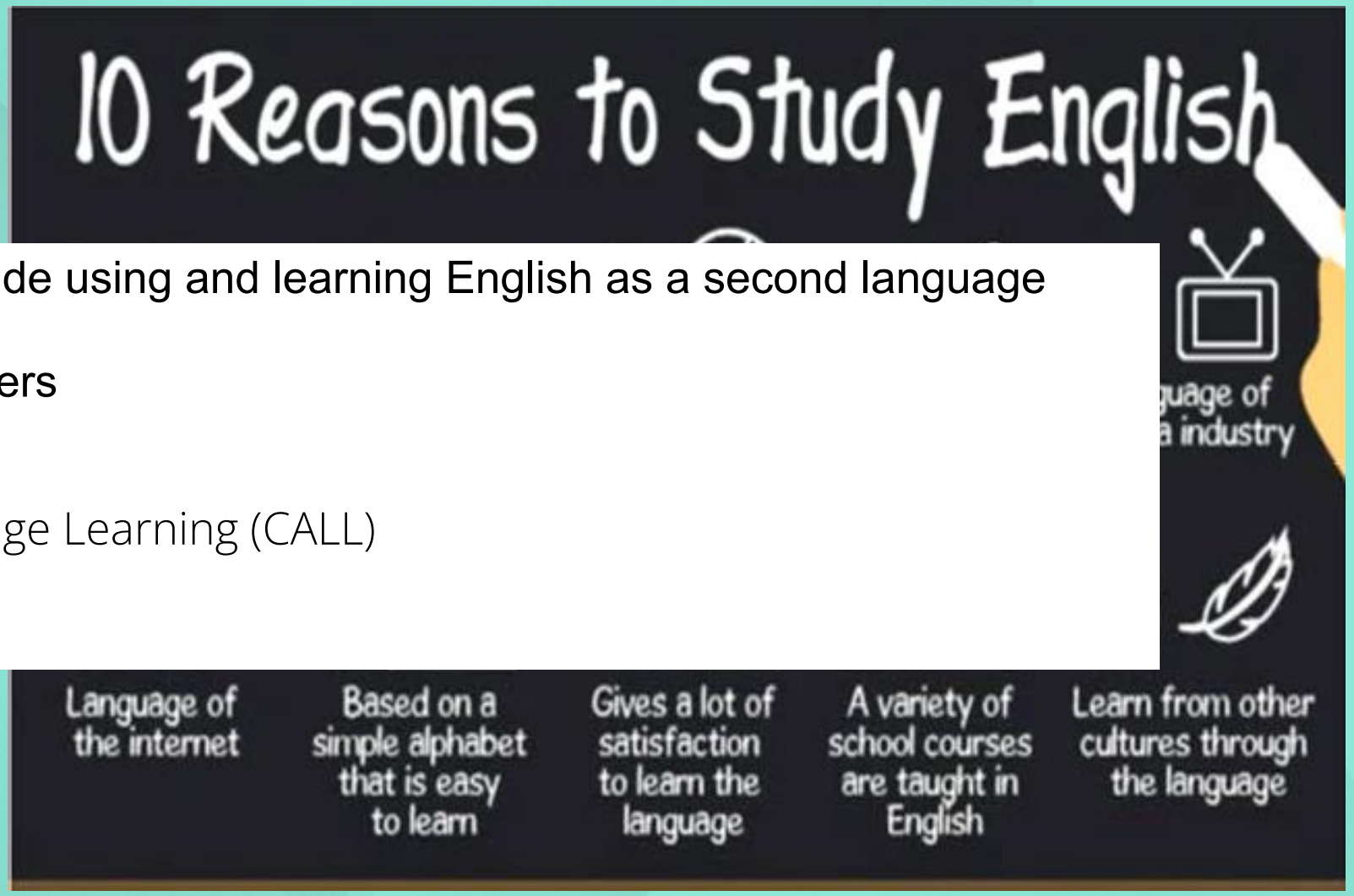
UNIVERSITY OF CAMBRIDGE
Department of Engineering

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

Natural Language-based Assessment of L2 Oral Proficiency using LLMs

Stefano Bannò
ALTA Institute & Machine Intelligence Lab
Cambridge University Engineering Department

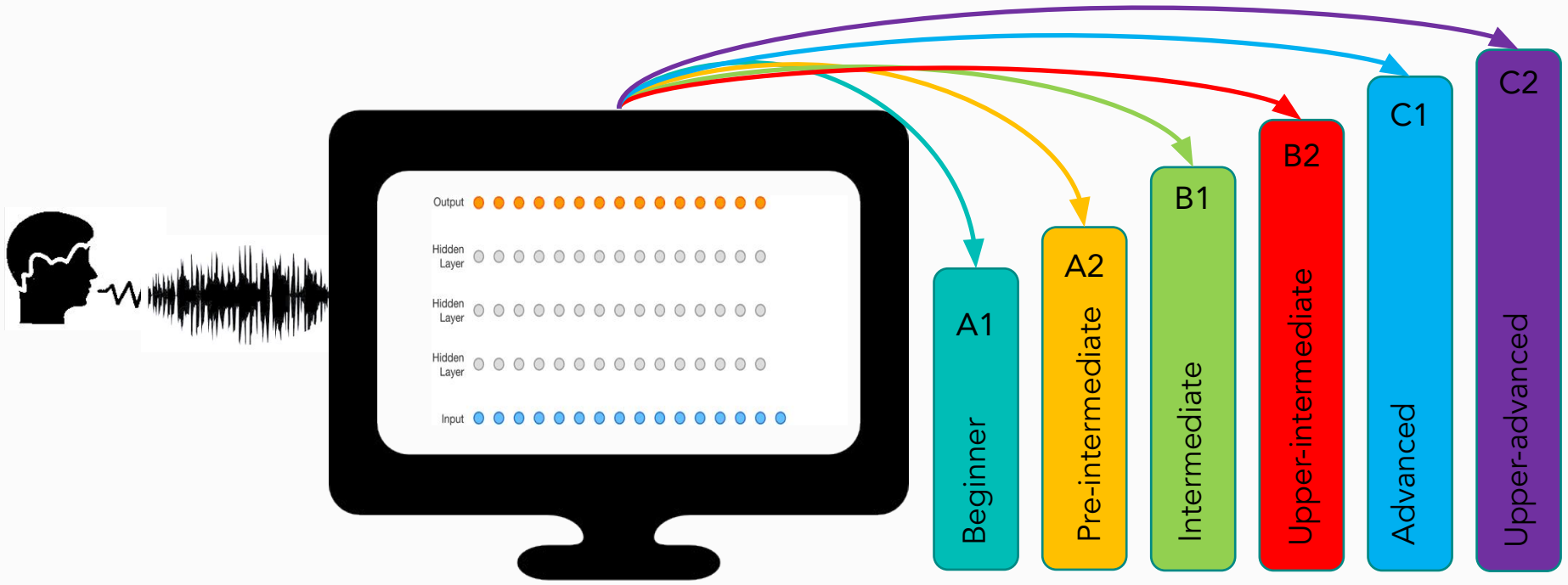
- Almost 2 billions people worldwide using and learning English as a second language
- Not enough teachers or examiners
  - Automated assessment
  - Computer-Assisted Language Learning (CALL)



