NLP and Generative AI for Language Learning and Assessment

Synergies between Research and Practice

Zheng Yuan
The University of Sheffield
UK

Mariano Felice
British Council
UK

Yunshi Lan
East China Normal University
China

Qiao Wang
Hosei University
Japan

AIED
26th July 2025

Tutorial Plan (Part 1)

- Introduction (09:00 - 09:20; 20 mins) - Yunshi
- Teaching and Learning (09:20 - 10:20; 60 mins)
  - Tutoring Systems (30 mins) - Yunshi
  - Automatic Item Generation (30 mins) - Mariano
- Q & A (10:20 - 10:30; 10 mins)
- Coffee Break (10:30 - 11:00; 30 mins)

Tutorial Plan (Part 2)

- Assessment (11:00 - 12:00; 60 mins)
    - Grammatical Error Correction (30 mins) - Zheng & Qiao
    - Automated Scoring (30 mins) - Qiao
- Q & A (12:00 - 12:10; 10 mins)
- Break (12:10 - 12:20; 10 mins)
- Ethics (12:20 - 12:40; 20 mins) - Mariano
- Moving forward (12:40 - 13:00; 20 mins) - Zheng & Guest Speaker

Information & Resources

- **Tutorial website:**

  `https://aied2025-lla-tutorial.github.io/`

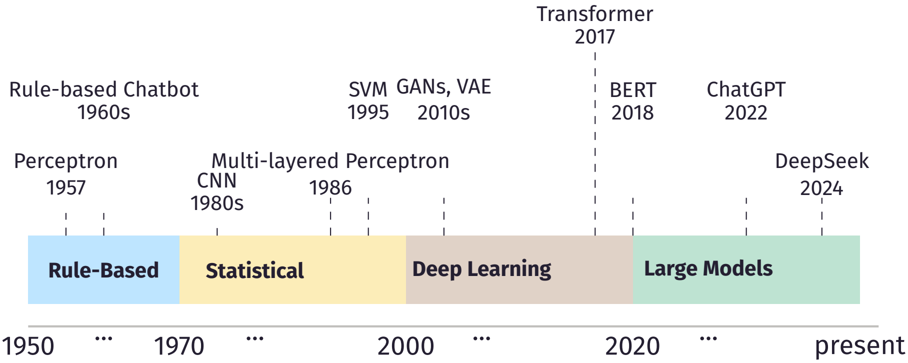
- Slides, materials, and further reading will be available online
- Please reach out with questions or feedback

1. Introduction

Outline

- Brief History of NLP/Generative AI
- NLP/Generative AI and Language Education
- Road Map

Brief history



Scenarios of Language Education

For teachers

- Fast quiz making
- Assess assignments
- Help answer questions out of classes
- Assistant customize education

For students

- Provide self-inspection for exercise
- Recommend learning materials
- Automate concepts noting and explanation

Learning with LLMs

I am an American, please teach me Chinese

Qwen3-235B-A22B 12:12

Great! Welcome, American friend! 😄

I'm happy to teach you **Chinese (Mandarin)** . Let's start from the beginning and learn step by step.

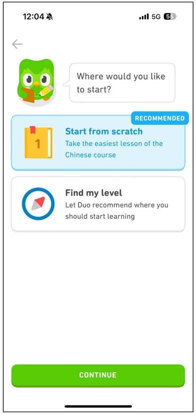
🌟 Lesson 1: Basic Greetings & Introductions

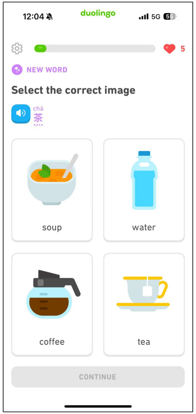
✅ Vocabulary

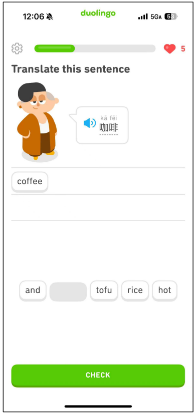
CHINESE	PINYIN	ENGLISH
你好	Nǐ hǎo	Hello
你叫什么名字?	Nǐ jiào shénme míngzì?	What's your name?
我叫...	Wǒ jiào ...	My name is ...
再见	Zàijiàn	Goodbye
谢谢	Xièxie	Thank you
不客气	Bù kèqì	You're welcome

`https://chat.qwen.ai/`

Learning with Duolingo

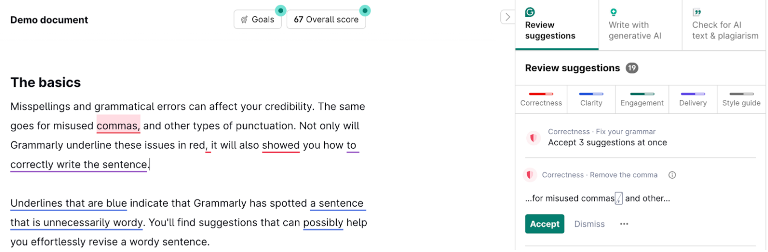






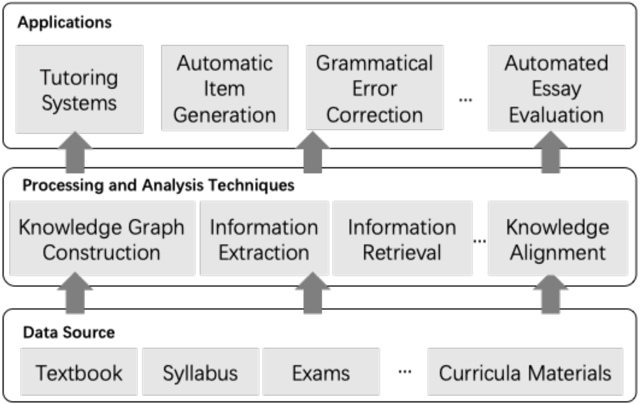
https://www.duolingo.com/

Learning with Grammarly



https://app.grammarly.com/

Road Map



Applications
Tutoring Systems
Automatic Item Generation
Grammatical Error Correction
...
Automated Essay Evaluation
Processing and Analysis Techniques
Knowledge Graph Construction
Information Extraction
Information Retrieval
...
Knowledge Alignment
Data Source
Textbook
Syllabus
Exams
...
Curricula Materials

Potential AI Technologies

- **Teaching and Learning**
  - Tutoring Systems
    - Not only provide tutorials for students but also save time for overloaded teachers by providing feedback on questions.
  - Automatic Item Generation
    - Generating questions as well as the distractors, which helps the teachers create fast quiz.
- **Assessment**
  - Grammatical Error Correction
    - Automatic correction is useful to improve language learning via explicating the correct edits.
  - Automated Essay Evaluation
    - Grading assignments for students automatically.
- **Ethics**

2. Teaching and Learning

2.1. Tutoring systems

Yunshi Lan

Outline

- Introduction
- Approaches
- Data Evaluation
- Challenges and Future Directions

Introduction

What is the difference between “is” and “are” in grammar?

Answer: Both "is" and "are" are present tense forms of the verb "to be". How do you know which one you are supposed to use? Look at the subject:
I am
You ***are***
He/she/it ***is***
We ***are***
They ***are***
To make things clearer for you, here are some sentences with the subject and verb highlighted in boldface type. (I also included an "I am" sentence, even though you didn't specifically ask about that one.)
In my family, ***I am*** *the oldest of four girls.*
Maria and Luis are *married.*
You are *taller than most kids your age.*
You are *supposed to turn in your paper by the end of the week.*
Jason is *Mrs. Smith's oldest son.*
Karen is *late for work almost every Wednesday.*
I think ***the dishwasher is*** *broken again.*
We are *going to Disneyland next month.*
Justin, Sidney, and Kyle are *coming to the party.*

Introduction

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

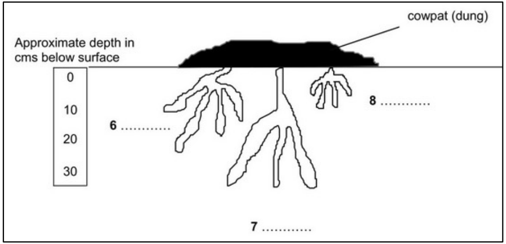
How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

Introduction

Dung beetles work from the inside of the past so they are sheltered from predators such as birds and foxes. Most species burrow into the soil and bury dung in tunnels directly underneath the pats, which are hollowed out from within. Some large species originating from France excavate tunnels to a depth of approximately 30cm below the dung pat. These beetles make sausage-shaped brood chambers along the tunnels. The shallowest tunnels belong to a much smaller Spanish species that buries dung in chambers that hang like fruit from the branches of a pear tree. South African beetles dig narrow tunnels of approximately 20cm below the surface of the past. Some surface-dwelling beetles, including a South African species, cut perfectly-shaped balls from the pat, which are rolled away and attached to the bases of plants.

Label the tunnels on the diagram below using words.



Answer: **6** South African
7 French
8 Spanish

Introduction

Tutoring System (TS) is a task that requires a system to comprehensively understand the question and the (multi-modal) information from the textbook curriculum. We formulate the general loss function of TS as follows for simplicity:

$$L_{\mathrm{TS}} = -\mathbb{E}_{(P,Q,A)\in\mathcal{D}} \log P_\theta(A|Q,P)$$

where P, Q, A are contexts, questions and answers from data collection $\mathcal{D}$, respectively. P may be presented in non-textual modalities.

