Designing a Speaking Assessment Module for the SELF Language Placement Test

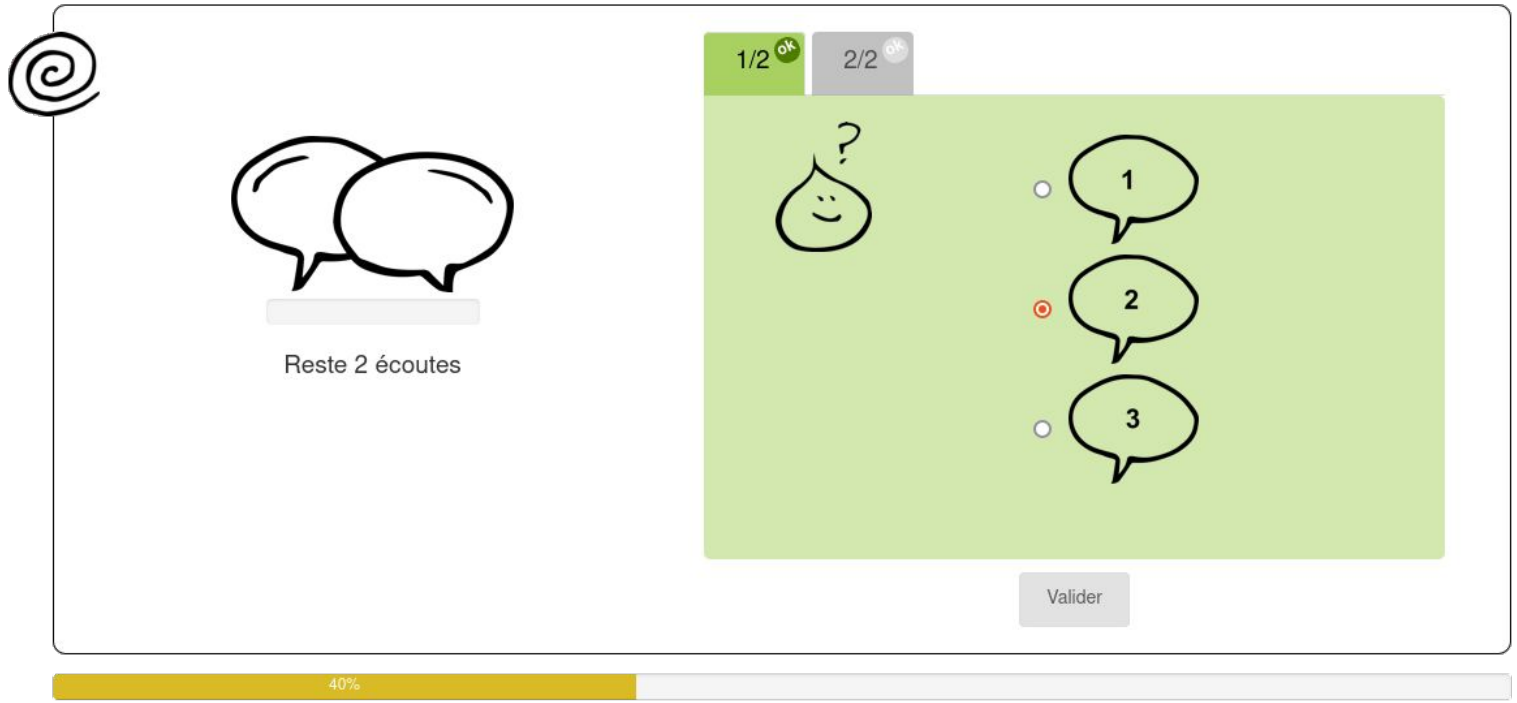
Pinxun HUANG, Eli STAFFORD, Sylvain COULANGE

LIDILEM *(Laboratory of Linguistics and Didactics of Foreign and Mother Tongues)*
LIG *(CNRS, Institute of Engineering, Grenoble Computer Science Laboratory)*
Université Grenoble Alpes

SELF - Formative Language assEssment System

Online placement test developed at Université Grenoble Alpes

SELF
INNOVA
LANGUES



SELF - Formative Language assEssment System

Online placement test developed at Université Grenoble Alpes

7 languages
(English, Italian, Mandarin, Japanese, Spanish, French, German)

~1h

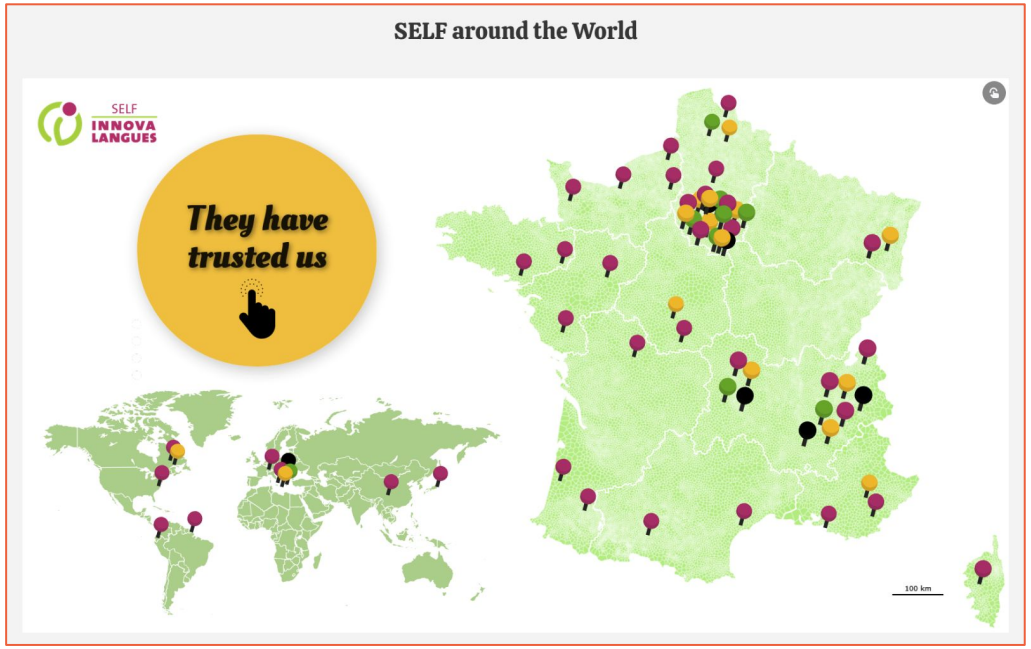
CECRL

A1-C1

~40 000
tests per year

~40
universities

adaptive
algorithm

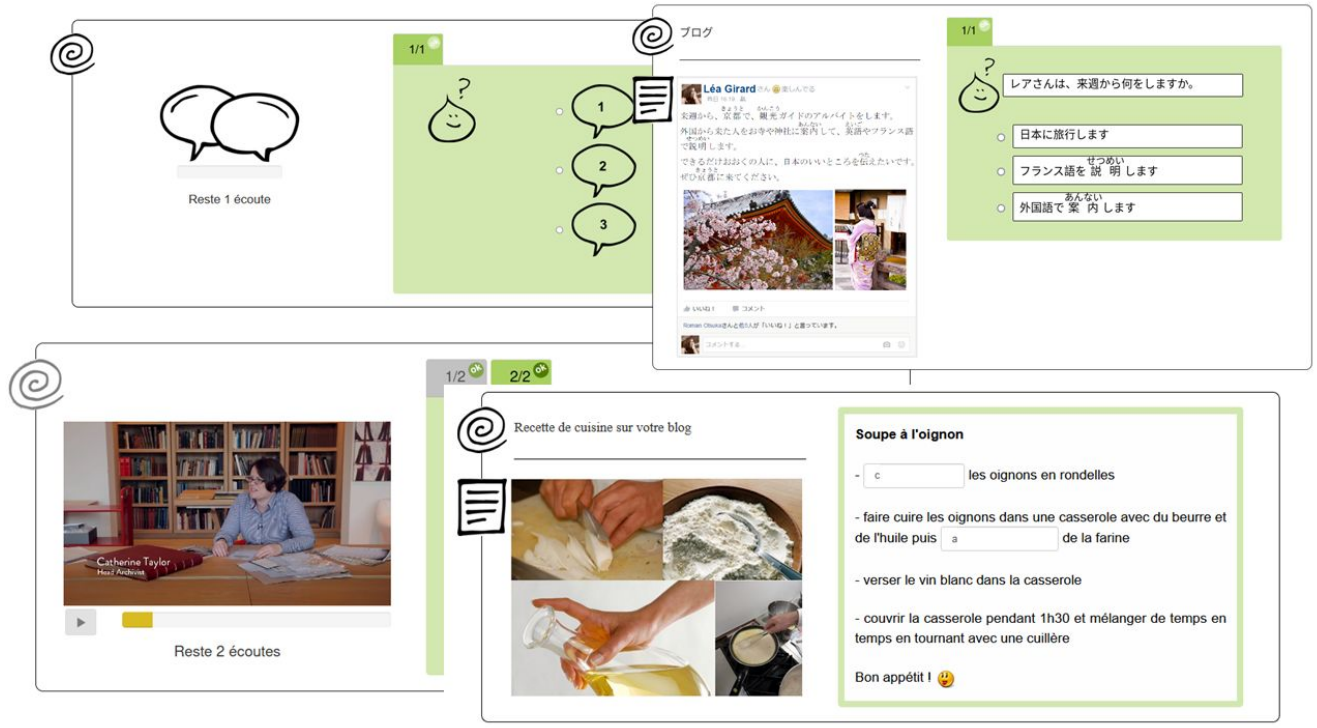


More info here: https://self.univ-grenoble-alpes.fr/

SELF - Formative Language assEssment System

SELF INNOVA LANGUES

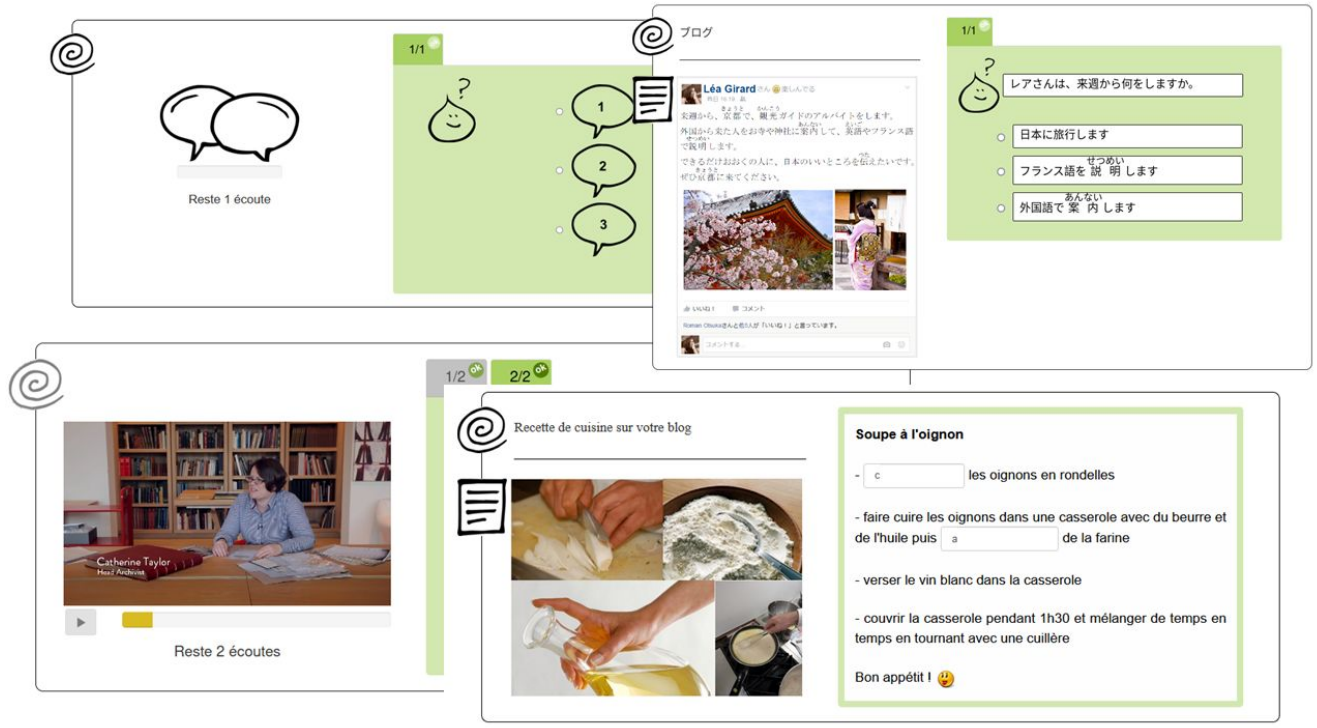
- Placement test
  - Listening
  - Reading
  - Writing



SELF - Formative Language assEssment System

SELF INNOVA LANGUES

- Placement test
  - Listening
  - Reading
  - Writing
  - Speaking



SELF - Speaking modules (English & French)

Developing a Speaking assessment module for SELF placement tests
May 2025 - Nov. 2026