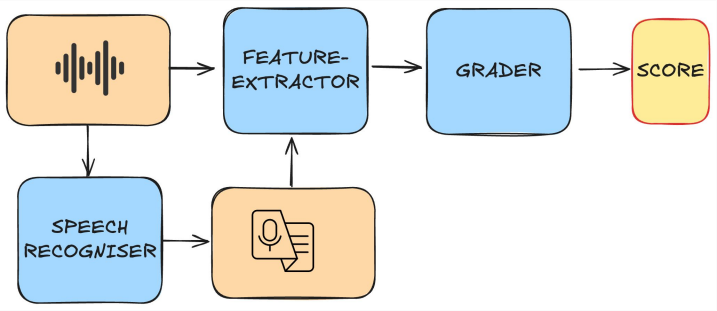
Overview

- **Natural Language-based Assessment**
  - Automated Language Assessment
  - Natural Language-Based Assessment
  - Experimental Setup
  - Experimental Results
  - Conclusions and Future Work

Automated Assessment Systems

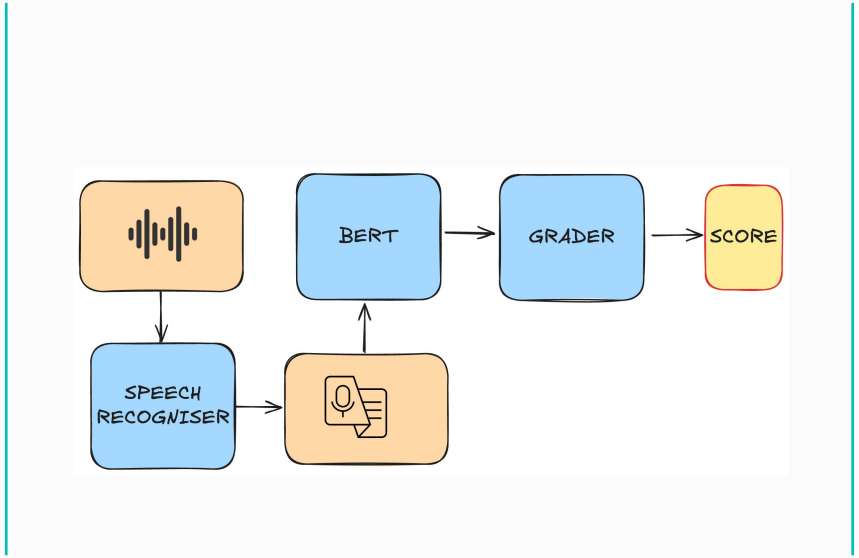


Automated speaking assessment systems



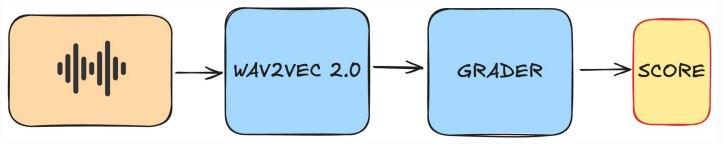
- PROS:
  - Rely on **interpretable features**;
- CONS:
  - Require **specifically designed and annotated training data**, which is an expensive process;
  - Efficacy of features relies heavily on their underlying assumptions and they risk discarding potentially salient information about proficiency;
  - Multiple modules can inject **more noise** into the pipeline.

Automated speaking assessment systems



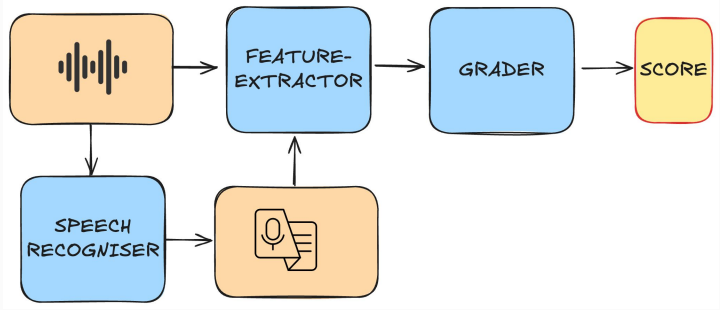
- PROS:
  - Features (embedding representations) are **extracted automatically** by BERT;
- CONS:
  - Still require **specifically designed and annotated training data**;
  - **No access to acoustic information**.

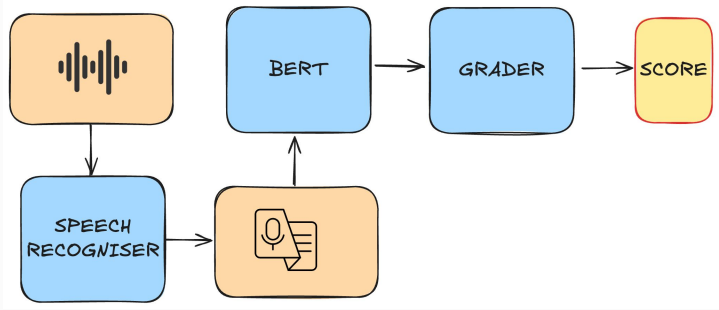
Automated speaking assessment systems

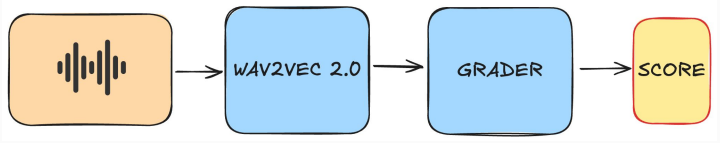


- PROS:
  - Features (embedding representations) are **extracted automatically** by wav2vec 2.0;
- CONS:
  - Still require **specifically designed and annotated training data**;
  - **No access to semantic information**.

