# Few-Shot Spoken Language Understanding Via Joint Speech-Text Models

Chung-Ming Chien[1], Mingjiamei Zhang[2], Ju-Chieh Chou[1], Karen Livescu[1]
TTIC[1], The University of Chicago[2]

# Highlight

# Highlight

- We use speech-text models for few-shot & zero-shot Spoken Language Under-standing (SLU).

# Highlight

- We use speech-text models for few-shot & zero-shot Spoken Language Under-standing (SLU).
  - Match the performance of previous models with 0-20% of speech data.

# Highlight

- We use speech-text models for few-shot & zero-shot Spoken Language Under-standing (SLU).
  - Match the performance of previous models with 0-20% of speech data.
- We analyze pre-trained & fine-tuned speech-text models.

# Highlight

- We use speech-text models for few-shot & zero-shot Spoken Language Under-standing (SLU).
  - Match the performance of previous models with 0-20% of speech data.
- We analyze pre-trained & fine-tuned speech-text models.
  - Explain the zero-shot text-to-speech transferability of speech-text models.

# Highlight

- We use speech-text models for few-shot & zero-shot Spoken Language Under-standing (SLU).
  - Match the performance of previous models with 0-20% of speech data.
- We analyze pre-trained & fine-tuned speech-text models.
  - Explain the zero-shot text-to-speech transferability of speech-text models.
  - Suggest fine-tuning with bottom layers frozen, which improves zero-shot performance.

# Background: Speech-Text Pre-Trained Models

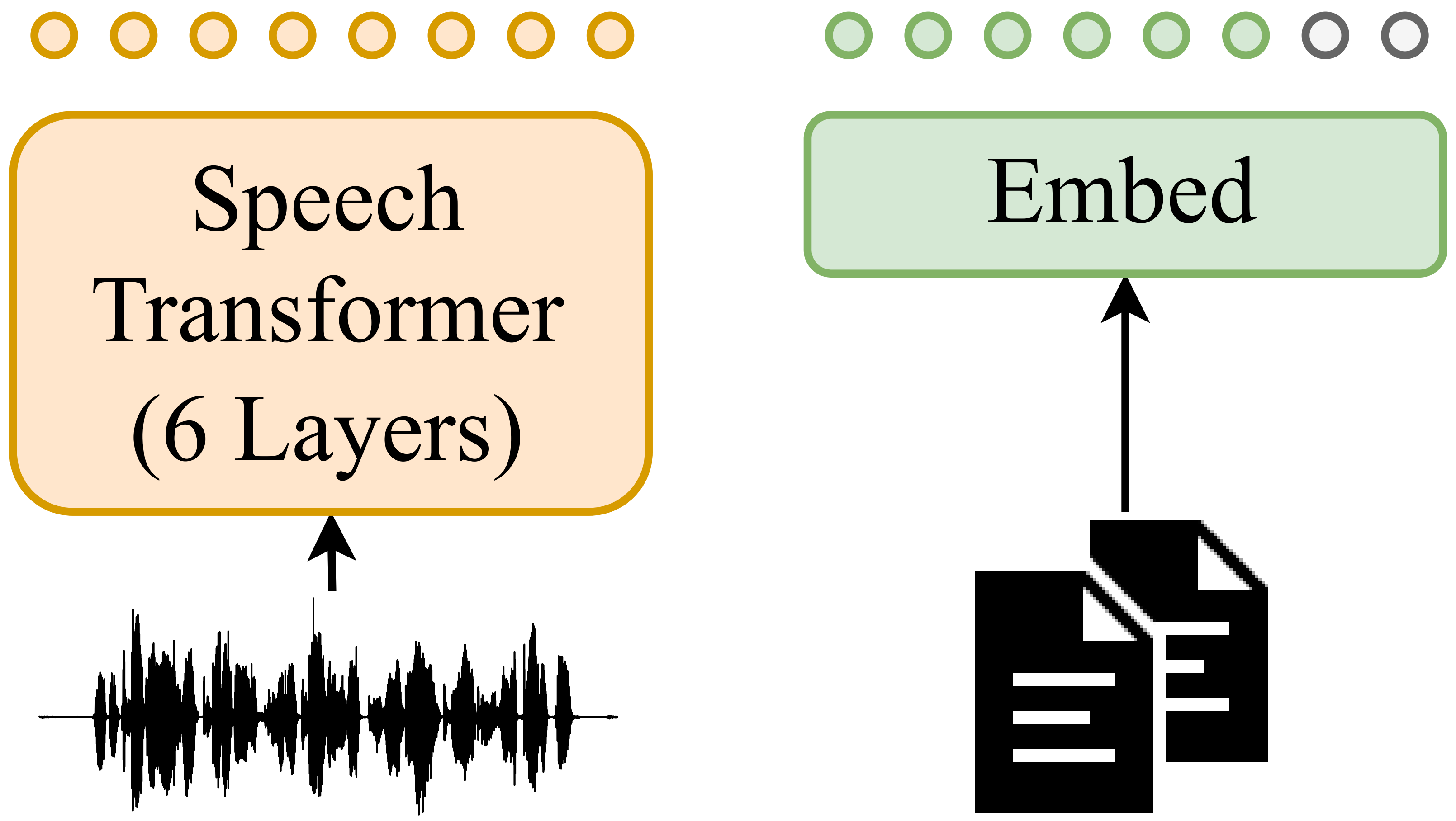
# Background: Speech-Text Pre-Trained Models

- Models studied:
  - SpeechLM [1]
  - SpeechUT [2]

[1] Z. Zhang, et al, “SpeechLM: Enhanced speech pre-training with unpaired textual data,” preprint arXiv:2209.15329, 2023.
[2] Z. Zhang, et al, “SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training,” in EMNLP, 2022.

# Background: Speech-Text Pre-Trained Models

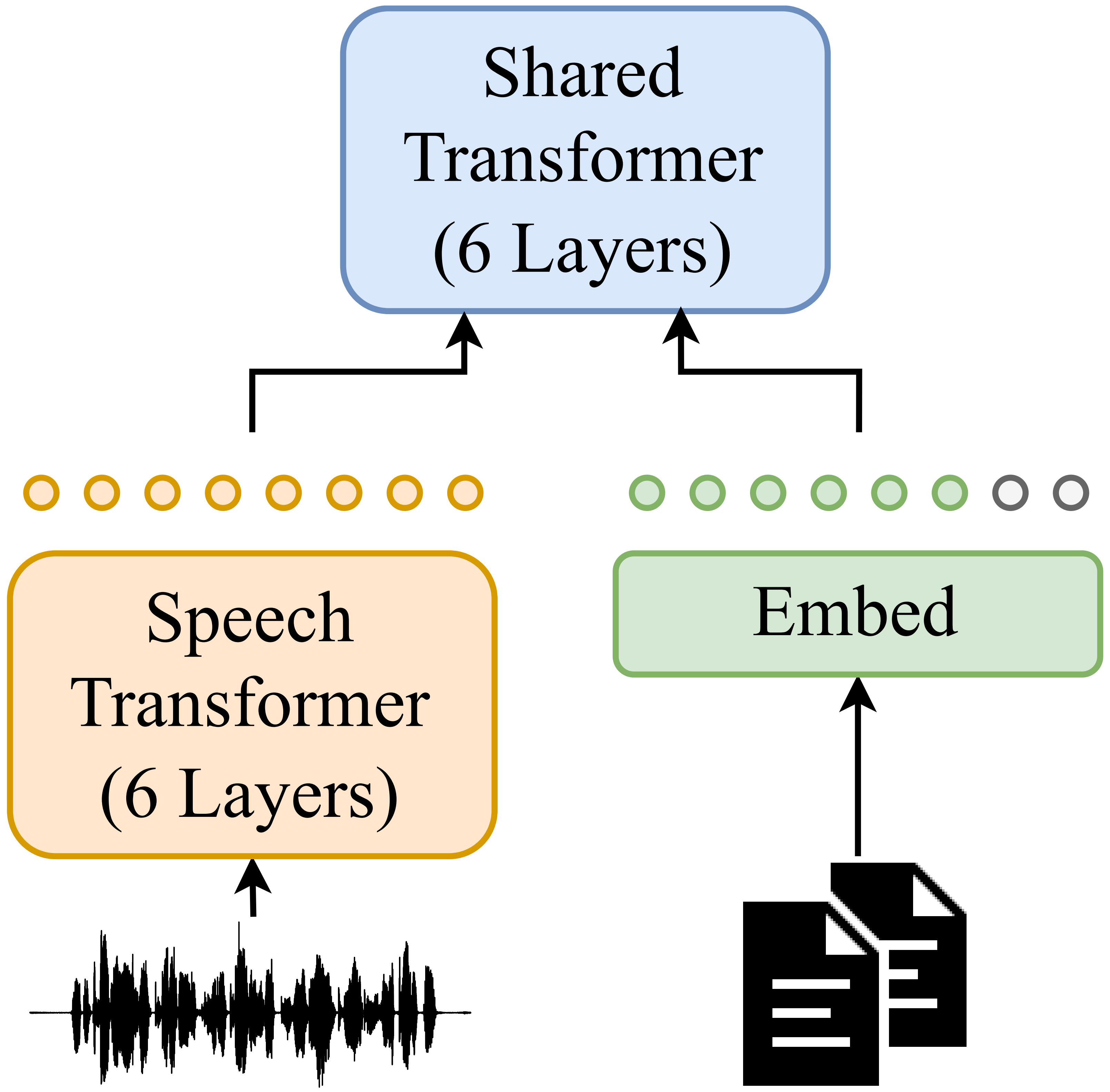
- Models studied:
  - SpeechLM [1]
  - SpeechUT [2]




[1] Z. Zhang, et al, “SpeechLM: Enhanced speech pre-training with unpaired textual data,” preprint arXiv:2209.15329, 2023.
[2] Z. Zhang, et al, “SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training,” in EMNLP, 2022.

# Background: Speech-Text Pre-Trained Models

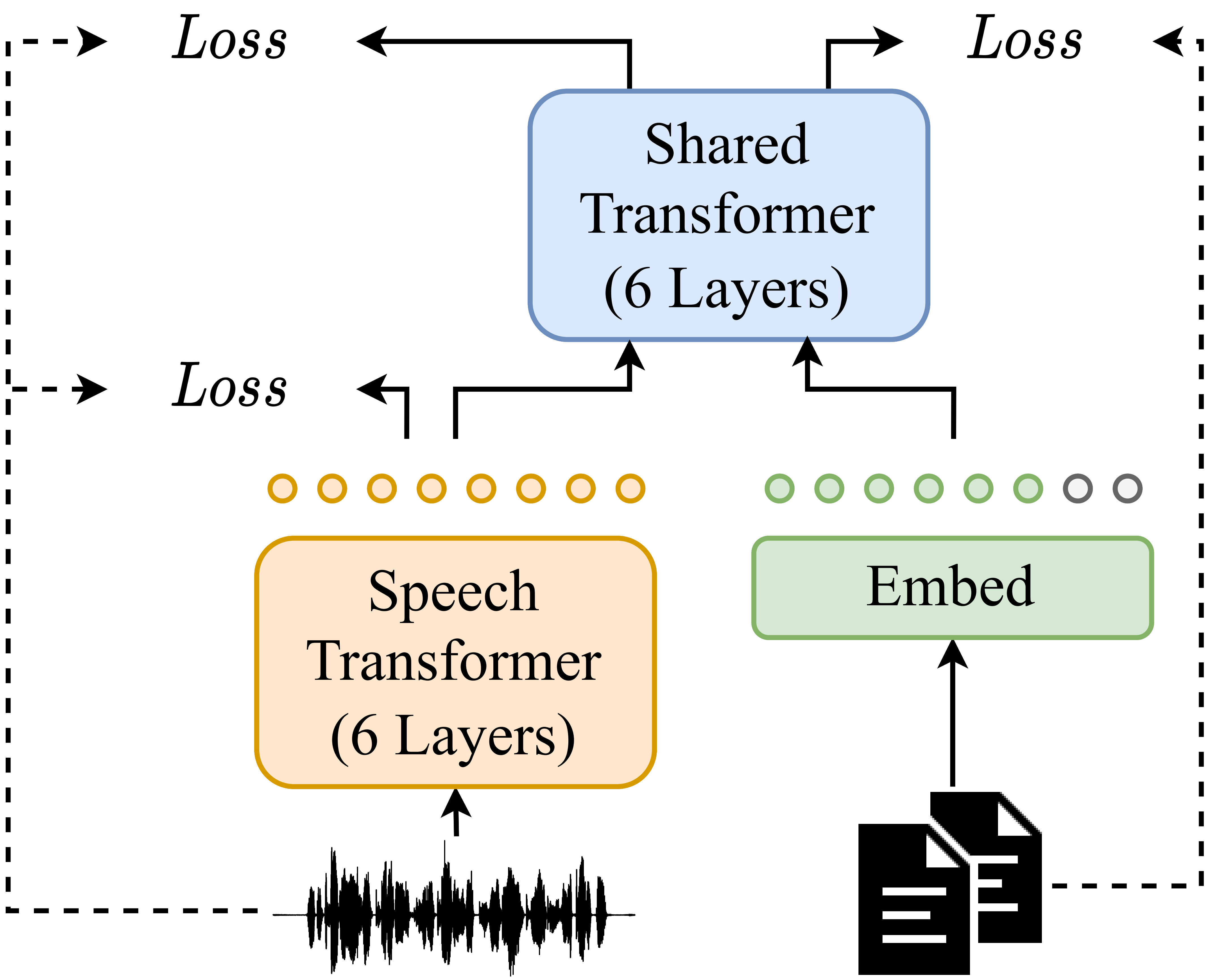
- Models studied:
  - SpeechLM [1]
  - SpeechUT [2]




[1] Z. Zhang, et al, "SpeechLM: Enhanced speech pre-training with unpaired textual data," preprint arXiv:2209.15329, 2023.
[2] Z. Zhang, et al, "SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training," in EMNLP, 2022.

