and-video-games-today/

Theories

Constructivist and social constructivist theories: Learners construct knowledge through meaningful experiences and reflection, social interaction, collaboration, and scaffolding support (Palincsar, 1998; Piaget, 1972; Vygotsky, 1978).

Affective theories:

- **Motivation** is strengthened when learning environments support autonomy, competence, and relatedness (Deci & Ryan, 2000).
- **Self-efficacy:** Feedback, achievable challenges, and visible progress build confidence and persistence (Bandura, 1977).
- **Social comparison:** Leaderboards and peer visibility can motivate engagement, though they should be used carefully (Festinger, 1954).

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215.
Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227-268.
Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*(2), 117-140.
Palincsar, A. S. (1998). Social constructivist perspectives on teaching and learning. *Annual Review of Psychology, 49*, 345-375.
Piaget, J. (1972). The psychology of the child. Basic Books.
Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

2. Games and Practices

COTS Games

Commercial Off-The-Shelf (COTS) games refer to readily available games produced **for entertainment purposes**.

Popular Genres	Representative Games
MMORPG (Massively Multiplayer Online Role-Playing Game)	🧙World of Warcraft
Simulation	🏡The Sims 4, Stardew Valley
Adventure	🗡️The Legend of Zelda: Tears of the Kingdom
Interactive Drama	🤖Detroit: Become Human
FPS (First-Person Shooter)	🔫Valorant, Counter-Strike
MOBA (Multiplayer Online Battle Arena)	⚔️ League of Legends
Virtual World / Sandbox	🧱Minecraft / Roblox

COTS Games

COTS Games in Language Learning

Designed for entertainment, not language learning

✓ Affective benefits

✗ Pedagogical limitations

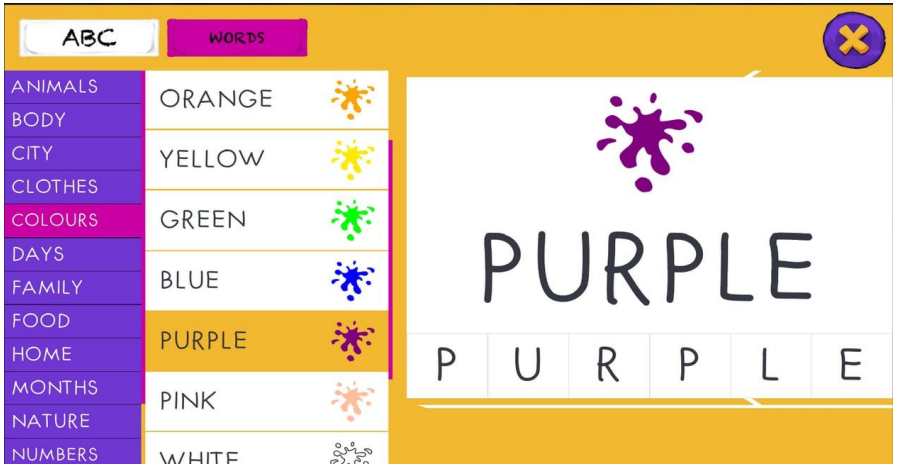
- Not aligned with learners' language proficiency or curricular objectives [2]
- incidental rather than intentional learning outside formal learning environments [1]
- Unpredictable language input and communicative opportunities [2]
- Low learning efficiency, especially for novice players [3]
- Potential exposure of learners to toxic language and negative social interactions [4]

[1] Calvo-Ferrer, J. R., & Belda-Medina, J. (2021). *The effect of multiplayer video games on incidental and intentional L2 vocabulary learning: The case of Among Us. Multimodal Technologies and Interaction, 5*(12), 80.
[2] Dykes, R. (2021). *Considerations for Using COTS Video Games as Language Learning Tools*. Sojo University.
[3] Andersson, S. (2025). *Language Learning Through Watching Gameplay: Merits for Extramural Participation. Frontiers in Foreign Language Education, 8*, 87–96.
[4] Sun, X., Yu, V., & Chen, V. H. H. (2025). *Toxic behavior in multiplayer online games: The role of witnessed verbal aggression, game engagement intensity, and social self-efficacy. Chinese Journal of Communication, 18*(4), 426–444. https://doi.org/10.1080/17544750.2024.2425662

Serious Games

Serious games: games designed with **a primary purpose other than entertainment**. In the context of language learning, serious games aim to facilitate the acquisition of language skills.

A shooter game for grammar learning (Chen & Hsu, 2020)



Antura and the Letters: An award-winning serious game designed to help children (5-9) learn languages.