Automated speaking assessment systems

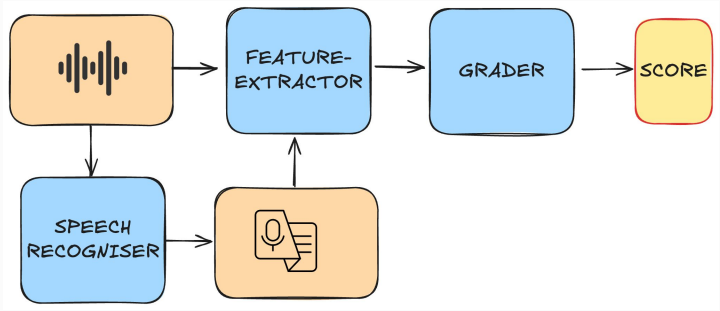


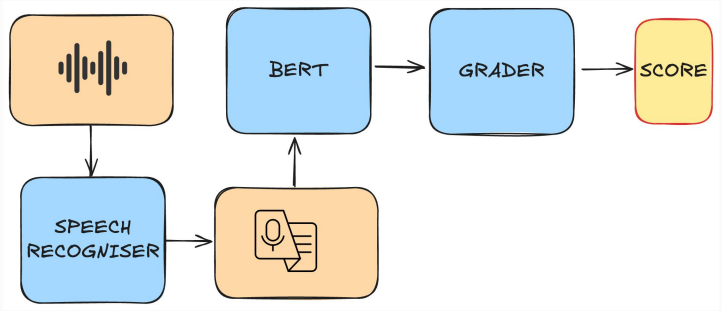


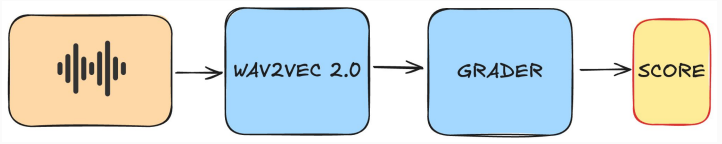


- Learners' productions are **mapped to numerical scores**, not to their underlying meanings.

Automated speaking assessment systems







- Systems show **great performance** in terms of Accuracy or correlation with human-annotated score

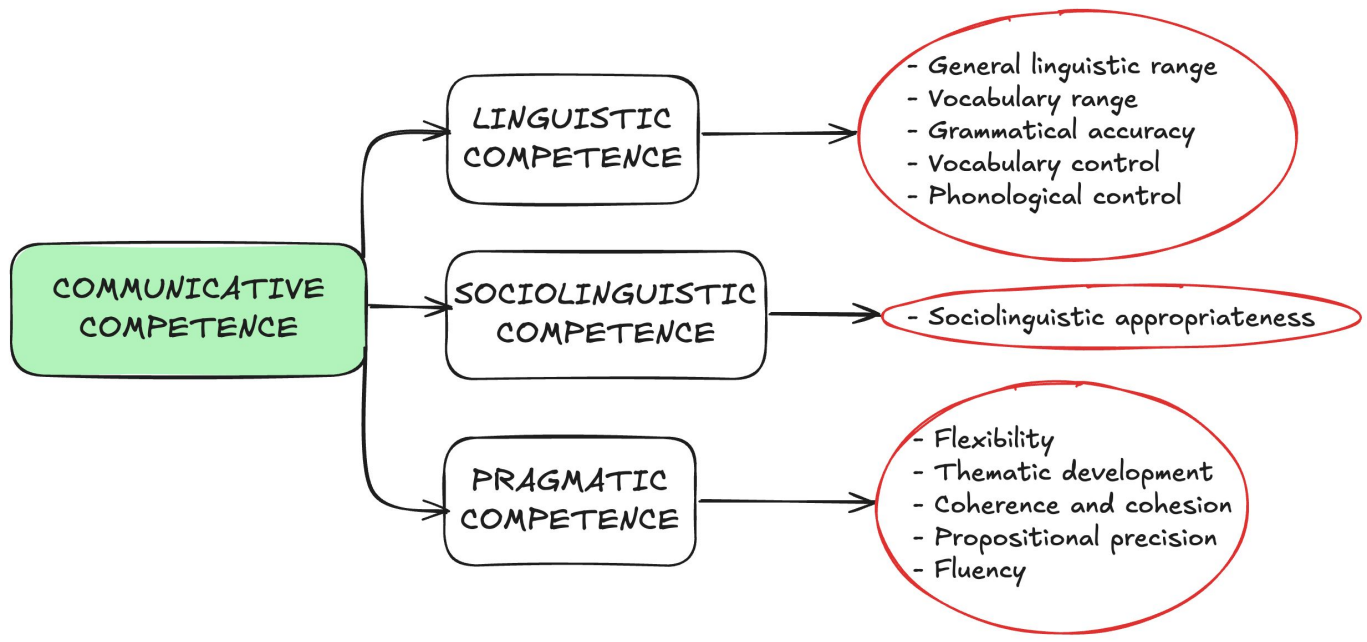
BUT

- **Interpretability** and **explainability of results** are extremely **limited**