Alex Carr
Instructional Design
SELF English

Alireza Pournouri
Instructional Design
SELF English

Pinxun Huang
Speech Processing

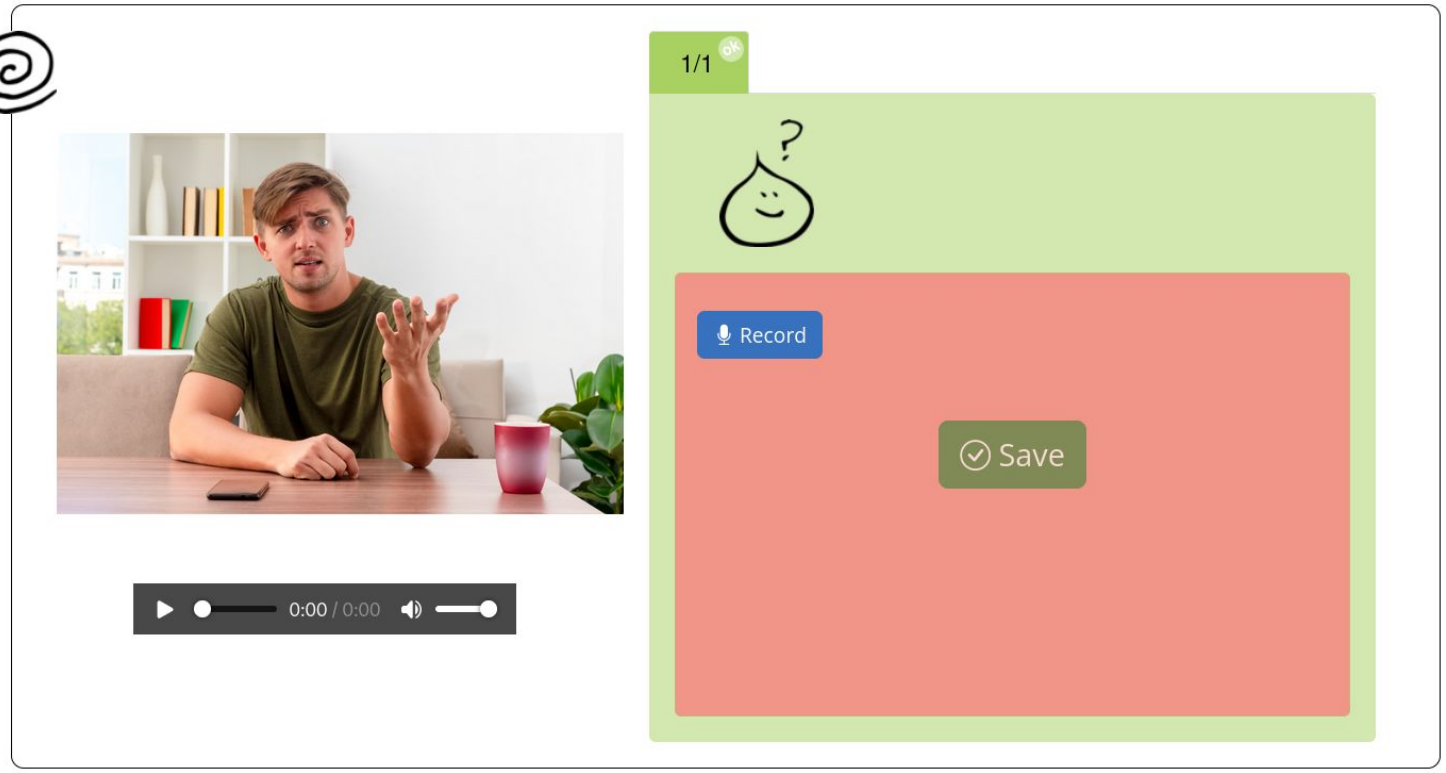
Eli Stafford
Speech Processing

Anne-Cécile Perret
Coordinator
SELF Français

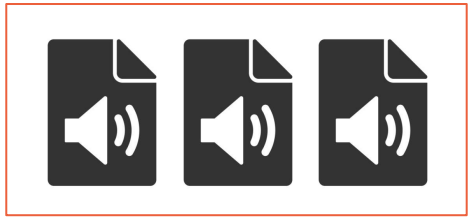
Sylvain Coulange
Technical and Scientific Coordinator

SELF - Speaking modules (English & French)

Developing a Speaking assessment module for SELF placement tests
May 2025 - Nov. 2026



student = Marine Dupont

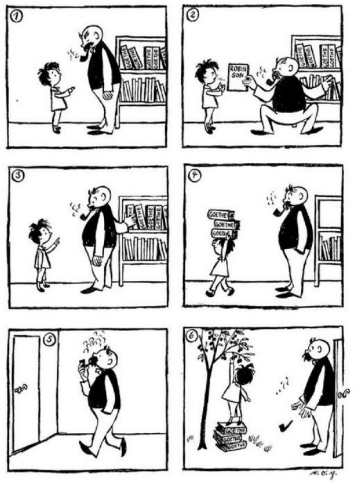


Speaking score: B1

CO + CE + EEC + **PO** = “On the way to B2.1”

Task Examples

You've just met Jonathan, describe him to your friend.



Take a look at these pictures and tell the story.

Send a voice message to your friend in response. Describe the activities you did. You can use the pictures below if you want.

0:00 / 0:00

"Excuse me... Hi! We're doing a report on what people think of what the Prime Minister said last night about climate change. He said quote "The government should not be held responsible for climate change. It's the individual citizens that must be held accountable."
Do you agree with this statement?"

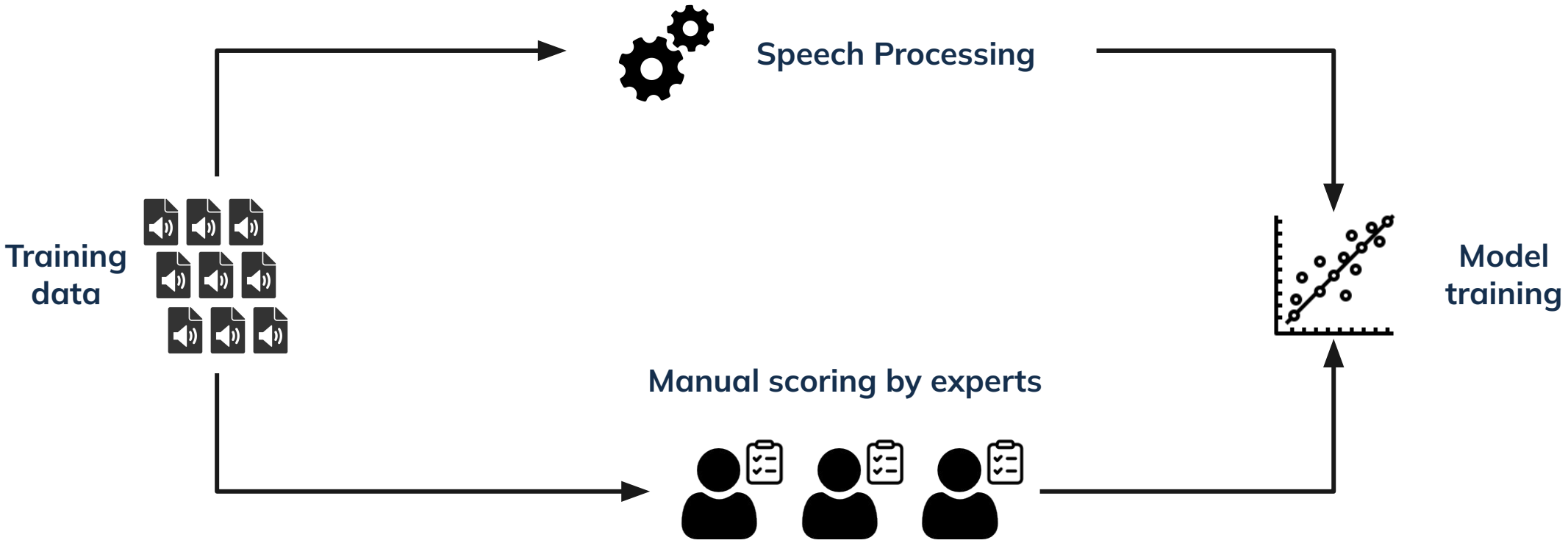
Answer the reporter's question and explain your reasons.

0:00 / 0:00

"Hey. So, I talked to my parents yesterday and they think it's not a good choice for me to go to art school next year. They said I should choose a profession that would be stable and that I should go to medical school to become a doctor like them. But I don't know, painting is my biggest passion, you know? So I don't really know what to do right now. What do you think?"

Help Alex making a choice, giving her advice about the pros and cons of both choices.

Training a Score prediction model



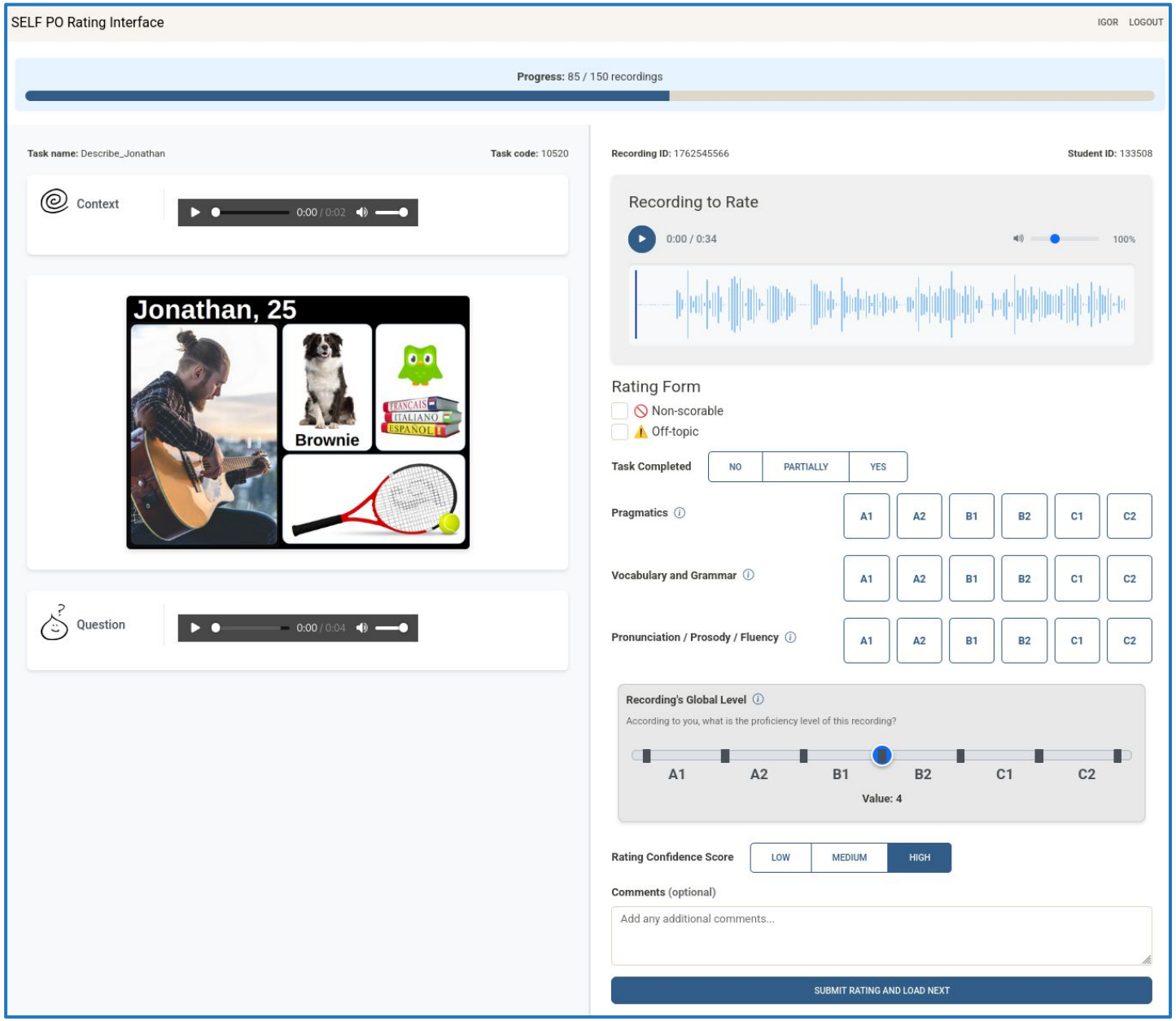
Using the model to predict scores on new data

Manual Scoring by Experts

- 12 raters recruited
- 150 recordings / rater

- 600 recordings
- ~3 ratings / recording
- = 1800 ratings

Human Scoring Interface



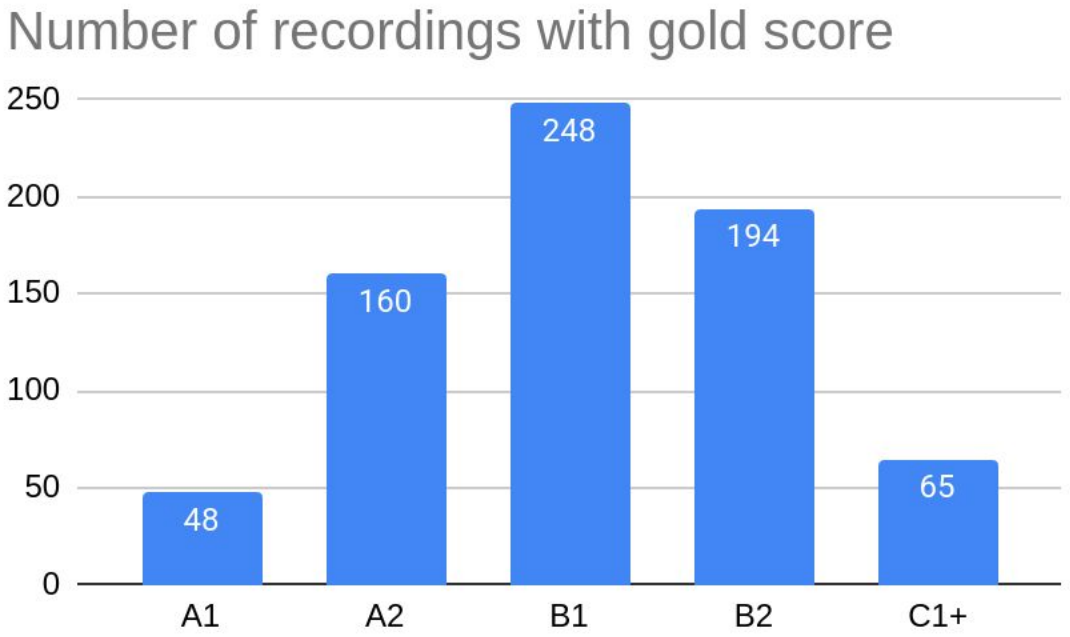
Final Dataset (Recordings with Human Scores)

~5000 recordings collected (~45h)

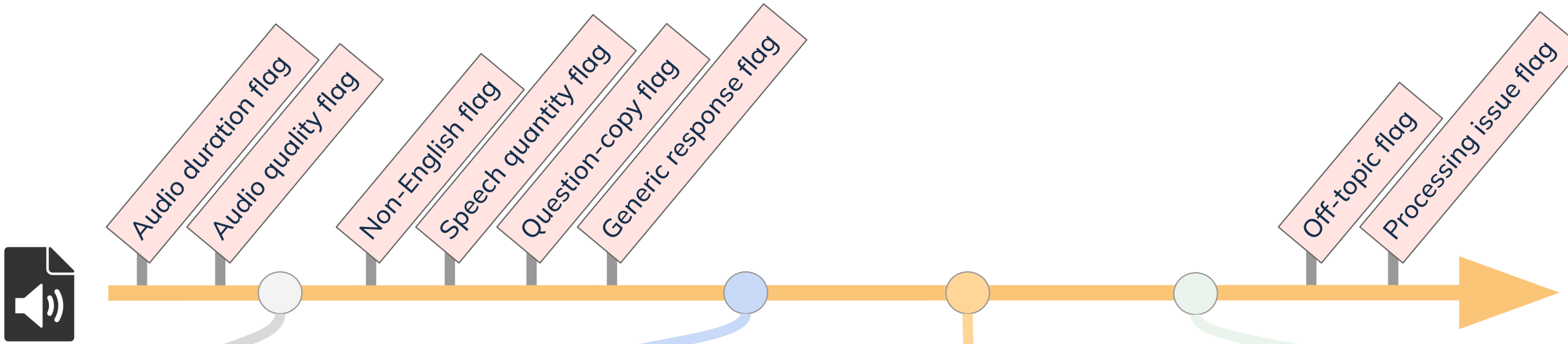
1800 ratings collected on 700 recordings

Filtering + Normalization

Total unique recordings: 715
Total unique speakers: 312 (~8h)



Evaluation Criteria (for English)



Self-supervised Features

23.21, 204.06, 2.1, 302.19, 294.0, 293.1, 405.11, 1.345, 30.191, 201.30, 291.1, 304.19, 11.5, 450.1, 59.5, 191.56, 234.5, 500.56, 505.49, 219.4, 19.4, 290.4, 304.19, 293.1, 405.11, 1.345, 30.191, 505.2, 18.01, 34.4, 292.18, 38.11, 294.11, 2.1, 234.5, 203.92, 192.3...