Data Evaluation

Annotation Challenges:
Answers are various for different types of questions, different evaluation metrics should be conducted.

- Single- and Multi-choice questions
- Reading comprehension
- Open-domain questions
- ...

Data Evaluation

- Accuracy of the retrieved tokens
  - For reading comprehension, employ BLEU and Rouge scores to measure the n-gram overlap between the retrieved tokens and the human annotated answers.
- Overlap to the reference answers
  - For single-choice and multi-choice questions, employ Recall, Precision, F score to measure the overlap between the retrieved answers and the ground truth answers.
- Fluency of the generated answers
  - For open-domain questions, employ Bert Score to measure the fluency of the responses.
- ...

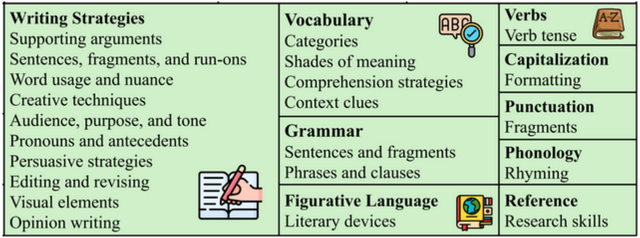
Data Evaluation

Many question answering datasets have been proposed beyond language education and extensive TS systems are developed based on them.

Dataset	Science subject	Source	Context	#Q
TQA△	life/earth/ physical science	grade 6-8 science curricula	image/diagram	26K
GeoSQA△	geography	high school	image/diagram	13K
AI2D△	science	grade 1-6	image/diagram	5K
SCIENCEQA△	natural/social/ language science	grade 1-12 science curriculum	image/diagram/ text	21K
MedQA△	medicine	professional medical board exam	text	40K
MedMCQA△	medicine	simulated exams	text	200K
TheoremQA△	math/physics/ EE/CS	university exam	text	800

Data Evaluation

ScienceQA are large-scale TS datasets with annotated lectures and explanations covering diverse science topics including natural science, social science, and language science.



Approaches

Knowledge-enhanced Tutoring Systems: Some studies[1,2] fine-tune the contextual knowledge in the dataset as language modeling:

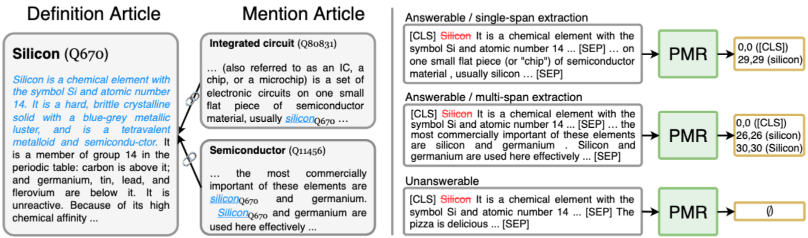
$$L_{\text{PreTrain}} = -\mathbb{E}_{P\in\mathcal{D}} \log P_\theta(P_m|Q, P_{\setminus m}),$$

where P_m, $P_{\setminus m}$ represent the masked tokens and unmasked tokens respectively.

[1] O. Ram et al. (2021). "Few-Shot question answering by pretraining span selection". In: *ACL*

[2] W. Xu et al. (2022). "From clozing to comprehending: Retrofitting pre-trained language model to pre-trained machine reader". In: *NeurIPS*

Knowledge-enhanced Tutoring Systems



Constructing a large volume of general-purpose and high-quality MRC-style training data by using Wikipedia hyperlinks.[3]

[3] W. Xu et al. (2022). “From clozing to comprehending: Retrofitting pre-trained language model to pre-trained machine reader”. In: *NeurIPS*

Knowledge-enhanced Tutoring Systems

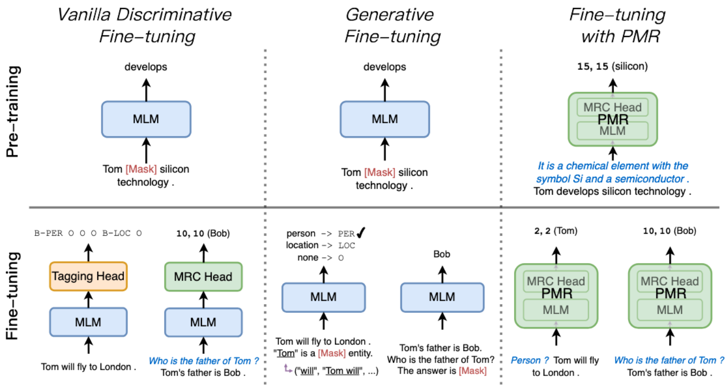


Figure: A Wiki Anchor Extraction task to guide the MRC-style pre-training.

Approaches

Multi-modal Tutoring Systems: To understand the diagrams and tables, graph-based parsing methods are developed to extract the concepts from diagrams[4,5]:

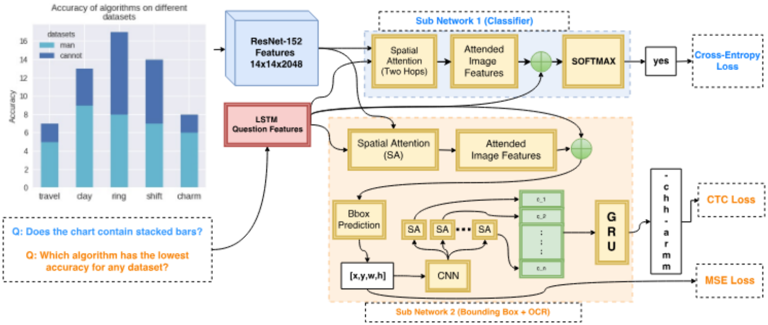
$$\tilde{P} = \mathcal{T}(P).$$

Here, $\mathcal{T}$ is a set of off-the-shelf toolkits for calling. After that, $\tilde{P}$ will be deemed as the contexts of Q.

[4] A. Kembhavi et al. (2016). "A diagram is worth a dozen images". In: *ECCV*

[5] K. Kafle et al. (2018). "Dvqa: Understanding data visualizations via question answering". In: *CVPR*

Multi-modal Tutoring Systems



Applying image encoder to process the contents in a diagram and fusing with text.[6]

[6] K. Kafle et al. (2018). "Dvqa: Understanding data visualizations via question answering". In: *CVPR*

Approaches

LLM-based dialogue systems Recently, researchers leverage Large Language Models as a tutoring system.

- Translate contexts to uni-modality[7]:

$$\tilde{P} = \text{Image2Text}(P); \quad A = \text{LLM}(Q, \tilde{P}).$$

- Multi-modal large language models[8]:

$$A = \text{MLLM}(Q, P).$$

- LLM-based agent as a planner to call a set of tools[9]:

$$A = \text{LLM}(Q, P, \mathcal{T}).$$

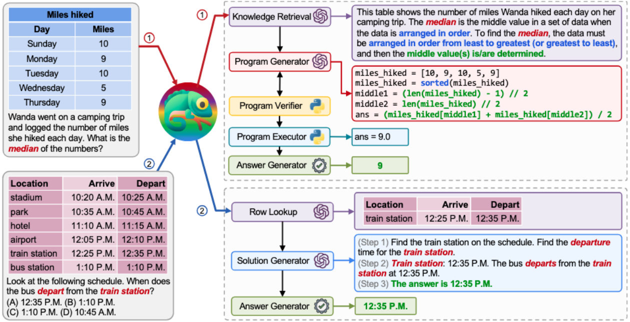
---

[7] P. Lu, S. Mishra, et al. (2022). "Learn to explain: Multimodal reasoning via thought chains for science question answering". In: *NeurIPS*

[8] H. Liu et al. (2023). "Visual Instruction Tuning". In: *NeurIPS*

[9] P. Lu, B. Peng, et al. (2023). "Chameleon: Plug-and-play compositional reasoning with large language models". In: *NeurIPS*.

LLM-BASED DIALOGUE SYSTEMS



An AI system that mitigates these limitations by augmenting LLMs with plug-and-play modules for compositional reasoning.[10]

[10] P. Lu, B. Peng, et al. (2023). "Chameleon: Plug-and-play compositional reasoning with large language models". In: *NeurIPS*

Challenges and Future Directions

- Answers are various for different types of questions, there is a lack of a unified evaluation metric for TS.
- The subject-specific QA data is limited for training.
  - Even though PLMs are frequently utilized as the backbone model, they are short of the subject knowledge.
- The contexts may be presented in diverse formats.
- Zero-shot issues frequently occur.
  - TS has a higher requirement for reasoning under zero-shot scenarios.

2.2. Automatic item generation

Mariano Felice

Outline

- Definition and benefits
- Content creation
- Multiple-Choice Questions (MCQs)
- Open cloze
- C-test
- Item difficulty prediction
- Challenges and future directions

Automatic Item Generation

Automatic/automated item generation (AIG) and automated question generation (AQG) are used synonymously to broadly refer to the process of generating items/questions from various inputs, including models, templates, or schemas.[11]

TL;DR: Item writing automation.