Chen, H. J. H., & Hsu, H. L. (2020). The impact of a serious game on vocabulary and content learning. *Computer Assisted Language Learning*, *33*(7), 811-832.
Guardiola, E., & Czauderna, A. (2018). Merging gameplay and learning in educational game design: The gameplay loop methodology in Antura and the Letters. In *Proceedings of the 12th European Conference on Games Based Learning (ECGBL 2018)* (pp. 154–161). Academic Conferences and Publishing International.

GBLL in Practice

Fun but pedagogically limited COTS games vs. Pedagogical relevant but unattractive serious games

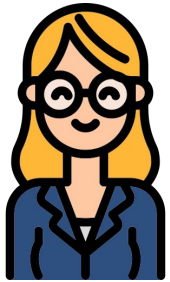
Language Teachers

Explore ways to enhance pedagogical relevance of COTS games

Game Developers

Make serious games more COTS-like and “playable”.

Teacher Intervention



01. Game selection

- pedagogical alignment
- learner proficiency
- Leaner gameplay experience

02. Instructions

- game mechanics
- key vocabulary and expressions

03. Structured gameplay

- Session planning
- Achievable tasks and quests

04. Learning activities

- vocabulary quizzes
- discussion activities
- Storytelling,presentation
- writing

Sims 4

Life Simulation Game

Players create and control virtual people, “Sims”, and guide their daily lives.

- build homes, develop careers, form relationships, and pursue personal goals.
- experience open-ended life choices.

Integration into the classroom

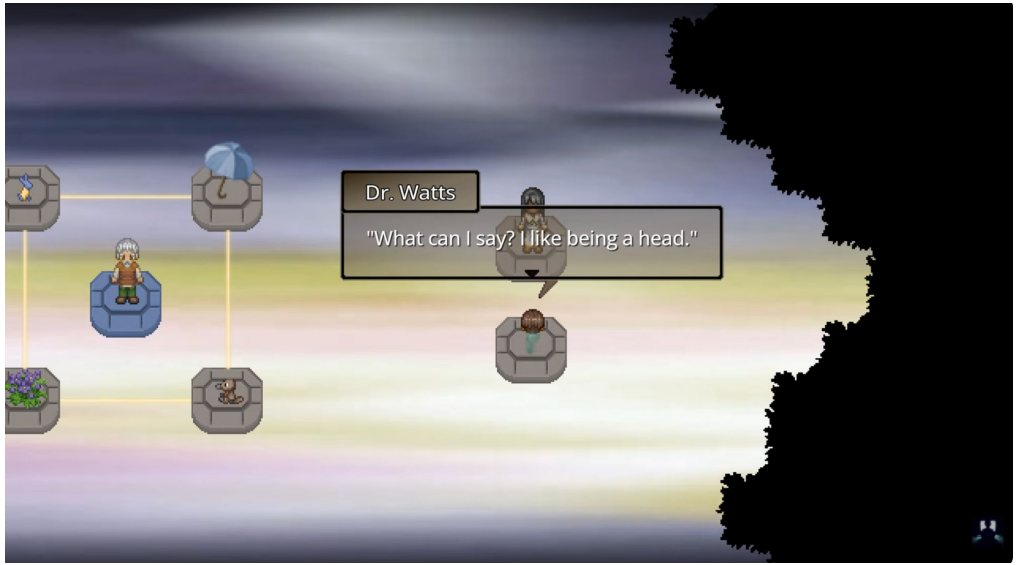
1. Courses aimed at daily communication
2. Theme-based gameplay: character building, home construction, work, study, etc.
3. Instruction: Gameplay mechanics, explanation of difficult language in the game
4. Learning activities: Pre- and post-gameplay vocabulary tests; storytelling of gameplay episodes; writing on events in the game

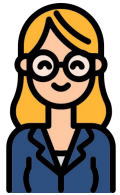
To the Moon

Visual novel that follows two doctors who use memory-altering technology to fulfill a dying man's final wish: to go to the moon. Through exploring the patient's memories, players uncover his life story, relationships, and regrets.