# Background: Speech-Text Pre-Trained Models

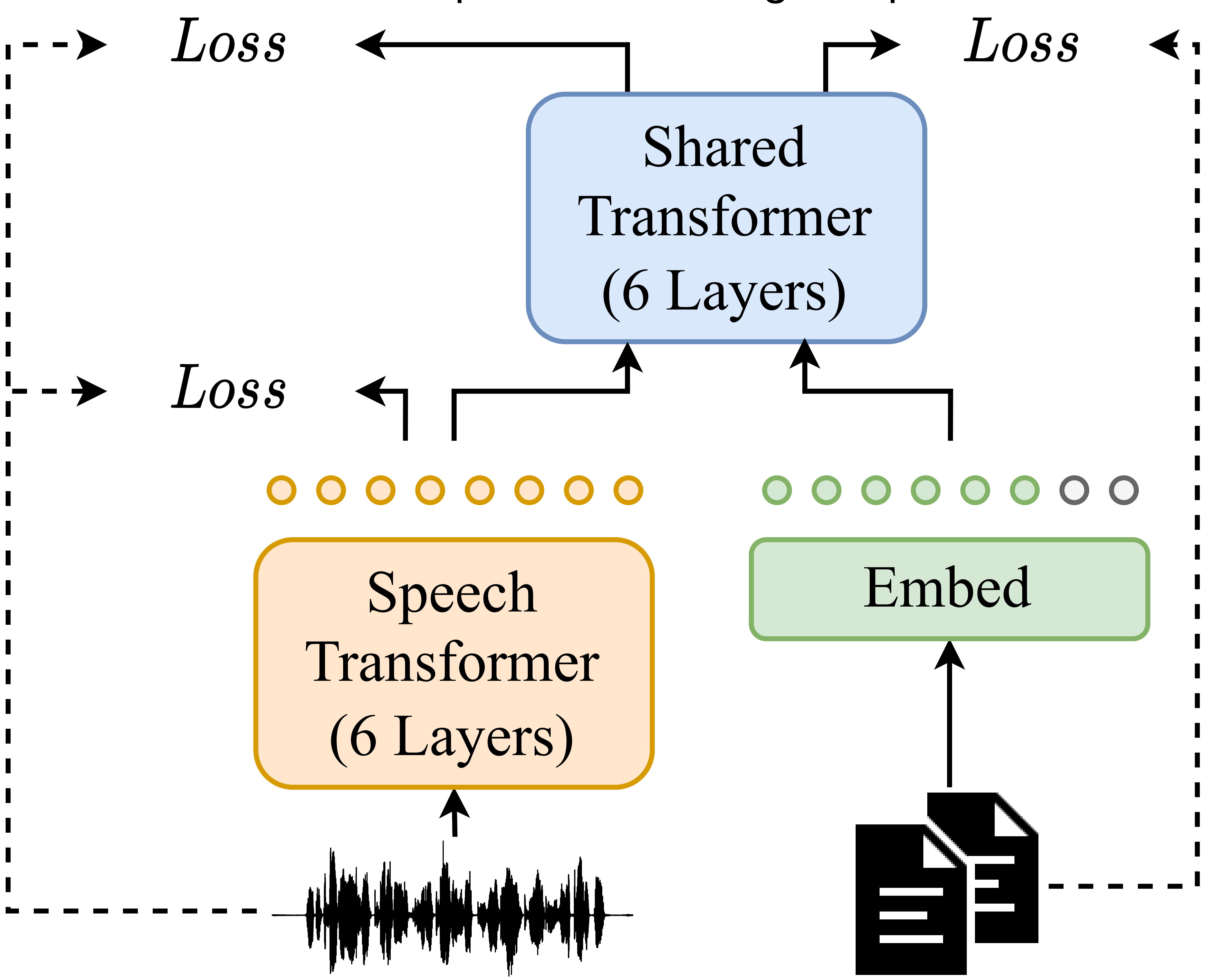
- Models studied:
  - SpeechLM [1]
  - SpeechUT [2]




[1] Z. Zhang, et al, "SpeechLM: Enhanced speech pre-training with unpaired textual data," preprint arXiv:2209.15329, 2023.
[2] Z. Zhang, et al, "SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training," in EMNLP, 2022.

# Background: Speech-Text Pre-Trained Models

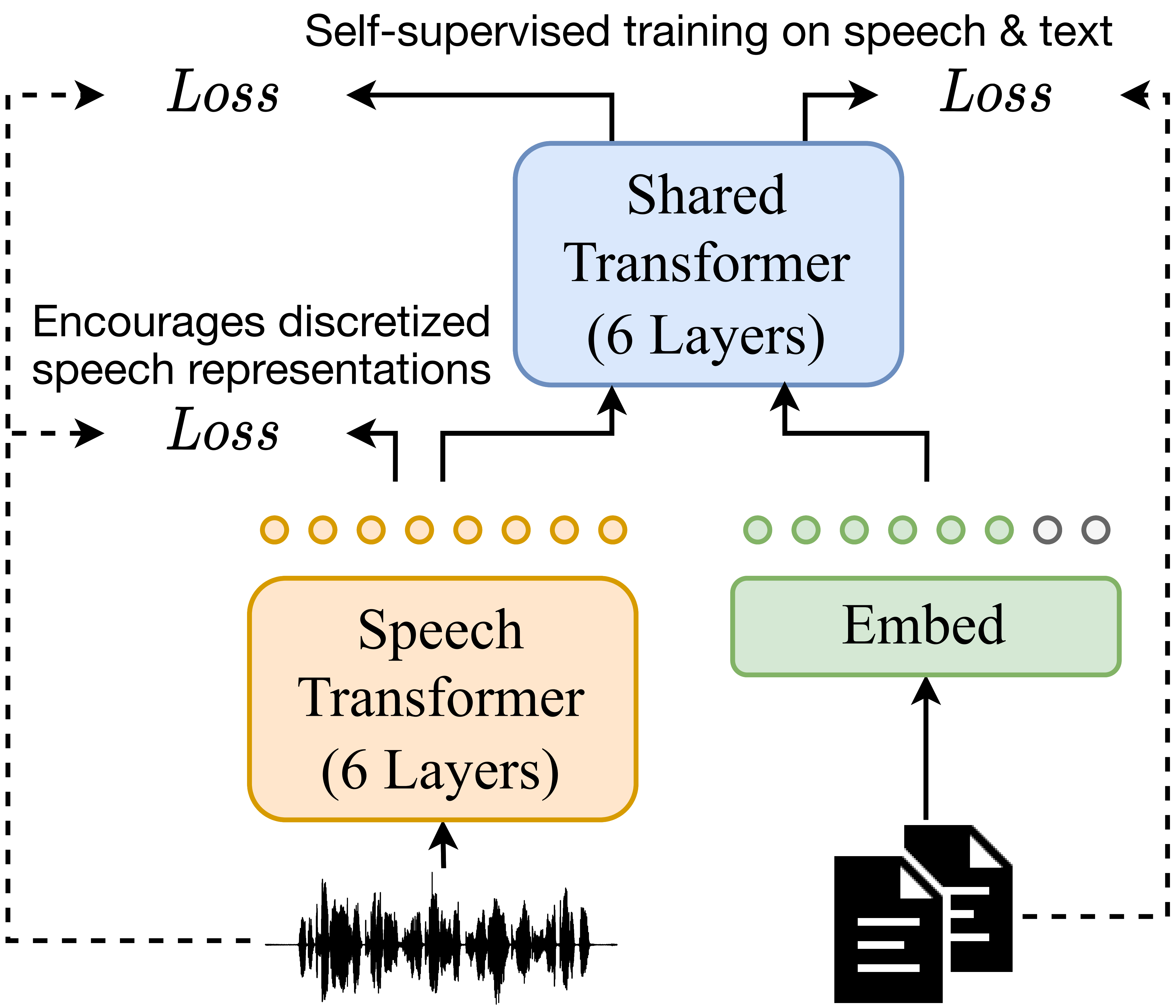
- Models studied:
  - SpeechLM [1]
  - SpeechUT [2]




[1] Z. Zhang, et al, “SpeechLM: Enhanced speech pre-training with unpaired textual data,” preprint arXiv:2209.15329, 2023.
[2] Z. Zhang, et al, “SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training,” in EMNLP, 2022.

# Background: Speech-Text Pre-Trained Models

- Models studied:
  - SpeechLM [1]
  - SpeechUT [2]




[1] Z. Zhang, et al, "SpeechLM: Enhanced speech pre-training with unpaired textual data," preprint arXiv:2209.15329, 2023.
[2] Z. Zhang, et al, "SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training," in EMNLP, 2022.

# Highlight

- We use speech-text models for few-shot & zero-shot Spoken Language Understanding (SLU).
  - Match the performance of previous models with 0-20% of speech data.
- We analyze pre-trained & fine-tuned speech-text models.
  - Explain the zero-shot text-to-speech transferability of speech-text models.
  - Suggest fine-tuning with bottom layers frozen, which improves zero-shot performance.

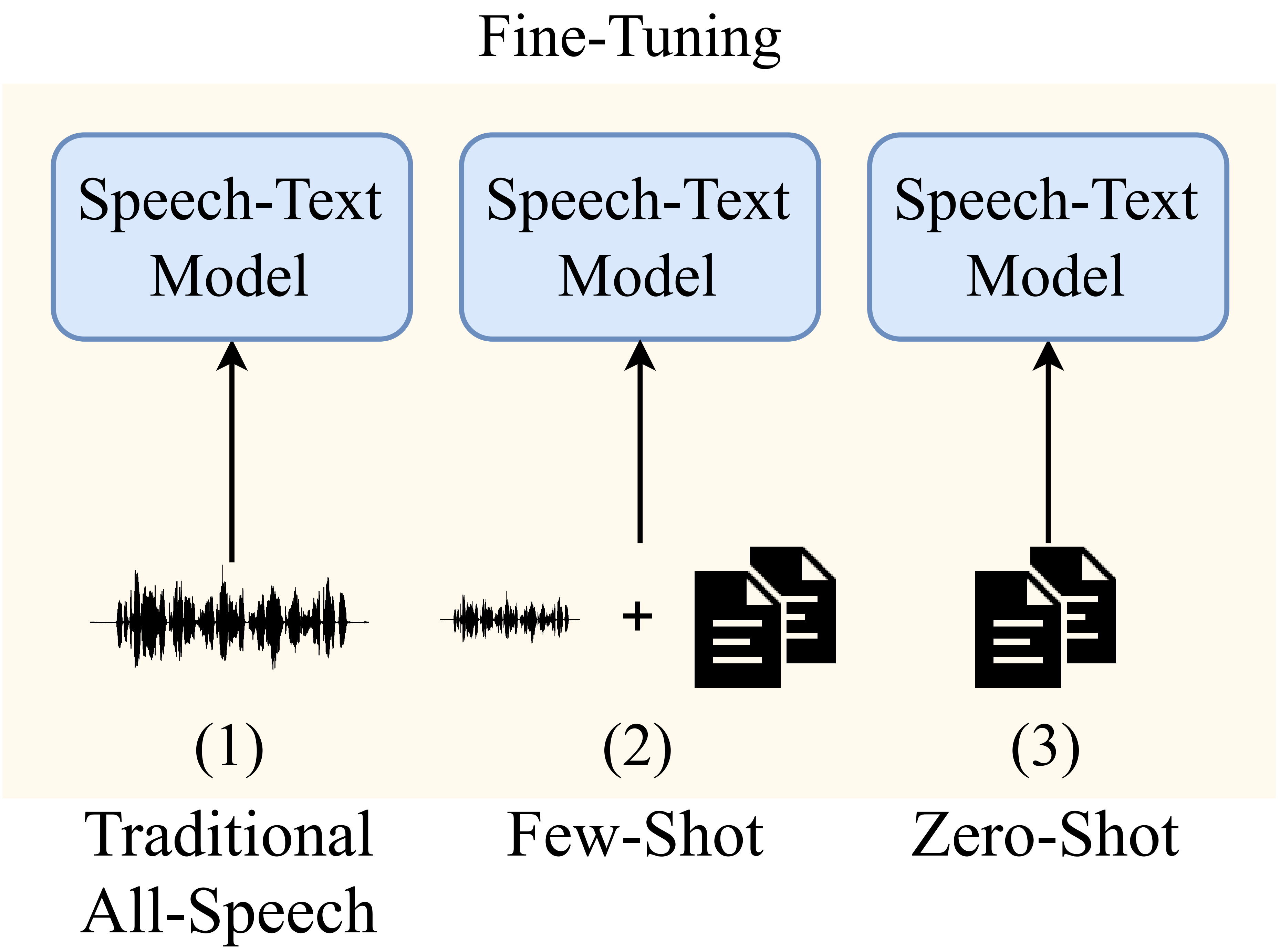
# Few-Shot & Zero-Shot Spoken Language Understanding