Natural Language-based Assessment

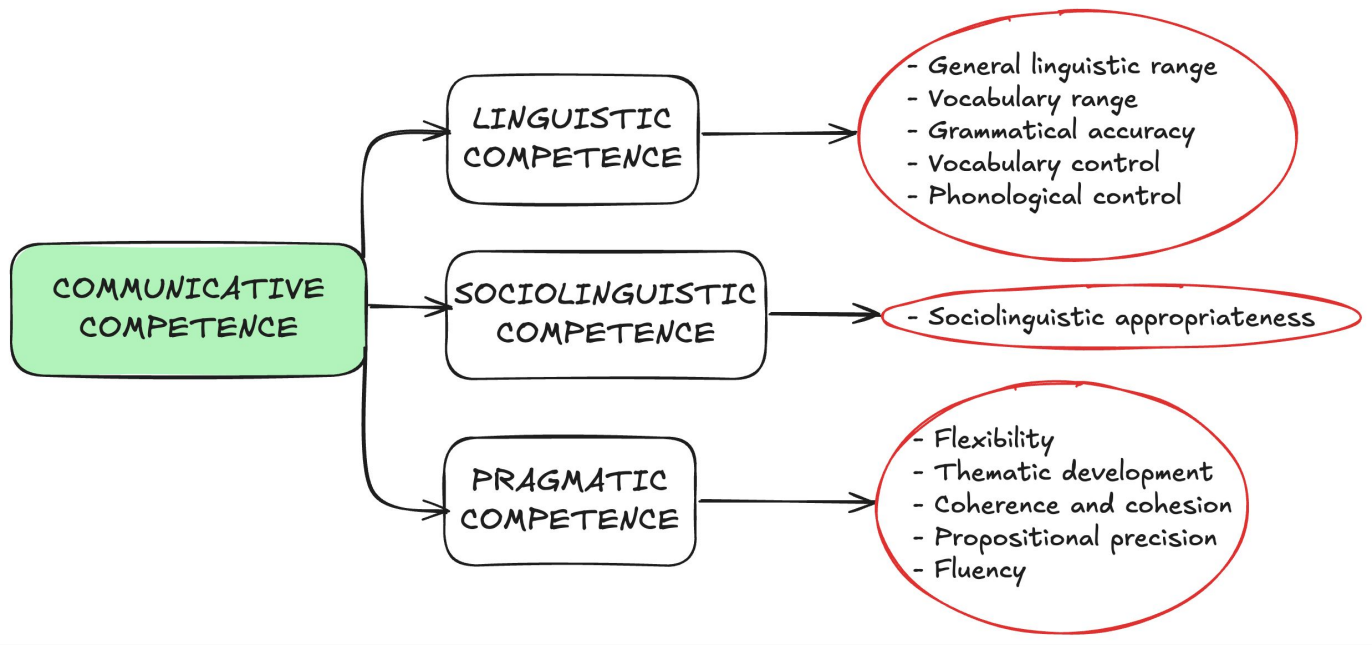
Natural Language-based Assessment (NLA)

- In the Common European Framework of Reference (CEFR) for Languages, the communicative competence is divided into multiple aspects represented by **natural language descriptors**



Natural Language-based Assessment (NLA)

- In the Common European Framework of Reference (CEFR) for Languages, the communicative competence is divided into multiple aspects represented by **natural language descriptors**



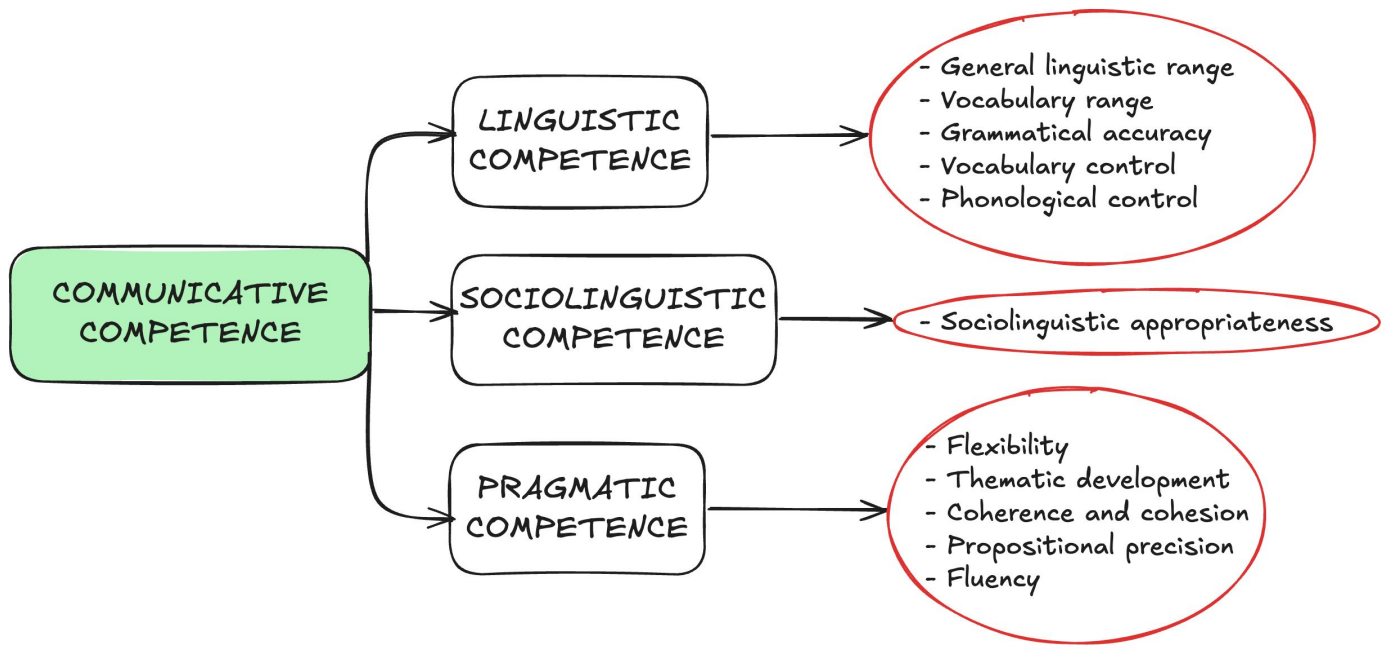
E.g.: **Grammatical accuracy B2**:
Uses some simple structures correctly, but still systematically makes basic mistakes; nevertheless, it is usually clear what they are trying to say.

- Human examiners are often **trained on these analytic criteria** → based on these, they provide a **holistic score**


LLM
LARGE LANGUAGE MODEL

Natural language-based Assessment (NLA)

- Can state-of-the-art LLMs:
  - **interpret analytic descriptors** originally intended for human examiners?
  - be leveraged to provide assessment **without being fine-tuned** for this purpose?