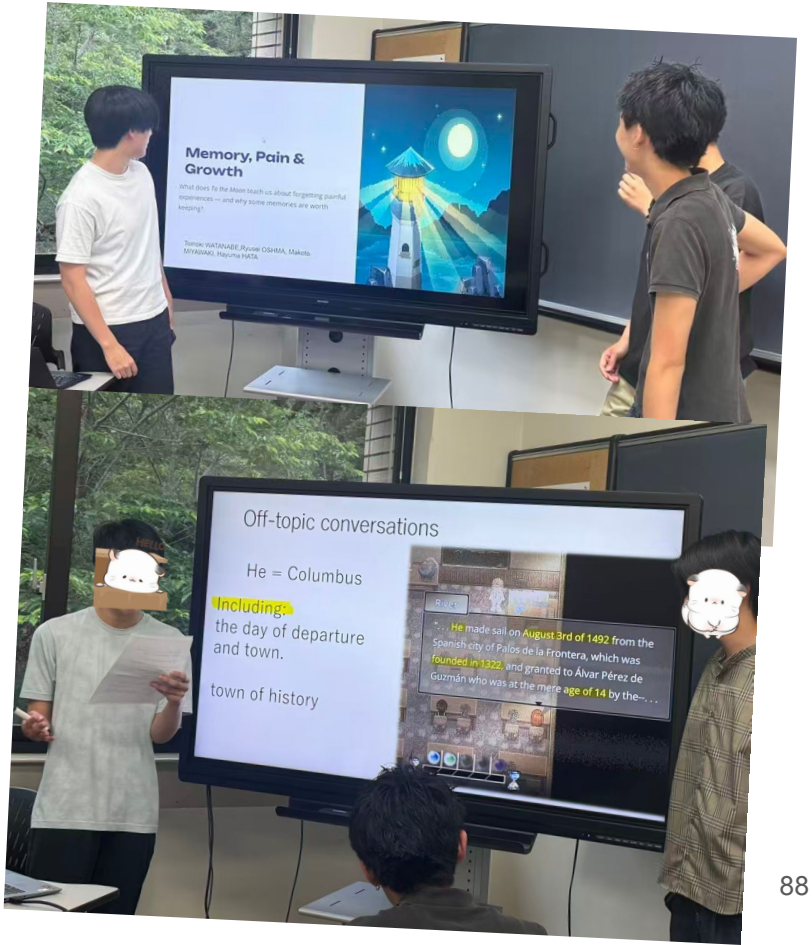
1. Reading
2. Gameplay divided into even durations based on walkthrough videos
3. Instruction on game language
4. Fully controllable pre- and post-gameplay vocabulary tests, discussion of ethical issues relevant to the game, group presentations and writing

To the Moon





Look at the scene above. What is a "pun"? And explain why "I like being a head" is a pun.



Death and Taxes

“Decisions matter” game

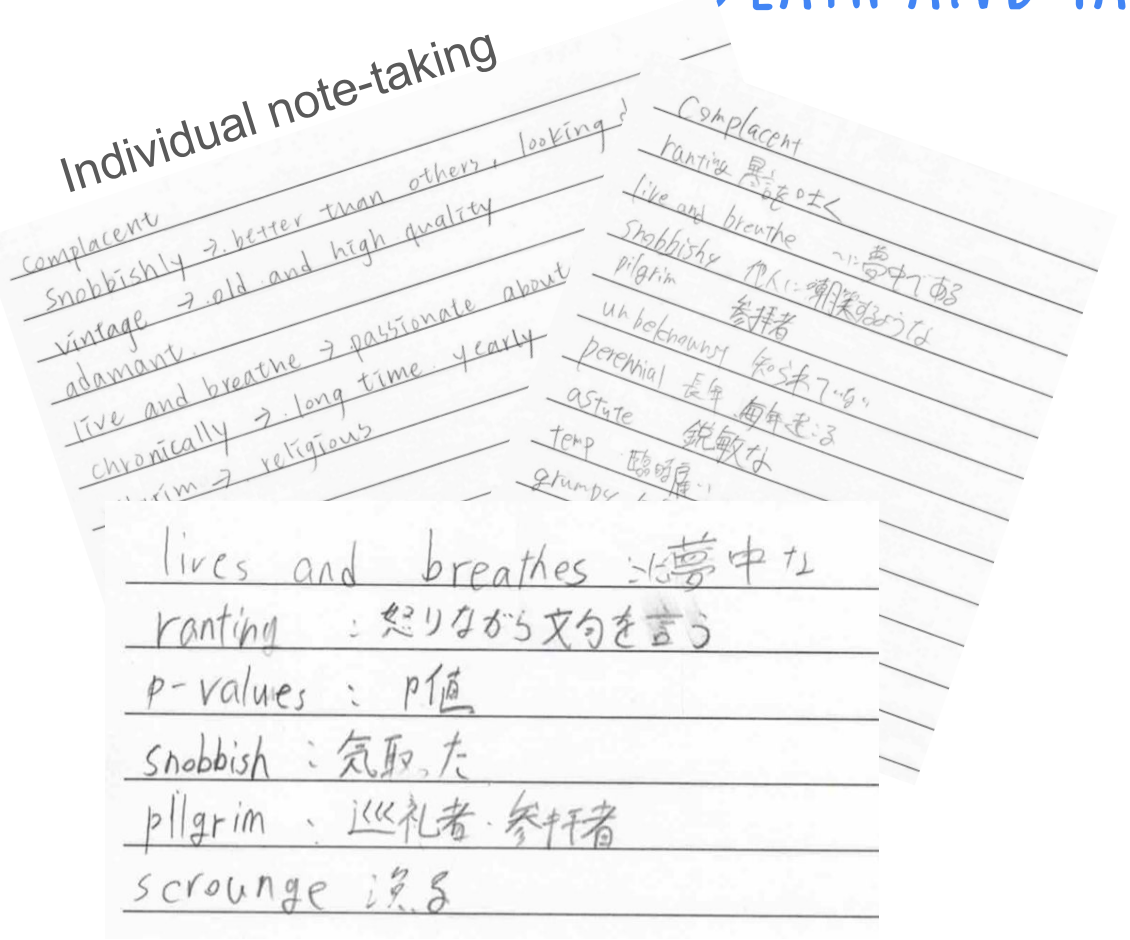
Players take the role of the Grim Reaper, deciding who lives and who dies each day following instructions from Fate, the boss of all Grim Reapers.