# Few-Shot & Zero-Shot Spoken Language Understanding

- Assumptions
  - Limited labeled speech data
  - More text data

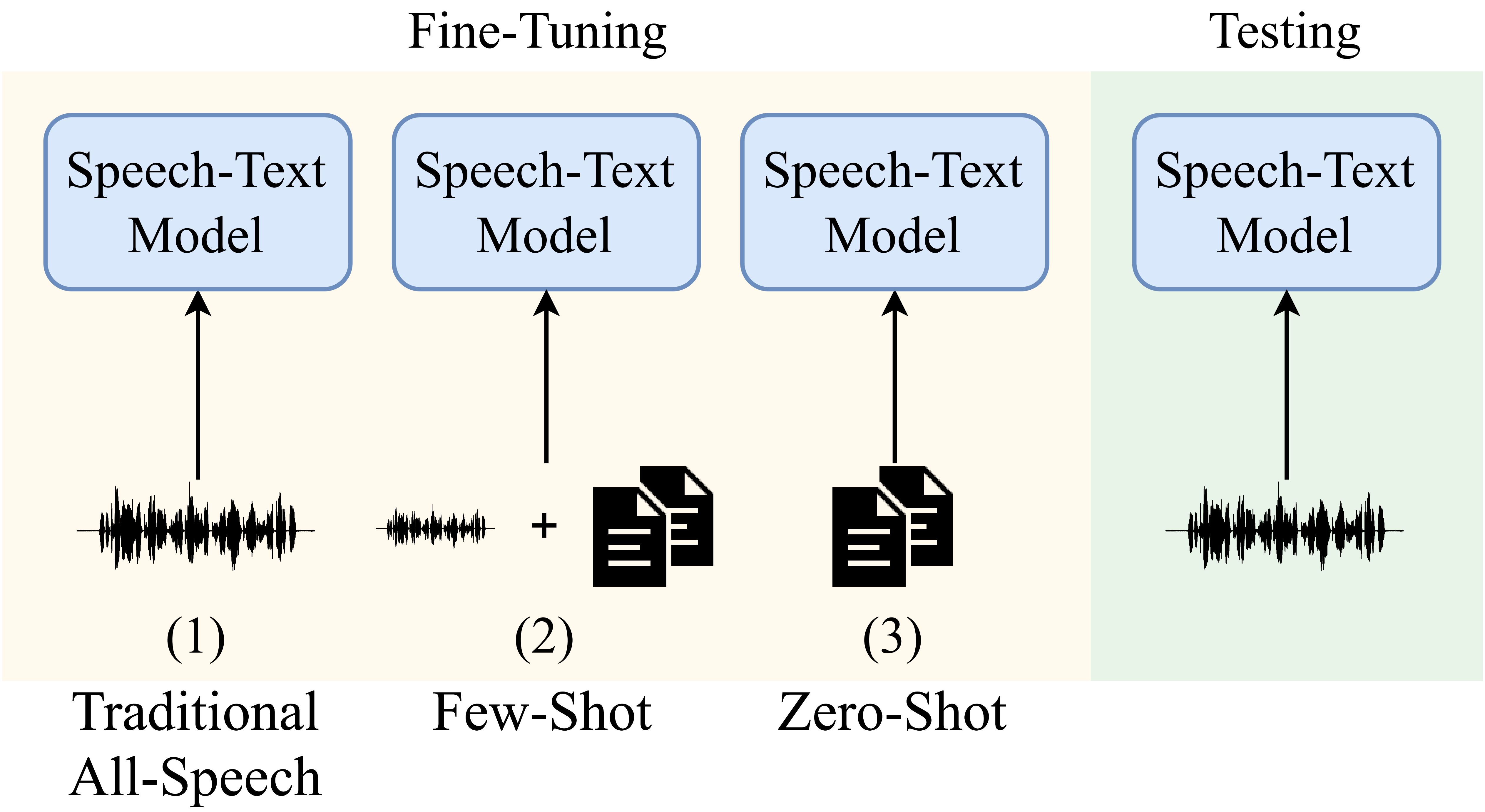
# Few-Shot & Zero-Shot Spoken Language Understanding

- Assumptions
  - Limited labeled speech data
  - More text data



# Few-Shot & Zero-Shot Spoken Language Understanding

- Assumptions
  - Limited labeled speech data
  - More text data



# Experimental Setups

- SLU tasks: SLUE Benchmark [3]
  - Sentiment Analysis (SA)
    - Classification: “positive,” “neutral,” or “negative” sentiments
  - Named Entity Recognition (NER)
    - Sequence labeling
- Speech-text models fine-tuned with labeled text data + different amounts of labeled speech data
- Other details follow the default setup of the SLUE benchmark

[3] S. Shon, et al, “SLUE: New benchmark tasks for spoken language understanding evaluation on natural speech,” in ICASSP, 2022.

# Sentiment Analysis

- Zero-shot performance comparable to models using full speech data.

| Sentiment Analysis Accuracy (%) | Labeled Data | | Prior work: Speech-Only | Speech-Text | |
|---|---|---|---|---|---|
| | Speech | Text | HuBERT | SpeechLM-P | SpeechLM-H |
| Baselines | 1 hr | - | | 36.9 | 37.7 |
| | 12.8 hrs | - | 43.0 | 45.6 | 45.3 |
| Proposed | - | full | | 45.2 | 45.2 |
| | 10 mins | full | | 45.2 | 38.3 |
| | 1 hr | full | | 46.4 | 43.4 |

# Sentiment Analysis

- Zero-shot performance comparable to models using full speech data.

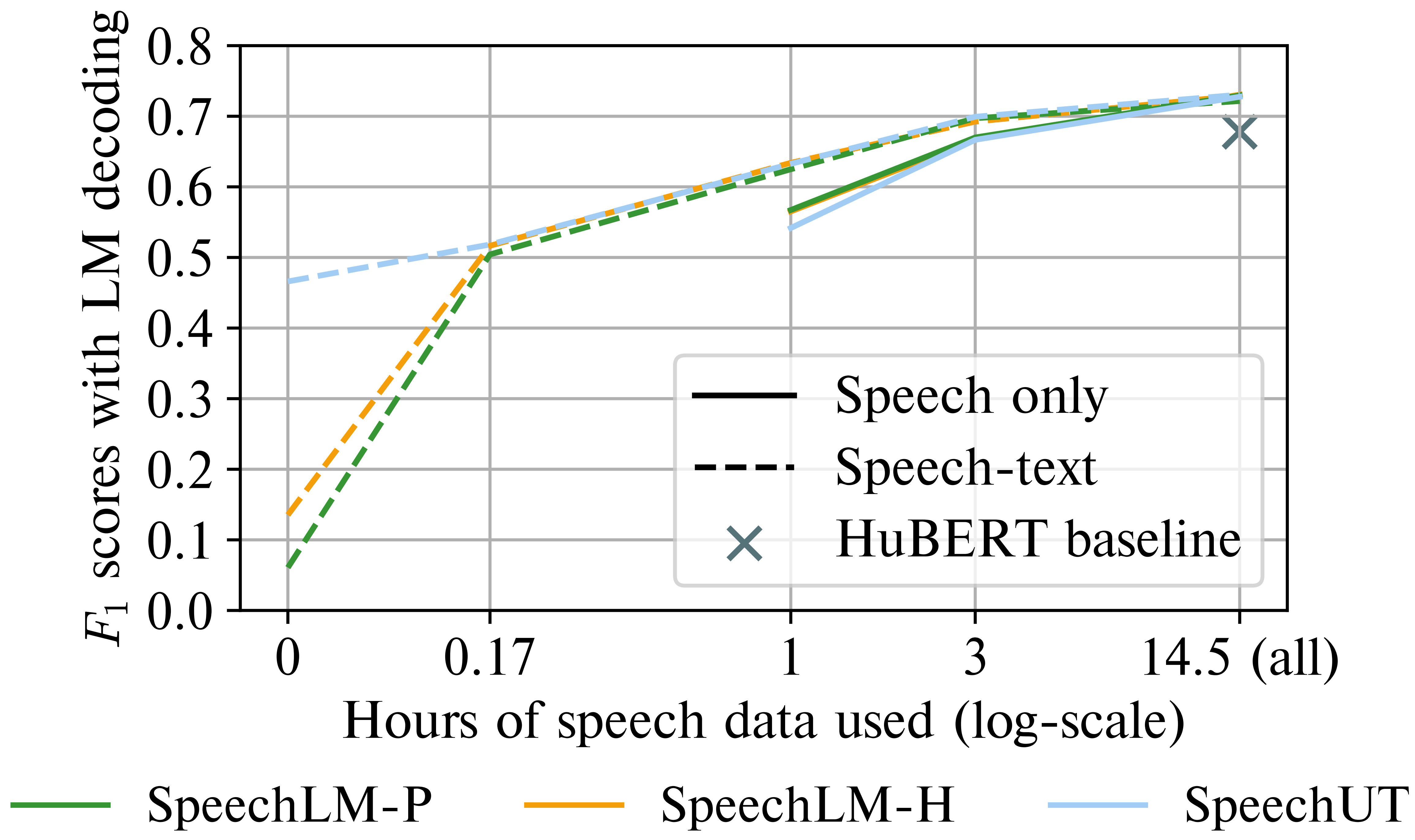
| Sentiment Analysis Accuracy (%) | Labeled Data | | Prior work: Speech-Only | Speech-Text | |
|---|---|---|---|---|---|
| | Speech | Text | HuBERT | SpeechLM-P | SpeechLM-H |
| Baselines | 1 hr | - | | 36.9 | 37.7 |
| | 12.8 hrs | - | 43.0 | 45.6 | 45.3 |
| Proposed | - | full | | 45.2 | 45.2 |
| | 10 mins | full | | 45.2 | 38.3 |
| | 1 hr | full | | 46.4 | 43.4 |

# Sentiment Analysis

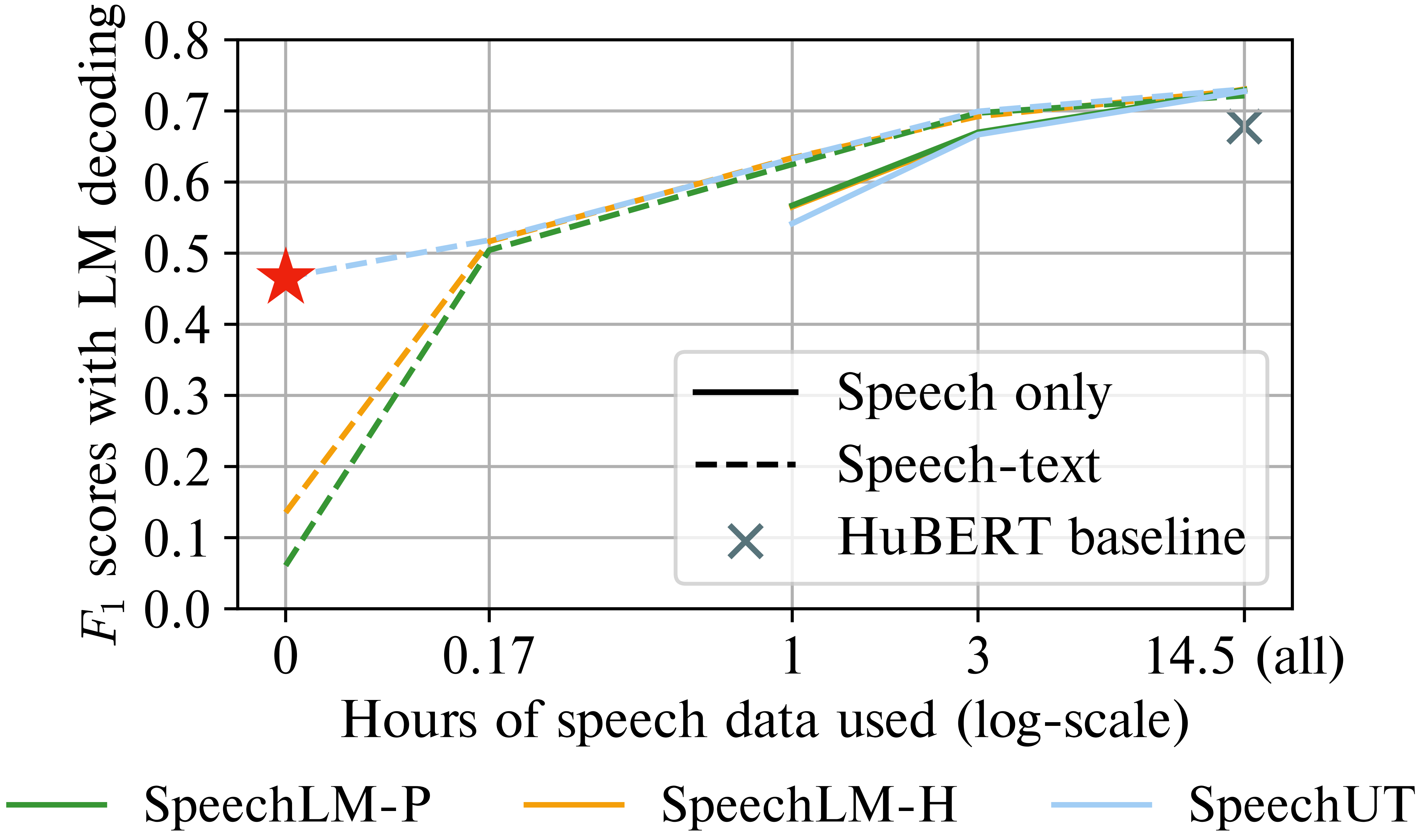
- Zero-shot performance comparable to models using full speech data.

| Sentiment Analysis Accuracy (%) | Labeled Data | | Prior work: Speech-Only | Speech-Text | |
|---|---|---|---|---|---|
| | Speech | Text | HuBERT | SpeechLM-P | SpeechLM-H |
| Baselines | 1 hr | - | | 36.9 | 37.7 |
| | 12.8 hrs | - | 43.0 | 45.6 | 45.3 |
| Proposed | - | full | | 45.2 | 45.2 |
| | 10 mins | full | | 45.2 | 38.3 |
| | 1 hr | full | | 46.4 | 43.4 |