[11] R. Circi et al. (2023). "Automatic item generation: foundations and machine learning-based approaches for assessments". In: *Frontiers in Education* Volume 8 - 2023. ISSN: 2504-284X. DOI: 10.3389/feduc.2023.858273

Key Benefits

- **Scalability:** Generate large numbers of items quickly.
- **Personalisation:** Adapt questions to learner ability.
- **Cost-efficiency:** Save time and resources.
- **Consistency:** Ensure standardised quality and difficulty.
- **Improved test security:** Reduce item exposure by continuously generating fresh questions.

Content creation

Last weekend, Sarah decided to go hiking in the mountains. She hadn't been hiking for several months, so she was excited to spend time in nature again. Before she left, she made sure to pack enough water, snacks, and a map of the trail. The weather forecast had predicted sunshine, but just as she reached the summit, dark clouds began to gather.

Despite the sudden change, Sarah remained calm. She would have turned back earlier if she had known the storm was coming, but now it was safer to continue down the other side of the mountain. Fortunately, the rain didn't start until she was almost back at her car. Soaked but satisfied, she promised herself not to trust the forecast so easily next time.

GPT-BASED GENERATION

Generate multiple item types for the Duolingo English Test.[12]

- Use GPT-3 to generate both:
  - Reading passages (short narrative or informational texts)
  - Comprehension questions (MCQs, cloze, etc.)
- Evaluate passages using readability metrics.
- Apply psychometric filtering (e.g., item discrimination, difficulty).
- Conduct manual vetting for language quality and fairness.

> Write a short informational passage (~80 words) for English learners about honey bees.

[12] Y. Attali et al. (2022). "The interactive reading task: Transformer-based automatic item generation". In: *Frontiers in Artificial Intelligence* 5, p. 903077. DOI: 10.3389/frai.2022.903077

GPT-based generation

Generate high-quality reading texts based on PIRLS-style prompts.[13]

- GPT-3 prompted with structured templates (e.g., genre, topic, etc.).
- Multiple versions generated per prompt.
- Scored using Lexile measures, Coh-Metrix features and CEFR-readability classifiers.
- Human evaluators assess content relevance, coherence and CEFR alignment.

> Generate a narrative text for 10-year-old readers about a child discovering a secret room.

[13] U. Bezirhan et al. (2023). "Automated Reading Passage Generation with OpenAI's Large Language Model". In: *arXiv preprint arXiv:2304.04616*

Multiple-Choice Questions (MCQs)

stem

What can be inferred about Sarah's preparation for the hike?

A) She left in a hurry without checking the weather.
B) She packed well but didn't expect the weather to change.
C) She forgot to bring food and water.
D) She expected it to rain during her hike.

distractors

key

Statistical and Semantic Methods (pre-2018)

- Choose sentence containing target word (with correct sense).
- Generate distractors using external resources (thesaurus, co-occurrence, word embeddings).
- Rank and filter distractors by similarity, collocation strength, word frequency and learner error frequency.

The company will __________ its operations internationally.
(A) extend (B) increase **(C) expand** (D) scale

V. Susanti et al. (2018). "Automatic distractor generation for multiple-choice English vocabulary questions". In: *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. ACL, pp. 231–240

C.-Y. Huang et al. (2018). "Personalized computer-aided question generation for English learners". In: *Educational Technology & Society* 21.4, pp. 248–261

Pretrained Models (2019–2023)

- Use of **pre-trained models** (T5, GPT-2/3, BERT) for MCQ and distractor generation
- Instruction + sentence/passage $\rightarrow$ MCQ stem, key and distractors
- **Distractor generation strategies:**
  - Reuse from corpora (e.g., exams, textbooks)
  - Generate via embeddings, masked LM or translation noise
  - Rank using semantic similarity (e.g. Sentence-BERT)

Example: Transformer-based Distractor Generation[14]

- Fine-tunes GPT-2 on RACE to generate distractors from passage and question.
- Filters results using a BERT QA model to remove weak options.

[14] J. Offerijns et al. (2020). "Better Distractions: Transformer-based Distractor Generation and Multiple Choice Question Filtering". In: *arXiv preprint arXiv:2010.09598*. Fine-tunes GPT-2 on RACE dataset to generate distractors; uses BERT QA model to filter

Joint Distractor Generation (2023–Now)

Use fine-tuned LMs to **jointly generate and discriminate** distractors.[15]

- **Generator stage:** A fine-tuned transformer model generates a set of candidate answers (including the correct key and distractors) from the question stem/context.
- **Discriminator stage:** A classifier distinguishes the correct key from distractors.
- **Selection & clustering:** Candidates are semantically clustered; top representatives from clusters become the final distractors.

Example: *The team **completed** the project ahead of schedule.* → [concluded, delivered, accomplished]

[15] S. Taslimipoor et al. (2024). "Distractor Generation Using Generative and Discriminative Capabilities of Transformer-based Models". In: *Proceedings of the 2024 International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*. ELRA, pp. 5052–5063

Open cloze

gap/blank

Too Good to be True

"Learn a language in 24 hours!" Sounds impressive — but is it realistic? Some courses promise fast results, while others suggest taking your **(1)** __________ and studying over several weeks. Believe it **(2)** __________ not, these timeframes refer to programmes that claim fluency faster than you can say 'Bonjour'.

But not all promises deliver. If you're not careful, you could end up with little **(3)** __________ than a big bill. WhizzLearn Systems had to drop its claim that its method is better than any **(4)** __________ technique, after admitting they should **(5)** __________ made things clearer. So before you sign up, take a closer look — flashy promises might hide more gaps than your language skills. Always read the **(6)** __________ print.

Rule-Based & Linguistic Generation (Pre-2020)

Use grammatical templates, collocation data, and patterns (often for prepositions and articles) to generate gap-fill (open cloze) items automatically.[16]

- Identify target grammar point (e.g. prepositions).
- Select sentences from learner and native corpora.
- Insert gaps via rule-based patterns or frequency-based collocation rankings.

Example: She arrived ________ London before midnight. (key: *in*)

[16] J. D. Lee and S. Seneff (2007). "Automatic generation of cloze items for prepositions". In: *Interspeech*, pp. 2173–2176

Masked Language Models (2020–2022)

Employ masked language models (e.g. BERT) and entropy/predictability metrics to select optimal gap positions that balance difficulty and exposure.[17]

- Score each token in context using a LM.
- Select tokens with high entropy or low predictability.
- Filter gaps for CEFR-level or pedagogical relevance.

Example: Tom [MASK] the decision. (key: *made*)

[17] S. Matsumori et al. (2022). "Mask and Cloze: Automatic Open Cloze Question Generation using a Masked Language Model". In: *RepL4NLP Workshop*

Transformers & LLMs (2022–2024)

Generate and discriminate:[18]

- **Generator:** Use a transformer (e.g. T5) to produce candidate gaps.
- **Discriminator:** Rank candidates based on model-based scoring or estimated difficulty.
- **Post-processing:** Remove ambiguous gaps; re-rank for learner appropriateness.

Personalized Cloze Test Generation with LLMs:[19] adaptive test creation tuned to learner proficiency, using LLM prompting and dynamic difficulty

---

[18] M. Felice, S. Taslimipoor, et al. (2022). "Constructing Open Cloze Tests Using Generation and Discrimination Capabilities of Transformers". In: *Findings of ACL*, pp. 1263–1273

[19] C.-H. Shen et al. (Sept. 2024). "Personalized Cloze Test Generation with Large Language Models: Streamlining MCQ Development and Enhancing Adaptive Learning". In: *Proceedings of the 17th International Natural Language Generation Conference*. Ed. by S. Mahamood et al. Tokyo, Japan: Association for Computational Linguistics, pp. 314–319

C-TEST

The Renaissance and its Impact on Europe

The Renaissance was a cultural movement that profoundly affected E u r o _ _ _ _ intellectual life in t h _ early modern period. Beginning in I t a _ _, and spreading to t h _ rest of Europe by the 1 6 _ _ century, its influence affected l i t e r _ _ _ _ _, philosophy, art, politics, science, r e l i _ _ _ _, and other aspects o _ intellectual inquiry.

Template-Based & Psychometric Approaches

- Select CEFR-aligned passages from learner corpora or curated sources.
- Apply a “damage” rule, e.g. delete second half of every second word.
- Predict item difficulty using regression/predictive models trained on reading complexity or word frequency.
- Rank and filter passages accordingly.

> Many students are trying to improve their vocabulary. →
> Many stud_ _ _ _ are tr_ _ _ _ to impr_ _ _ their voca_ _ _ _ _ _ .

J.-U. Lee, E. Schwan, et al. (July 2019). “Manipulating the Difficulty of C-Tests”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. Korhonen et al. Florence, Italy: Association for Computational Linguistics,

Item Difficulty Prediction

Automatically predicting the difficulty level of test items (e.g., multiple choice, reading comprehension) based on the textual content alone, without human pretesting.

Benefits:

- Avoids expensive pretesting (e.g., IRT calibration).
- Enables adaptive test design and scalable item generation.
- Supports filtering low-quality or overly hard/easy items.

Typical Setup

Input: question stem and answer options.
Output: difficulty score (e.g., IRT parameters) or class (easy/medium/hard).

Key Approaches

Feature-Based[20]

- Linguistic features (e.g., CEFR level, sentence length, word frequency).
- Model: linear regression, random forest.

Transformer-Based[21]

- Fine-tune BERT on labeled items (difficulty from operational data).
- More accurate for complex item types.

