UNIVERSITY OF CAMBRIDGE

Automated Learning Teaching and Assessment Spoken Language Processing Technology Project

Dr Mengjie Qian

ALTA Institute, Machine Intelligence Lab, Cambridge University Engineering Department

18th June 2024

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

Cambridge Automated Language Teaching and Assessment Institute

CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF ENGINEERING

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

- Virtual Institute for

  cutting-edge research on second language (L2) English assessment

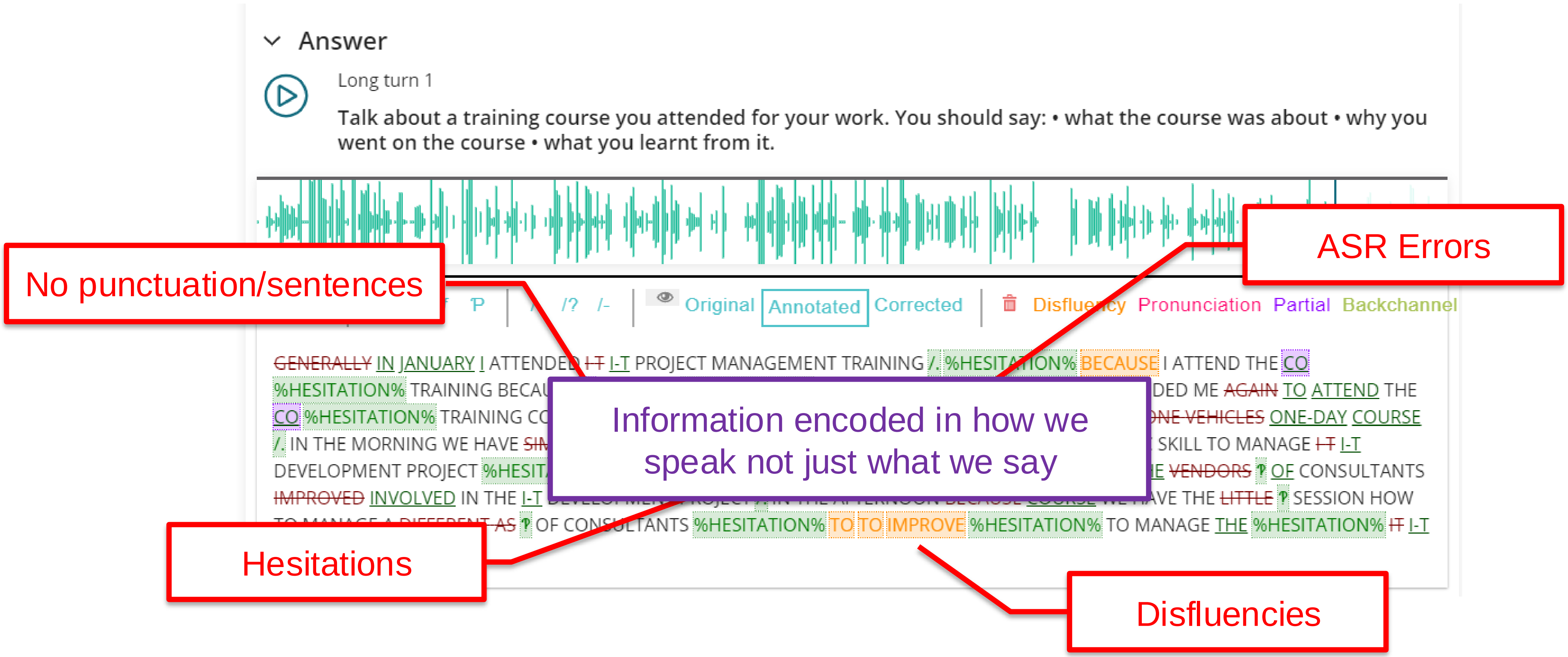
  - Machine Learning and Natural Language Processing
  - Develop technology to enhance assessment and learning
  - Look to benefit learners and teachers worldwide

ALTA SLP Project Team

- Principal Investigators: **Dr Kate Knill**, **Prof Mark Gales**
- Postdocs: **Dr Mengjie Qian, Dr Stefano Bannò**, **Dr Simon McKnight, Dr Hari Vydana**
- Research Assistant: **Siyuan Tang**
- PhD students: Charles McGhee, **Rao Ma**, Yassir Fathullah, **Adian Liusie**, Potsawee Manakul, Vatsal Raina, **Vyas Raina**
- 4th year Engineering students
- Public webpage: http://mi.eng.cam.ac.uk/~mjfg/ALTA/index.html

Bold = (part)-funded by ALTA

L2 learner speech data is challenging!