# Named Entity Recognition

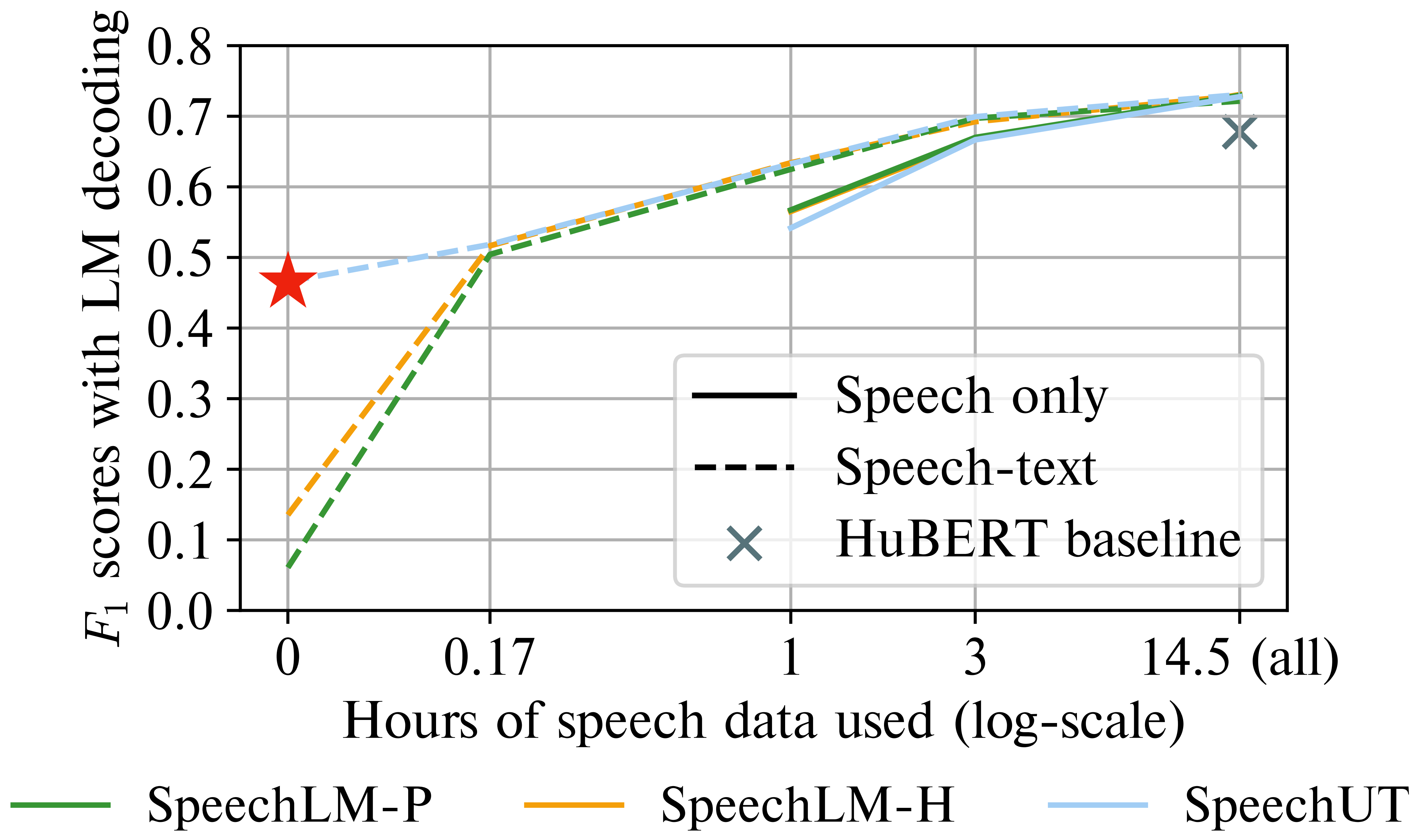


# Named Entity Recognition



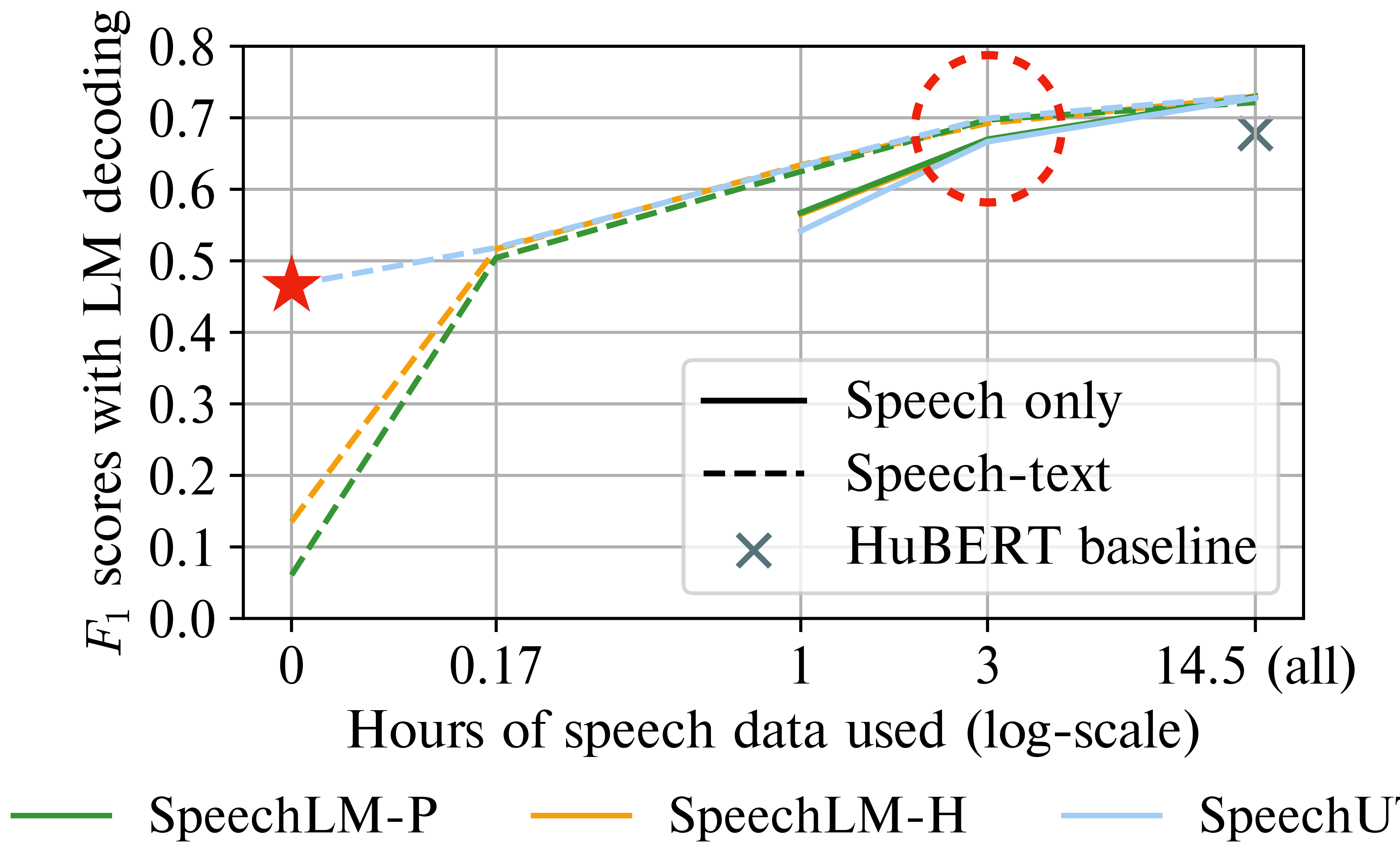
# Named Entity Recognition

- SpeechUT has great zero-shot performance.



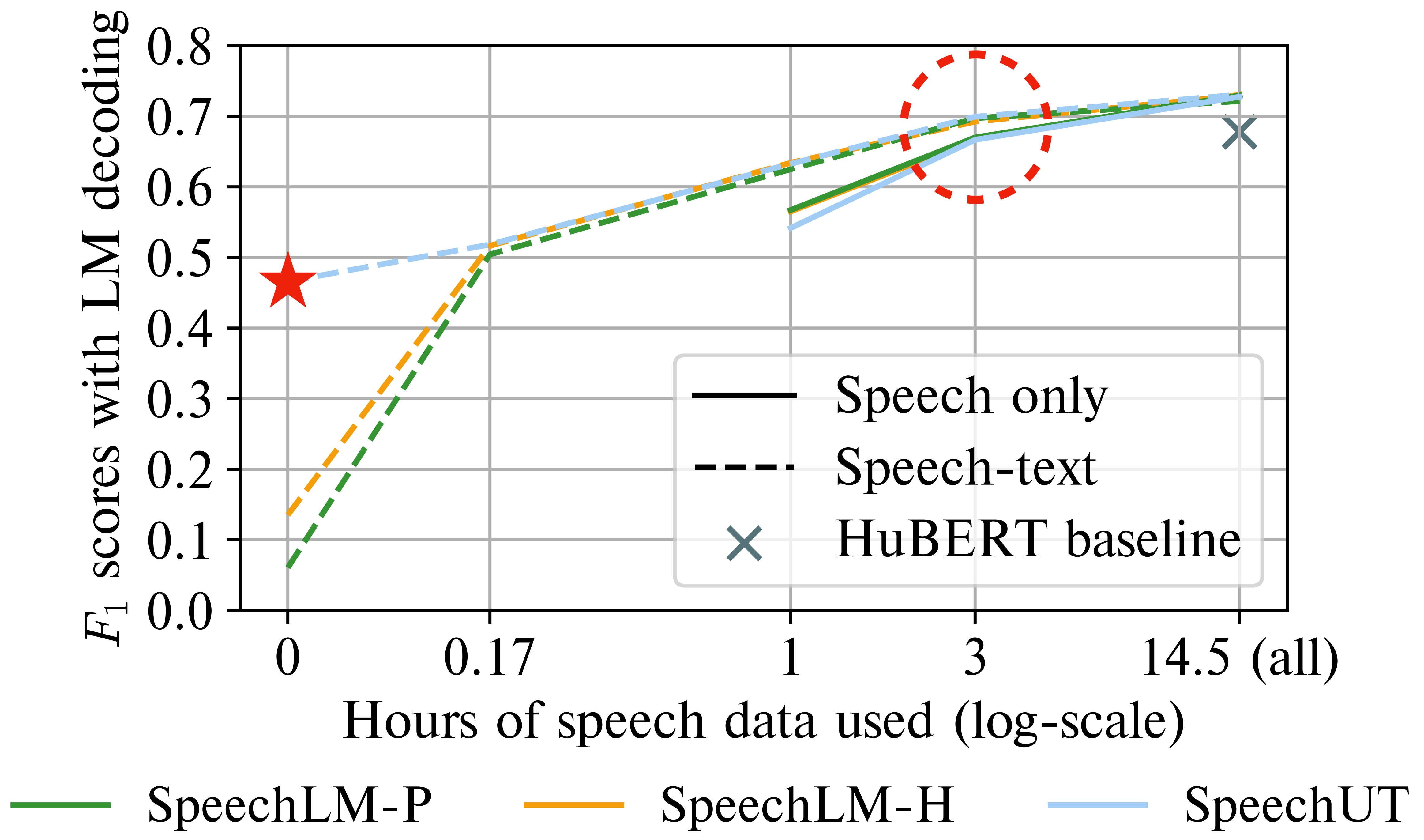
# Named Entity Recognition

- SpeechUT has great zero-shot performance.



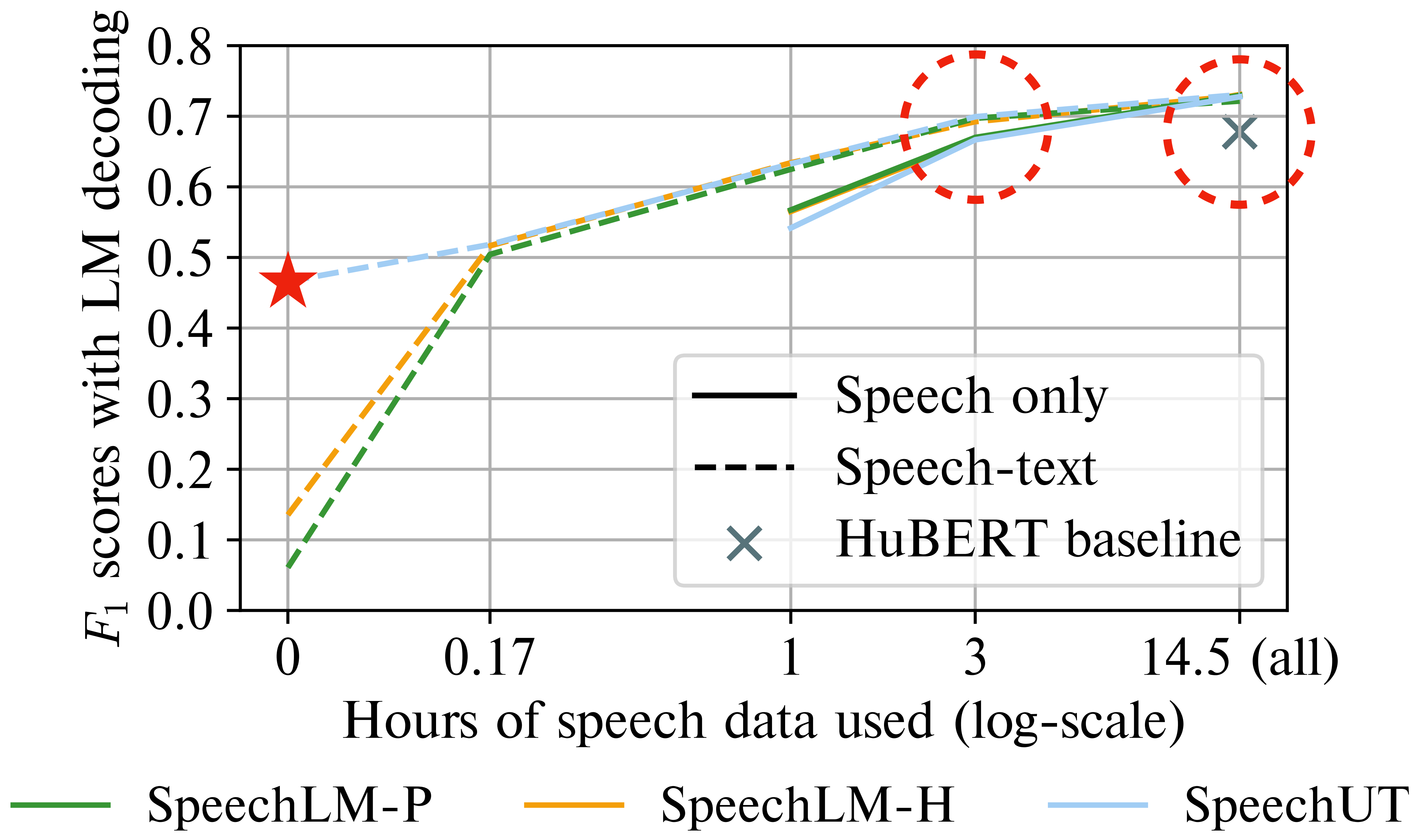
# Named Entity Recognition

- SpeechUT has great zero-shot performance.
- Speech+text fine-tuning is better than speech-only fine-tuning.