- moral dilemmas
- hidden stories
- dark humor and witty and sarcastic tones
- multiple possible endings through their choices.

1. Speaking and argumentative writing
2. Game days-based sessions
3. Instruction: class reading aloud, discussion and voting led by the teacher;
4. Activities: peer gameplay, collective glossary building; writing of decision-making reasoning

Death and Taxes

Individual note-taking



Shared glossary on Moodle

A

acclimate

Definition: to become used to a new place, situation, or climate / 新しい環境・状況・気候に慣れる

Example: The new grim reaper slowly acclimated to the strange office.

alleviate

Definition: to make pain, difficulty, or a problem less serious / 痛み・問題・困難を和らげる、軽減する

Example: The new policy was meant to alleviate the city's financial problems.

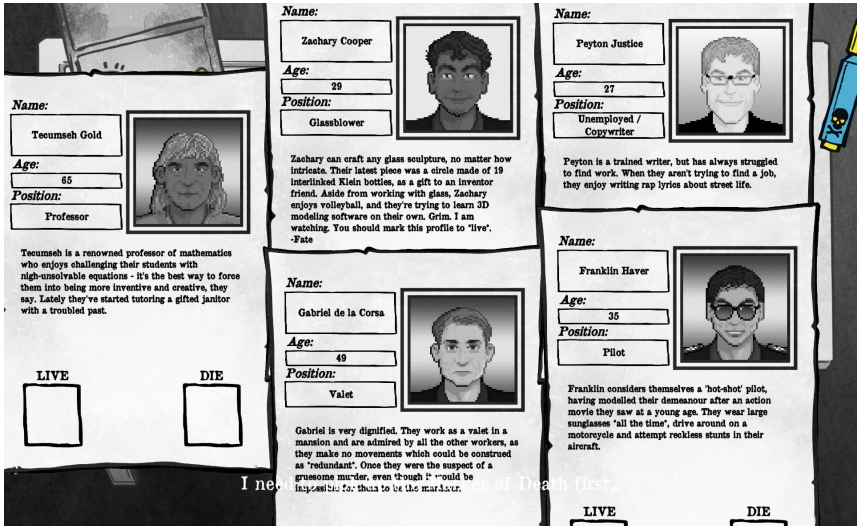
avid fan

Definition: someone who likes or supports something very much / 熱心なファン、大ファン

Example: He is an avid fan of dark comedy games.

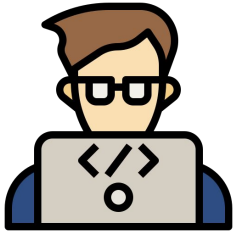
Death and Taxes

Authentic student writing without AI contamination.



> *We decided to kill three people, Peyton Justice, Gabriel de la Corsa, Franklin Haver. First, we chose Peyton because he has not worked, and he makes his lyrics while he is not trying to find a job, which is regarded as low(no) productivity. Second, Gabriel de la Corsa. He was suspect of gruesome murder, but it would be impossible for him. However, we think where there is smoke, there is fire, and so he might be related to that accident. Third, Franklin Haver. He considers himself as a competent pilot. In fact, he often attempts reckless stunts in his aircraft. For these reasons, we decided to kill these people.*