Surface Features

- Speech rate
- Silence ratio
- Pause frequency
- Syntactic Pause Ratio
- Stress Contrast
- ASR confidence
- ...

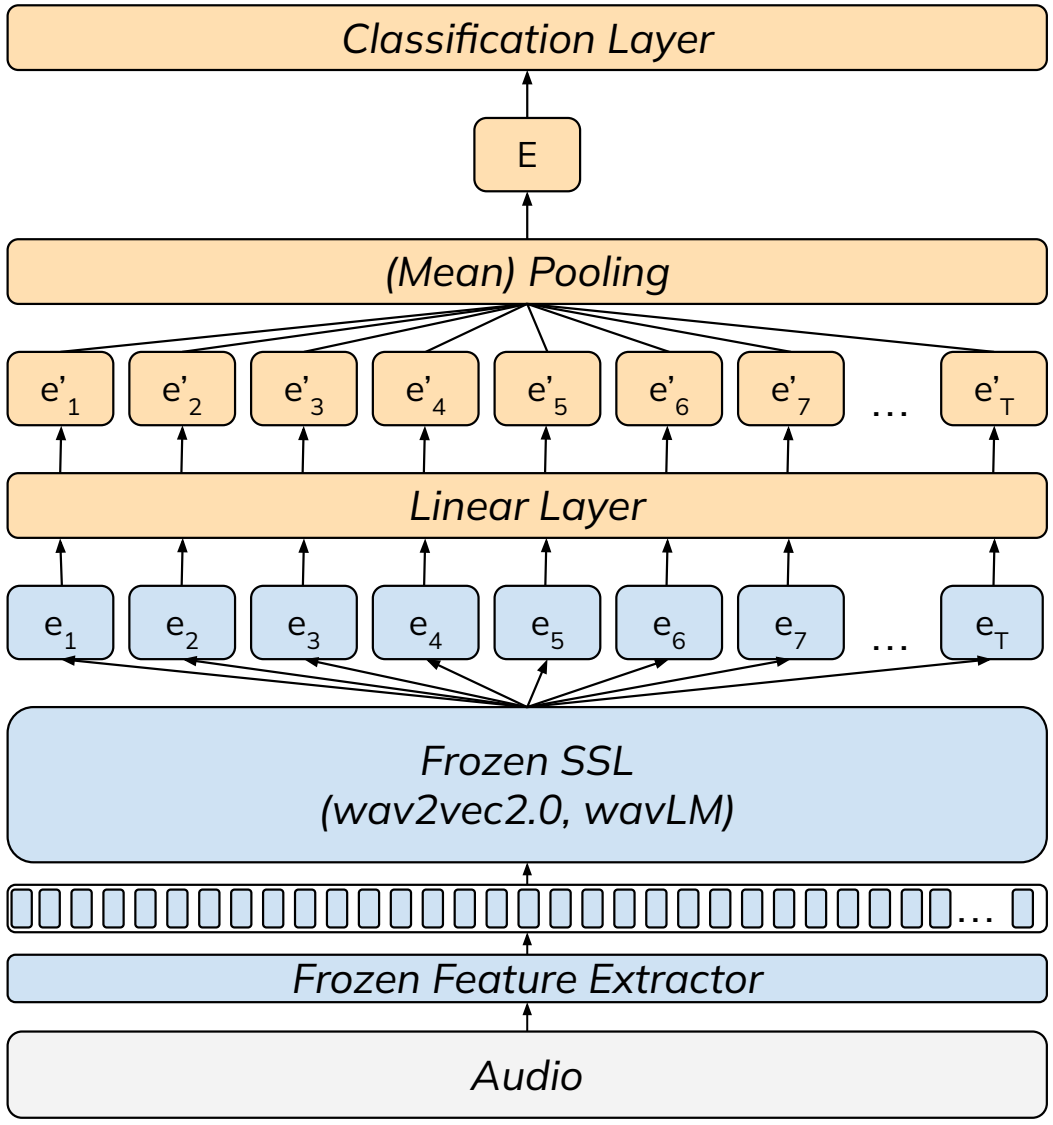
Linguistic Features

- Lexical diversity
- Lexical frequency
- LLM perplexity
- Syntactic complexity
- Syntactic similarity with ref. recordings
- ...

Discourse Features

- Task completion score
- Coherence markers
- Topic consistency
- Idea density
- ...

Using Pre-Trained Speech Embeddings



Pre-Trained Embeddings

- Models like wav2vec2 and wavLM, which produce sequences of vectors representing slices of audio, trained on unlabeled specs
- Low explicability

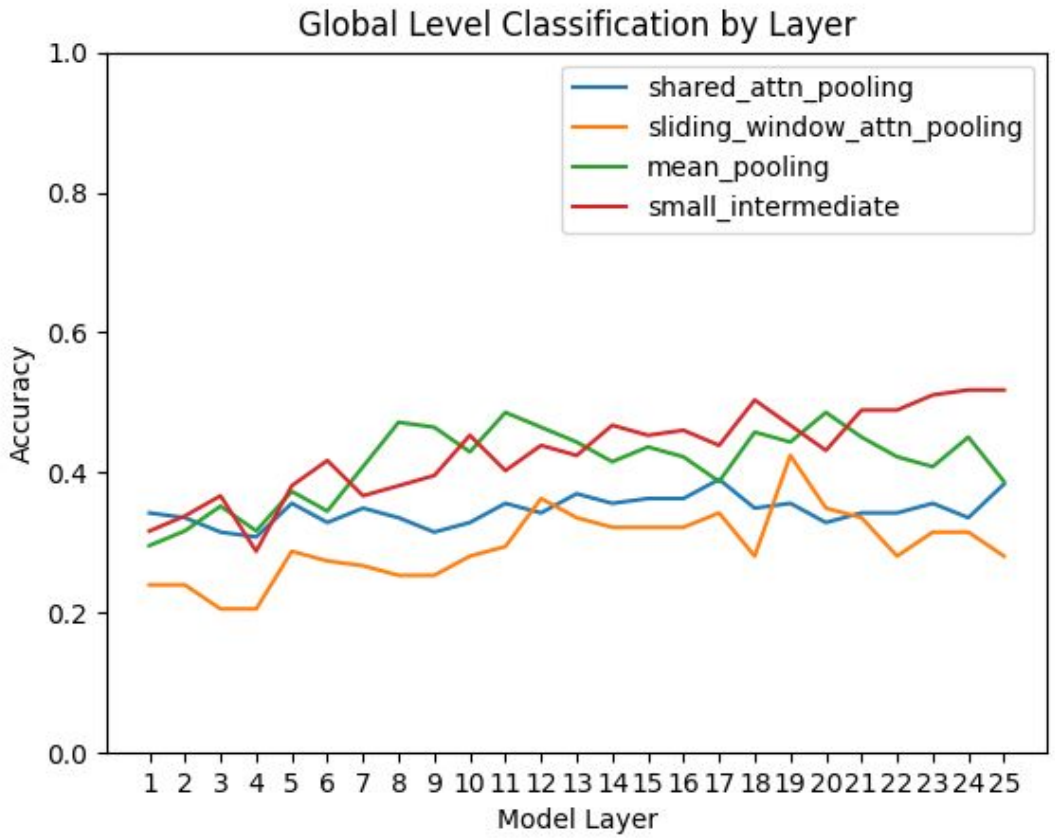
End-to-End Task Architecture

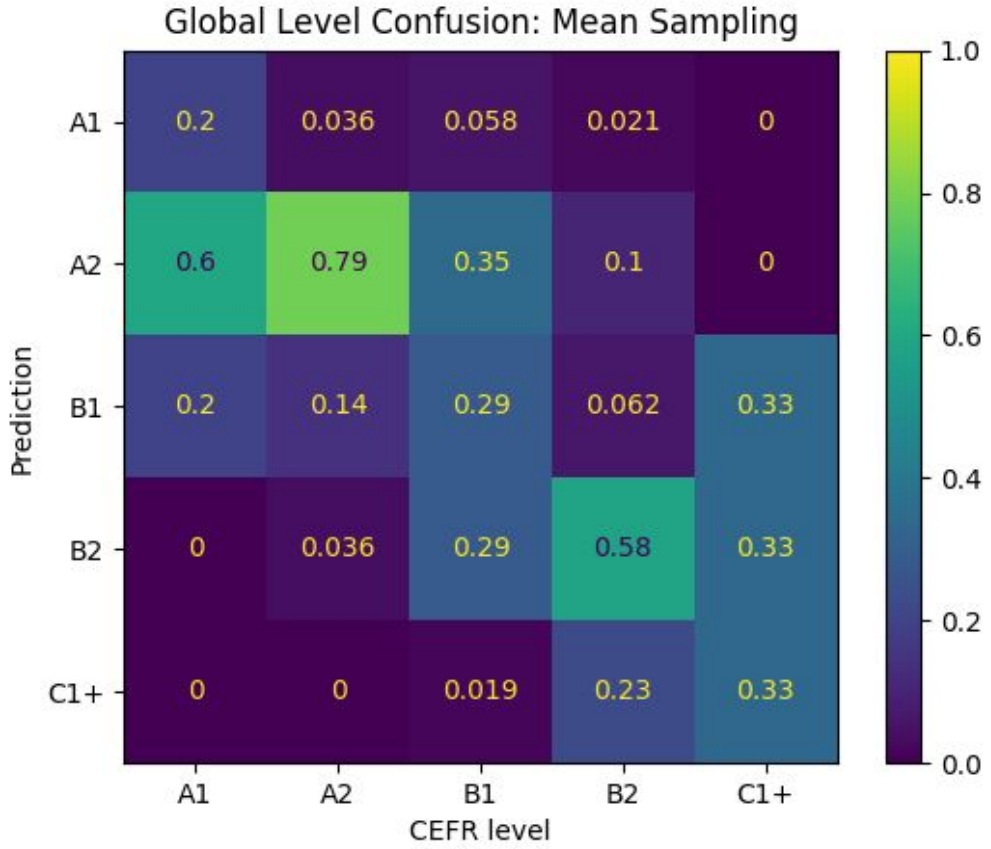
- Frozen SSL + (pooling +) classifier head
  - Leverage SSL embeddings with lightweight classifier layers
- Utterance level classification (~Emotion Recognition)

Implications of pooling

- Lack of generic sentence-vector-like encoders for speech, usually task specific
- Unlikely to retain word level linguistic information (pronunciation, stress, semantics)
- Proven performance on non-linguistic tasks (Emotion recognition, speaker recognition)

Global Level Results - wavlm-large





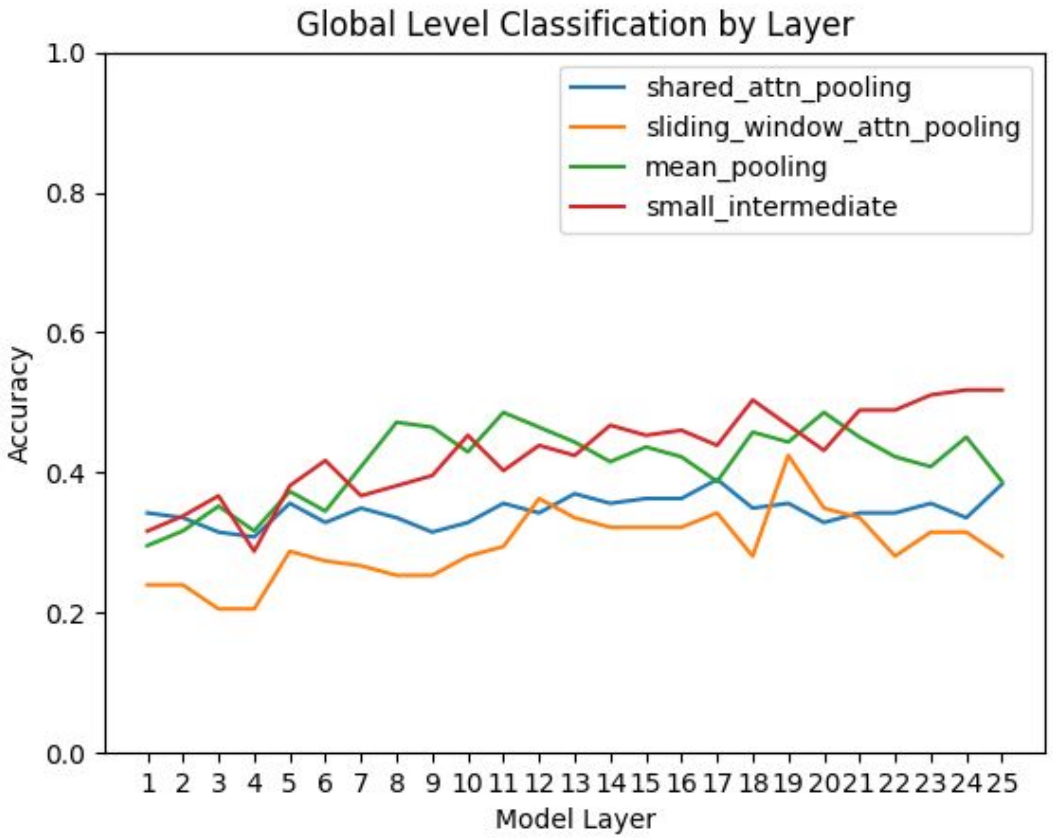
Best Results: (naive baseline: 39.5%)