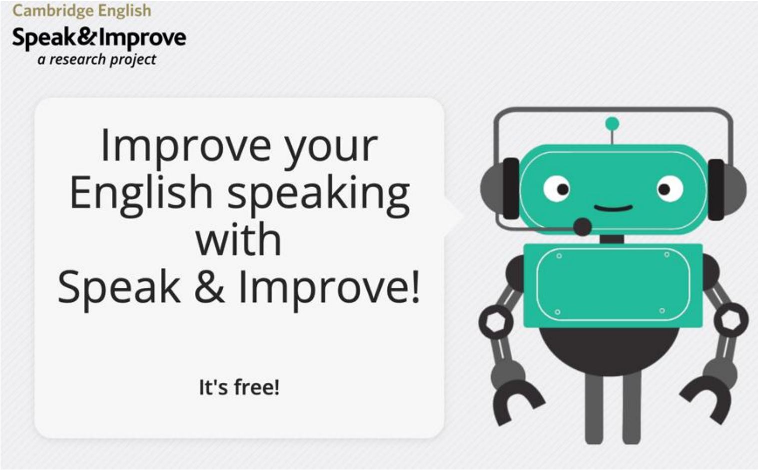


ALTA Spoken Language Processing Technology Project

Linguaskill

Upskill from Cambridge

>300k SUBMISSIONS
April 2023



https://speakandimprove.com

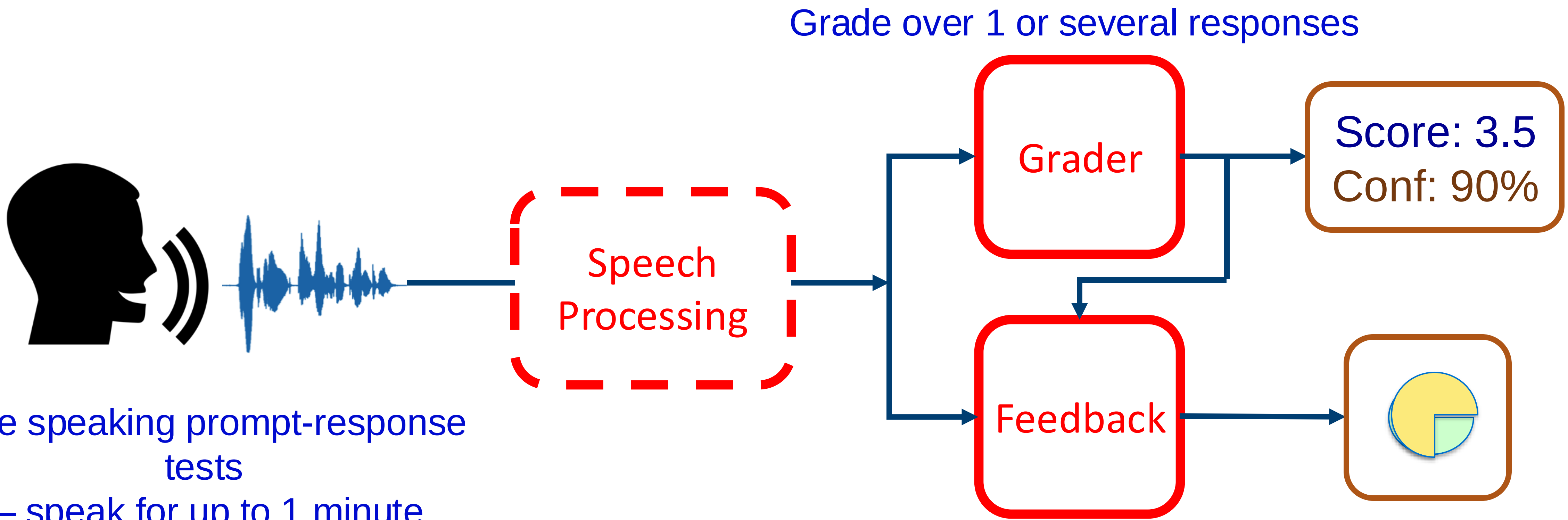
> 150 COUNTRIES

> 400k CANDIDATES / VISITORS

>9M SUBMISSIONS
June 2022

- Achieved through medium to long-term research at ALTA SLPTP
  - with technology transfer and collaboration with CUP&A and technology partners

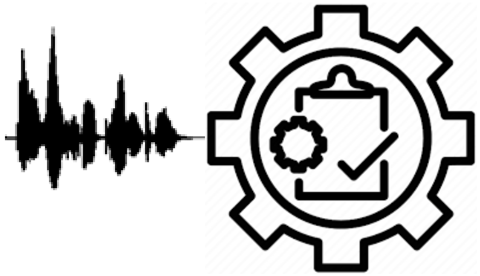
Spoken Language Assessment and Feedback Pipeline



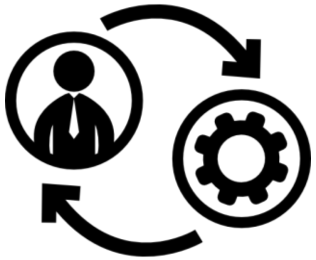
Free speaking prompt-response tests
– speak for up to 1 minute

Analytic – holistic feedback across all speech
Fine-grained – feedback on specific errors in words/phrases