More Playable Serious Games

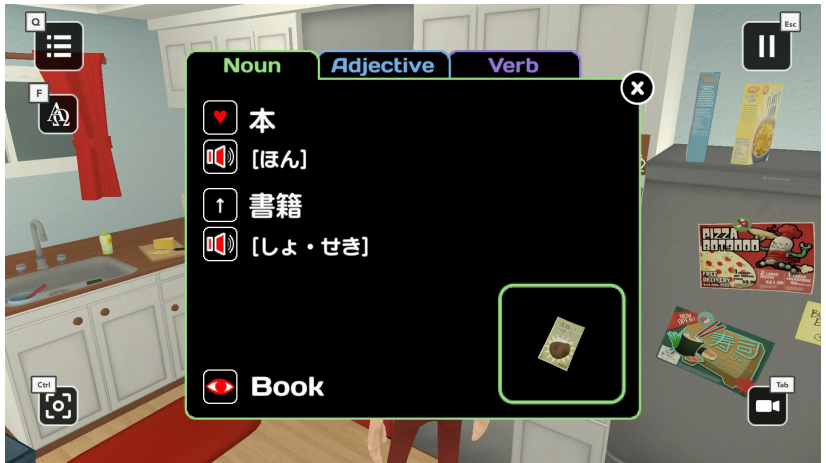


Make serious games more COTS-like and playable
Preset language content, labor and resource-intensive, unscalable.

INFLUENT

Influent

- Limited scenes
- All objects and corresponding vocabulary are preset.
- Unsustainable, unscalable and non-customizable.



The developers are said to have given up on further development…

Duolingo Adventure

3. AI and GBLL

AI in GBLL Classes

AI as a language assistant

AI for creating learning materials

AI in Game Development

Large Language Models enable adaptive, conversational, and scalable language learning experiences while dramatically reducing game development costs (AI-generated visual resources, vibe coding)

- **Scalable content creation:** AI generates dialogue, scenarios, and practice materials, making serious games more replayable and less labor-intensive to design
- **Richer language interaction:** real-time conversation, and contextual response generation within the game world
- **Adaptive support:** AI personalizes difficulty, feedback, and scaffolding to learners' needs during gameplay

AI in COTS Game Development

Many developers are using AI for creating COTS games: 2320 games tagged AI on Steam as of June 22nd, 2026: very few are suitable for language learning.

Whispers from the star

AI in COTS Game Development

Players are calling for AI integration: Custom AI “mods”

AI in Serious Game Development

- Create multi-modal assets such as audios, videos, pictures
- Create learning content
- Still very limited and not so playable compared to COTS games