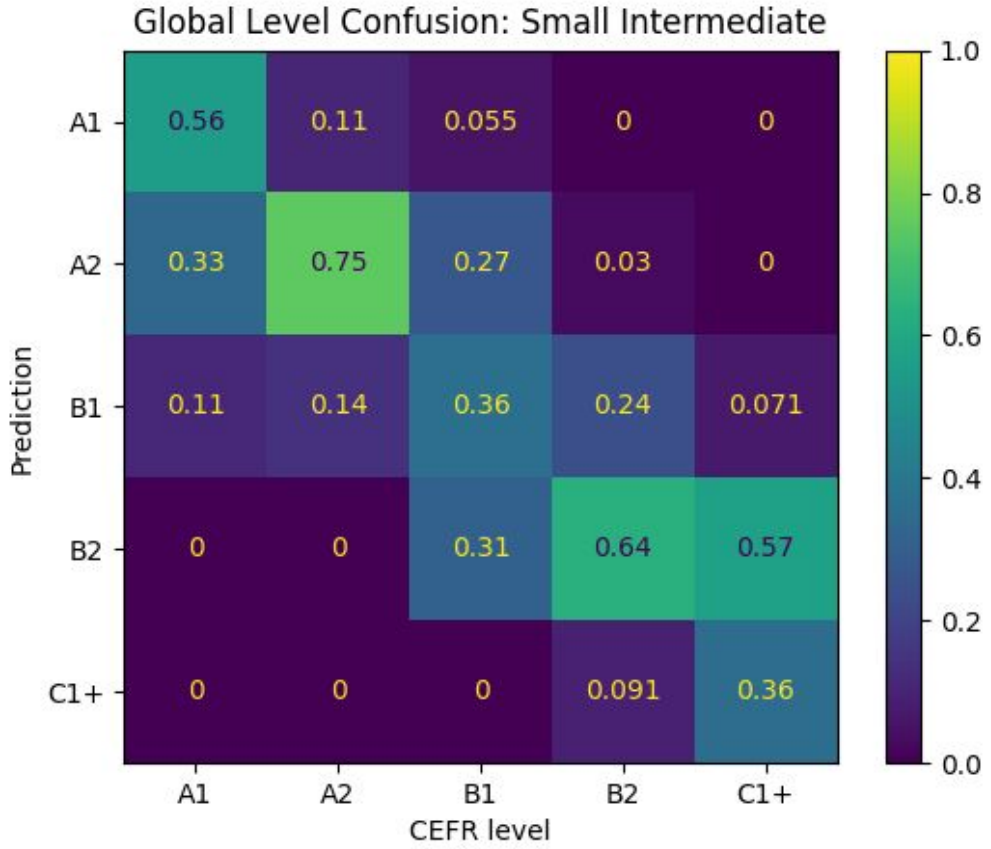
Layer 11 with weighted sampling

Accuracy: 48.5% Macro (unweighted) F1: .397

W/ ± 1-level: 82.0% Train Acc : ~85%

Global Level Results - wavlm-large



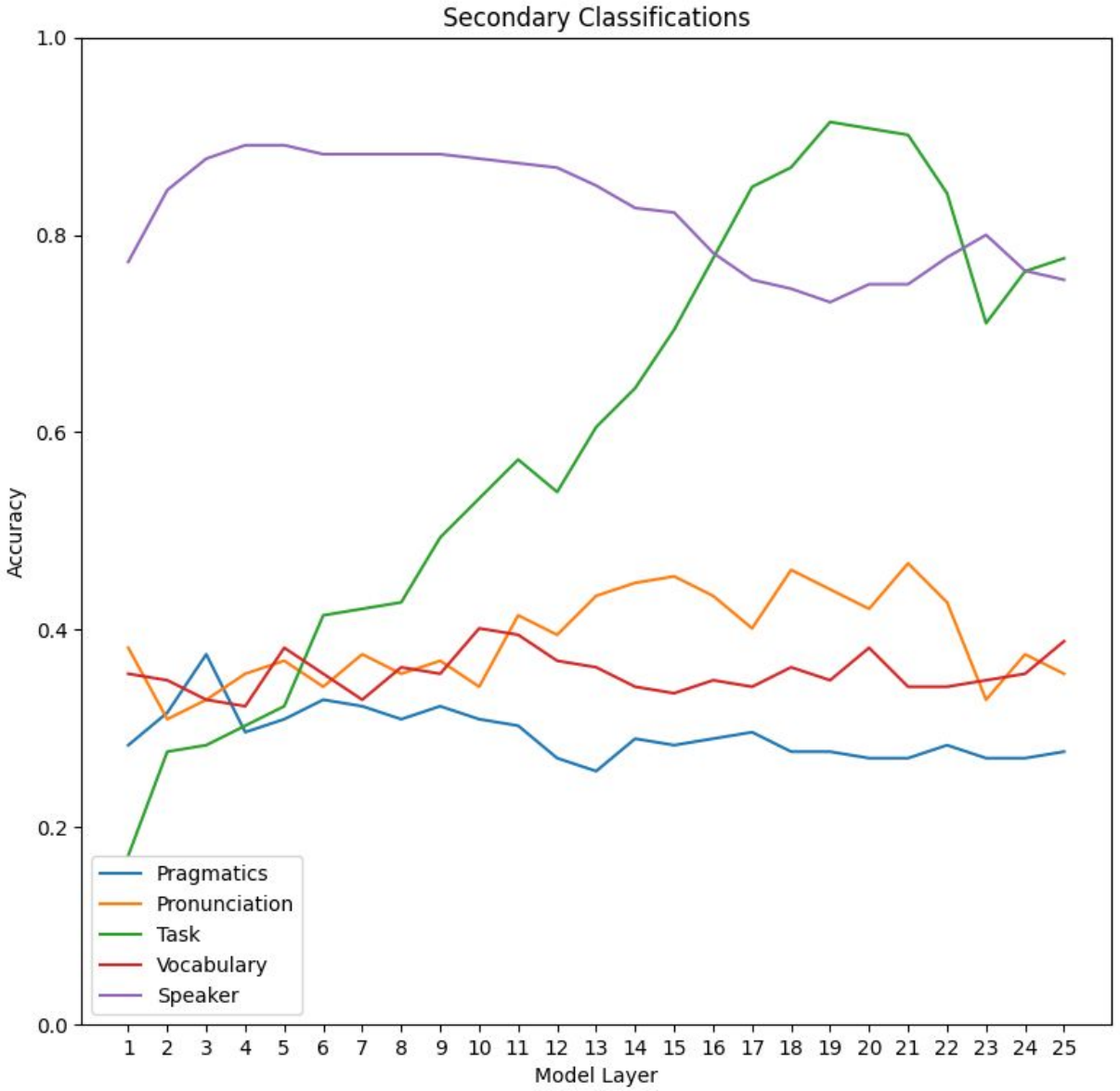


Best Results:

Layer 25 with small intermediate

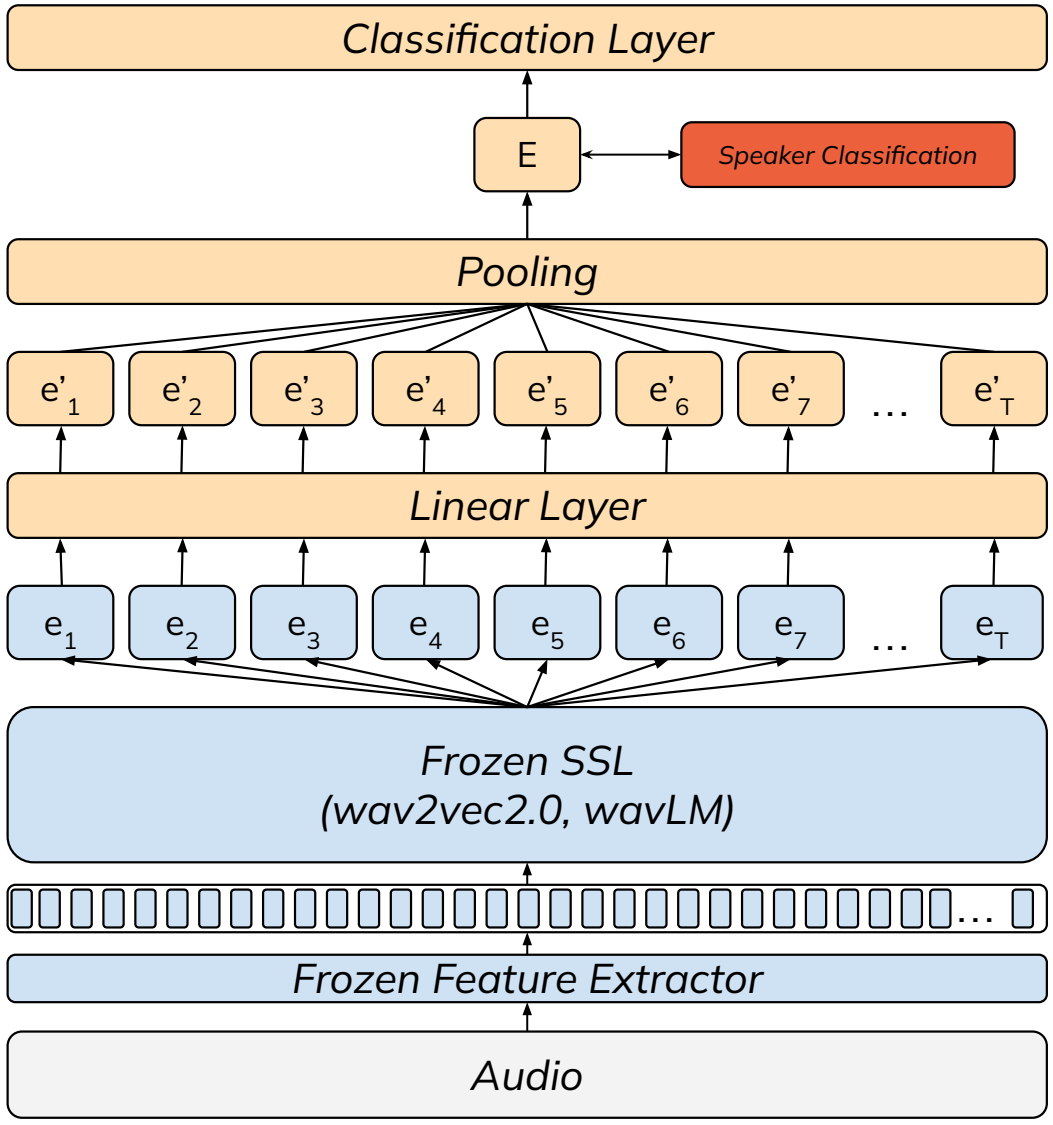
Accuracy:	51.8%	Macro (unweighted) F1: .511
W/ ± 1-level:	89.4%	Train Acc: ~75%

Secondary Characteristics



- Stronger predictions on pronunciation than pragmatics or vocabulary
- No significant strengths compared to the score based classifier (with this architecture)
- Much better at speaker identification
  - Can rely on this for level identification
  - With many single example speakers, this ability can hurt generalization
- Task identification
  - Keyword identification

Adversarial Training



Penalize intermediate representations which allow for speaker classification

Algorithm

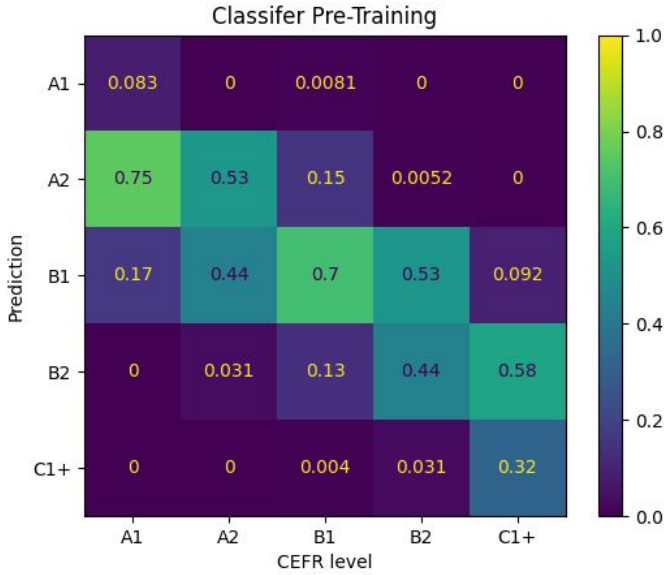
- Pre-train base classification
- Pre-train a speaker classification head on the intermediate embedding
- Perform adversarial training (Hardt et al, 2016)
- Fine-tune classification layer on new intermediate embeddings

Ensures both Equality of Opportunity (as described in Hardt et al) and better generalization across speakers

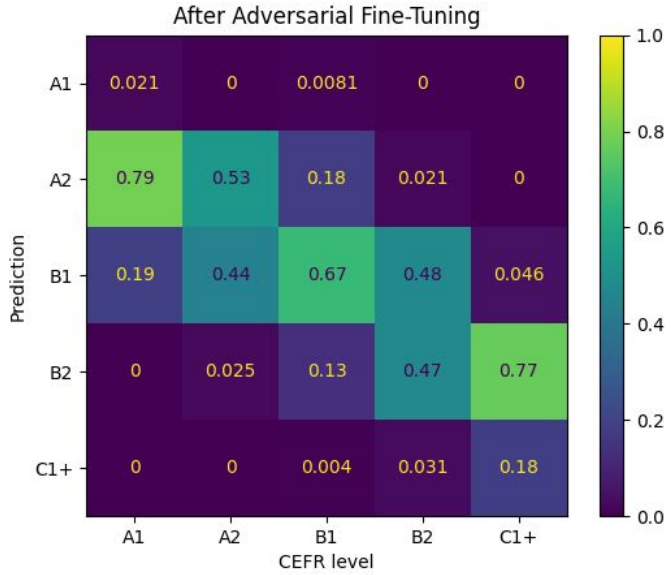
Preliminary Results

10-fold cross validation
wavLM layer 25
Intermediate embedding of size 64
Heavily parameter dependant

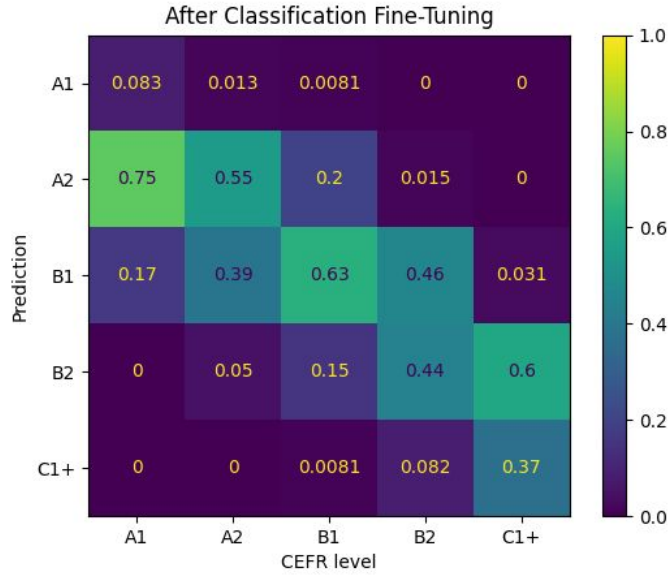
No change in generalization power
Strong theoretical guarantees