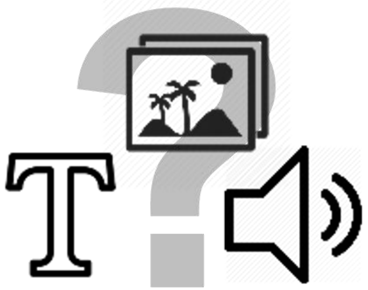
ALTA SLPTP Research Strands



SPEAKING ASSESSMENT



LEARNER ORIENTED FEEDBACK



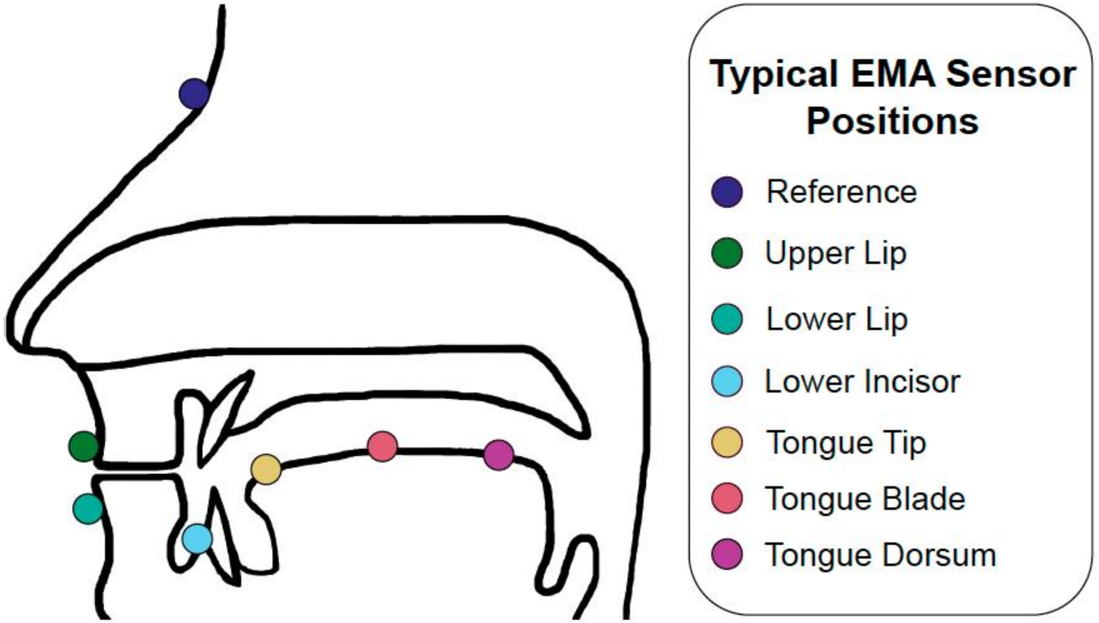
CONTENT CREATION

CORE TECHNOLOGY

Learner Oriented Feedback

Pronunciation Training

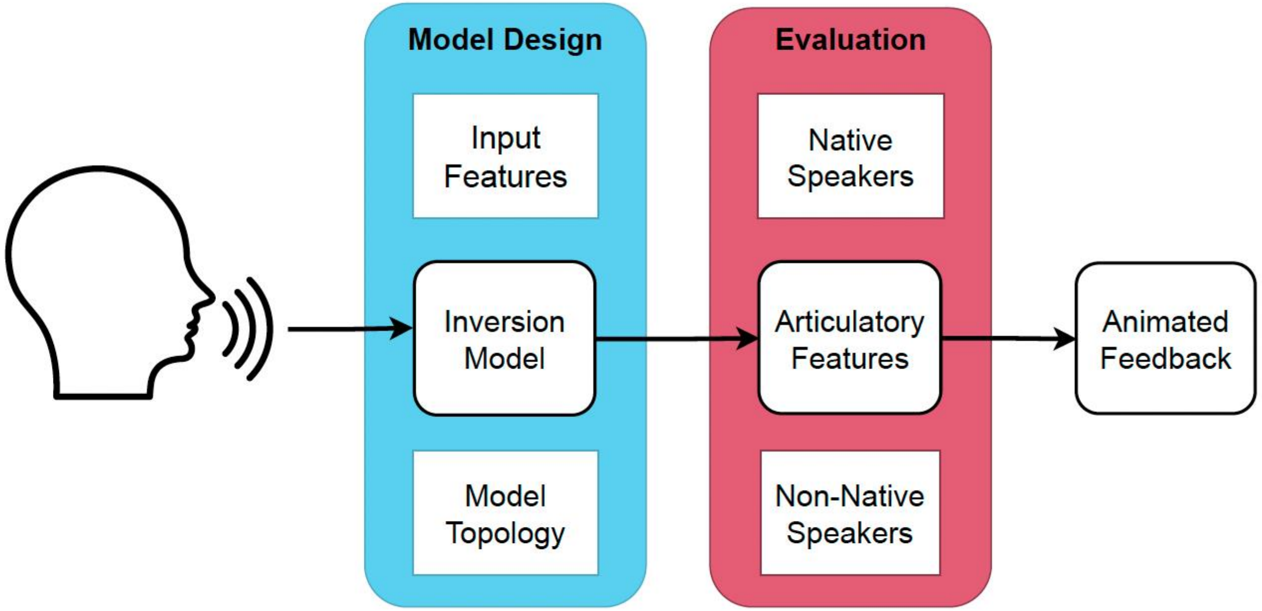
- Objective
  - Show an English language learner movement of their tongue, lips and jaw to aid non-native (L2) speech sound acquisition
- Problem
  - Measuring articulatory movements with sensors, as in Electromagnetic Articulography (EMA), can be invasive and expensive
  - EMA, ultrasound etc not suitable for general practice e.g. through web-based app



Pronunciation Training

- Solution (Charlie McGhee)
  - Use Acoustic-to-Articulatory Inversion (AAI) to predict articulatory features, such as EMA positions, from speech
  - Provide learner with animated feedback

- What we would like to learn about:
  - How best to animate?
  - What is most useful?
  - What to avoid?
  - Real-time or on playback?



McGhee, Charles, Kate Knill, and Mark Gales. "Towards Acoustic-to-Articulatory Inversion for Pronunciation Training." in *Proc. of Speech and Language Technology in Education (SLaTE)*. Workshop 2023.

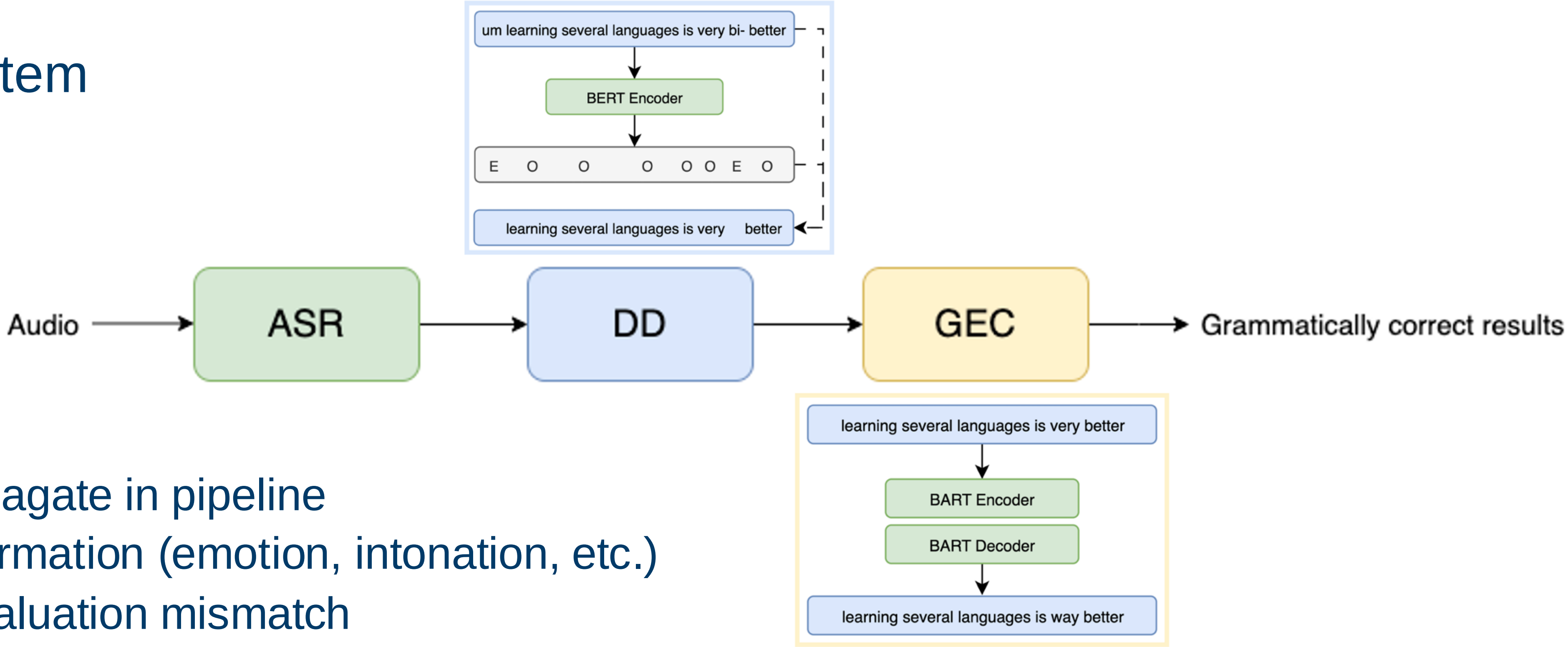
Spoken Grammar Error Correction (Spoken GEC)