Natural Language-based Assessment

Natural Language-based Assessment of L2 Oral Proficiency Using LLMs

10th Workshop on Speech and Language Technology in Education (SLaTE)
22-24 August 2025, Nijmegen, Netherlands

Natural Language-based Assessment of L2 Oral Proficiency using LLMs

Stefano Bannò[1], *Rao Ma*[1], *Mengjie Qian*[1], *Siyuan Tang*[1], *Kate Knill*[1], *Mark Gales*[1]

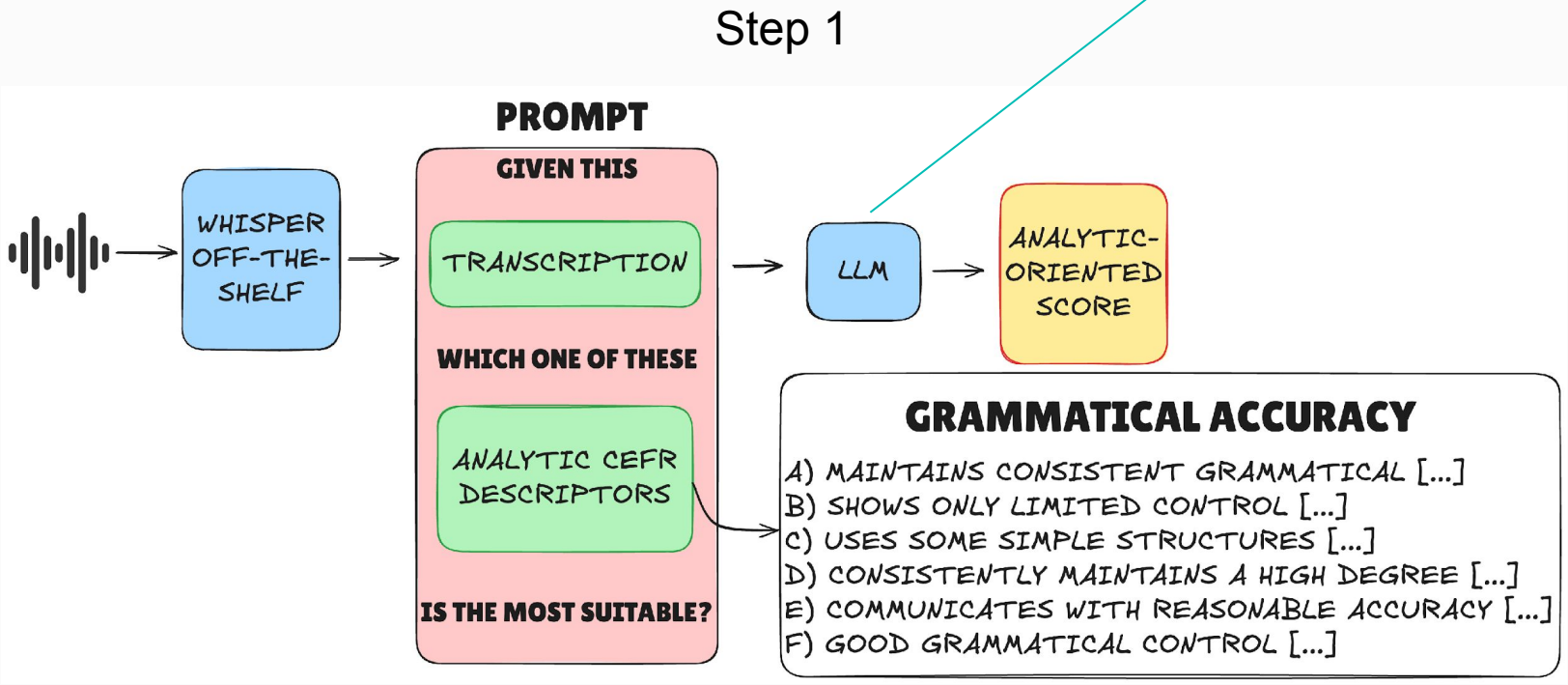
[1]ALTA Institute, Machine Intelligence Lab, Department of Engineering, Cambridge University, UK

{sb2549,rm2114,mq227,st941,kmk1001,mjfg100}@cam.ac.uk

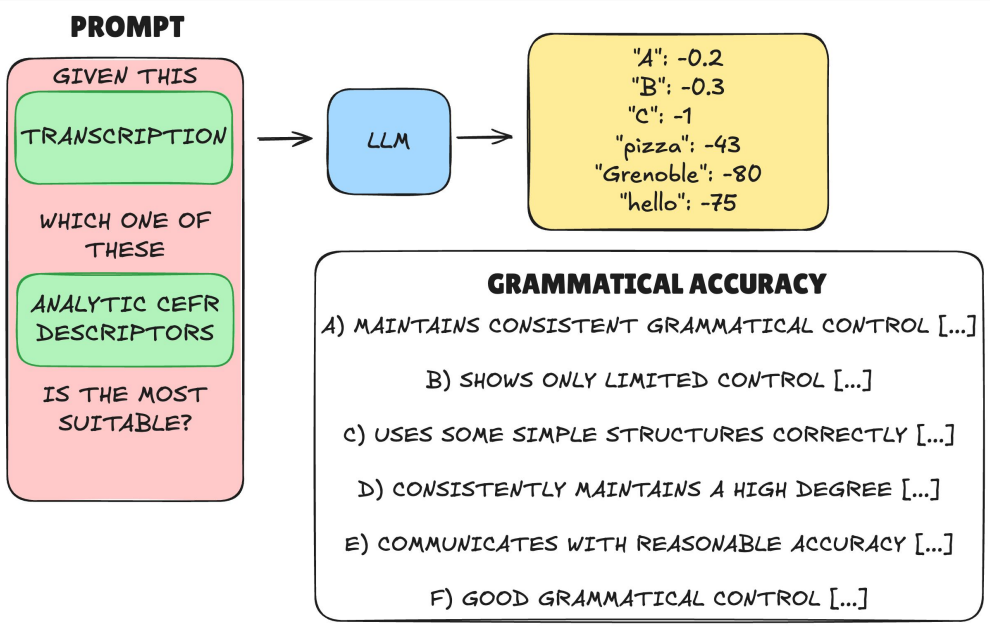
Experimental Setup

The NLA pipeline

- We used an **open-source** LLM, **Qwen 2.5 72B** (4-bit quantised)



The NLA pipeline



Step 1

1. **We bias the LLM to output A, B, C, D, E, or F.**

2. **To obtain the score:**

a) we extract the logit probabilities of the top predicted tokens (i.e., A-F);