Accuracy: 51.6%
Macro F1: .436

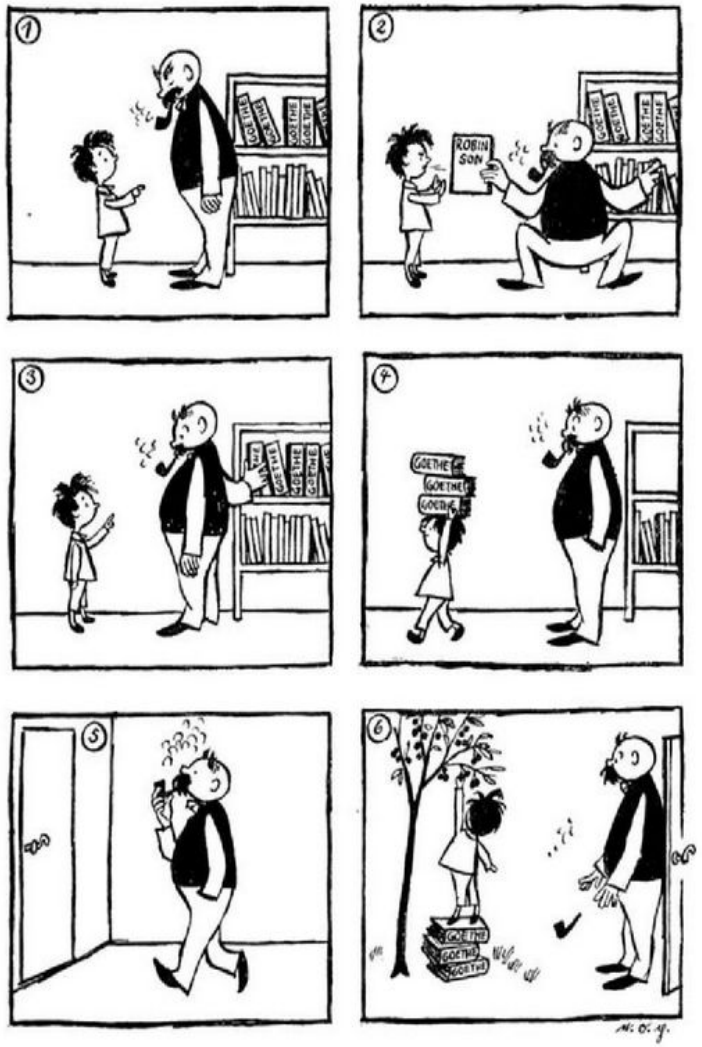


Accuracy: 49.8%
Macro F1: .378



Accuracy: 50.2%
Macro F1: .428
Train Acc: ~63%

On-Topic or Off-Topic ?



Task:

Telling a story. Take a look at these pictures and tell the story.

A) Student answer:

A child asks some books to a man to put them in the garden and reach a branch of a tree.

B) Student answer:

A kid walks up to a man asking for a book the man hands him what looks like a comic book the kid refuses saying he wants bigger books the man surprised gives him three books when he walks out he sees that the kid was only using them to reach a higher place.

C) Student answer:

It's a children and his parents and the parents give some knowledge to his children and the children use this knowledge to arrive at his objective of thinking.

Approaches to Off-Topic Detection

1) GPT2 Likelihood

- Convert Task Question and Image into Text Prompt
- Measure how likely the Student's Answer is given the Prompt for a Large Language Model-GPT2

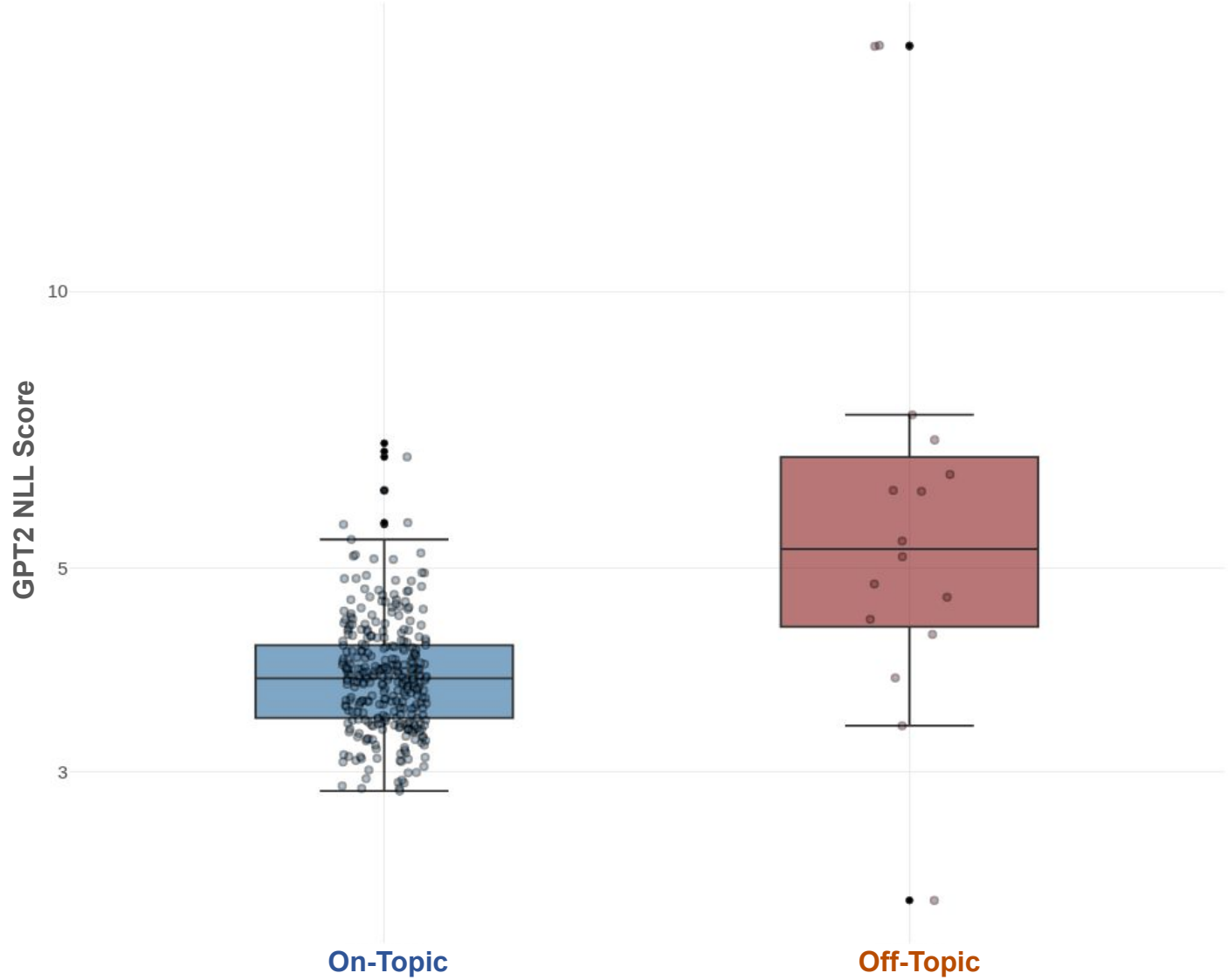
2) Sentence-BERT Semantic Similarity

- Measure the Semantic Similarity between the Prompt and the Student's Response

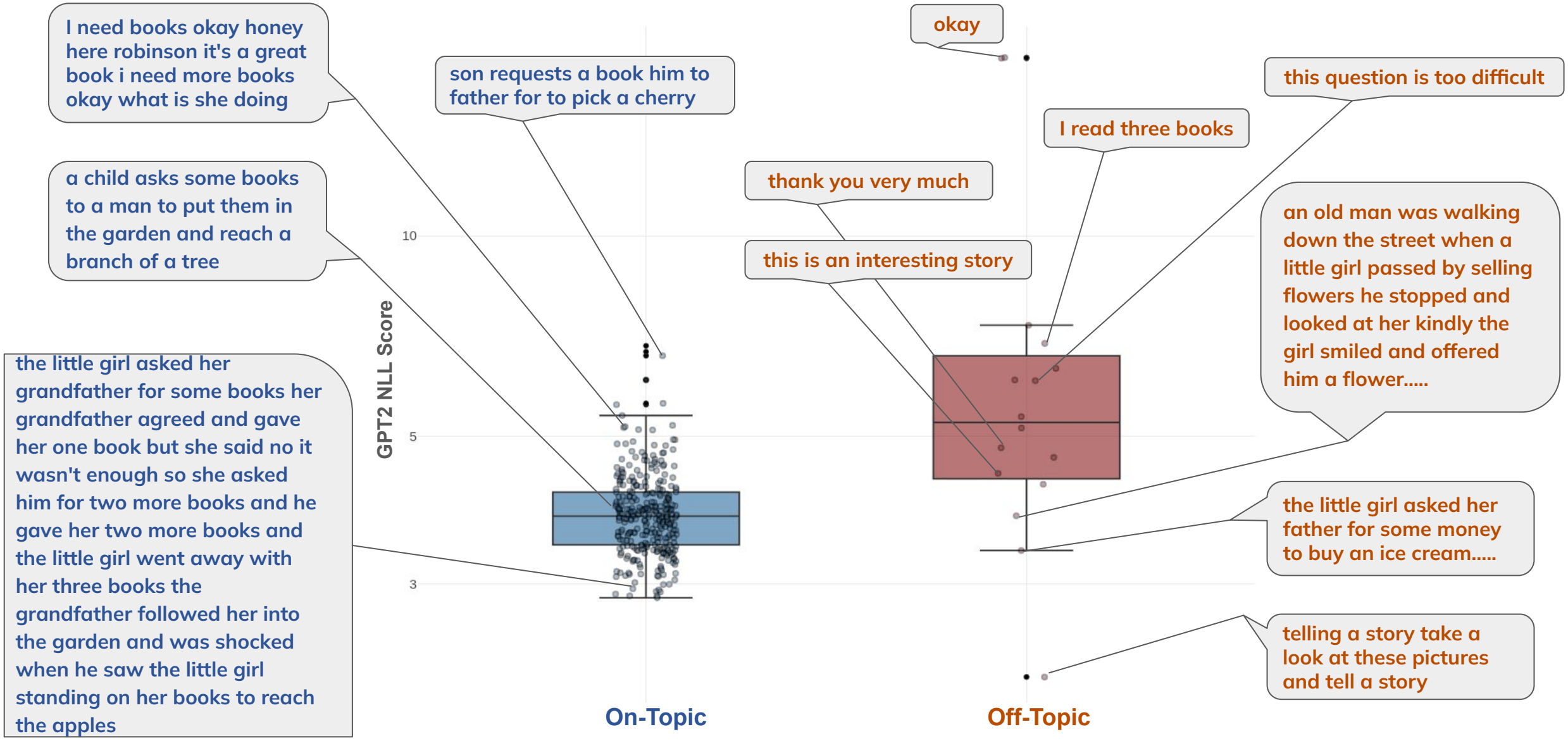
3) LLM as a Judge

- Give Task Question and Student's Response to LLM-Mistral
- Use LLM to judge Response Relevance

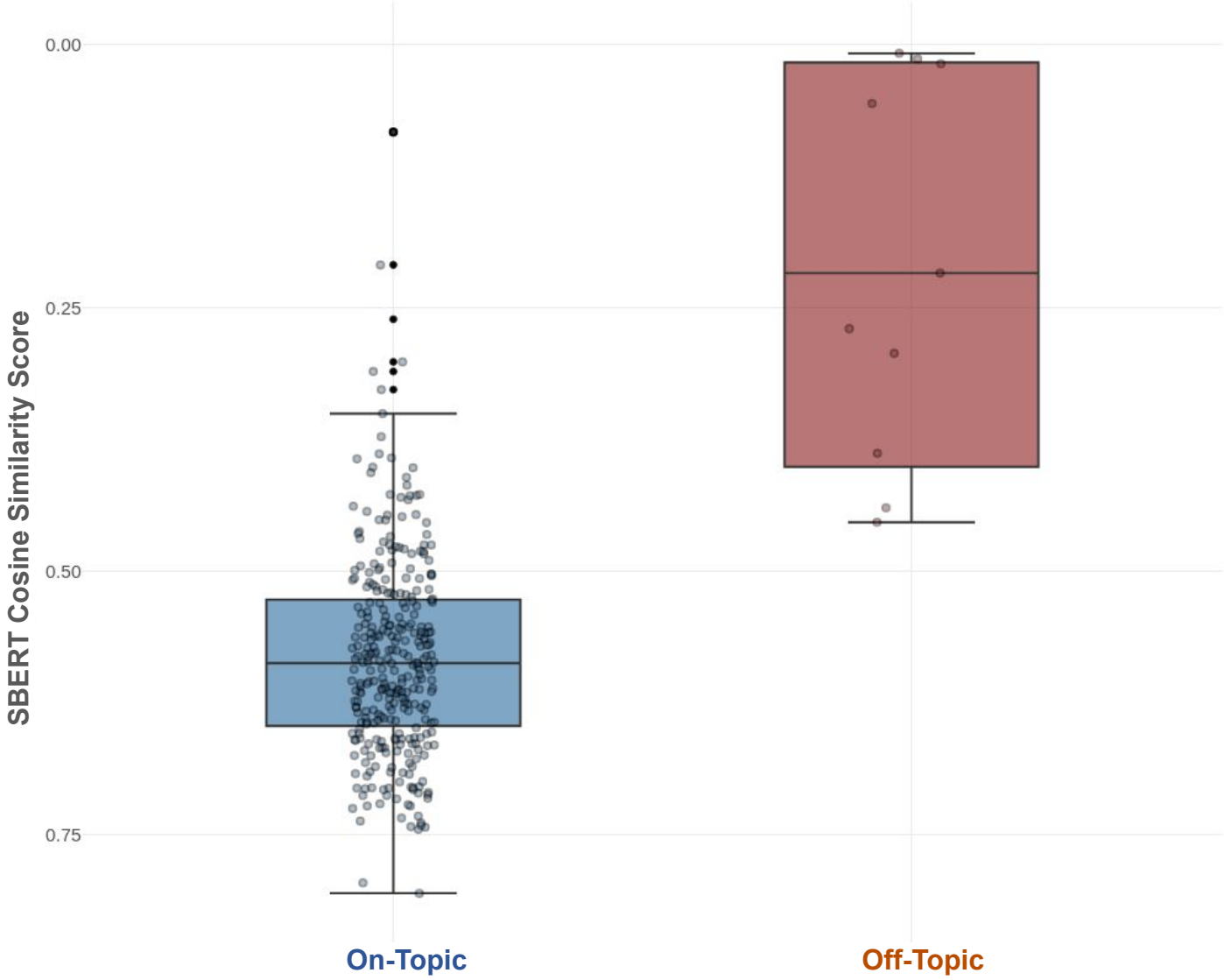
1) GPT2 Likelihood (how likely the response given the prompt)