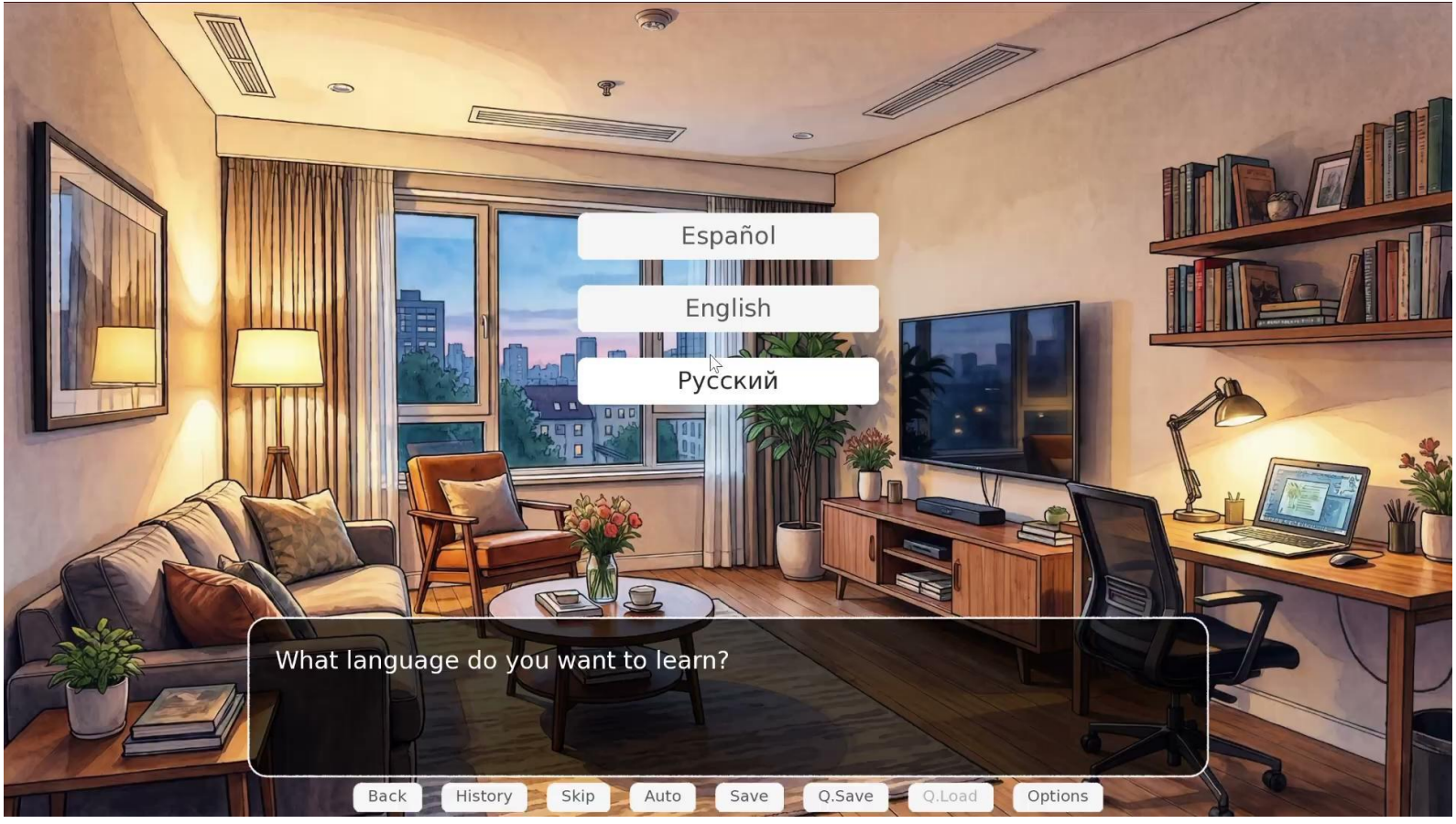
AI for creating game assets

Ask Mom: Find Objects, Learn Words

AI for creating learning content

GenQeust

AI in Serious Game Development



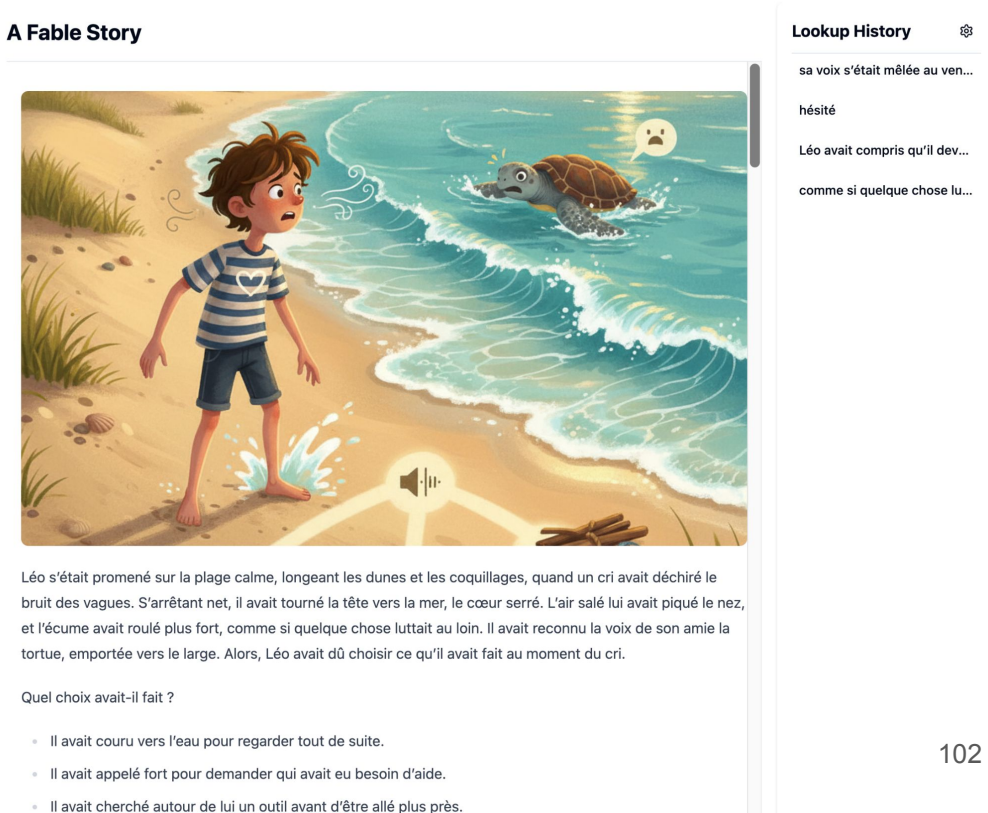
GenQuest

A multi-modal narrative game with player-determined branching stories.