Check out a comprehensive survey.[22]

[20] L. Benedetto et al. (2020). “R2DE: a NLP approach to estimating IRT parameters of newly generated questions”. In: *LAK 2020*

[21] A. D. McCarthy et al. (2021). “Jump-Starting Item Parameters for Adaptive Language Tests”. In: *EMNLP 2021*, pp. 842–857

[22] L. Benedetto (2023). *A Quantitative Study of NLP Approaches to Question Difficulty Estimation*.

BEA 2024 Shared Task[23]

Goal: "... to advance the state-of-the-art in item parameter prediction"

Given an item's text and metadata...

- predict the item's difficulty (Track 1)
- predict the item's response time (Track 2)

Data: 667 retired MCQs from the United States Medical Licensing Examination (USMLE) [466 training, 201 test]. Additional data allowed. Difficulty values between 0 (easiest) and 1.4 (most difficult).

[23] V. Yaneva et al. (June 2024). "Findings from the First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple-Choice Questions". In: *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*. Ed. by E. Kochmar et al. Mexico City, Mexico: Association for Computational Linguistics, pp. 470–482

Results

Rank	Team Name	Run	RMSE
1	EduTec	electra	0.299
2	UPN-ICC	run1	0.303
3	EduTec	roberta	0.304
4	ITEC	RandomForest	0.305
5	BC	ENSEMBLE	0.305
6	Scalar	Predictions	0.305
7	BC	FEAT	0.305
8	BC	ROBERTA	0.306
9	UnibucLLM	run1	0.308
10	EDU	Run3	0.308
11	EDU	Run1	0.308
12	ITEC	Ensemble	0.308
13	UNED	run3	0.308
14	Rishikesh	1	0.310
15	Iran-Canada	run2	0.311
16	**Baseline**	**DummyRegressor**	**0.311**
⋮	⋮	⋮	⋮
43	ITEC	BERT-ClinicalQA	0.393

- 17 teams, 43 submissions (up to 3 per team).
- Coefficient of Variation = 6.38%
- Relative Improvement (top vs baseline) = 3.9%

Top teams

- **1st Place: EduTec (electra)**
  - ELECTRA-based encoder
  - *Scalar mixing* (= weighted mean of all hidden layers)
  - Two-layer classification head with *rational activation* (= a learnable activation function)
  - Multitask learning setup
- **2nd Place: UPN-ICC (run1)**
  - GPT-3.5-generated answers and justifications
  - Multiple prompt setups (full text, some text removed, varying temperature, etc.)
  - Ridge regression on 40 engineered features from the answers
- **3rd Place: ITEC (RandomForest)**
  - Linguistic features + Bio_ClinicalBERT embeddings
  - Random Forest regression model

Duolingo's S2A3[24]

Goal: Score new test items in real time, without pretesting.

1. **SPICE** estimates item difficulty using a Bayesian engine fusing NLP features and learner responses.
2. **Soft Scoring (S2)** reduces the impact of uncertain items on test scores.
3. **Adaptive Admin (A3)** selects when and to whom new items are shown to optimize calibration.
4. Item parameters are updated continuously as more data arrives.

100k items calibrated in under 5 hours using live test data.

[24] S. Nydick et al. (2025). *What if new test items didn't need to wait?* https://blog.englishtest.duolingo.com/what-if-new-test-items-didnt-need-to-wait/. Duolingo English Test Blog

Challenges and Future Directions

- **Sparse & noisy labels:** Difficulty annotations are limited, inconsistent, and domain-specific.
- **Model limitations:** Surface features and LLM confidence don't reliably reflect cognitive challenge. LLMs cannot simulate student responses reliably.
- **Generalization across item types is hard:** NLP models struggle to transfer across formats (e.g., MCQs, cloze, prompts).
- **Explainable, hybrid models:** Combine LLMs, cognitive design features and psychometric estimates.

Q&A

BREAK

3. Assessment

3.1. Grammatical Error Correction

Zheng Yuan and Qiao Wang

Outline

- Introduction
- Approaches
- Data and evaluation
- Challenges and future directions

What is Grammatical Error Correction

Consider the following text written by an ESL learner. Can you detect the errors and suggest corrections?

- *Nowadays, there are many people that are learning foreign language. Is it worth to learn a foreign language? [...] people who know how to speak a foreign language have more opportunities to get a job in important companies [...] It could allow you to communicate with people, know different cultures ...*

What is Grammatical Error Correction

One possible correction:

- *Nowadays, there are many people ~~that~~who are learning foreign ~~language~~languages. Is it worth ~~to learn~~learning a foreign language? [...] people who know how to speak a foreign language have more opportunities to get a job in ~~important~~big companies [...] It could allow you to communicate with people, get to know different cultures ...*

Grammatical Error Correction (GEC) is the task of automatically detecting and correcting errors in learner text

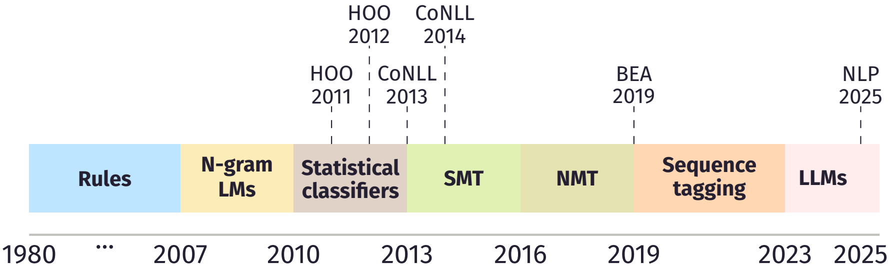
Challenges

- Alternative corrections are possible
  - *In conclude* $\rightarrow$ *In conclusion* OR *To conclude*
- Errors may interact
  - *Book is good*
    - Option 1: $\rightarrow$ *The Book is good* $\rightarrow$ *The book is good*
    - Option 2: $\rightarrow$ *A Book is good* $\rightarrow$ *A book is good*
    - Option 3: $\rightarrow$ *Books is good* $\rightarrow$ *Books are good*

Challenges

- Some error types are harder to correct than others
  - Closed-class errors (e.g., articles, prepositions)
    - *in home* → *at home*
  - Open-class errors (e.g., content words)
    - *look at TV* → *watch TV*
- Error distributions differ significantly among users and domains
  - Learners from different L1 backgrounds make different types of errors
    - E.g., missing articles are more common among speakers of article-less languages
  - Domain-specific writing shows distinct error patterns
    - E.g., scientific papers vs. social media posts

Timeline



- Paradigm shifts approximately every 3 years
- International competitions have significantly accelerated progress

Approaches: Rule-based methods

- Use hand-coded rules; e.g.,
  - *informations* $\rightarrow$ *information*
  - *in the other hand* $\rightarrow$ *on the other hand*
- Advantages
  - 👍 Precise and easy to customise
  - 👍 Interpretable and explainable
  - 👍 Do not require annotated data
- Disadvantages
  - 👎 Rules can become complex and hard to maintain
  - 👎 Rule order matters and maintenance is complicated
  - 👎 Ineffective for certain error types
  - 👎 Requires language-specific expertise

Approaches: N-gram language models

- Certain words are more probable in context than others
- Use this to detect and correct improbable sequences
- Example: *I often work in home.*
    1. Train an N-gram language model on native text
    2. Generate a confusion set: {in, at, from, on, ...}
    3. Score candidate sentences

I often work in home.	284.1275
I often work at home.	98.49942
I often work from home.	55.42596
I often work on home.	315.6587

Approaches: N-gram language models

- Advantages
  - 👍 Require only large native text corpora (e.g., Wikipedia)
  - 👍 Can detect a range of error types, including some semantic errors
- Disadvantages
  - 👎 Probability is not grammaticality (e.g., *I is the ninth letter of the alphabet.*)
  - 👎 Struggle with rare words (e.g., *paraklausithyron*)
  - 👎 Confusion set generation can be difficult (e.g., *I ate a big* ______ .)

Approaches: Statistical classifiers

- Example: Predict the correct form of every verb
  - *They were eat ice-cream when I arrive.*
- Learn to predict correct forms from features
- Train a model on labelled data
- Model learns feature importance and outputs corrected form labels
- Logistic Regression, Decision Trees, Support Vector Machines (SVMs), etc

Approaches: Statistical classifiers

- Advantages
  - 👍 More flexible than rule-based approaches
  - 👍 Can work with only native data (though annotated data helps)
- Disadvantages
  - 👎 Feature engineering is complex
  - 👎 Perform best with small confusion sets (e.g., function words)
  - 👎 Typically target single error types
  - 👎 Classifier order affects results

Approaches: Machine translation

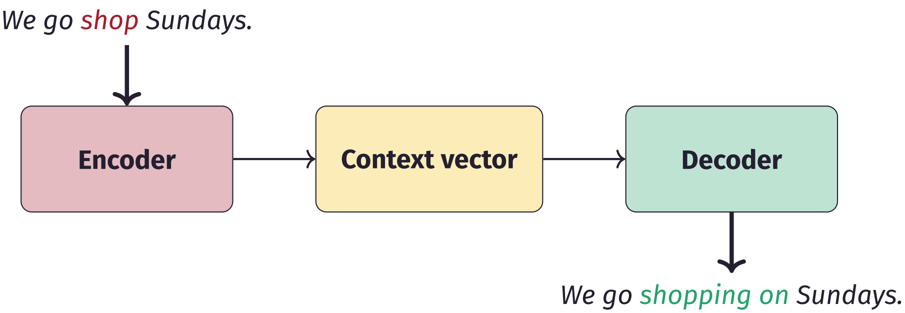
- Translate from “bad” English to “good” English
- Similar principle for both Statistical Machine Translation (SMT) and Neural Machine Translation (NMT)
- Requires parallel data
- Train a model to produce corrected output given erroneous input

Approaches: SMT

There is hundred of ways that an idea can originate from .