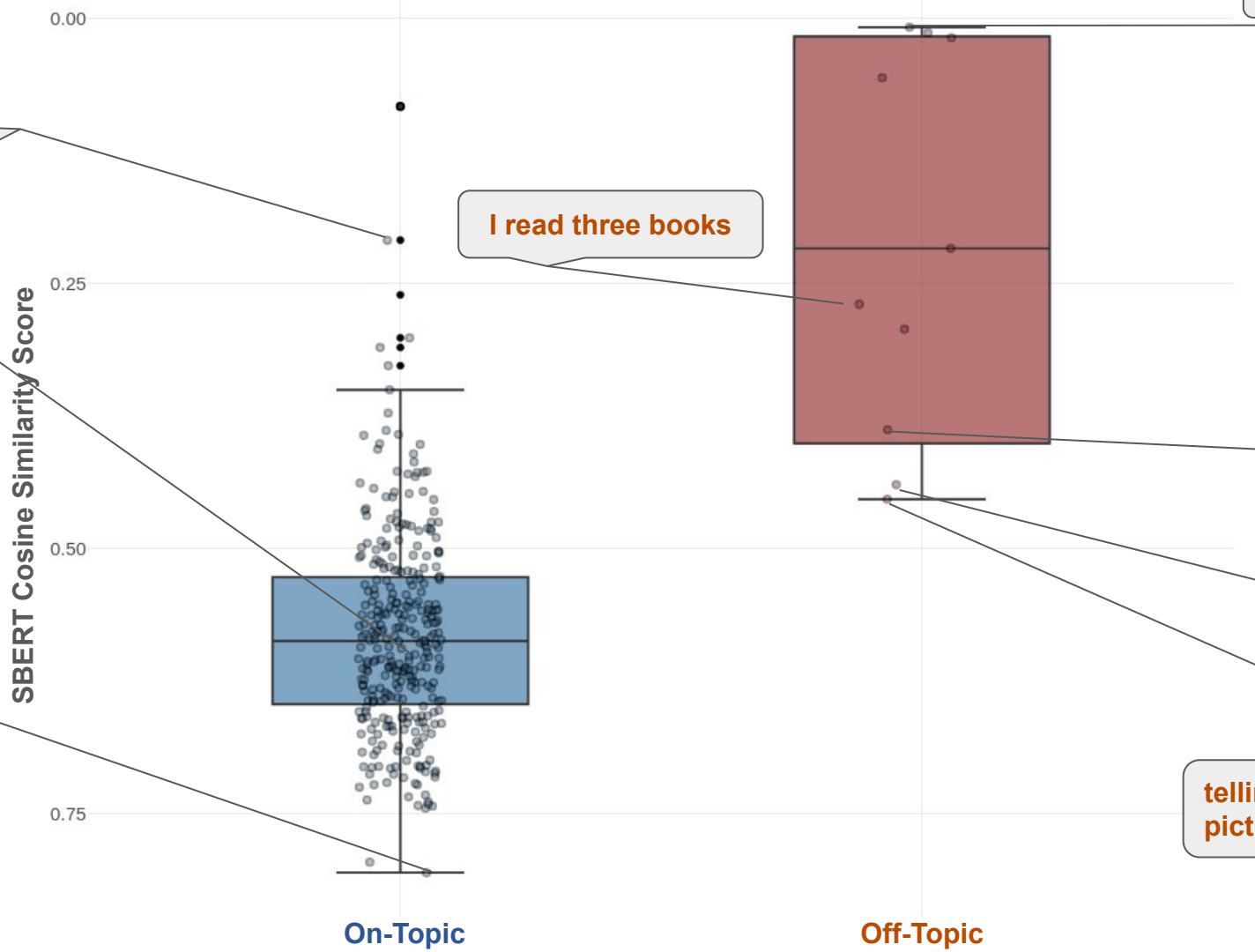
1) GPT2 Likelihood (how likely the response given the prompt)



2) Sentence-BERT Semantic Similarity between Prompt and Answer



2) Sentence-BERT Semantic Similarity between Prompt and Answer



I need books okay honey here robinson it's a great book i need more books okay what is she doing

a child asks some books to a man to put them in the garden and reach a branch of a tree

in the story you can see there is a children who go to the library with an old librarian and he's asking for some book and after this the old man look happy because the children who want to read and culture them but after the children go with the three very big books the old man sees the children take those three books only to use them like a stair and up to take a fruit from a tree and he was surprised and shocked

I read three books

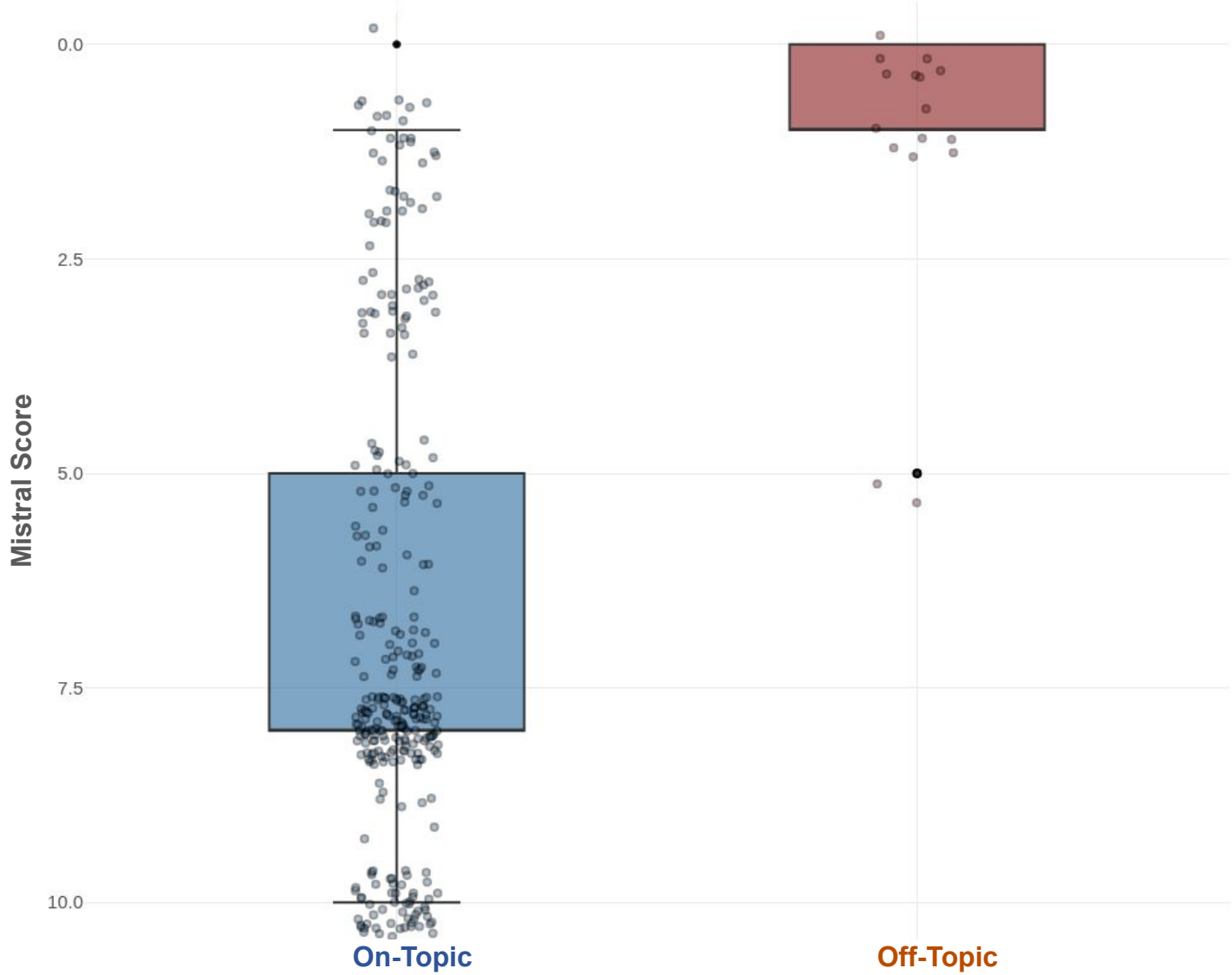
this question is too difficult

an old man was walking down the street when a little girl passed by selling flowers he stopped and looked at her kindly the girl smiled and offered him a flower…..

the little girl asked her father for some money to buy an ice cream.....

telling a story take a look at these pictures and tell a story

3) LLM as a Judge

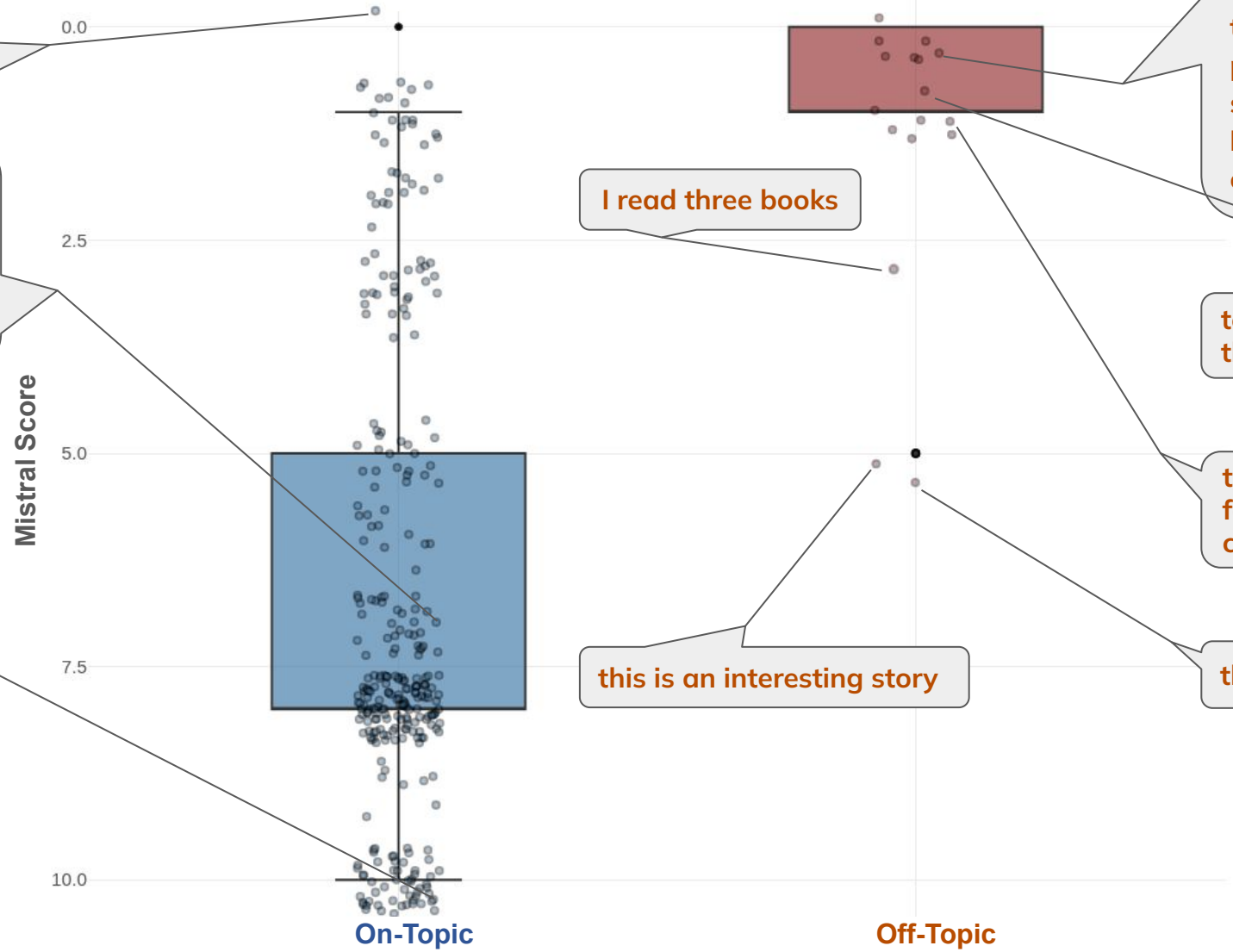


3) LLM as a Judge

he's asking to her father if she can take books so his father takes give her a book but he he understands that

the kid is asking a book from his father but he wants more so his father give him three books later his father went outside and then he see the kids using the books for climbing on a tree

the little girl asked her grandfather for some books her grandfather agreed and gave her one book but she said no it wasn't enough so she asked him for two more books and he gave her two more books and the little girl went away with her three books the grandfather followed her into the garden and was shocked when he saw the little girl standing on her books to reach the apples



I read three books

this is an interesting story

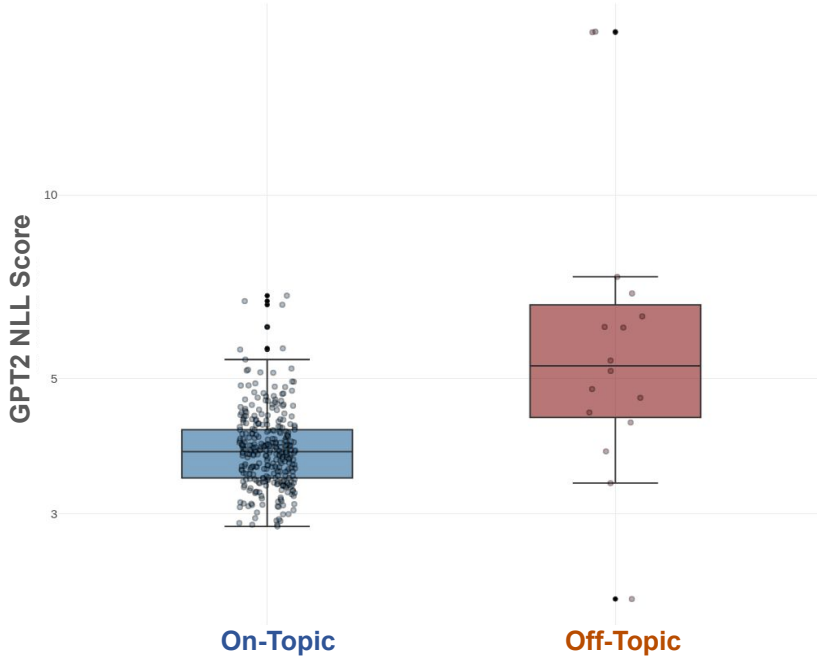
an old man was walking down the street when a little girl passed by selling flowers he stopped and looked at her kindly the girl smiled and offered him a flower.....