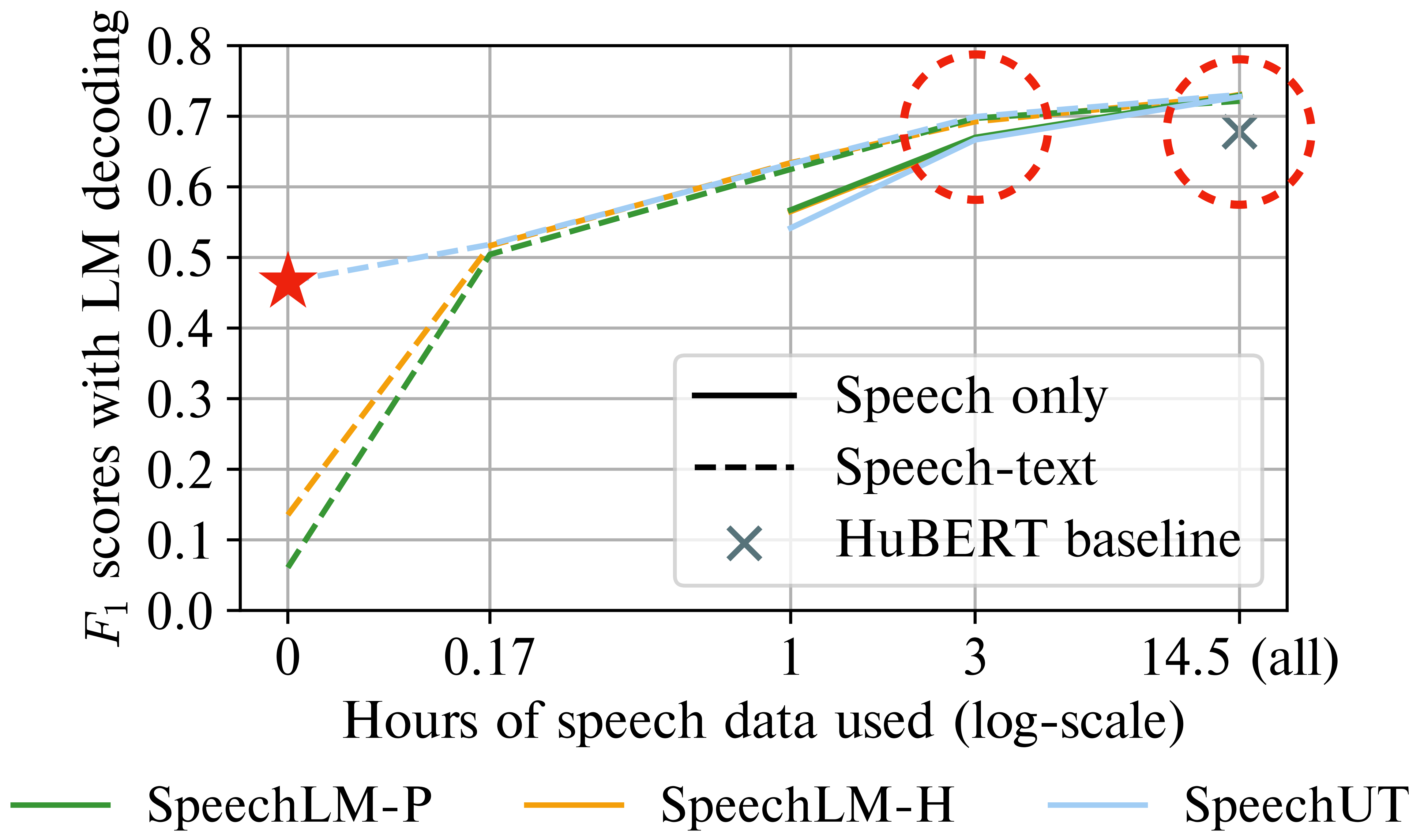
# Named Entity Recognition

- SpeechUT has great zero-shot performance.
- Speech+text fine-tuning is better than speech-only fine-tuning.



# Named Entity Recognition

- SpeechUT has great zero-shot performance.
- Speech+text fine-tuning is better than speech-only fine-tuning.
  - Outperforms HuBERT (speech-only) with 20% of speech data.



# Highlight

- We use speech-text models for few-shot & zero-shot Spoken Language Under-standing (SLU).
  - Match the performance of previous models with 0-20% of speech data.
- We analyze pre-trained & fine-tuned speech-text models.
  - Explain the zero-shot text-to-speech transferability of speech-text models.
  - Suggest fine-tuning with bottom layers frozen, which improves zero-shot performance.

# Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) [4]

$$\frac{1}{d}\sum_{i=1}^{d} corr(X_i, Y_i)$$

- with $X, Y \in \mathbb{R}^d$ representing different views (e.g. text & speech) of the same data instance.

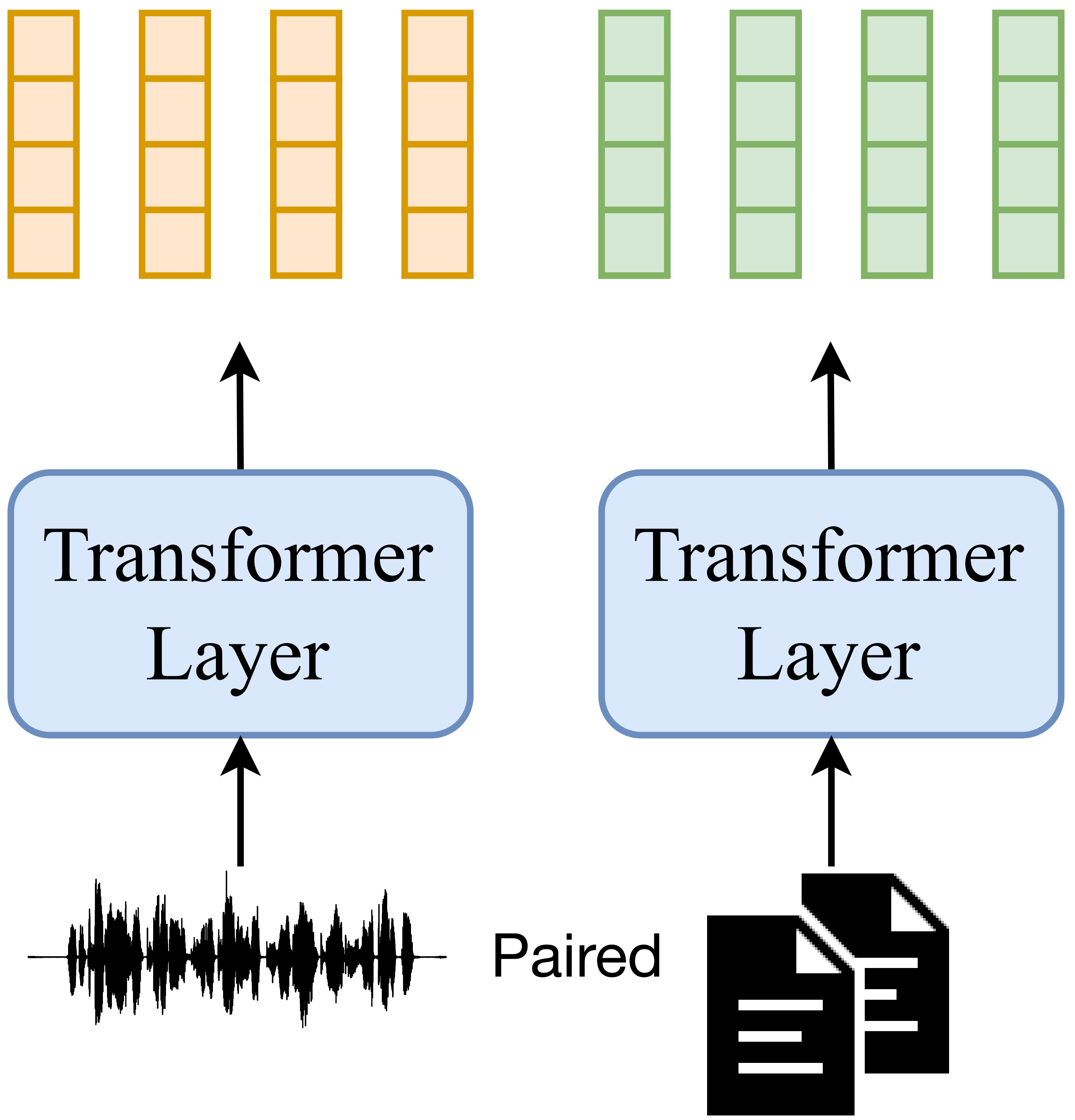
[4] M. Del and M. Fishel, “Cross-lingual similarity of multilingual representations revisited,” in AACL, 2022.

# Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) [4]

$$\frac{1}{d}\sum_{i=1}^{d} corr(X_i, Y_i)$$

- with $X, Y \in \mathbb{R}^d$ representing different views (e.g. text & speech) of the same data instance.



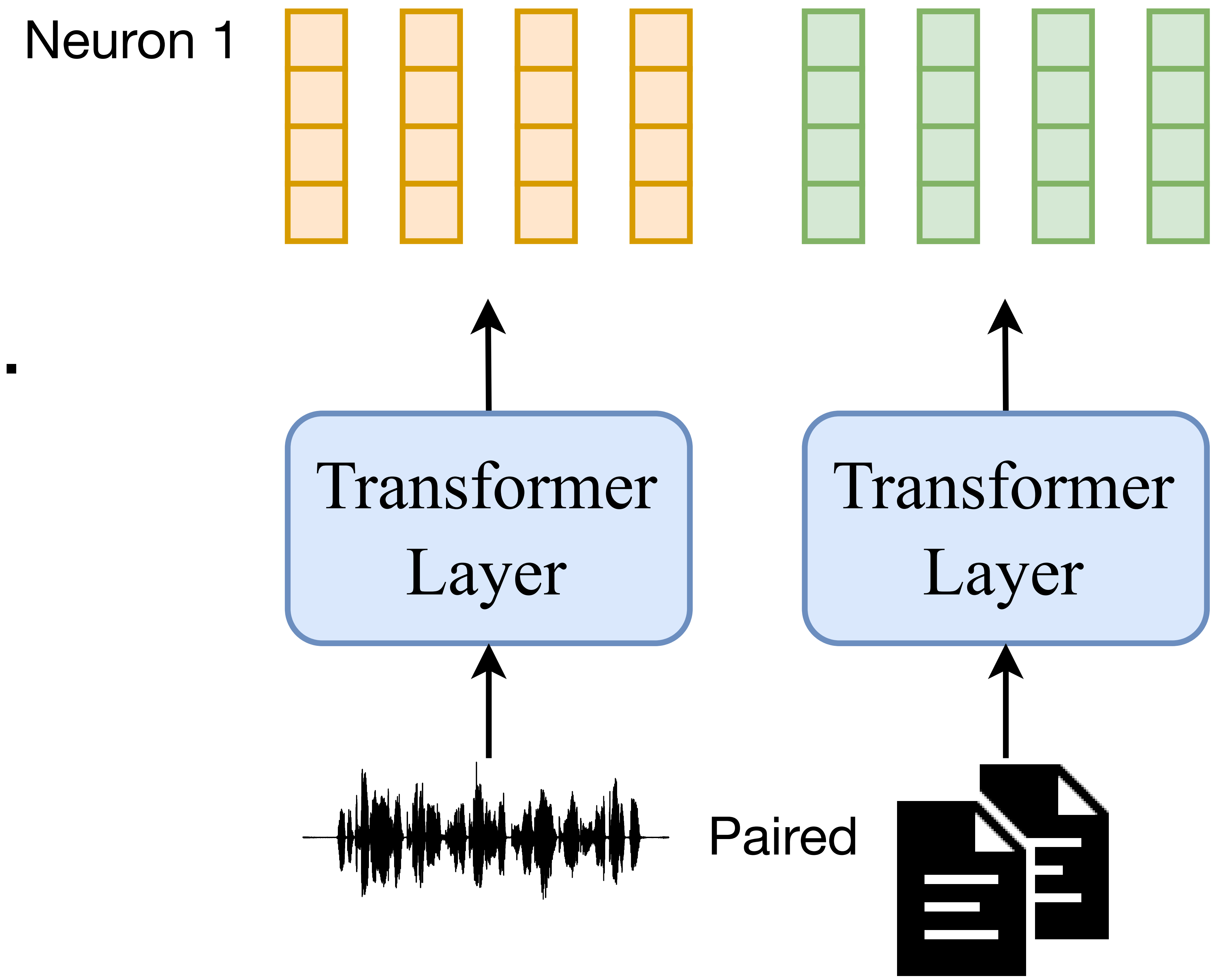

[4] M. Del and M. Fishel, "Cross-lingual similarity of multilingual representations revisited," in AACL, 2022.

# Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) [4]

$$\frac{1}{d}\sum_{i=1}^{d} corr(X_i, Y_i)$$

- with $X, Y \in \mathbb{R}^d$ representing different views (e.g. text & speech) of the same data instance.



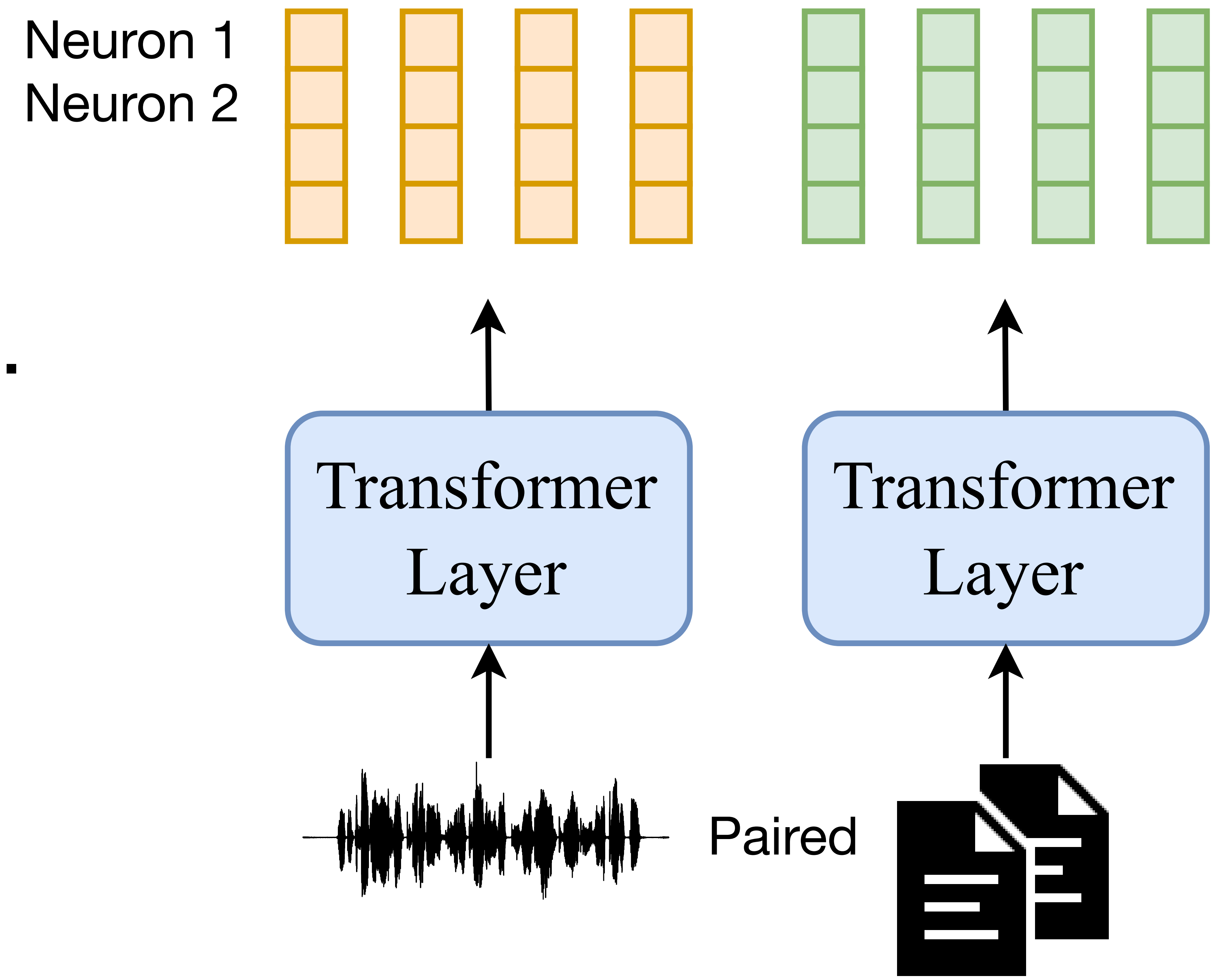

[4] M. Del and M. Fishel, "Cross-lingual similarity of multilingual representations revisited," in AACL, 2022.

# Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) [4]

$$\frac{1}{d}\sum_{i=1}^{d} corr(X_i, Y_i)$$

- with $X, Y \in \mathbb{R}^d$ representing different views (e.g. text & speech) of the same data instance.



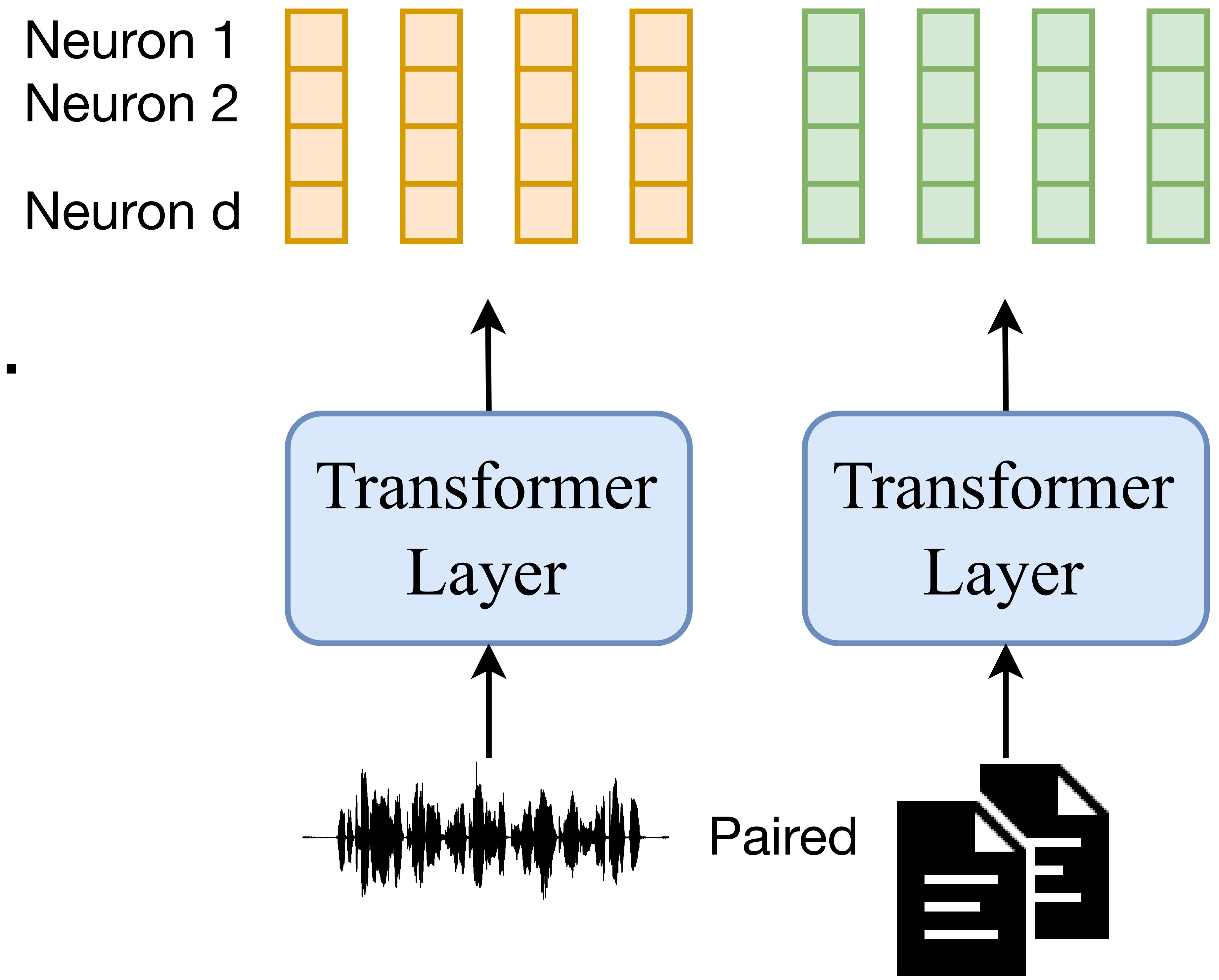

[4] M. Del and M. Fishel, "Cross-lingual similarity of multilingual representations revisited," in AACL, 2022.

# Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) [4]

$$\frac{1}{d}\sum_{i=1}^{d} corr(X_i, Y_i)$$

- with $X, Y \in \mathbb{R}^d$ representing different views (e.g. text & speech) of the same data instance.



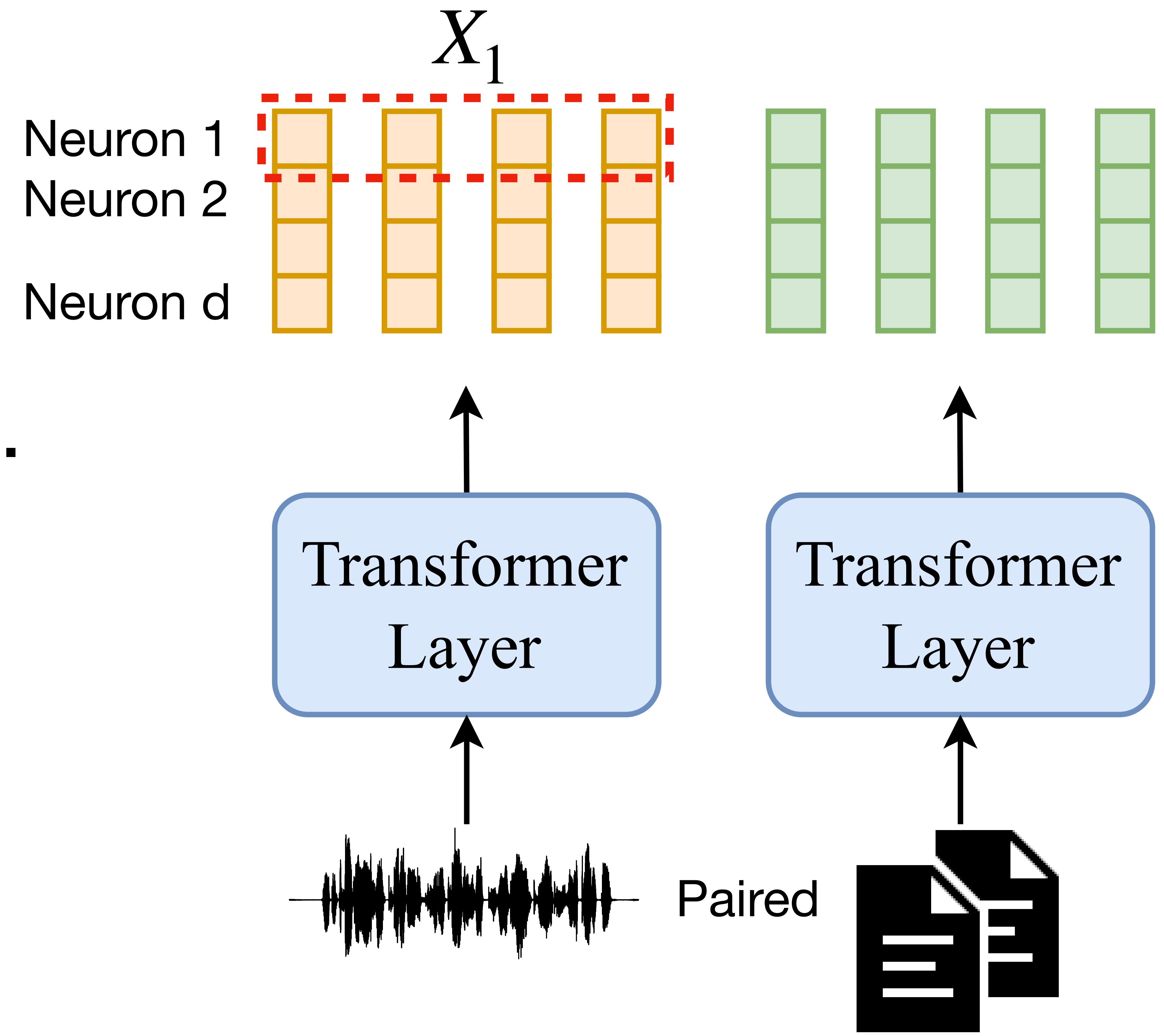

[4] M. Del and M. Fishel, "Cross-lingual similarity of multilingual representations revisited," in AACL, 2022.

# Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) [4]

$$\frac{1}{d}\sum_{i=1}^{d} corr(X_i, Y_i)$$

- with $X, Y \in \mathbb{R}^d$ representing different views (e.g. text & speech) of the same data instance.