There are hundreds of ways in which an idea can originate ∅ .

APPROACHES: NMT[25]



[25] Z. Yuan et al. (n.d.). "Grammatical error correction using neural machine translation". In: *Proceedings of NAACL 2016*

APPROACHES: MACHINE TRANSLATION

- Advantages
  - 👍 Corrects all error types simultaneously, including interacting errors
  - 👍 No feature engineering or expert knowledge required
  - 👍 Single end-to-end model
  - 👍 State-of-the-art performance (i.e., Transformer-based NMT)
- Disadvantages
  - 👎 Requires large amounts of parallel training data
  - 👎 Computationally expensive and time-consuming to train
  - 👎 Often uninterpretable
  - 👎 Difficult to customise

Approaches: Sequence tagging

- Predict a tag for every word

They	*likes*	*to*	*eat*	*the*	*ice-cream*	*.*
KEEP	REPLACE (like)	KEEP	KEEP	DELETE	KEEP	KEEP

- Essentially a word-level classifier
- Similar principle to Part-of-Speech tagging
  - Requires labelled data
  - Fine-tune various pretrained neural language models
  - Choice of label set is an open question
    - E.g., binary (correct/incorrect) vs. detailed labels ($>$5,000)

Approaches: Sequence tagging

- Advantages
  - 👍 Handles most error types (depending on label set)
  - 👍 Single end-to-end model
  - 👍 More efficient than NMT
  - 👍 Somewhat interpretable
  - 👍 State-of-the-art performance
- Disadvantages
  - 👎 Requires large amounts of parallel training data
  - 👎 Careful design of label set needed
  - 👎 May struggle with multi-token or interacting errors

Approaches: LLM-based approaches

- Prompting (zero-shot), e.g.,[26]

> Make minimal changes to the following text such that it is grammatically correct. {text}

[26] C. Davis et al. (n.d.). "Prompting open-source and commercial language models for grammatical error correction of English learner text". In: *Findings of ACL 2024*

Approaches: LLM-based approaches

- Prompting (zero-shot), e.g.,[27]

> You are a grammatical error correction tool. Your task is to correct the grammaticality and spelling in the input sentence. Make the smallest possible change in order to make the sentence grammatically correct. Change as few words as possible. Do not rephrase parts of the sentence that are already grammatical. Do not change the meaning of the sentence by adding or removing information. If the sentence is already grammatically correct, you should output the original sentence without changing anything.
> Input sentence: {text}
> Output sentence:

[27] C. Davis et al. (n.d.). "Prompting open-source and commercial language models for grammatical error correction of English learner text". In: *Findings of ACL 2024*

Approaches: LLM-based approaches

- In-Context Learning (ICL) (few-shot or Chain-of-Thought), e.g.,[28]

> You are an English language teacher. A student has sent you the following text.
> {text}
> Provide a grammatical correction for the text, making only necessary changes. Do not provide any additional comments or explanations. If the input text is already correct, return it unchanged.
> Examples:
> {input 1} $\rightarrow$ {output 1}
> {input 2} $\rightarrow$ {output 2}
> {input 3} $\rightarrow$ {output 3}

[28] C. Davis et al. (n.d.). "Prompting open-source and commercial language models for grammatical error correction of English learner text". In: *Findings of ACL 2024*

Approaches: LLM-based approaches

- Supervised Fine-Tuning (SFT)
  - Update model weights using labelled data
  - Achieves state-of-the-art performance

Data and evaluation: Precision and Recall

GEC evaluation centers on a fundamental trade-off between precision and recall.

- **Precision:** Make just the right correction; i.e., avoid making unnecessary corrections.
- **Recall:** Make enough corrections, i.e., avoiding neglecting errors.
- Metrics are often weighted to favor precision: making an incorrect change is typically considered worse than missing a potential one.

DATA AND EVALUATION: ERRANT

ERRANT (ERRor ANnotation Toolkit[29]): mostly widely used tool in GEC evaluation.

- Aligns an original sentence with its corrected version.
- Extracts the specific changes, or ”edits”.
- Classifies edits into specific categories (e.g., ‘R:VERB:TENSE‘, ‘M:PUNCT‘).

Example:

- She go to school everyday. $\rightarrow$ She goes to school every day.
- **ERRANT Output:**
  S She go to school everyday .
  A 1 2|||R:VERB:SVA|||goes|||REQUIRED|||-NONE-|||0
  A 4 5|||R:ORTH|||every day|||REQUIRED|||-NONE-|||0

[29] C. Bryant, M. Felice, and T. Briscoe (July 2017). “Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 793–805. DOI: 10.18653/v1/P17-1074

Reference-Based Metrics

These metrics compare a system's output to a "gold standard" correction.

- **ERRANT** $F_{0.5}$:
  - Weighs precision **twice as much** as recall to penalize incorrect edits:

$$\mathit{ERRANTF}_{0.5} = \frac{1.25 \times \text{precision} \times \text{recall}}{0.25 \times \text{precision} + \text{recall}}$$

  - Widely used in shared tasks (e.g., BEA 2019[30]).
  - **Limitations:**
    - **Single-reference evaluation**: Only one gold reference is used. Valid alternative corrections not in the reference are penalized.
    - **Ignores meaning preservation**: A grammatically correct edit that changes the intended meaning can still receive full credit.

[30] C. Bryant, M. Felice, Ø. E. Andersen, et al. (June 2019). "The BEA-2019 Shared Task on Grammatical Error Correction". In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Florence, Italy: Association for Computational Linguistics, pp. 52–75

Reference-Based Metrics

GLEU[31]:

- Balances **precision and recall of n-grams** by punishing:
  - **Undercorrection**: failing to produce correct n-grams present in the reference.
  - **Overcorrection**: producing n-grams that are not present in any reference.
- Typically computed as:

$$\text{GLEU} = \min\left(\text{precision}, \text{recall}\right)$$

  for each n-gram level, averaged across n = 1 to 4.
- Applicable to multiple reference corrections.

[31] C. Napoles et al. (July 2015). "Ground Truth for Grammatical Error Correction Metrics". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*. Beijing, China: Association for Computational Linguistics, pp. 588–593. DOI: 10.3115/v1/P15-2097

Reference-Based Metrics

PT-ERRANT (Preserved-Token ERRANT[32]):

- Extension of ERRANT metrics to evaluate **semantic fidelity**.
- Measures how well a system **preserves the meaning** of tokens not meant to be edited.
- Especially useful in avoiding unnecessary or harmful changes.

Key Steps:

- Identify tokens in the original sentence that are not part of any gold edits: **preserved tokens**.
- Computes the precision and recall of **preservation**

[32] P. Gong et al. (2022). "Revisiting Grammatical Error Correction Evaluation and Beyond". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 6891–6902

Reference-Free Human Evaluation

Human evaluation remains the best method for accurately assessing correction quality because people can account for the diversity of valid grammatical constructions.

- Studies have shown that system scores can improve significantly when humans decide the valid correction[33].
- Despite its accuracy, comprehensive human evaluation is often impractical at scale.

[33] Q. Wang et al. (May 2024). “Assessing the Efficacy of Grammar Error Correction: A Human Evaluation Approach in the Japanese Context”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by N. Calzolari et al. Torino, Italy: ELRA and ICCL, pp. 1666–1672

AI-Human Collaboration

A hybrid ”LLM-as-a-Judge” framework can be used to bridge the gap between accuracy and scale[34].

- **Example with reference-based evaluation:**
  - Two LLMs as primary judges to assess corrections.
  - Classifying corrections that differ from the gold standard into three categories: model preferred, gold preferred, or equally valid.
  - Human evaluators to resolve disagreement.

[34] M. Kobayashi et al. (2024). “Large Language Models Are State-of-the-Art Evaluator for Grammatical Error Correction”. In: *EMNLP 2024: Findings*

Challenges and Future Directions

- **Interactive and Pedagogically-Informed Feedback**
  - Provide pedagogically meaningful feedback (e.g., error explanations, revision options).
  - Move beyond static correction to multi-turn, dialog-based revision support.
  - Encourage self-correction and metalinguistic awareness via guided prompts.
- **Multilingual and Low-Resource GEC**
  - Extend GEC systems to non-English languages (e.g., Japanese, Arabic, Chinese).
  - Develop benchmarks and datasets for low-resource settings.
- **Integration with AWE Systems**
  - Embed GEC into holistic writing evaluation platforms.
  - Support longitudinal tracking of learner revisions and progress.

3.2. Automated Essay Evaluation

Qiao Wang

Outline

- **Introduction:** AWE and its importance
- **Evolution** From rule-based to deep learning to LLMs
- **Datasets** Key datasets in building AWE systems
- **System Evaluation** Benchmark-based validation, human review, and user studies
- **Showcase Tools:** Walk through major systems: Write & Improve, Criterion, WrAFT.
- **Future Directions**

Introduction

- What is AWE?
- The Importance of AWE

What is AWE

AWE (automated writing evaluation) refers to the process of evaluating and scoring written prose via computer programs[35].