telling a story take a look at these pictures and tell a story

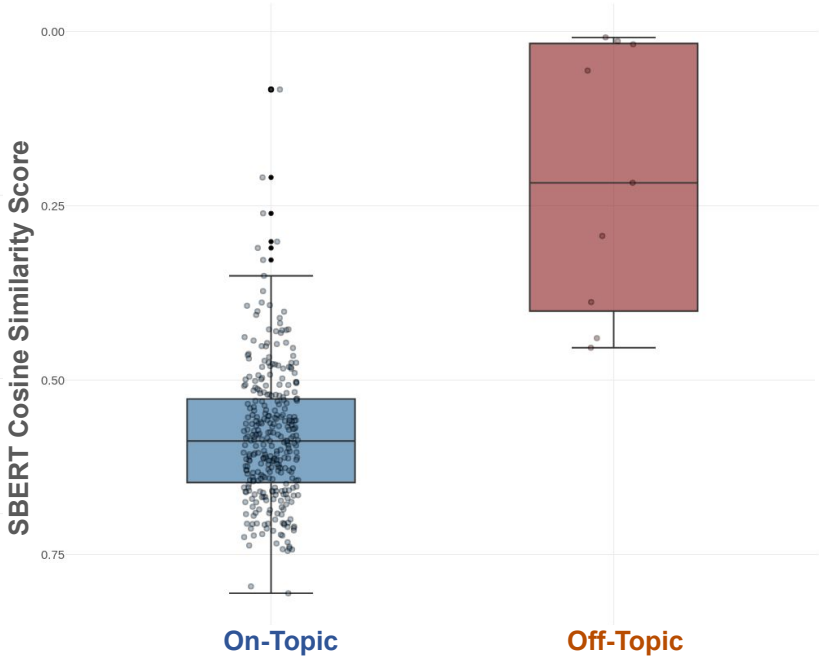
the little girl asked her father for some money to buy an ice cream.....

thank you very much

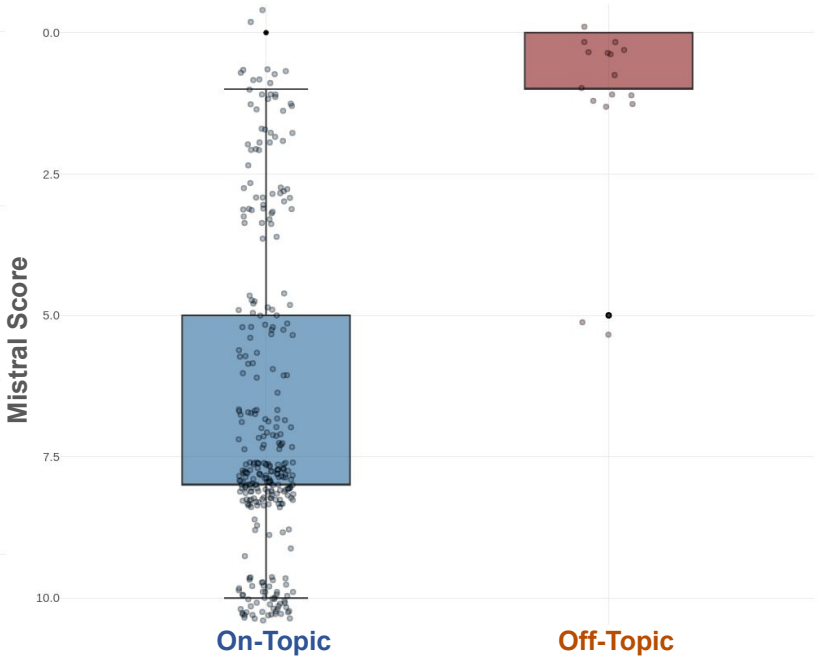
Comparison across GPT2 likelihood, Sbert similarity and LLM judgement



F1 score: 0.722

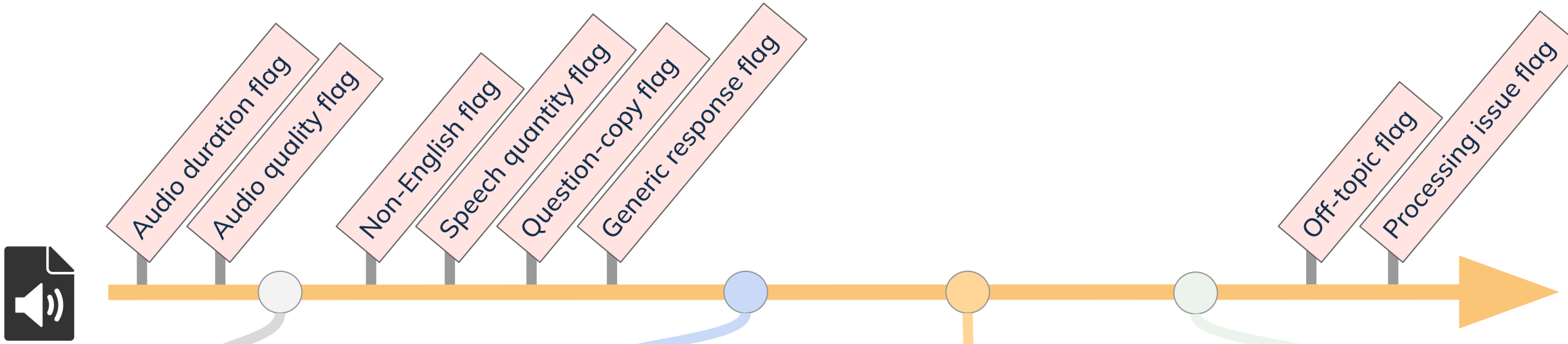


F1 score: 0.909



F1 score: 0.75

Evaluation Criteria (for English)



Self-supervised Features

23.21, 204.06, 2.1, 302.19, 294.0, 293.1, 405.11, 1.345, 30.191, 201.30, 291.1, 304.19, 11.5, 450.1, 59.5, 191.56, 234.5, 500.56, 505.49, 219.4, 19.4, 290.4, 304.19, 293.1, 405.11, 1.345, 30.191, 505.2, 18.01, 34.4, 292.18, 38.11, 294.11, 2.1, 234.5, 203.92, 192.3...

Surface Features

- Speech rate
- Silence ratio
- Pause frequency
- Syntactic Pause Ratio
- Stress Contrast
- ASR confidence
- ...

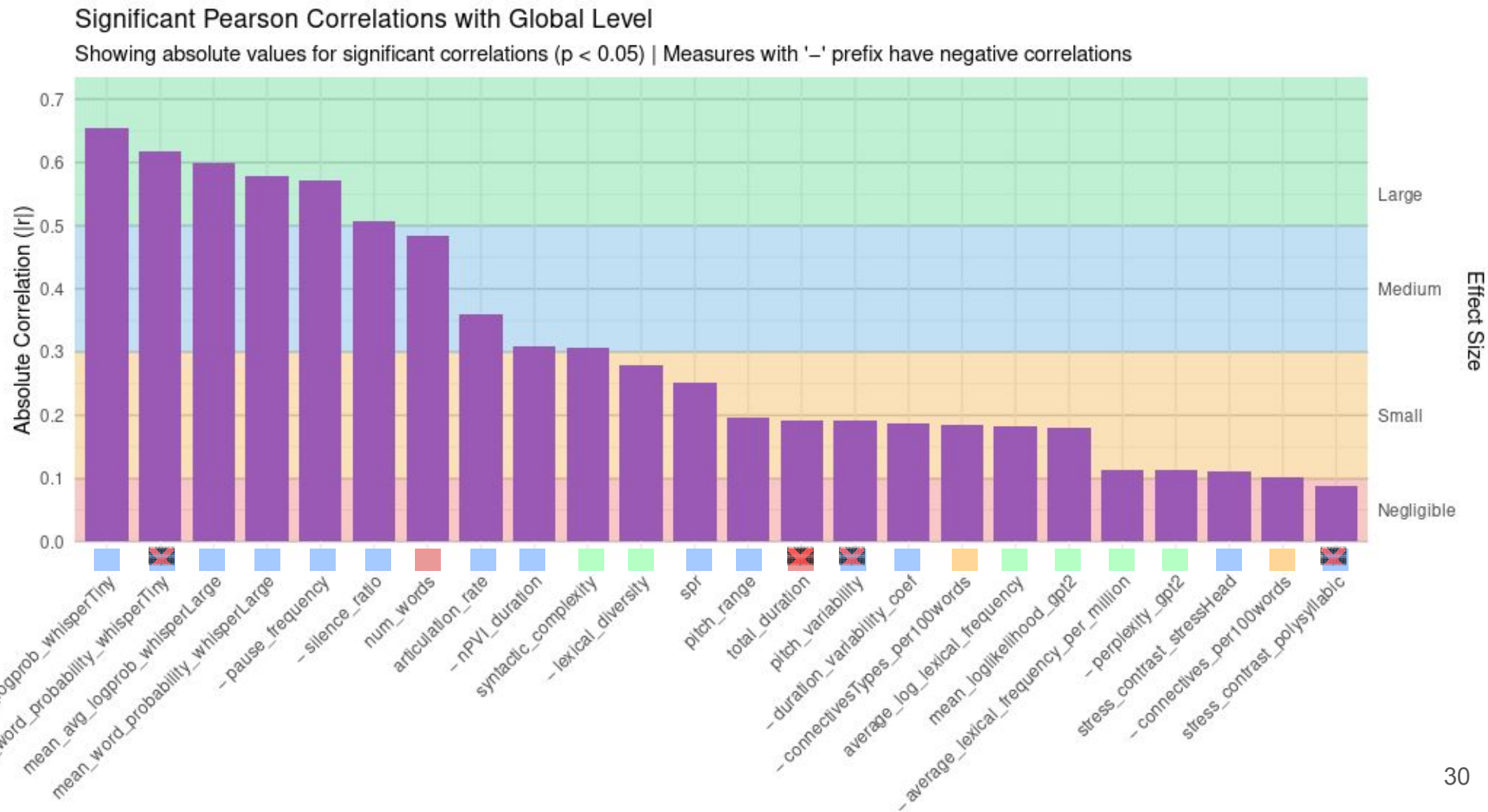
Linguistic Features

- Lexical diversity
- Lexical frequency
- LLM perplexity
- Syntactic complexity
- Syntactic similarity with ref. recordings
- ...

Discourse Features

- Task completion score
- Coherence markers
- Topic consistency
- Idea density
- ...

Correlations between Features and Human Scores



- Surface
- Linguistic
- Discourse
- Added features
- Removed predictor after collinearity check

Correlations between Features and Human Scores

Category	Sub-category	Measure	r	Effect Size	Sig	Description
surface	Pronunciation	mean_avg_logprob_whisperTiny	0.654	Large	***	Mean segment-level log-probability score from Whisper tiny.en
surface	Pronunciation	mean_word_probability_whisperTiny	0.618	Large	***	Mean word-level confidence score from Whisper tiny.en
surface	Pronunciation	mean_avg_logprob_whisperLarge	0.599	Large	***	Mean segment-level log-probability score from Whisper large
surface	Pronunciation	mean_word_probability_whisperLarge	0.577	Large	***	Mean word-level confidence score from Whisper large
surface	Fluency	pause_frequency	-0.57	Large	***	Number of pauses / number of words
surface	Fluency	silence_ratio	-0.508	Large	***	Total silence duration / total audio duration
surface	Meta	num_words	0.485	Medium	***	Absolute number of words
surface	Fluency	articulation_rate	0.359	Medium	***	Number of words / (duration - pauses)
surface	Prosody	nPVI_duration	-0.309	Medium	***	Rhythm (Normalized Pairwise Variability Index for vowel duration)
surface	Fluency	spr	0.251	Small	***	Syntactic Pause Ratio (Coulange & de Jong, 2025)
surface	Prosody	pitch_range	0.196	Small	***	5th and 95th percentile of semitone pitch values after bandpass filtering
surface	Meta	total_duration	0.193	Small	***	Duration in seconds
surface	Prosody	pitch_variability	0.193	Small	***	Standard deviation of semitone pitch values after bandpass filtering
surface	Prosody	duration_variability_coef	-0.187	Small	***	Coefficient of variation for vowel durations (SD/Mean, Dellwo, 2006)
surface	Prosody	stress_contrast_stressHead	0.112	Small	**	S score for words with initial and medial stress only
surface	Prosody	stress_contrast_polysyllabic	0.088	Negligible	*	S score for polysyllabic words only (Coulange, 2025)
surface	Prosody	content_function_stress_contrast	-0.055	Negligible		Stress contrast between content and function words: mean of expected primary stressed syllable of content words / mean of syllables of function words
surface	Prosody	stress_contrast	-0.047	Negligible		S score for all words (Coulange, 2025)
surface	Prosody	nPVI_f0	0.019	Negligible		Normalized Pairwise Variability Index for vowel F0 (pitch pattern metric)
linguistic	Syntax	syntactic_complexity	0.308	Medium	***	Mean number of closing and opening constituents between adjacent words
linguistic	Vocabulary	lexical_diversity	-0.28	Small	***	Type-token ratio
linguistic	Vocabulary	average_log_lexical_frequency	0.182	Small	***	Lexical types average log word frequency based on the COCA corpus
linguistic	Syntax	mean_loglikelihood_gpt2	0.18	Small	***	Overall English "Likeliness" based on GPT2
linguistic	Vocabulary	average_lexical_frequency_per_million	-0.114	Small	**	Lexical types average per-million word frequency based on the COCA corpus
linguistic	Syntax	perplexity_gpt2	-0.114	Small	**	Perplexity score from GPT2
linguistic	*Syntax*	*syntactic_similarity*	*(pending)*			*Similarity of syntactic structures between response and model answers*
linguistic	*Vocabulary*	*target_identification*	*(pending)*			*Percentage of target words identified in the response based on task-specific keywords*
discourse	Coherence	connectivesTypes_per100words	-0.184	Small	***	Number of conjunction types (e.g., "however", "therefore") per 100 words
discourse	Coherence	connectives_per100words	-0.103	Small	**	Number of conjunction tokens (e.g., "however", "therefore") per 100 words
discourse	*Topic*	*task_completion*	*(pending)*			*Percentage of task requirements met in the response*
discourse	*Topic*	*topic_consistency*	*(pending)*			*Topic consistency measure (definition pending)*
discourse	*Topic*	*idea_density*	*(pending)*			*Number of content words / total number of words*