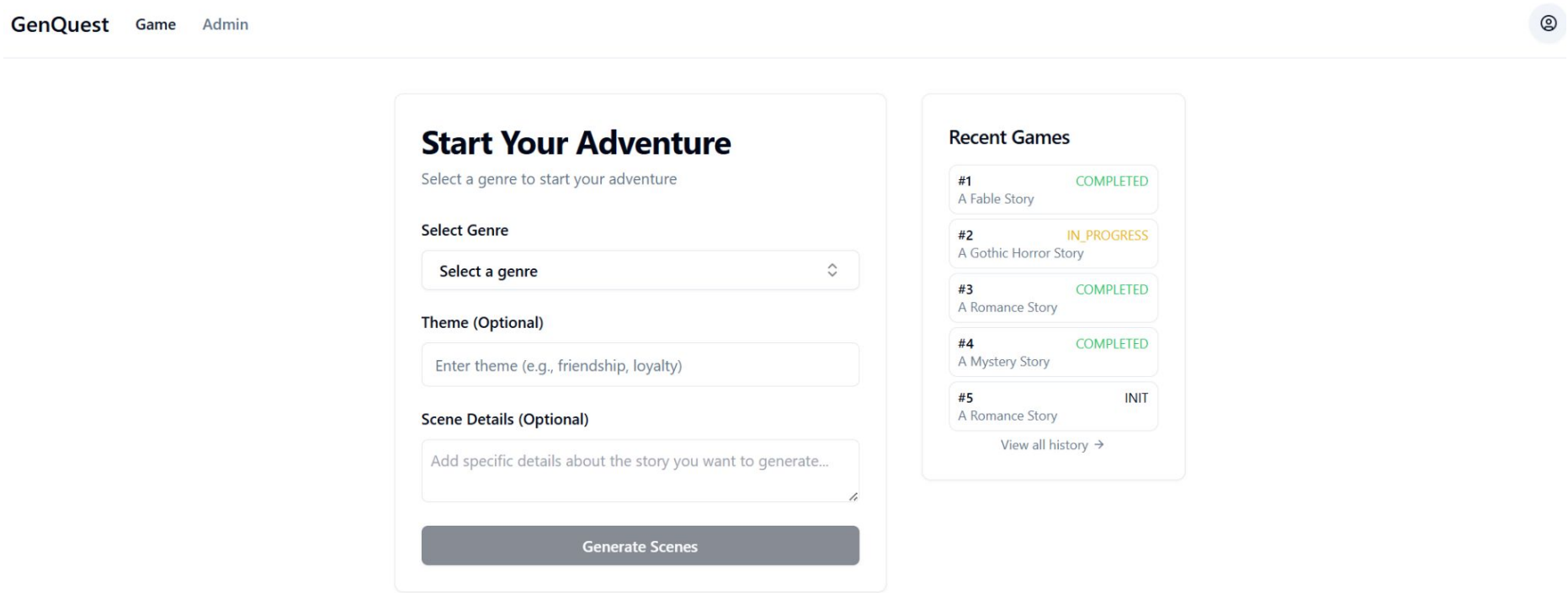
“Choose Your Own Adventure”

Differences:

- On-the-fly generation of story progression and illustrations
- Language support
- Proficiency-aligned stories
- Review quizzes



Demo of GenQuest



AI in Serious Game

LIMITATIONS

Not “game” enough. More like “gamification”:

> The intentional use of game elements for a gameful experience of non-game tasks and contexts (Krath et al., 2021).