- Objective
  - Correcting errors within spoken language
  - Typical approach:
    - step1: automatic speech recognition (**ASR**) system
    - step2: disfluency detection (**DD**) module
    - step3: **GEC** model

- Written GEC:
  - **Original:** Learning several languages is very better.
  - **Corrected:** Learning several languages is way better.
- Spoken GEC:
  - **Original:** um learning several languages is very bi- better
  - **Fluent:** learning several languages is very better
  - **Corrected:** learning several languages is way better

Spoken Grammar Error Correction (Spoken GEC)

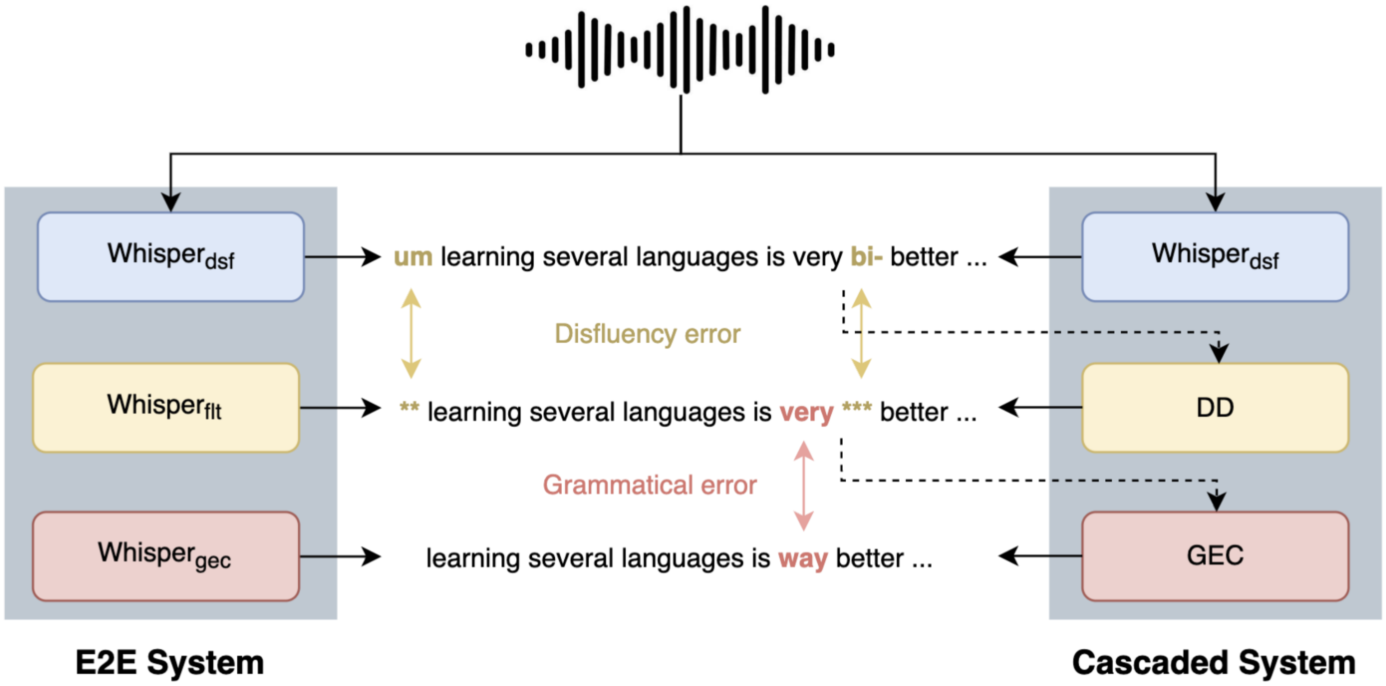
- Cascaded system



- Problem
  - errors propagate in pipeline
  - loss of information (emotion, intonation, etc.)
  - training-evaluation mismatch

Spoken Grammar Error Correction

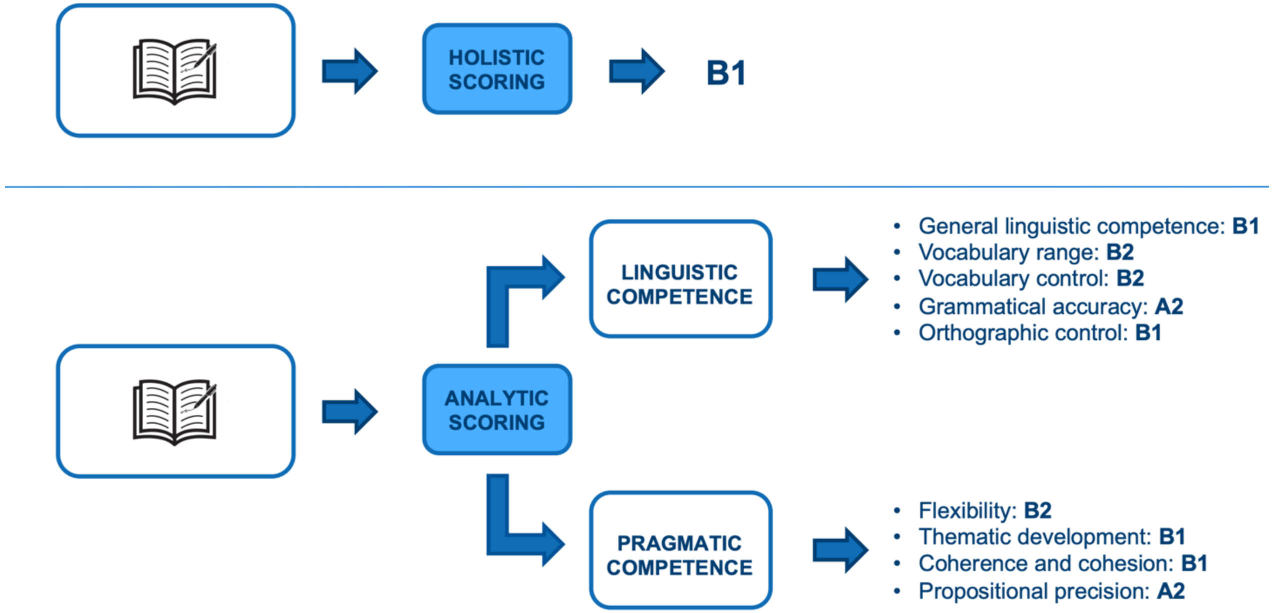
- Solution (Dr Stefano Bannò, Rao Ma, Mengjie Qian)
  - Whisper foundation model
    - Fine-tune to target targets
  - End-to-end spoken GEC
    - Translate audio to GEC text
  - Also
    - E2E disfluency detection and correction model
    - Disfluent speech recognition



Bannò, Stefano, et al. "Towards end-to-end spoken grammatical error correction." in *ICASSP*.