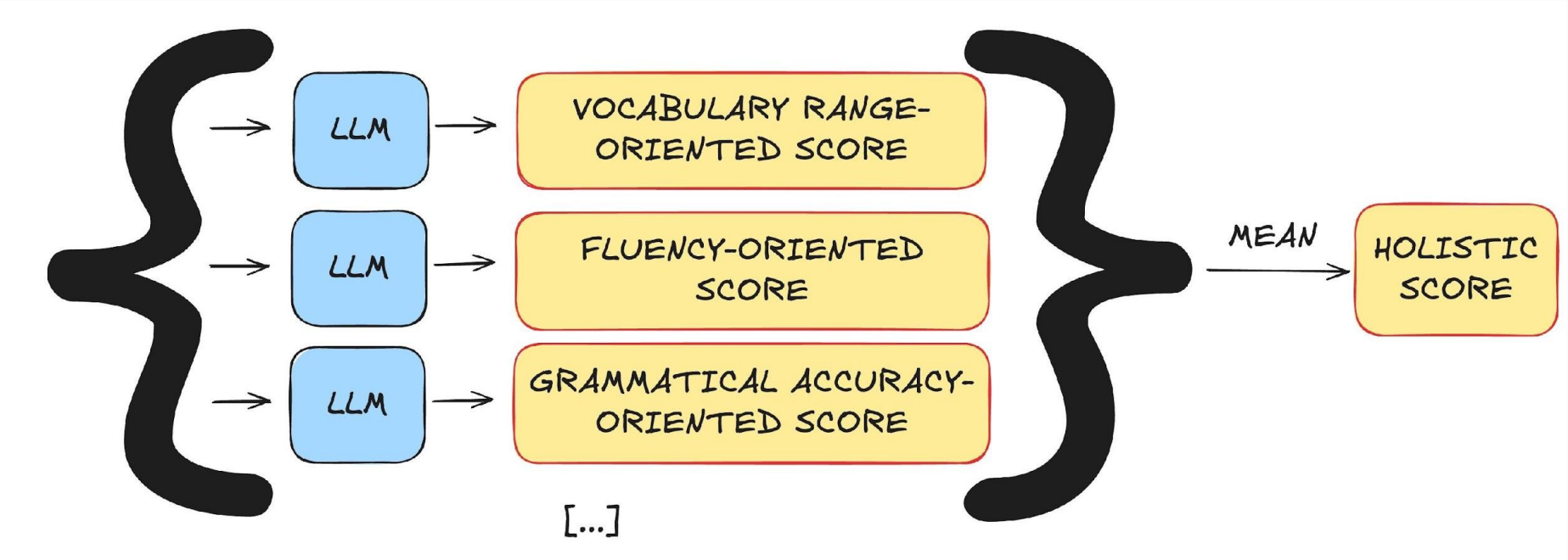
b) apply softmax;

c) weigh each probability by its respective level (A1: 1, A2: 2, B1: 3, etc.).

3. Repeat this process three times shuffling the order of the options to limit positional bias.

The NLA pipeline

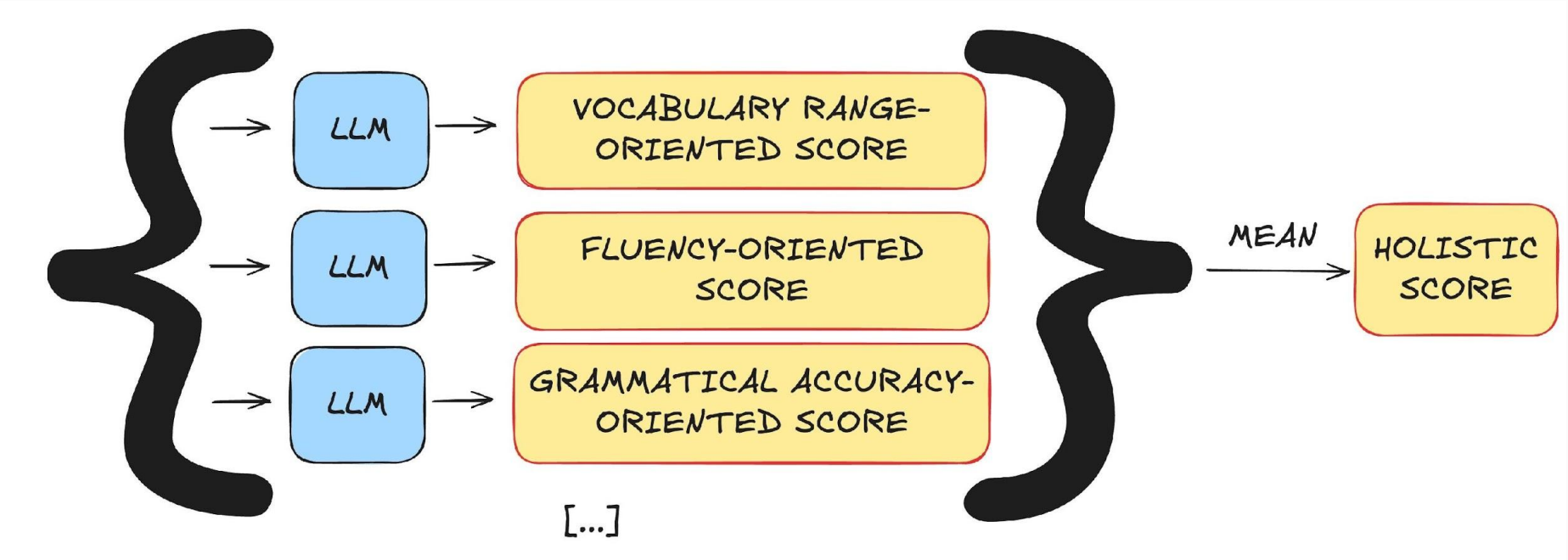
Step 2



- The **average** of all the analytic-oriented scores give us the **holistic score**.