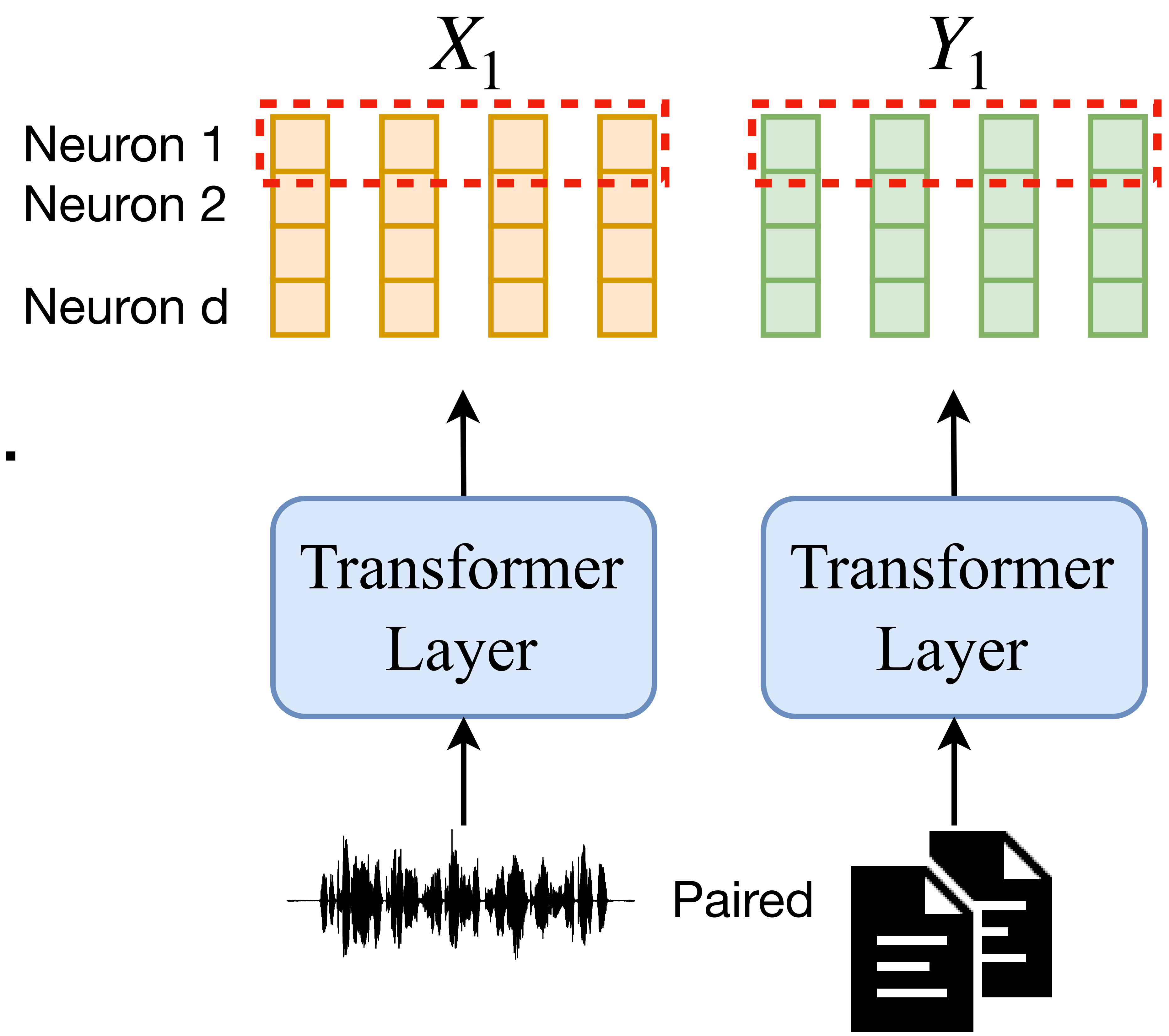

[4] M. Del and M. Fishel, "Cross-lingual similarity of multilingual representations revisited," in AACL, 2022.

# Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) [4]

$$\frac{1}{d}\sum_{i=1}^{d} corr(X_i, Y_i)$$

- with $X, Y \in \mathbb{R}^d$ representing different views (e.g. text & speech) of the same data instance.




[4] M. Del and M. Fishel, "Cross-lingual similarity of multilingual representations revisited," in AACL, 2022.

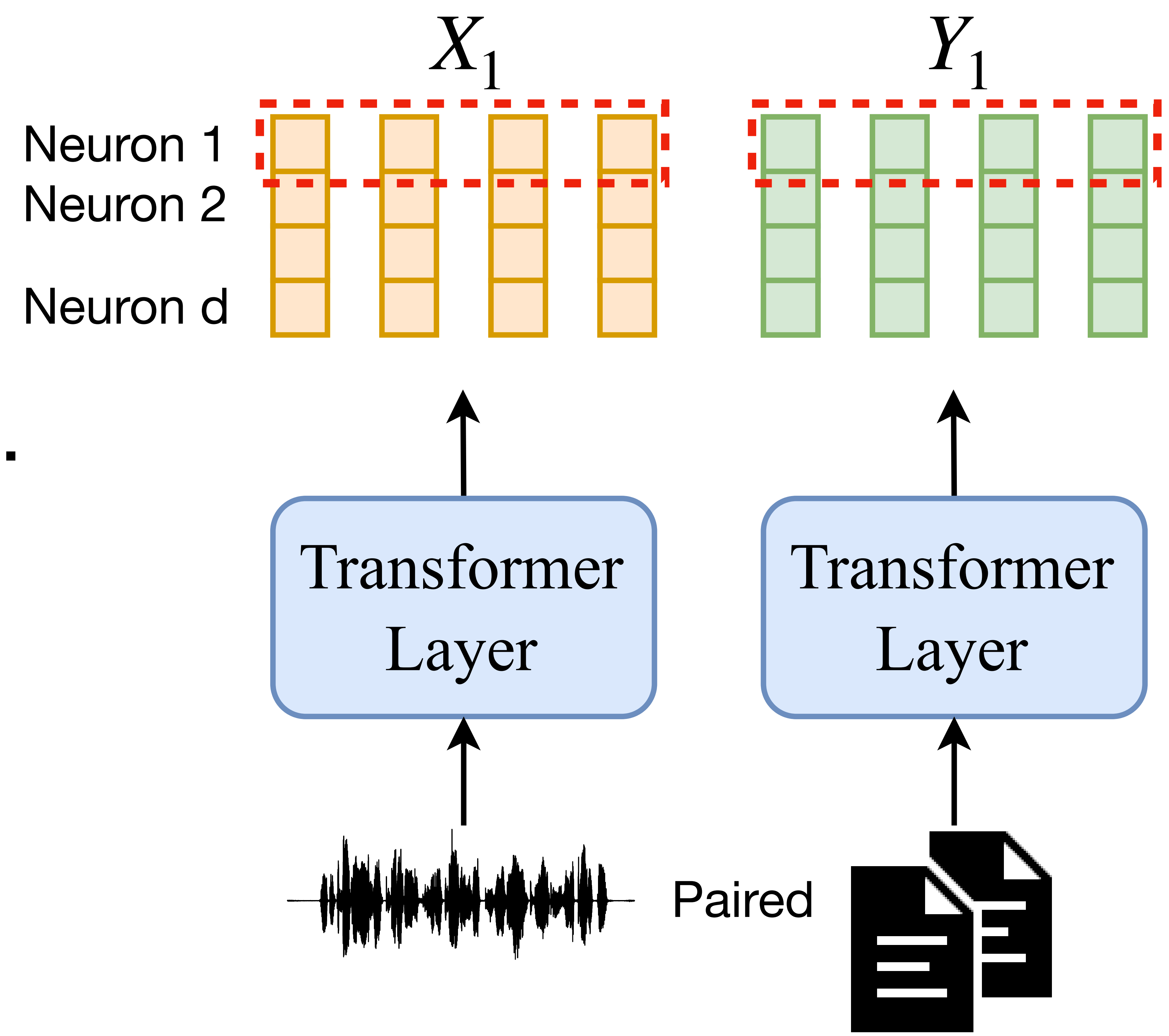
# Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) [4]

$$\frac{1}{d}\sum_{i=1}^{d} corr(X_i, Y_i)$$

- with $X, Y \in \mathbb{R}^d$ representing different views (e.g. text & speech) of the same data instance.

$corr(X_i, Y_i)$: whether speech & text representations are aligned.




[4] M. Del and M. Fishel, "Cross-lingual similarity of multilingual representations revisited," in AACL, 2022.

# Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) [4]

$$\frac{1}{d}\sum_{i=1}^{d} corr(X_i, Y_i)$$

- with $X, Y \in \mathbb{R}^d$ representing different views (e.g. text & speech) of the same data instance.

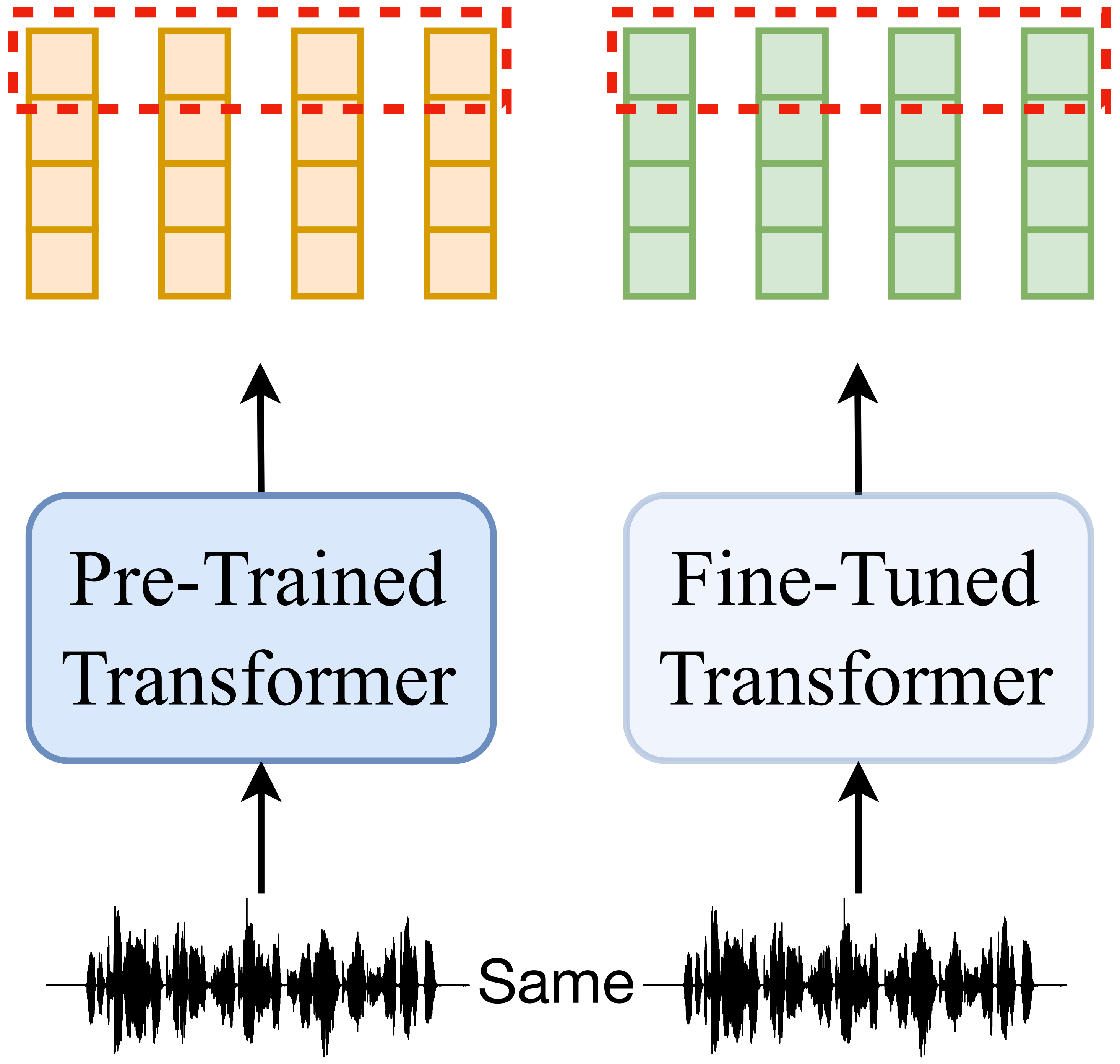
[4] M. Del and M. Fishel, “Cross-lingual similarity of multilingual representations revisited,” in AACL, 2022.

# Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) [4]

$$\frac{1}{d}\sum_{i=1}^{d} corr(X_i, Y_i)$$

- with $X, Y \in \mathbb{R}^d$ representing different views (e.g. text & speech) of the same data instance.




[4] M. Del and M. Fishel, "Cross-lingual similarity of multilingual representations revisited," in AACL, 2022.