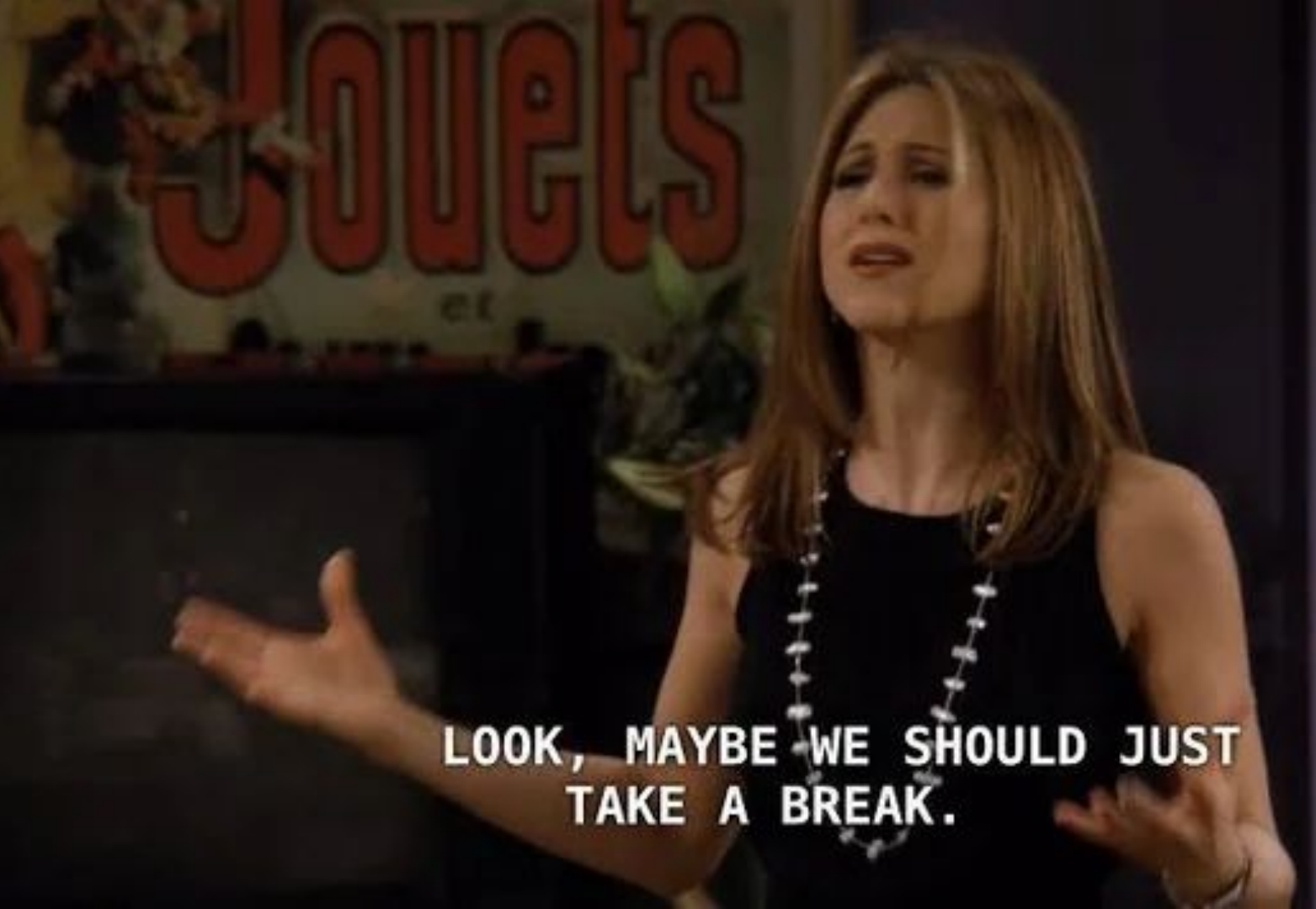
Krath, J., Schürmann, L., & von Korflesch, H. F. O. (2021). Revealing the theoretical basis of gamification: A systematic review and analysis of theory in research on gamification, serious games and game-based learning. *Computers in Human Behavior, 125*, 106963. https://doi.org/10.1016/j.chb.2021.106963

AI in Serious Game

Most recent work is still dominated by prototypes, pilot studies, and short interventions.

- narrative coherence over time: consistent storylines over extended gameplay
- pedagogical reliability: integrating learning goals
- Learner adaptation: language-level control
- Responsible AI: safety, privacy, and bias
- Long-term learning effects: durable learning gains beyond novelty effects

THANK YOU


LOOK, MAYBE WE SHOULD JUST
TAKE A BREAK.

5. Demos

Please feel free to get a demo account from the presenters

GenQuest

atg.judywang.jp

Conversation system

conv.judywang.jp

6. Challenges and Future Directions

What’s Still Hard

- Pedagogical reliability, assessment validity, and language-level control
- Narrative & multimodal coherence over a full learning sequence
- Fairness across non-native speakers, dialects, and low-resource languages
- Safety, privacy, and proving durable learning gains beyond novelty effects

Principles for Responsible Adoption

Be intentional

Automate high-impact stages with purpose; not every step needs it.

Use the right tool

Rules, Statistical, ML, or GenAI — match the technique to the problem.

Keep humans in the loop

AI suggests; teachers and experts decide and stay accountable.

Design for fairness

Build in bias & fairness checks; test continuously across groups.

Monitor always

Language and populations shift — re-validate; “set and forget” is risky.

Design for inclusion

Plurilingual, neurodivergent, and low-resource learners by default.

Automation is not about replacing humans — it is about combining the right technology with human judgment.

Open Discussion

- Where should the human stay firmly in the loop & where are you comfortable letting AI act alone?
- What does AI genuinely change about how people learn vs. what it just makes faster and/or cheaper?
- Should AI in education optimise for measurable outcomes, or for harder-to-measure goals like curiosity and agency?
- Are we building tools that adapt to learners, or quietly training learners to adapt to the tools?
- Who should own and govern the learner data these systems depend on and/or collect?
- What is the single biggest barrier to adopting these tools in your own context?

A few questions to take into the room & we'd love your perspective.

Further Information & Resources

- Tutorial website:

https://aied2026-hacc-tutorial.github.io/

- Slides, materials, and further reading will be available online
- Please reach out with questions or feedback

Thank you!