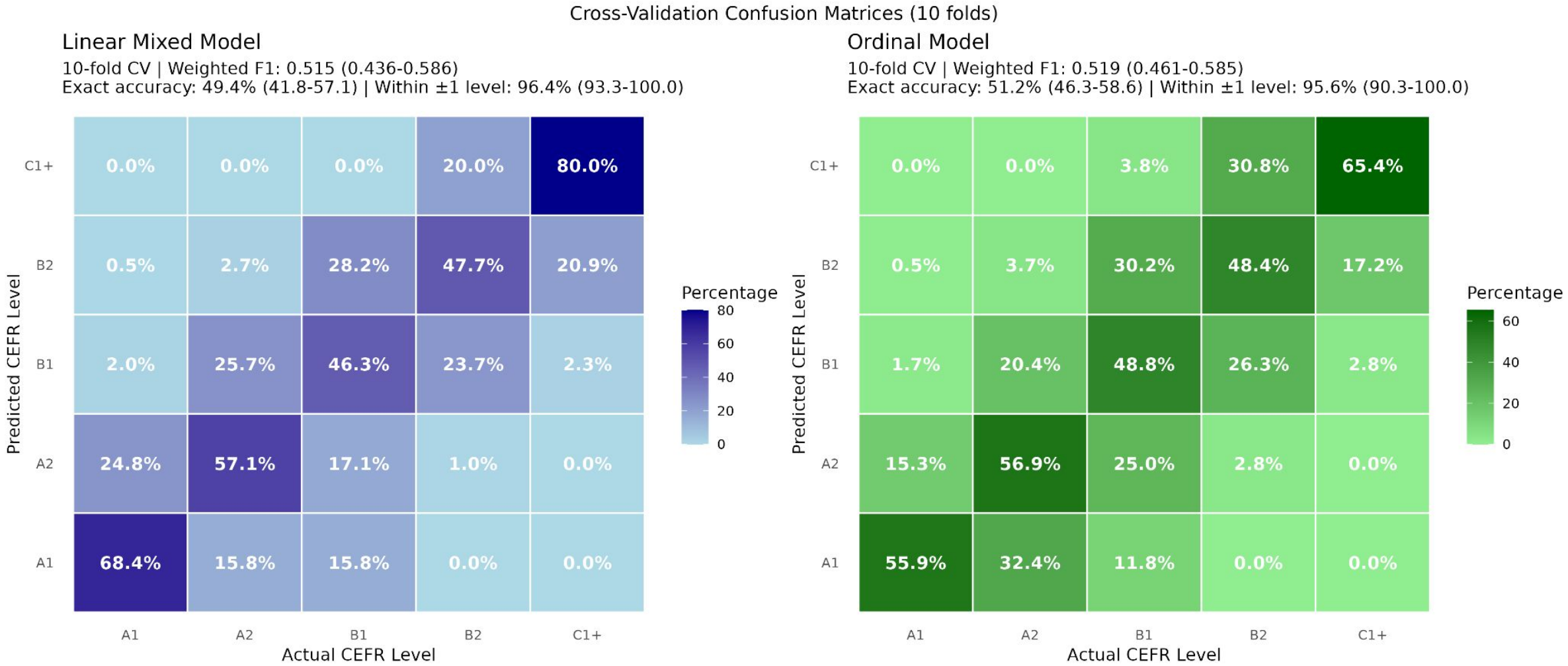
Training Simple Logistic Classifiers based on Features

A) Linear Mixed Model (lmm)

Trained on continuous scores (1-7)

B) Cumulative Link Mixed Model (clmm)

Trained on CEFR levels (A1-C2)



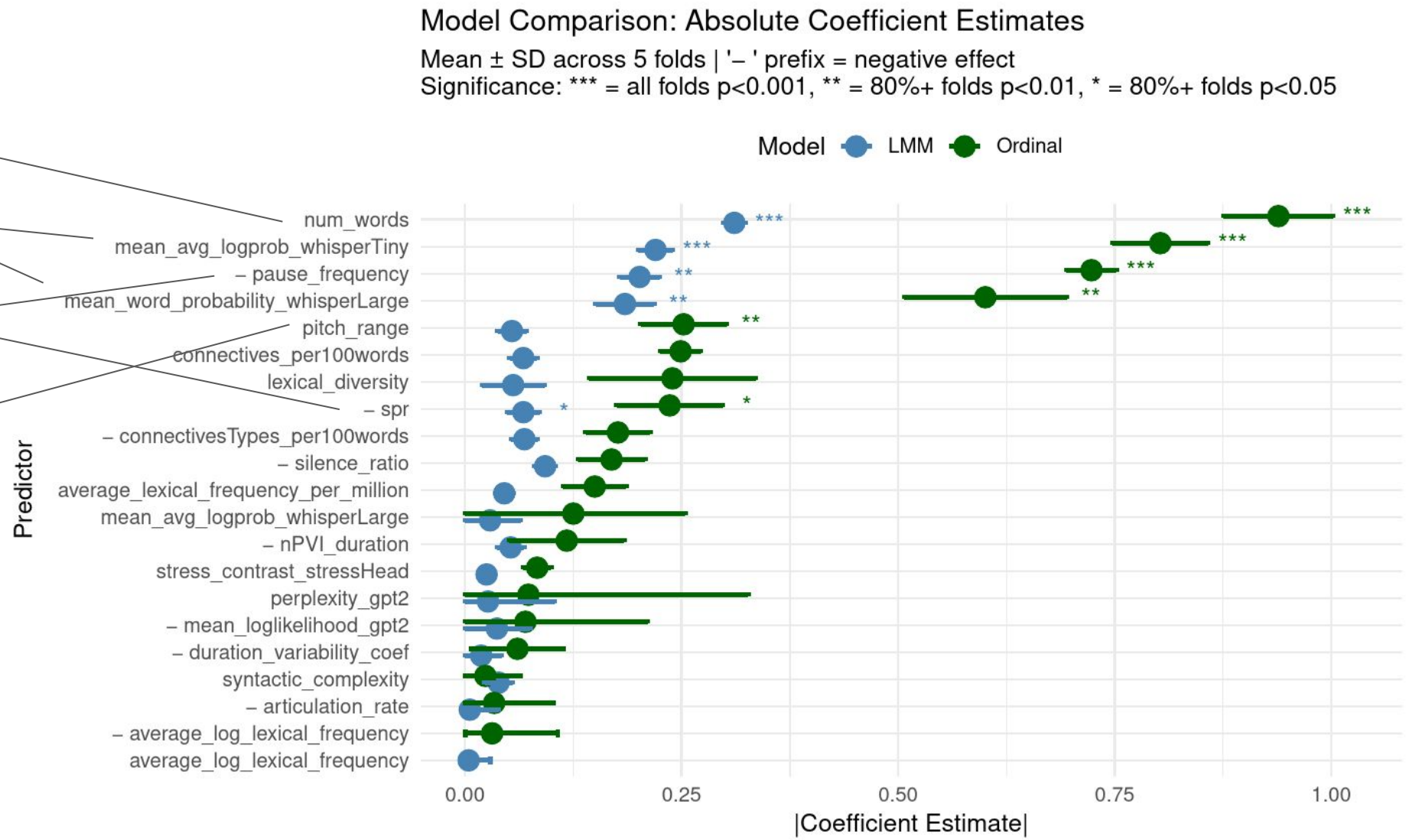
Features used by the Classifier

meta

pronunciation

fluency

prosody



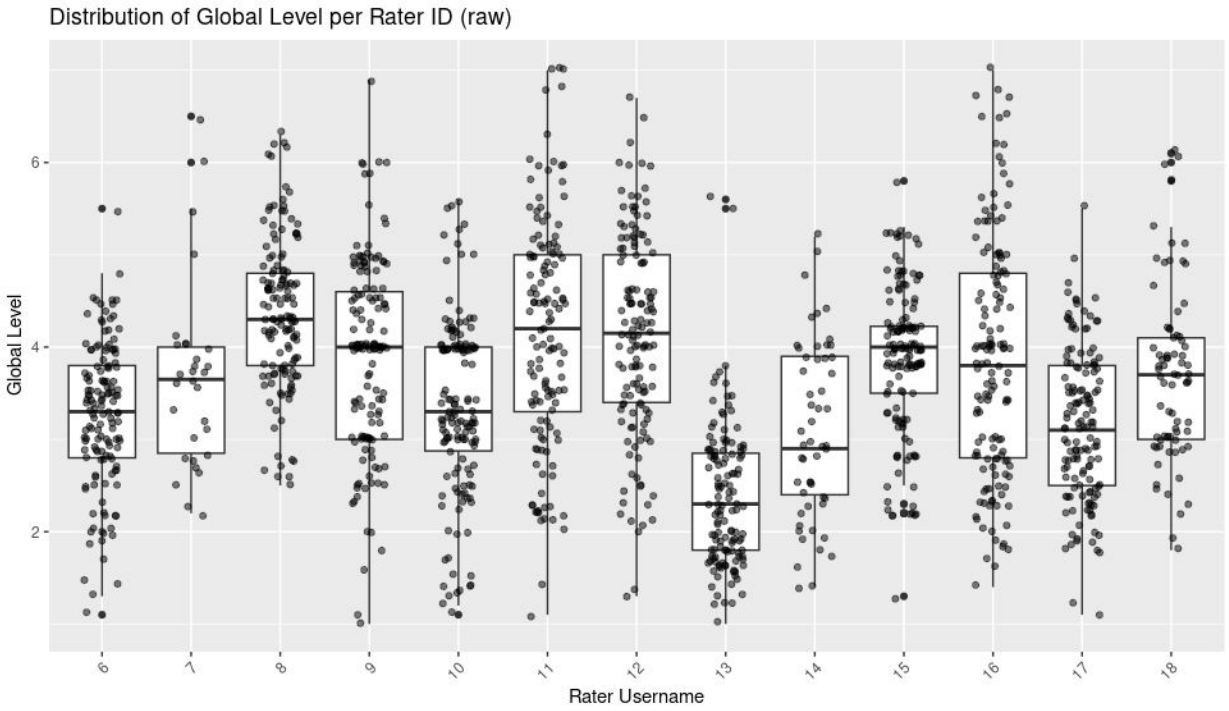
Merci !

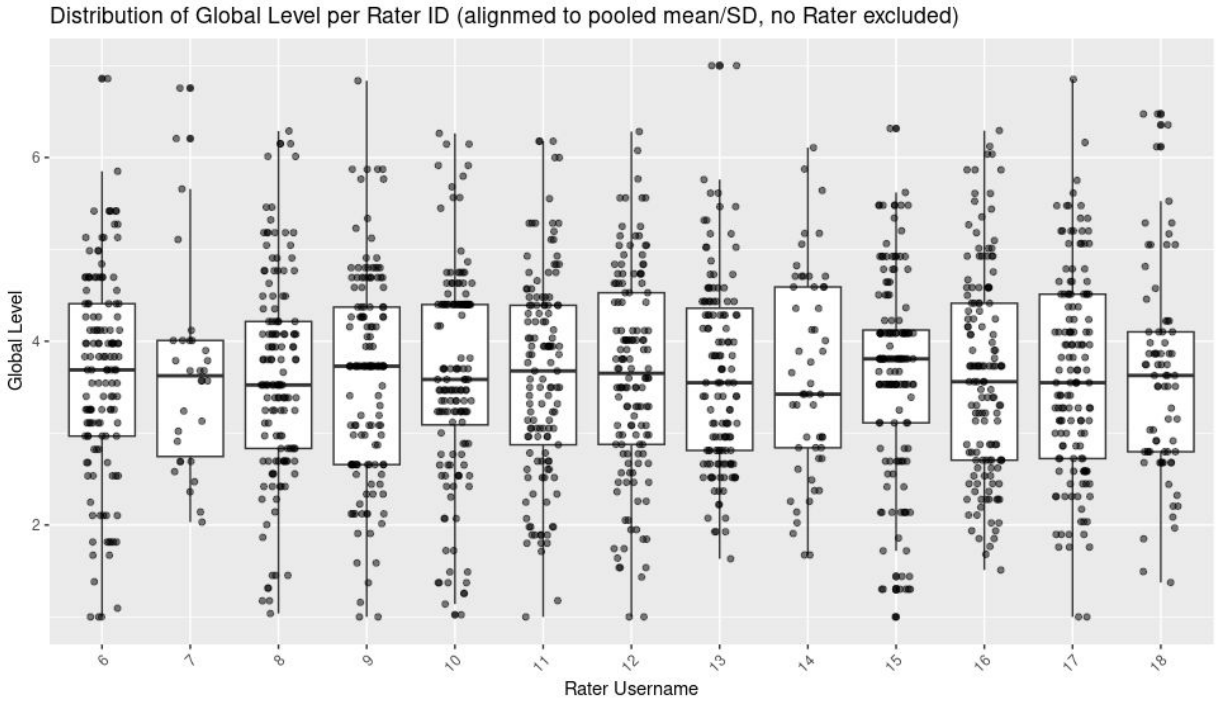
Pinxun HUANG, Eli STAFFORD, Sylvain COULANGE

LIDILEM *(Laboratory of Linguistics and Didactics of Foreign and Mother Tongues)*
LIG *(CNRS, Institute of Engineering, Grenoble Computer Science Laboratory)*
Université Grenoble Alpes

Fundings: ANR France 2030

Simple CEFR prediction model

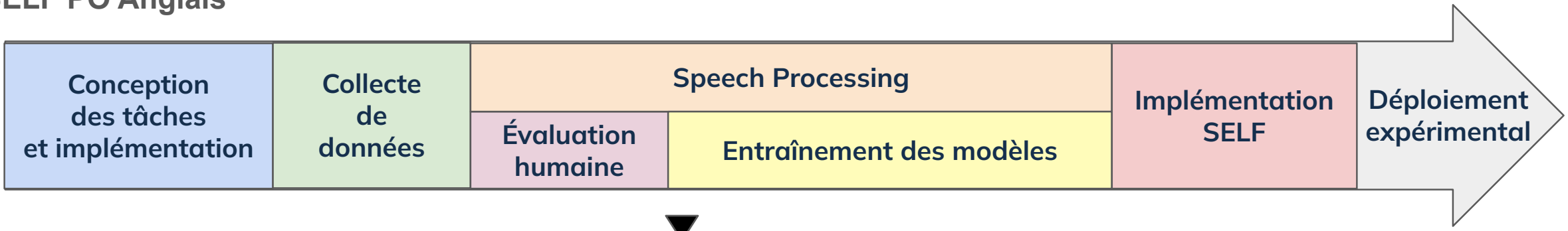




Rater normalization method: Linear equating
(linear alignment to pooled mean/SD)
+ clipping [1.0, 7.0]

Calendrier du projet

SELF PO Anglais



SELF PO Français