- **Two main tasks:**
  - *Scoring* – a holistic or analytic score based on rubric criteria.
  - *Feedback* – grammar, mechanics, and genre-specific aspects such as content, organization, coherence, etc.
- **Applications:**
  - High-stakes exams (e.g., TOEFL, GRE)
  - Classroom writing practice and revision
  - Online learning platforms and self-access tools

[35] M. D. Shermis et al. (2024). "Introduction to automated essay evaluation". In: *The Routledge international handbook of automated essay evaluation*. Routledge, pp. 3–22

Why is AWE Important?

- **Teacher perspective:** Reduces teacher workload for routine scoring and corrections, esp. in repetitive work such as :
  - grading fact-based writing, e.g., summaries or IELTS Academic Writing Task 1
  - correcting grammar and mechanical errors
- **Learner perspective:** Promotes learner autonomy and reduces educational disparity.
- **Institutional perspective:** Enables large-scale, consistent evaluation across diverse learners.
  - Reduces human biases resulting from human fatigue and/or halo effect.

Historical Background and Evolution

- Early rule-based systems
- Statistical & ML-based systems
- Deep learning-based systems
- LLM-powered systems

Era 1: Rule-based & Feature Engineering

Early systems relied on hand-crafted features to predict scores.

- **Core Idea:** Extract surface-level and syntactic properties of the text that correlate with writing quality.
  - *Fluency*: Word/sentence length, spelling/grammar errors.
  - *Lexical Sophistication*: Type-token ratio (TTR), use of academic word lists.
  - *Syntactic Complexity*: Parse tree depth, clause-level metrics.
- **Models:** Typically linear regression on these features.
- **Pros:** Interpretable, efficient.
- **Cons:** Lacked understanding of content or organization; Could be "gamed" by increasing essay length or stuffing sophisticated but nonsensical words.

Examples of Rule-Based Systems

- **PEG (Project Essay Grade)** – Ellis Page, 1966[36]
  First AES system: regression on hand-engineered features to predict holistic scores (e.g., word count, sentence length and lexical sophistication).
- **Writer's Workbench** - Bell Labs, 1980s[37]
  Grammar and style checker: passive voice, sentence fragments, spelling errors; rudimentary feedback.
- **Early versions of e-rater** - ETS, 1990s–2000s[38]
  Deployed in operational tests (e.g., GMAT, TOEFL).

[36] E. B. Page (1966). "The Imminence of... Grading Essays by Computer". In: *Phi Delta Kappan* 47.5. Introduced Project Essay Grade (PEG), first AES system, pp. 238–243

[37] N. H. MacDonald et al. (1982). "The Writer's Workbench: Computer Aids for Text Analysis". In: *IEEE Transactions on Communications* COM-30.1. Early Bell Labs grammar and style checker, pp. 105–110

[38] Educational Testing Service (ETS) (1999). *Computer Analysis of Essays: The e-rater® Automated Essay Scoring System*. Tech. rep. First deployment of early e-rater in GMAT and TOEFL AWA scoring. ETS

Era 2: Deep Learning

The focus shifted from hand-crafted features to learned representations.

- **Core Idea:** Use neural networks to automatically learn relevant features from text[39].
  - Supervised learning: essays (feature vectors) → human-assigned scores as labels.
  - Essays are represented as sequences of word embeddings (e.g., Word2Vec, GloVe) [40].

[39] D. Alikaniotis et al. (2016). "Automatic Text Scoring Using Neural Networks". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 715–725

[40] T. Firoozi et al. (2023). "The Effect of Fine-Tuned Word Embedding Techniques on the Accuracy of Automated Essay Scoring Systems Using Neural Networks". In: *Journal of Applied Testing Technology* 23, pp. 21–29

Era 2: Deep Learning

- **Models:**
  - *Recurrent Neural Networks (LSTMs):* Process text sequentially to capture coherence and long-range dependencies[41].
  - *Convolutional Neural Networks (CNNs):* Identify local patterns, e.g., N-grams and phrases relevant to quality[42].
  - *Attention Mechanisms:* Allow the model to focus on the most important parts of the text for scoring[43].
- **Pros:** Better capture of semantic meaning, less feature engineering required.
- **Cons:** Requires large labeled datasets for each prompt; ”black box” nature makes interpretation difficult.

[41] K. Taghipour et al. (2016). “Neural Automated Essay Scoring and Coherence Modeling for Cross-Prompt Evaluation”. In: *Proceedings of EMNLP*, pp. 1065–1071

[42] Alikaniotis et al. 2016

[43] F. Dong et al. (2017). “Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring”. In: *Proceedings of CoNLL*, pp. 153–162

Era 3: Large Language Models (LLMs)

Pre-trained LLMs such as the GPT series and LLaMA brought a new paradigm of transfer learning to AWE.

- **Approach:**
  - *Few/Zero-Shot Prompting:* Instructing a general-purpose LLM (e.g., GPT-4) with a rubric to score an essay with or without exemplar essays.
  - *Fine-tuning:* Fine-tune, or further train, a pre-trained LLM on a smaller set of essays.
- **Pros:** State-of-the-art performance, strong grasp of context and meaning, capable of generating high-quality, human-like feedback.
- **Cons:** High computational cost, data privacy issues, potential for generating ”hallucinated” or biased feedback.

Key Datasets for AWE Research

High-quality, publicly available datasets are crucial for building and benchmarking AWE systems.
Ideal datasets: A large number of essays with reliable scores and feedback. Reality: essays with scores only; severe shortage of datasets with human feedback.

- **ASAP (Automated Student Assessment Prize)**[44]
  - The *de facto* benchmark for AWE research, released in a 2012 Kaggle competition.
  - Contains 13,000 essays across 8 different prompts (argumentative, narrative, source-based).
  - Each essay is double-scored by human raters.

[44] The Hewlett Foundation (2012). *Automated Student Assessment Prize (ASAP)*. Kaggle competition. Released student essays scored by human raters; used for advancing AES research

Key Datasets for AWE Research

- **TOEFL11**[45]
  - A corpus of 12,100 essays written by non-native English speakers for the TOEFL test.
  - The original scores from 0-5 points with 0.5 increments are turned into three bands: High, Medium and Low.
- **Proprietary Datasets**
  - ETS provides proprietary datasets consisting test-taker writing from TOEFL and GMAT exams.
  - Limited number of essays (480 samples from TOEFL independent writing)
  - Applications can be sent to the external data request center at ETS.

[45] D. Blanchard et al. (Dec. 2013). "TOEFL11: A Corpus of Non-Native English". In: *ETS Research Report Series* 2013.2. Describes a corpus of 12,100 TOEFL iBT essays for NLI and AES research, pp. i–15. DOI: 10.1002/j.2333-8504.2013.tb02331.x

System Validation Methods

- Quantitative metrics
- Human expert evaluation
- User experiments

Evaluation 1: Benchmark-based Validation

This approach uses quantitative metrics to measure how closely machine scores align with human rater scores on a common dataset.

- **Quadratic Weighted Kappa (κ)**
  - Assessing inter-rater agreement on an ordinal scale.
  - Correcting for agreement by chance and penalizes large disagreements (e.g., a score of 1 vs. 4 is penalized more than 1 vs. 2).
- **Pearson Correlation (r)**
  - Measures the strength and direction of the *linear relationship* between machine and human scores.
  - Not sensitive to the absolute score values (suitable for scores on different scales).

Evaluation 1: Benchmark-based Validation

- **Root Mean Squared Error (RMSE)**
  - Measures the average magnitude of the error between machine scores and human scores.
  - An intuitive measure of the difference between machine and human scores.
- **Percentage Agreement**
  - An intuitive measure of how often machine and human scores match.
  - *Exact Agreement*: The percentage of essays where the machine score is identical to the human score.
  - *Adjacent Agreement*: The percentage of essays where the machine score is within an acceptable discrepancy from the human score (e.g., 1 point).

Evaluation 2: Human Review

Human review is used to qualitatively assess the generated feedback.

- **Methodology:**
  - Experts (writing instructors) review system output against criteria such as:
    - **Precision:** Is the feedback necessary? (e.g., not flagging correct grammar as an error)
    - **Recall:** Are there any missing feedback? (e.g., failing to identify a grammar error)
    - **Effectiveness:** Is the suggestion for revision effectively addresses the identified issue?
- **Importance:** Crucial for moving beyond scoring to create tools that support learning through actionable feedback.

Evaluation 3: User Studies

The ultimate test: Does the AWE tool lead to writing proficiency gains?

- **Goal:** Measure the impact of the AWE system on student writing behavior and performance in a real-world setting.
- **Methodology: Controlled Experiments**
  - *Treatment Group:* Uses the AWE tool to write and revise an essay.
  - *Control Group:* Uses a standard word processor or receives no feedback.
  - Compare the quality of final drafts between groups.
- **Data to Collect:**
  - **Learning Gains:** Pre-test/post-test improvement in writing scores.
  - **User Perceptions:** Surveys and interviews on perceived usefulness, trust, and satisfaction.
  - **System Data:** usage time, analysis of revision behavior (do students accept/ignore feedback?).