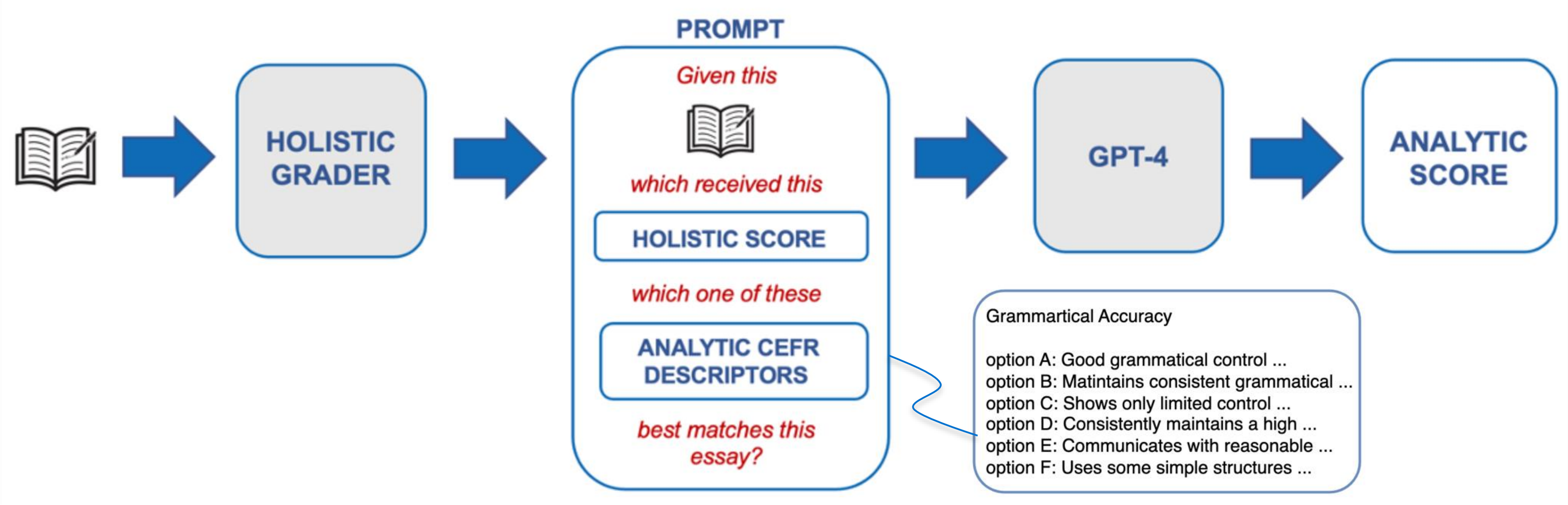
Can GPT-4 do L2 analytic assessment?

- Objective
  - Analytic assessment allows for a more detailed evaluation and more informative feedback
  - Can enhance scoring validity
- Problem
  - Less time efficient and more cognitively demanding than holistic assessment
  - Halo effect: raters may fail to distinguish between different aspects
  - No L2 learner datasets annotated with analytic scores available



Can GPT-4 do L2 analytic assessment?

- Solution (Dr Stefano Bannò)
  - Extract information about analytic aspects from L2 learner essays and their assigned holistic scores using GPT-4?
- What we would like to learn about
  - Can GPT-4 perform L2 analytic assessment?



Bannò, Stefano, et al. "Can GPT-4 do L2 analytic assessment?." *arXiv preprint arXiv:2404.18557* (2024).

Speaking Assessment

Comparative Assessment

- Objectives
  - Natural language generative assessment
  - Automatic Speech Recognition: Single reference

 → the cat sat on the mat

  - Neural Machine Translation: many valid references

你还好吗? → How are you? / Are you okay?

  - Summarization: Vast number acceptable summaries

 → Man enters burning house to save child / Heroic person saves young boy / Home burns, child makes it out safe / ...

Comparative Assessment

News article: A G4S security van has been robbed outside a branch of royal bank of Scotland in Glasgow city centre. Police said three armed men took a five-figure sum from the vehicle in the city's Sauchiehall street on Monday at about 21:45. A spokesman said no-one had been injured [...]

→ Summary Generation System →

Summary: Two security guards have been threatened during a bank robbery in Scotland.