The NLA pipeline

Step 2



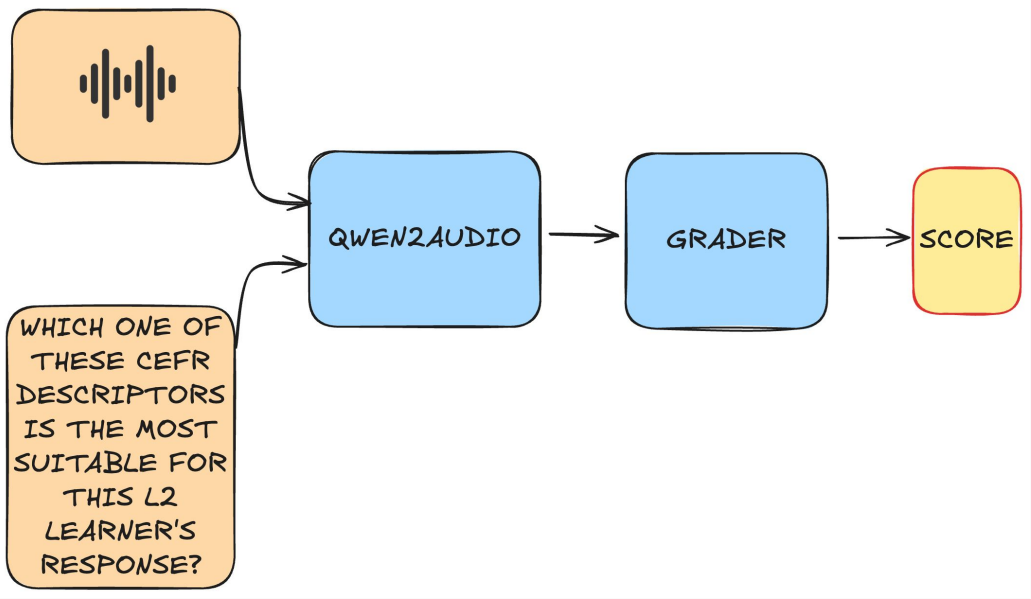
- The **average** of all the analytic-oriented scores give us the **holistic score**.

*we only used language aspects that can be decoded from transcriptions

Speech LLM-based grader

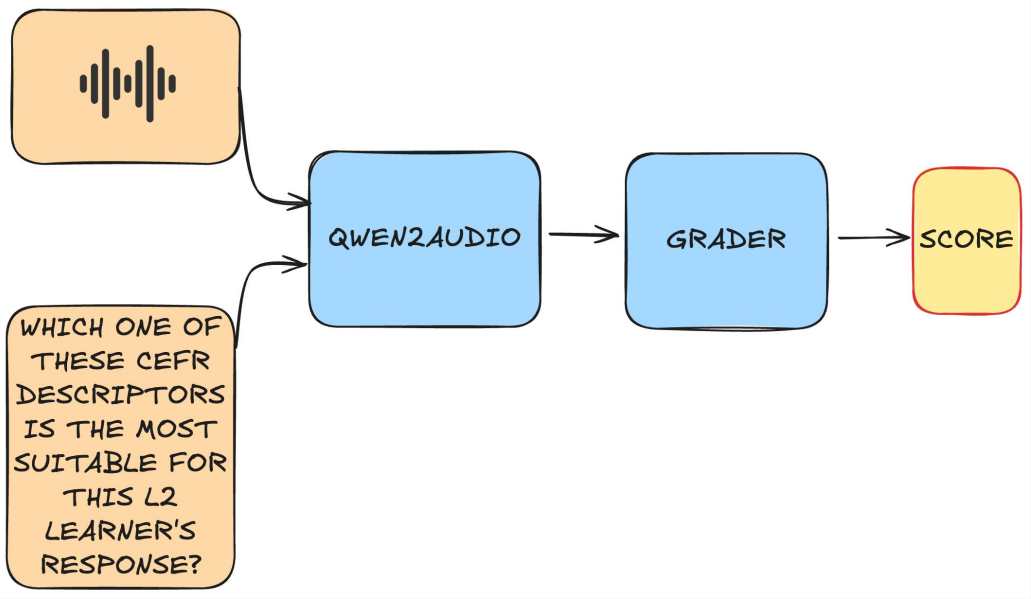
Ma et al., 2025

- **Pros:**
  - Can access **both acoustic and semantic** information
  - **Doesn't rely on assumptions** underlying selected features
  - **Doesn't rely on ASR**
  - Trained to map learners' responses **to holistic CEFR natural language descriptors**, not just their numerical representations
- **Cons:**
  - Still requires annotated **training data**
  - Expensive in terms of **computational resources**



Speech LLM-based grader

Ma et al., 2025



- **Pros:**
  - Can access **both acoustic and semantic** information
  - **Doesn't rely on assumptions** underlying selected features
  - **Doesn't rely on ASR**
  - Trained to map learners' responses **to holistic CEFR natural language descriptors**, not just their numerical representations
- **Cons:**
  - Still requires annotated **training data**
  - Expensive in terms of **computational resources**

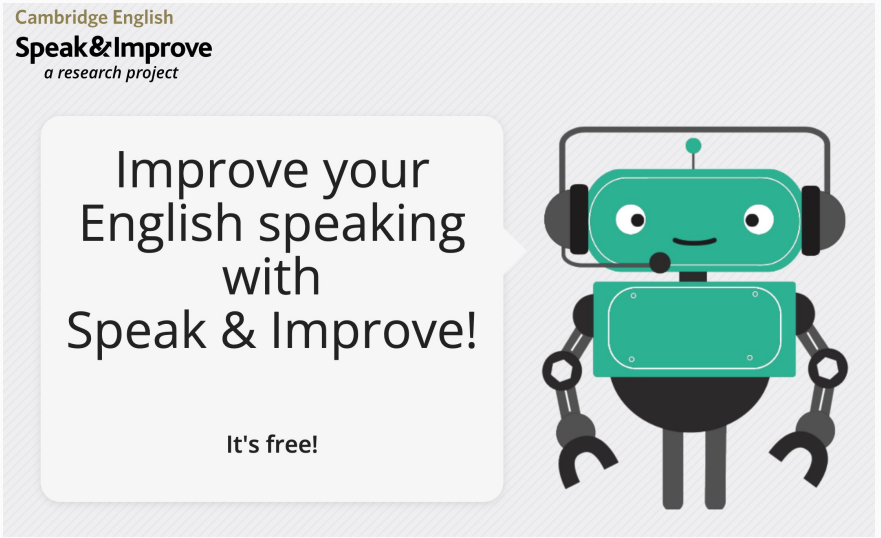
*we also tried using Qwen2Audio in a zero-shot fashion, but it didn't yield good results

Speak & Improve (S&I) Corpus 2025

Data obtained from www.speakandimprove.com:

- **Part 1:** Answer 8 personal questions (each 10-20 seconds long)
- **Part 3, 4:** Speak about a topic or diagram (1 minute long)
- **Part 5:** Answer 5 follow-up questions (each 20 seconds long)

The scores derived from each part are averaged into an **overall score**.