Existing Tools

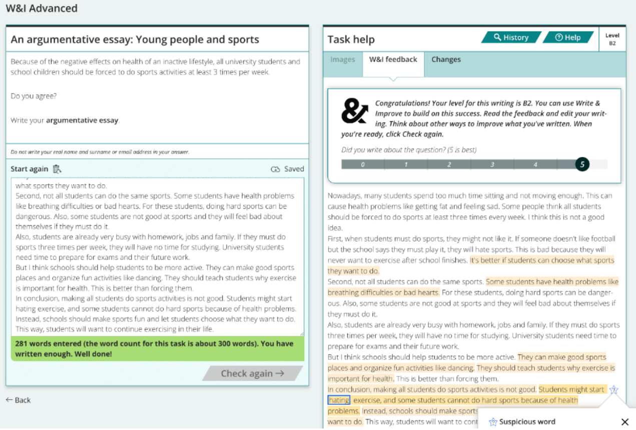
- Write & Improve
- Criterion
- WrAFT

Showcase: Write & Improve (Cambridge)[46]

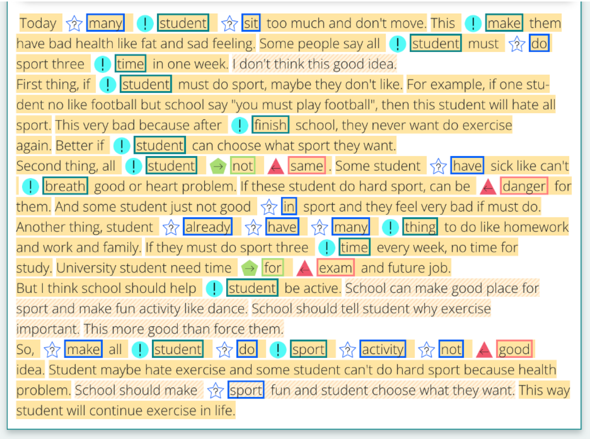
- **Focus:** Formative feedback to support practice and revision.
- **Key Features:**
  - Aligns scores to the Common European Framework of Reference (CEFR).
  - Highlights potential error spans in the text.
  - Encourages multiple revision attempts to improve the score.
  - Task-based: Users choose from dozens of prompts at different levels.

[46] https://writeandimprove.com/

Write & Improve Screenshots



Write & Improve Screenshots



Today many student sit too much and don't move. This make them have bad health like fat and sad feeling. Some people say all student must do sport three time in one week. I don't think this good idea.

First thing, if student must do sport, maybe they don't like. For example, if one stu-dent no like football but school say "you must play football", then this student will hate all sport. This very bad because after finish school, they never want do exercise again. Better if student can choose what sport they want.

Second thing, all student not same. Some student have sick like can't breath good or heart problem. If these student do hard sport, can be danger for them. And some student just not good in sport and they feel very bad if must do.

Another thing, student already have many thing to do like homework and work and family. If they must do sport three time every week, no time for study. University student need time for exam and future job.

But I think school should help student be active. School can make good place for sport and make fun activity like dance. School should tell student why exercise important. This more good than force them.

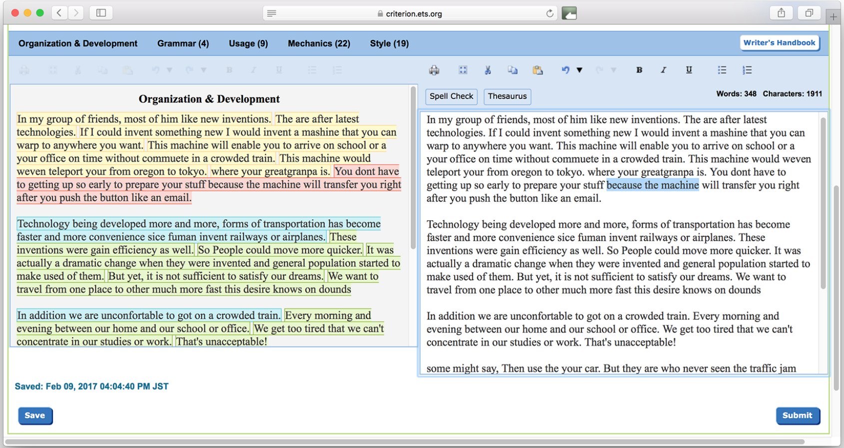
So, make all student do sport activity not good idea. Student maybe hate exercise and some student can't do hard sport because health problem. School should make sport fun and student choose what they want. This way student will continue exercise in life.

Showcase: Criterion (ETS)[47]

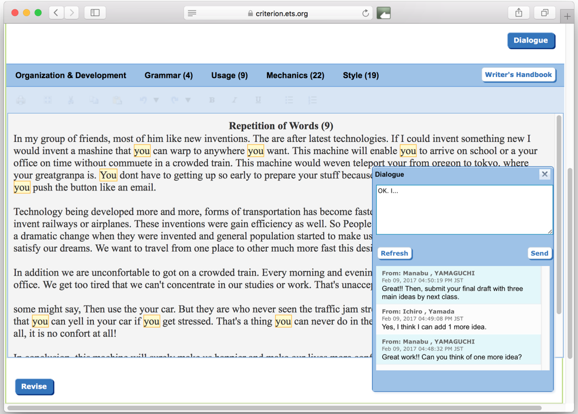
- **Focus:** Summative and diagnostic feedback for instructors and students.
- **Technology:** Powered by the *e-rater* scoring engine.
- **Key Features:**
  - Provides a holistic score and detailed diagnostic feedback on traits like:
    - Organization & Development
    - Word Choice & Style
    - Grammar, Usage, Mechanics

[47] https://www.ets.org/criterion.html

Criterion Screenshots



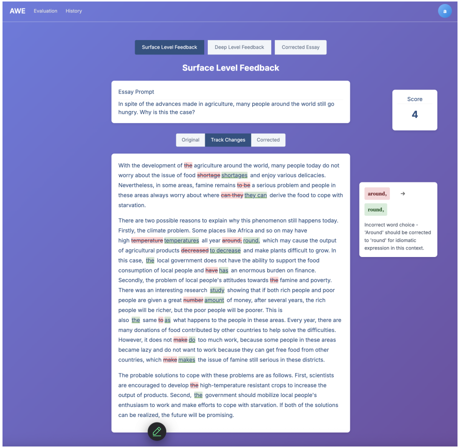
Criterion Screenshots



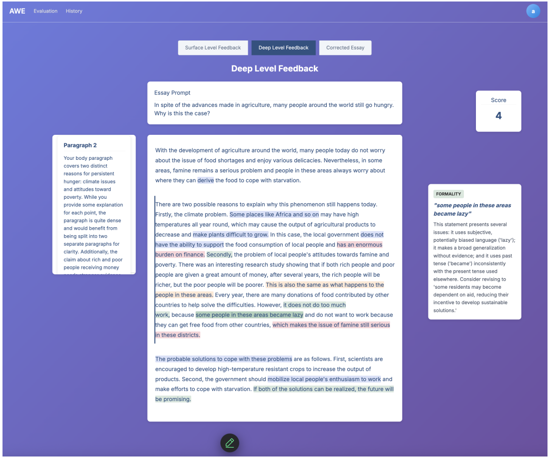
Showcase: WrAFT

- **Focus:** Holistic scoring and detailed formative feedback for argumentative essays.
- **Technology:** Modular LLM-powered architecture, separating tasks:
  - Scoring module: Fine-tuned GPT-4o (on a proprietary dataset of 480 TOEFL independent writing tasks.)
  - Surface-level feedback (grammar and mechanics): Direct prompting with GPT-4o
  - Deep-level feedback (argumentation organization, coherence, etc.): Few-shot prompting with Claude 3.7
- **Key Features**
  - In-line edits in surface-level feedback
  - Anchored comments in deep-level feedback

WRAFT SCREENSHOTS



WrAFT Screenshots



Future Directions

- **Genre Expansion**
  - Move beyond prompt-based essays to cover summaries, source-based writing, and narratives.
  - Tailor scoring rubrics and feedback to genre-specific features.
- **Alignment with High-Stakes Assessments**
  - Benchmark AWE systems on real-world exam datasets (e.g., IELTS, TOEFL, GRE).
  - Provide valid, reliable and fair test preparation strategies.
- **Integration of Other Modules**
  - Support pre-writing stages such as AI-facilitated brainstorming or peer discussion.
  - Use learner analytics to monitor individual growth over time.
  - Provide personalized feedback based on past performance and common errors.