Manual assessment

1	2	3	4	5	6	7

costly

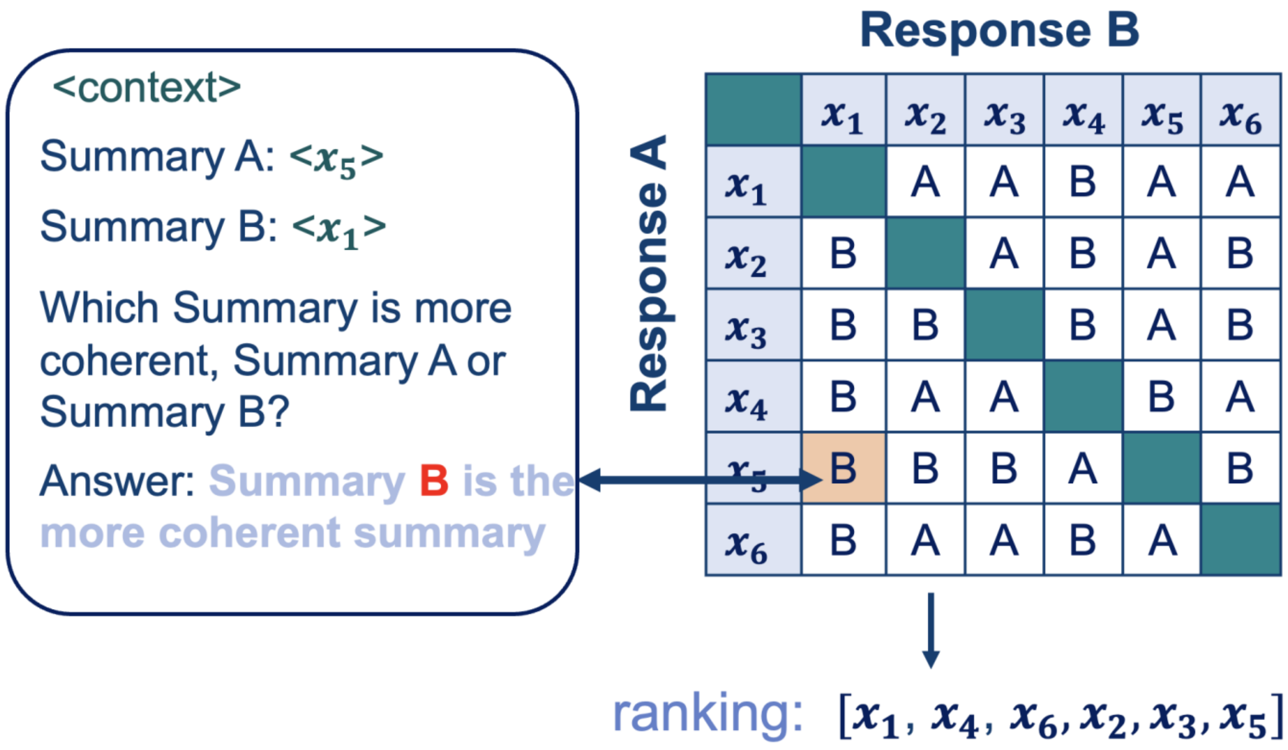
manual

time-intensive

Can we replace manual evaluation with effective automatic methods?

Comparative Assessment

- Solutions (Adian Liusie, Potsawee Manakul)
  - Prompt LLM to make pairwise comparisons for NLG assessment
    - Debias
    - Win-ratio / average probabilities



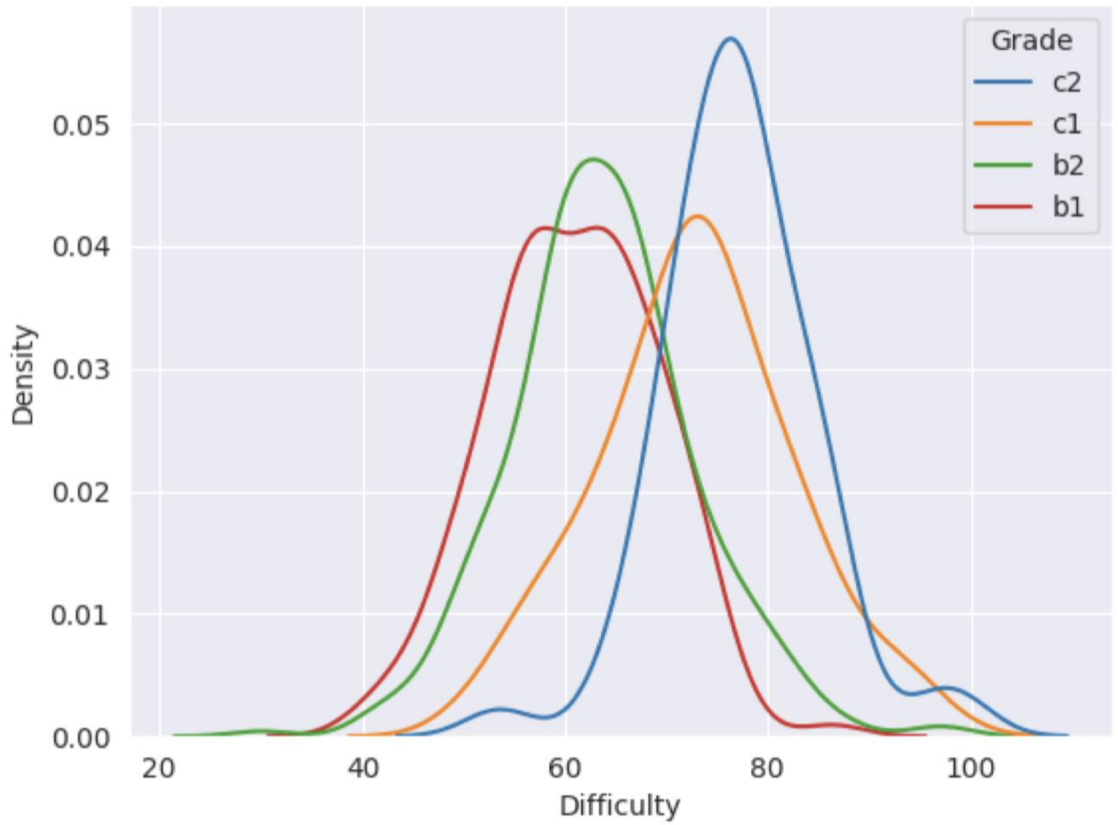
	x_1	x_2	x_3	x_4	x_5	x_6
x_1		A	A	B	A	A
x_2	B		A	B	A	B
x_3	B	B		B	A	B
x_4	B	A	A		B	A
x_5	B	B	B	A		B
x_6	B	A	A	B	A	

Liusie, Adian, et al. "LLM comparative assessment: Zero-shot nNLG evaluation through pairwise comparisons using large language models." In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 139-151. 2024.

Core Technology

Question Difficulty Ranking

- Objectives
  - Multiple-choice (MC) tests are efficient to assess English learners
  - Rank candidate MC questions by difficulty
- Problems
  - Determining the difficulty level of questions with human test taker trials is expensive and not scalable