Split	Train	Dev	Eval
#Submissions	6,640	438	300
#Hours	244	35	23

More about the corpus here:

Evaluation metrics

Results are evaluated in terms of:

- Pearson's Correlation Coefficient (**PCC**)
- Spearman's Rank Coefficient (**SRC**)

between the **predictions** and the **human-annotated** ground truth holistic scores.

Experimental Results

Experimental Results: holistic speaking proficiency assessment

Model	S&I dev		
	Training	PCC	SRC
BERT	✅	0.753	0.728
NLA	❌	0.806	0.812
Qwen2Audio	✅	**0.833**	**0.837**

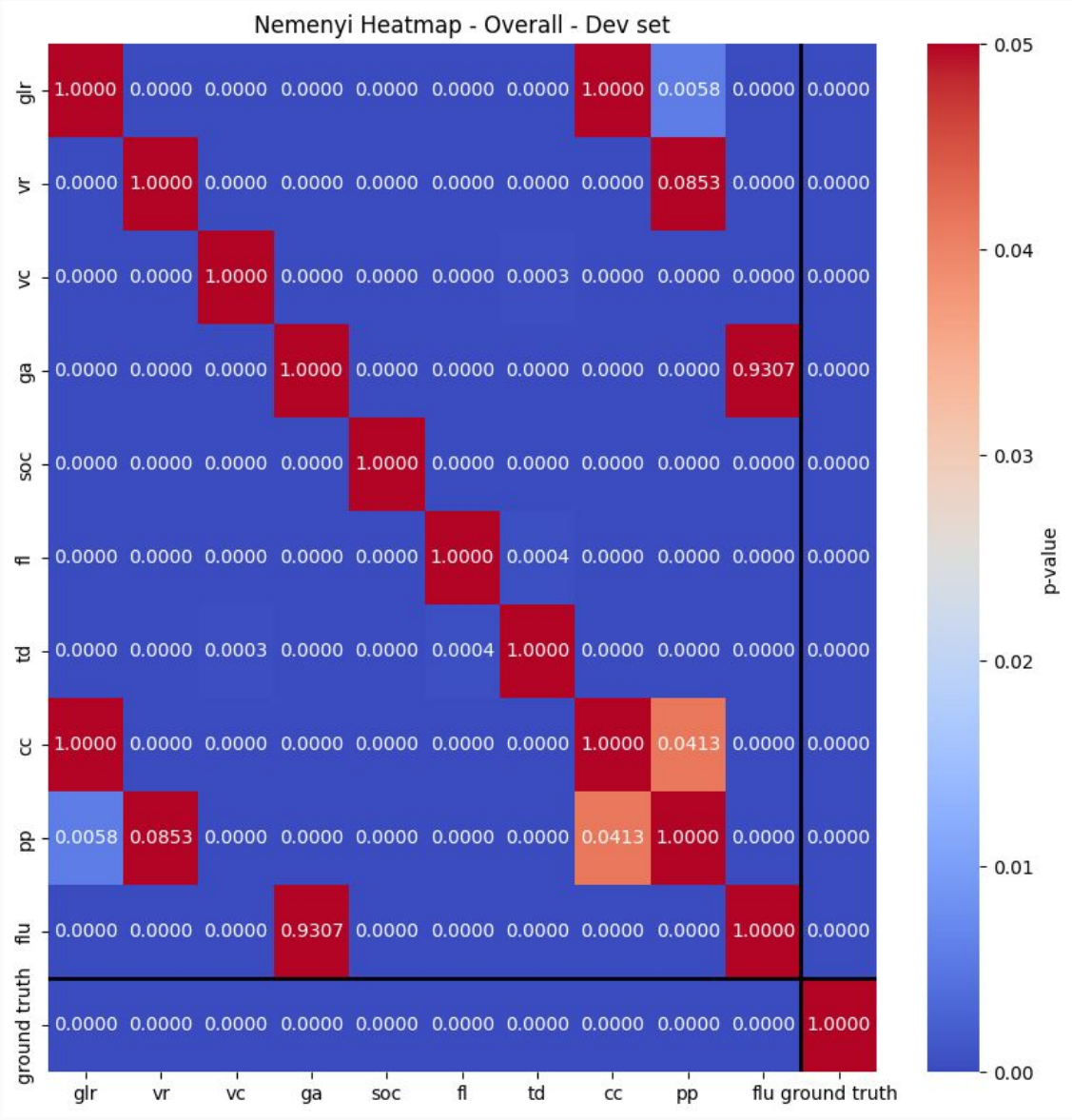
- **Natural Language-based Assessment**
  - outperforms fine-tuned BERT-based grader
  - is only a few points behind fine-tuned Qwen2Audio

- Note that NLA does not explicitly leverage acoustic information

Experimental results: analytic scoring

- **Analytic scores used for holistic score**
  - For 10 individual CEFR proficiency aspects
- **Friedman + Nemenyi tests between predicted analytic and ground truth holistic scores**
  - Significant *p*-value for Friedman: significant differences in rankings
  - Nemenyi test: most paired comparisons show significant differences (*p*-value < 0.05)

Results suggest that **analytic-oriented scores** may be **capturing some distinct aspects** of language proficiency.



Natural Language-based Assessment

Conclusions and future work

- NLA approach achieves **competitive performance**:
  - outperforms a specifically fine-tuned BERT-based system
  - inherently **portable** to other data types, descriptors, and languages
  - offers **greater interpretability** as it is anchored in clearly explainable descriptors

- Next steps:
  - More **thorough analysis** of individual **analytic** scores predicted by the LLM
  - More efficient ways to **aggregate analytic-oriented scores**
  - Investigating ways to incorporate acoustic information better (e.g., few shot learning)

UNIVERSITY OF CAMBRIDGE Department of Engineering Cambridge ALTA Institute for Automated Language Teaching and Assessment

This presentation reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

If you want to know more about the ALTA SLP Project Team:

sb2549@cam.ac.uk

6 March 2026