Q&A

BREAK

4. Ethics

Mariano Felice

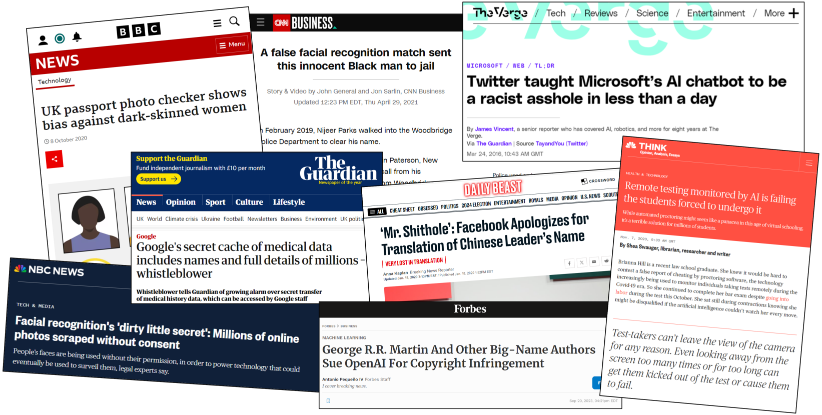
Outline

- The importance of Ethics in AI
- Ethics by design
- Legislation
- Ethical AI and assessment organisations
- Recommendations

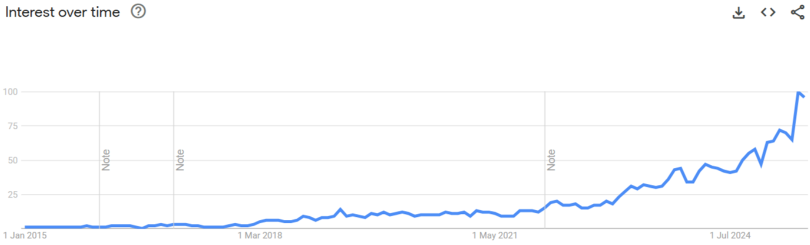
The importance of Ethics in AI

- AI systems can introduce bias, lack transparency, and raise privacy issues.
- These concerns affect fairness and trust in educational outcomes.

The importance of Ethics in AI



THE IMPORTANCE OF ETHICS IN AI



Ethics of Artificial Intelligence (2015-2025)

Google Trends – web search

Ethics by Design

The aim of Ethics by Design is to make people think about and address potential ethics concerns, while they are developing a system.[48]

[48] European Commission (Nov. 2021). *Ethics By Design and Ethics of Use Approaches for Artificial Intelligence.* Version 1.0. Version 1.0, published 25 November 2021. European Commission, Directorate-General for Research and Innovation

Ethics by Design

1. **Respect for Human Agency**
   AI systems should preserve human autonomy, dignity, and freedom. They must empower users in their decision-making rather than undermine it.
2. **Privacy, Personal Data Protection & Data Governance**
   Systems must incorporate privacy-by-design principles, ensure responsible data governance, and protect personal data throughout the AI lifecycle.
3. **Fairness**
   AI systems must avoid bias and discrimination, delivering equitable outcomes and treating all individuals and groups impartially.

Ethics by Design

4. **Individual, Social & Environmental Well-being**
   AI should contribute positively to individuals, communities, and the environment, promoting well-being and sustainability.
5. **Transparency**
   Systems should be explainable and intelligible, allowing users to understand how decisions are made and enabling traceability.
6. **Accountability & Oversight**
   Ethical responsibilities must be clearly defined, supported by governance structures and auditability to ensure responsible development and deployment.

Legislation

- **EU AI Act**[49]: Classifies educational AI as high-risk, requiring transparency, safety and ongoing oversight.
- **UNESCO Guidelines**[50]: Promote human rights, inclusiveness and ethical AI use worldwide.
- **OECD Recommendations**[51]: Stress trustworthy, accountable and robust AI across members.
- **EU Guidelines for Educators**[52]: Offer practical advice for ethical AI integration in classrooms.

---

[49] European Parliament and Council (July 2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts.* Official Journal of the European Union, L 2024/1689, 12.7.2024

[50] UNESCO (Nov. 2021). *Recommendation on the Ethics of Artificial Intelligence.*

[51] *OECD Principles on Artificial Intelligence* (2019). Tech. rep. OECD

[52] European Commission (2022). *Ethics Guidelines for Trustworthy AI in Education.* Tech. rep. Practical advice for ethical AI use in classrooms. European Commission

EU AI Act: Implications for Education

- In force since 1st August 2024, phased into 2026.
- Aims to ensure AI systems are safe, ethical, and respect fundamental rights.
- Risk categories:
  - **Unacceptable** (banned): Social scoring, emotion recognition in schools, predictive policing, subliminal manipulation.
  - **High-risk** (strict compliance): Automated grading, admissions, proctoring, learner profiling—requires transparency, oversight, and risk controls.
  - **Limited-risk** (transparency obligations only): Chatbots, virtual tutors must clearly disclose AI use.
  - **Minimal/no risk** (no restrictions): Spell-checkers, basic tutoring apps, productivity tools with negligible impact.

National AI policies

- Policies aim to balance innovation with ethical and social safeguards.
- Public consultation and stakeholder input are becoming more common.

United States	Sector-specific guidance and voluntary standards.
China	Strict rules on data governance and algorithm transparency.
Canada	Risk-based AI and Data Act targeting high-impact systems.
Japan	Ethical guidelines promoting human-centric, responsible AI.

Ethical AI and assessment organisations

- Major testing organizations are defining responsible AI standards to uphold fairness, transparency, validity and accountability.
- These principles guide the design, deployment, and oversight of their AI-powered language assessment solutions.

Duolingo English Test (DET)[53]

- Defines a set of responsible AI standards, combining expert review, bias mitigation, fairness evaluation and security practices.
- Invites public feedback on their standards, encouraging transparency and dialogue.
- This year, Duolingo declared itself an **“AI-first”** company, sparking public debate.

[53] J. Burstein et al. (2024). “Responsible AI for Test Equity and Quality: The Duolingo English Test as a Case Study”. In: *arXiv preprint arXiv:2409.07476*

ETS[54]

- Emphasises fairness, bias reduction, privacy, accountability and test integrity.
- Emphasises principles-driven policies, stakeholder governance and continuous ethical oversight.

[54] Educational Testing Service (ETS) (2024). *Highlights: Responsible Use of AI in Assessment – ETS Principles AI.* Tech. rep. ETS

Cambridge University Press & Assessment[55]

- AI can improve or harm assessment quality — depending on ethical use.
- Stresses the importance of **human-centred** AI for accuracy, access, and security.
- Highlighted the need for greater **AI literacy** among educators and test developers.
- Supports transparency and multi-stakeholder dialogue in global education contexts.

[55] Cambridge University Press & Assessment (May 2025). *'Ethical AI' essential to future of assessment—presentation at NAFSA 2025.*

British Council[56]

- Advocates a **human-centred**, **learner-first** AI approach.
- Work is built on four pillars:
  - AI literacy
  - Expertise
  - Responsible use
  - Human leadership and oversight
- Ensure AI systems support assessment goals — not replace them.

[56] M. Felice, R. Spiby, et al. (Mar. 2025). *Human-centred AI: lessons for English learning and assessment.* Tech. rep. British Council

Recommendations

1. Follow established ethical principles.
2. Treat ethics assessment as an ongoing process.
3. Stay focused on purpose and pedagogy.
4. Build AI literacy to empower stakeholders and enable critical evaluation.
5. Prioritize humans over technology.
6. Ensure solutions are inclusive, accessible, and diverse.
7. Foster collaboration across stakeholders.

5. Moving forward

Zheng Yuan

Adaptive NLP Techniques for Education

- From one-size-fits-all to **personalised learning**
- Dynamically adapt content and feedback to individual learner profiles (e.g., proficiency, learning style, learning aims)
- Enhance learner engagement and accelerate progress
- Examples
  - Adaptive dialogue systems that adjust difficulty
  - Personalised feedback generation

Interpretable Educational NLP

- **Transparency** is crucial in educational contexts
- Need for **explainable** predictions in scoring, feedback, and content generation
- Supports fair, accountable, and pedagogically sound use of NLP
- Benefits
  - Greater trust and acceptance among educators and learners
  - Improved debugging and model refinement

Human-AI Co-creation

- Position AI as a **collaborator**, not just a tool
- Examples
  - Co-writing essays and creative stories
  - Interactive language learning companions
- Challenges
  - Avoiding over-reliance
  - Preserving learner agency and creativity

Multilingual and Cross-cultural Educational NLP

- Move beyond English-centric systems
- Address challenges in low-resource and underrepresented languages
- Create opportunities for **inclusive, equitable** education worldwide
- Support cross-cultural communication and global learning

Integration with Human-Centred Pedagogy

- Align NLP innovations with human values and pedagogical goals
- Promote collaboration, critical thinking, and creativity
- Example in assessment
  - AI can now score written responses using detailed rubrics with high consistency
  - But AI cannot decide
    - What or how to measure
    - How to interpret scores
    - Ethical boundaries
  - These decisions remain the responsibility of educators, researchers, and policymakers

Educational AI for Collective Intelligence

- *"A framework for Design Based Research into AI to support education for Collective Intelligence"*
    - Professor Li Yuan

Reflections & Open Challenges

We warmly invite all participants to share their reflections and contribute to the discussion

- What are the key open challenges to address?
- What should be our next steps as a research and practitioner community?

Let's shape the future of learning together!

Poll: Which Area Excites You Most?

Please select the areas you believe hold the most promise or need the most attention

- Adaptive NLP Techniques for Education
- Interpretable Educational NLP
- Human-AI Co-creation
- Multilingual and Cross-cultural Educational NLP
- Integration with Human-Centred Pedagogy
- **Other:** *(please type your own idea in the chat or poll box)*

Further Information & Resources

- **Tutorial website:**
  `https://aied2025-lla-tutorial.github.io/`
- Slides, materials, and further reading will be available online
- Please reach out with questions or feedback

Thank you!

Zheng Yuan
The University of Sheffield, UK
zheng.yuan1@sheffield.ac.uk

Mariano Felice
British Council, UK
mariano.felice@britishcouncil.org

Yunshi Lan
East China Normal University, China
yslan@dase.ecnu.edu.cn

Qiao Wang
Hosei University, Japan
judy.wang@hosei.ac.jp