Question Difficulty Ranking

- Solutions (Vatsal Raina)
  - Task transfer

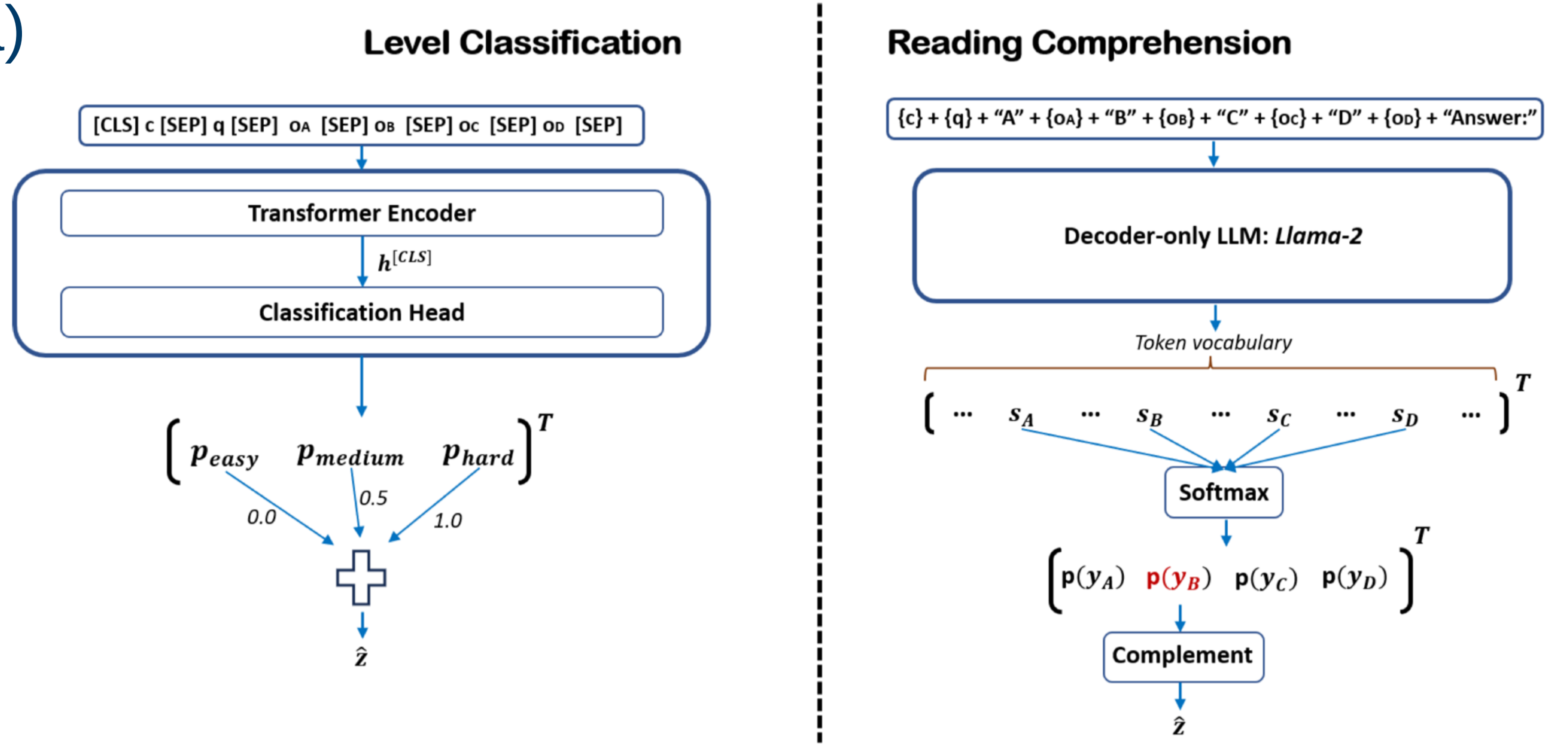
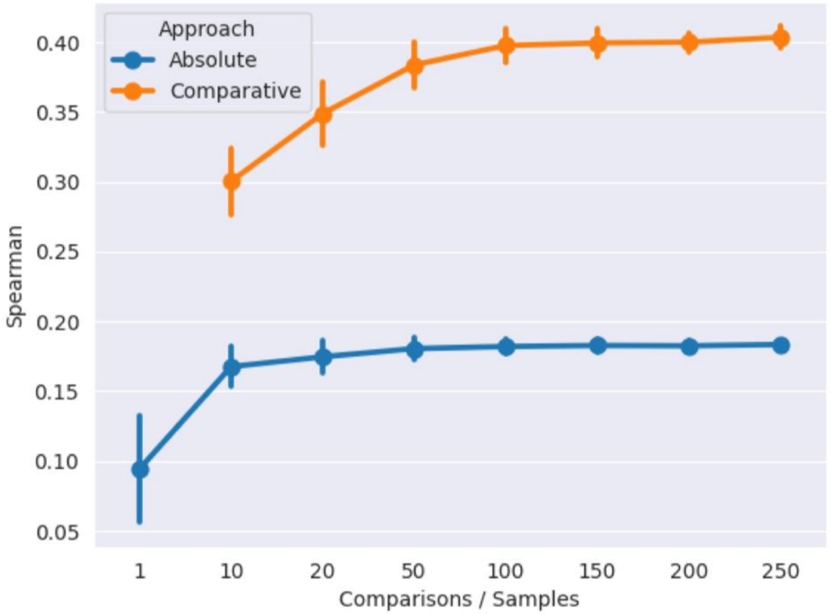


Figure 1: Task transfer for difficulty estimation with context, c, question, q and options, o.

Raina, Vatsal, and Mark Gales. "Question Difficulty Ranking for Multiple-Choice Reading Comprehension." *arXiv preprint arXiv:2404.10704* (2024).

Question Difficulty Ranking

- Solutions (Vatsal Raina)
  - Zero-shot with ChatGPT



Absolute

{context}

{question}
A) {option_A}
B) {option_B}
C) {option_C}
D) {option_D}

Provide a score between 1 and 10 that measures the difficulty of the question. Return only a single score. ”

Comparative

1:
{context_1}

{question_1}
A) {option_A_1}
B) {option_B_1}
C) {option_C_1}
D) {option_D_1}

2:
{context_2}

{question_2}
A) {option_A_2}
B) {option_B_2}
C) {option_C_2}
D) {option_D_2}

Which reading comprehension question is more difficult, 1 or 2? Return only 1 or 2. ”

Conclusions

- ALTA SLP Technology Project aims to advance language assessment using Machine Learning and Natural Language Processing techniques

- Research on speaking assessment, learner-oriented feedback, and core technology

- On-going work leverages foundation models to develop more robust and efficient approaches

Questions?

Thanks to:

Diane Nicholls and the Humannotator team at ELiT for Linguaskill Speaking annotations.

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

ALTA SLPT Project publications can be found at: http://mi.eng.cam.ac.uk/~mjfg/ALTA/index.html

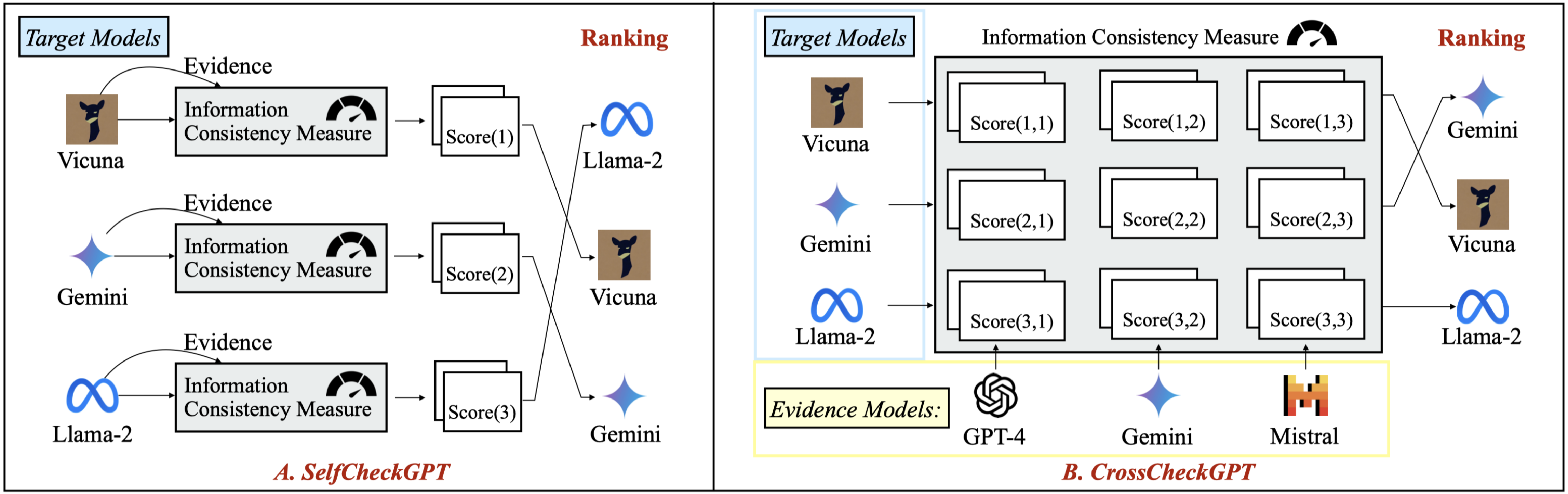
Appendix

SelfCheckGPT and CrossCheckGPT

- Objectives
  - Foundation models “hallucinate”
    - the generated outputs, while seemingly credible, are either inconsistent with the provided context or contradict established factual knowledge
  - Quantify a system’s susceptibility to hallucination
- Problems
  - Current benchmarks are designed for particular tasks
  - Assume access to gold-standard labels

SelfCheckGPT and CrossCheckGPT

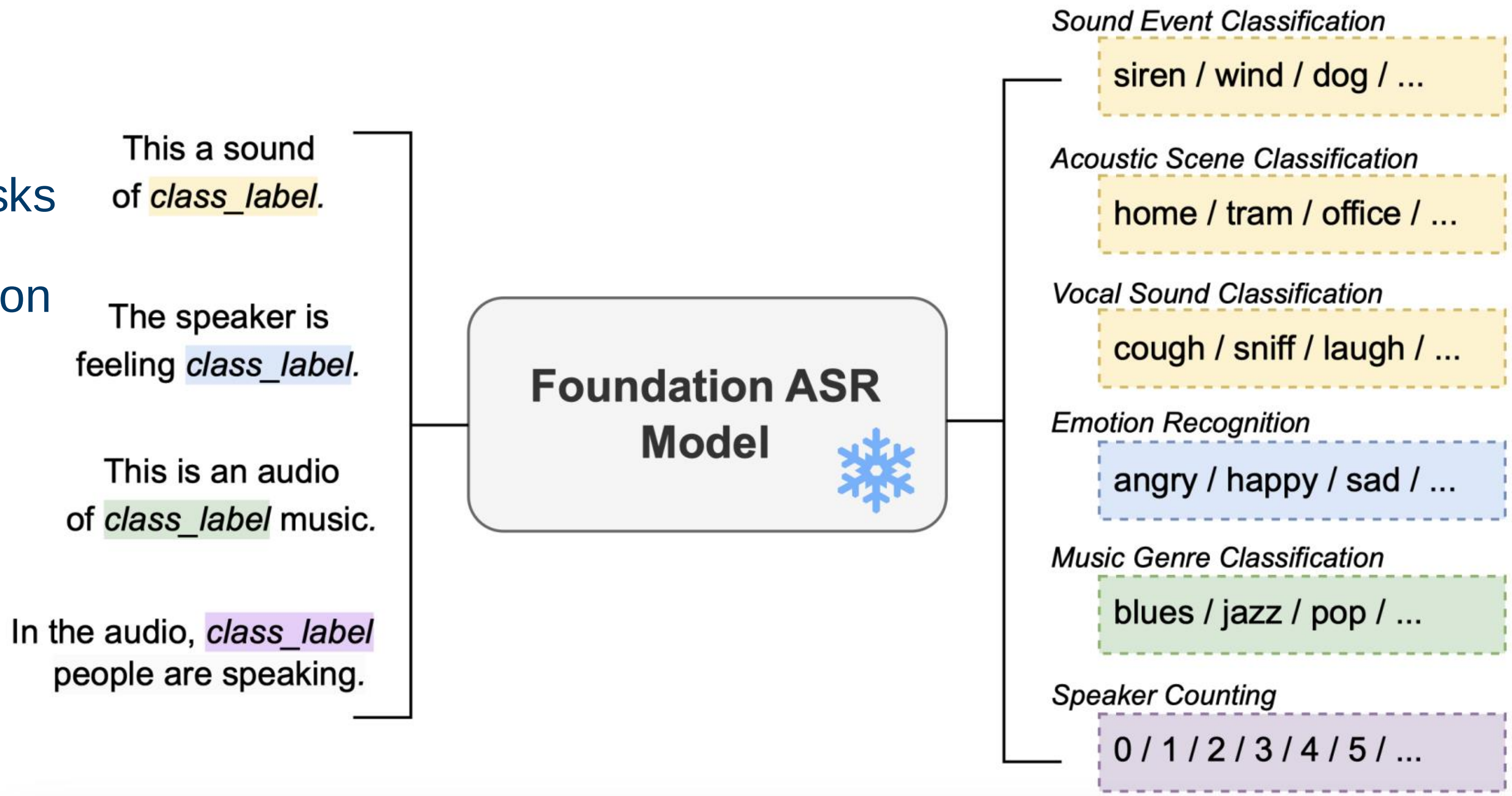
- Solution (Potsawee Manakul)



Sun, Guangzhi, Potsawee Manakul, et al. "CrossCheckGPT: Universal Hallucination Ranking for Multimodal Foundation Models." *arXiv preprint arXiv:2405.13684* (2024).

Emergent Audio Classification Ability of Whisper

- Objective
  - OpenAI whisper trained on ASR, speech translation tasks
  - Emergent ability of foundation speech models?



- Solution (Rao Ma)
  - Zero-shot prompting of Whisper models

Ma, Rao, et al. "Investigating the Emergent Audio Classification Ability of ASR Foundation Models." *arXiv preprint arXiv:2311.09363* (2023).

Emergent Audio Classification Ability of Whisper

- Solution (Rao Ma)
  - Zero-shot prompting of Whisper models