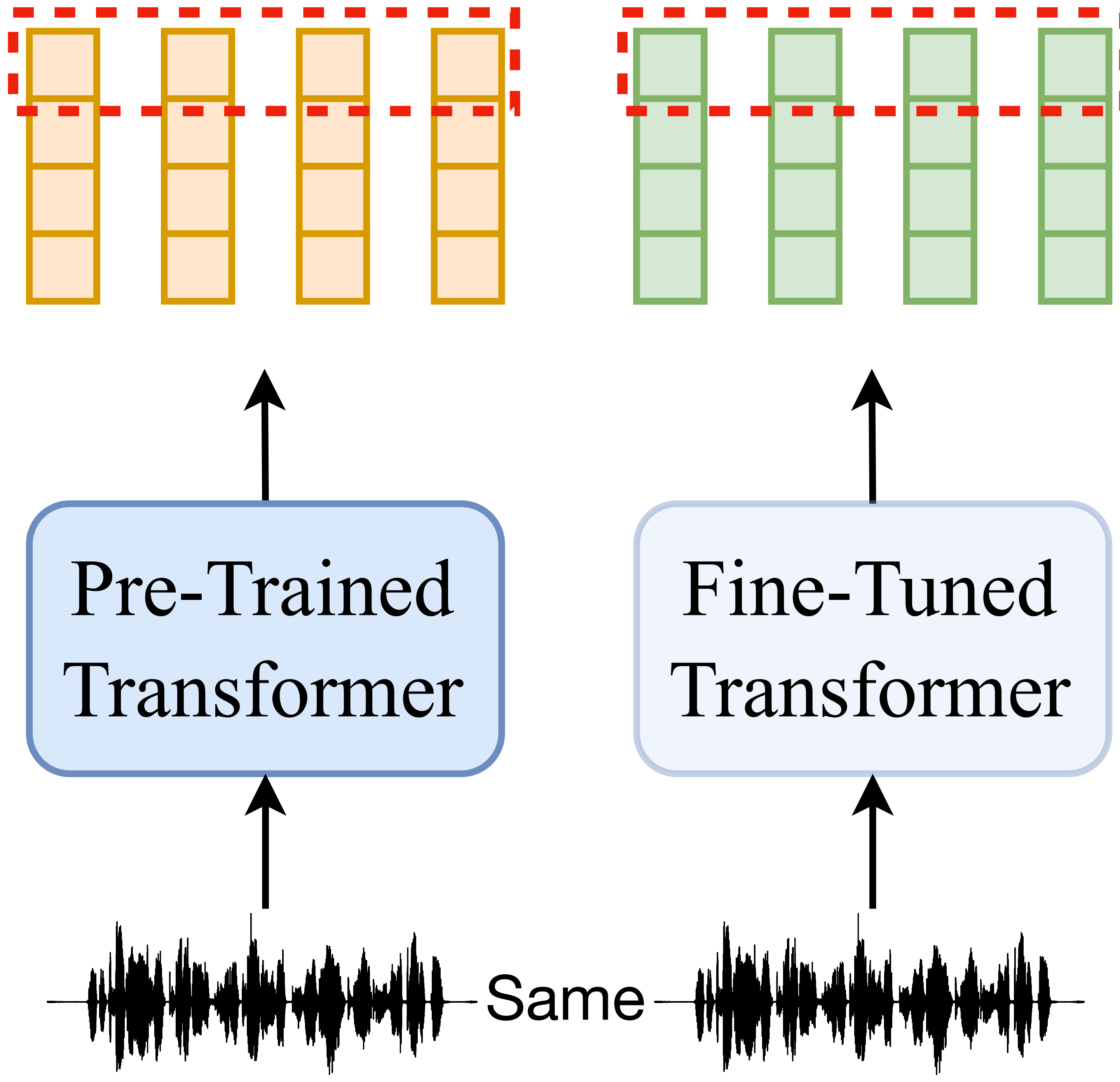
# Analysis Method: Average Neuron-Wise Correlation

- Average Neuron-Wise Correlation (ANC) [4]

$$\frac{1}{d}\sum_{i=1}^{d} corr(X_i, Y_i)$$

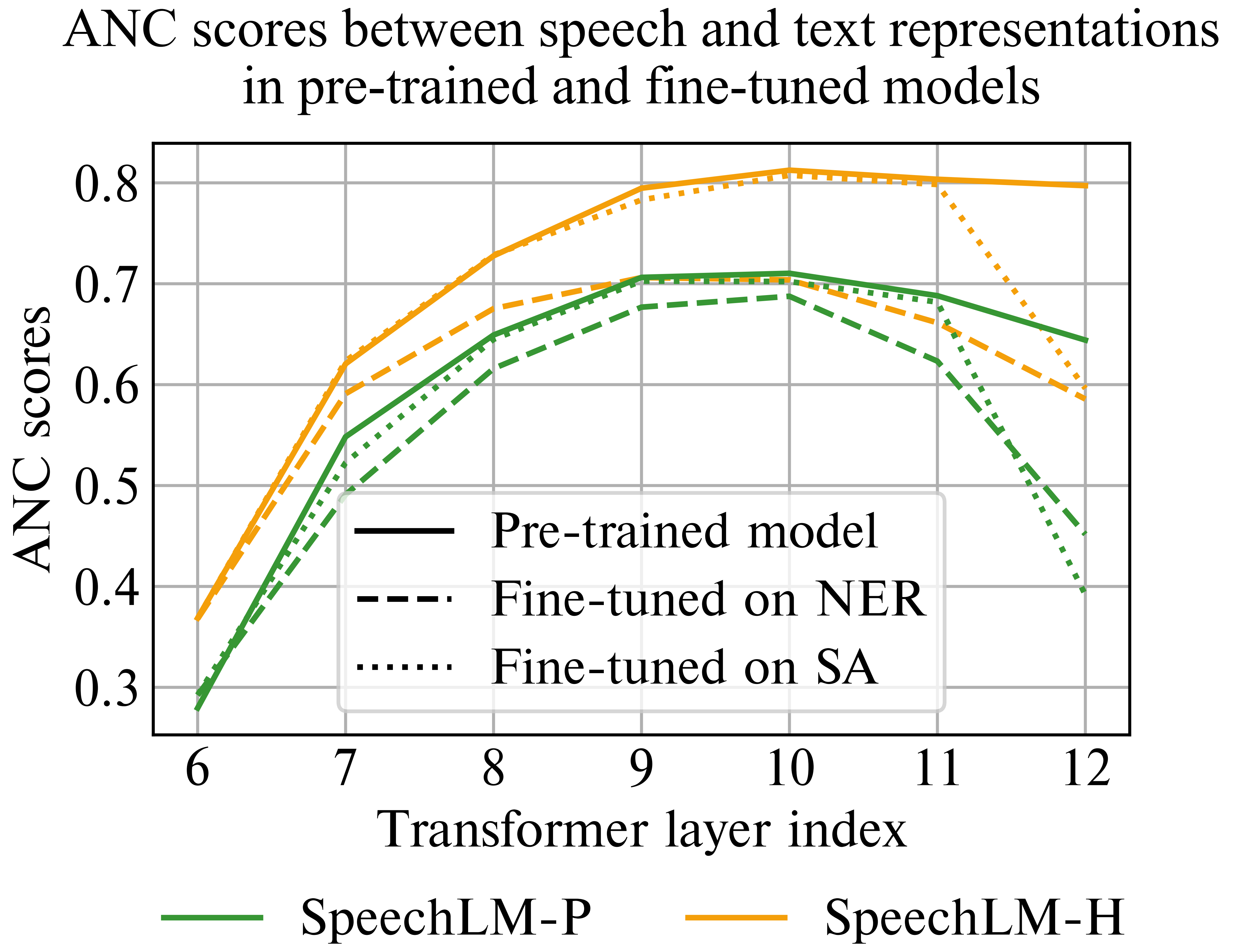
- with $X, Y \in \mathbb{R}^d$ representing different views (e.g. text & speech) of the same data instance.

$corr(X_i, Y_i)$: how much pre-trained and fine-tuned models differ.



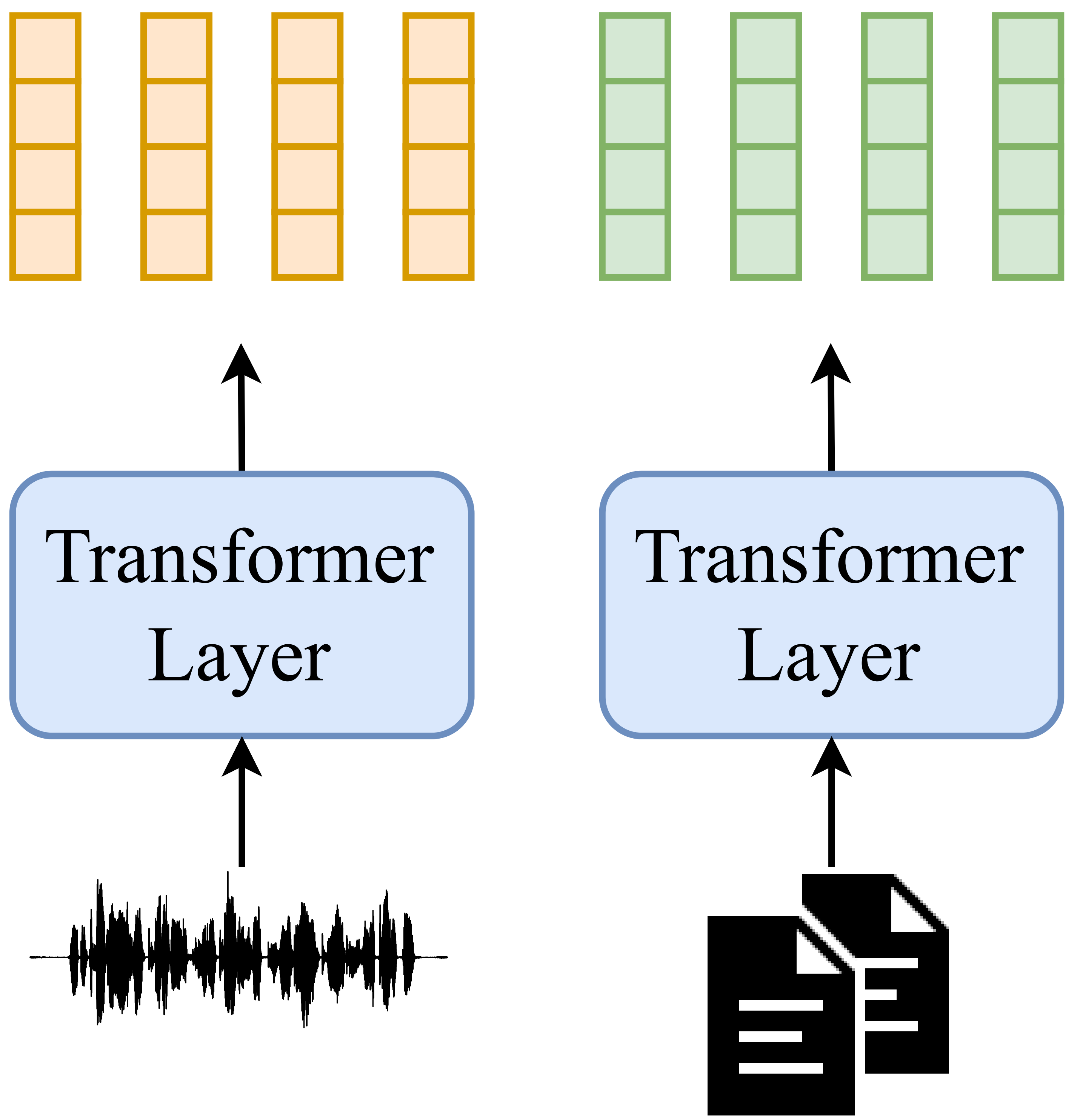

[4] M. Del and M. Fishel, "Cross-lingual similarity of multilingual representations revisited," in AACL, 2022.

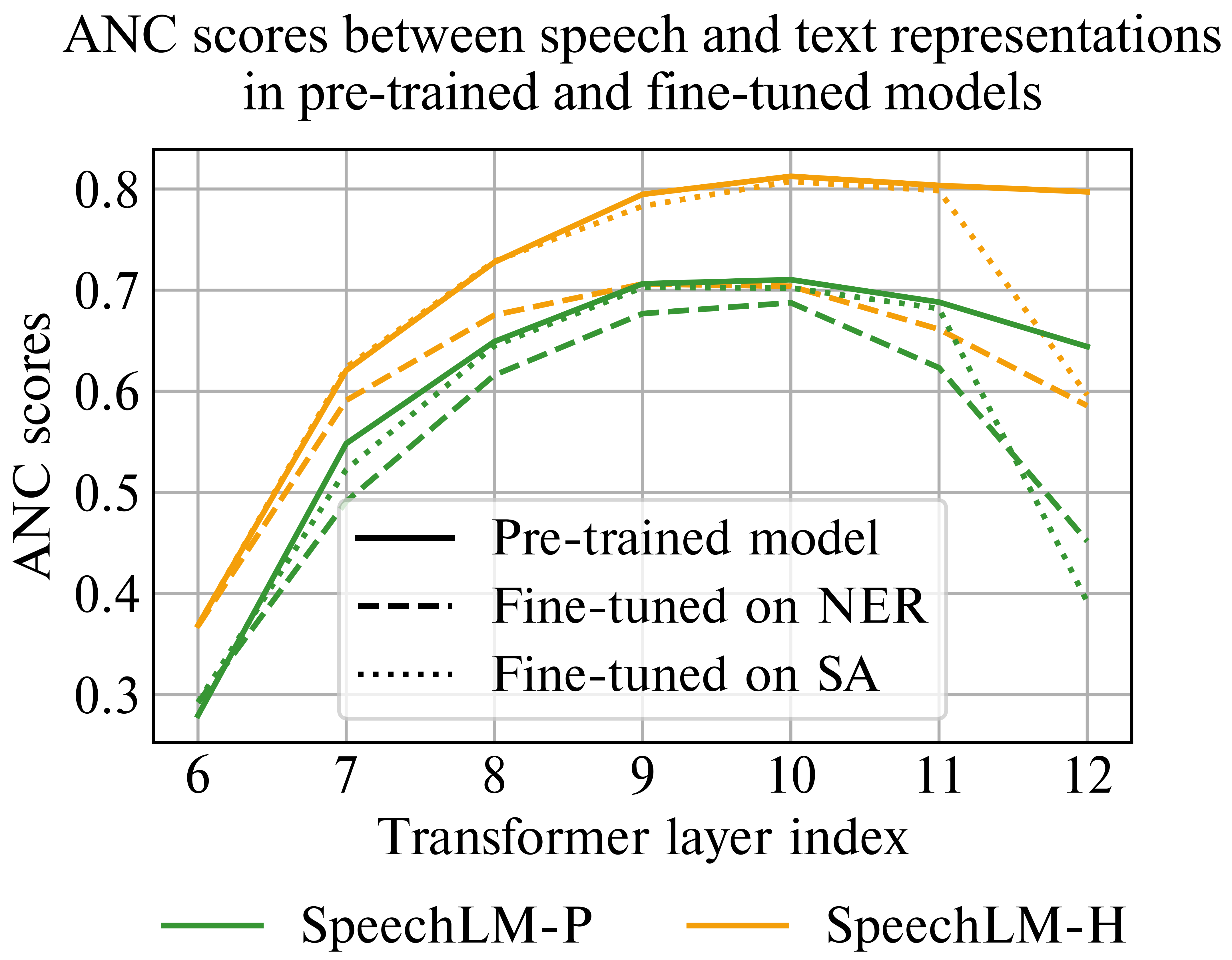
# Representation Alignment in Pre-Trained & Fine-Tuned Models



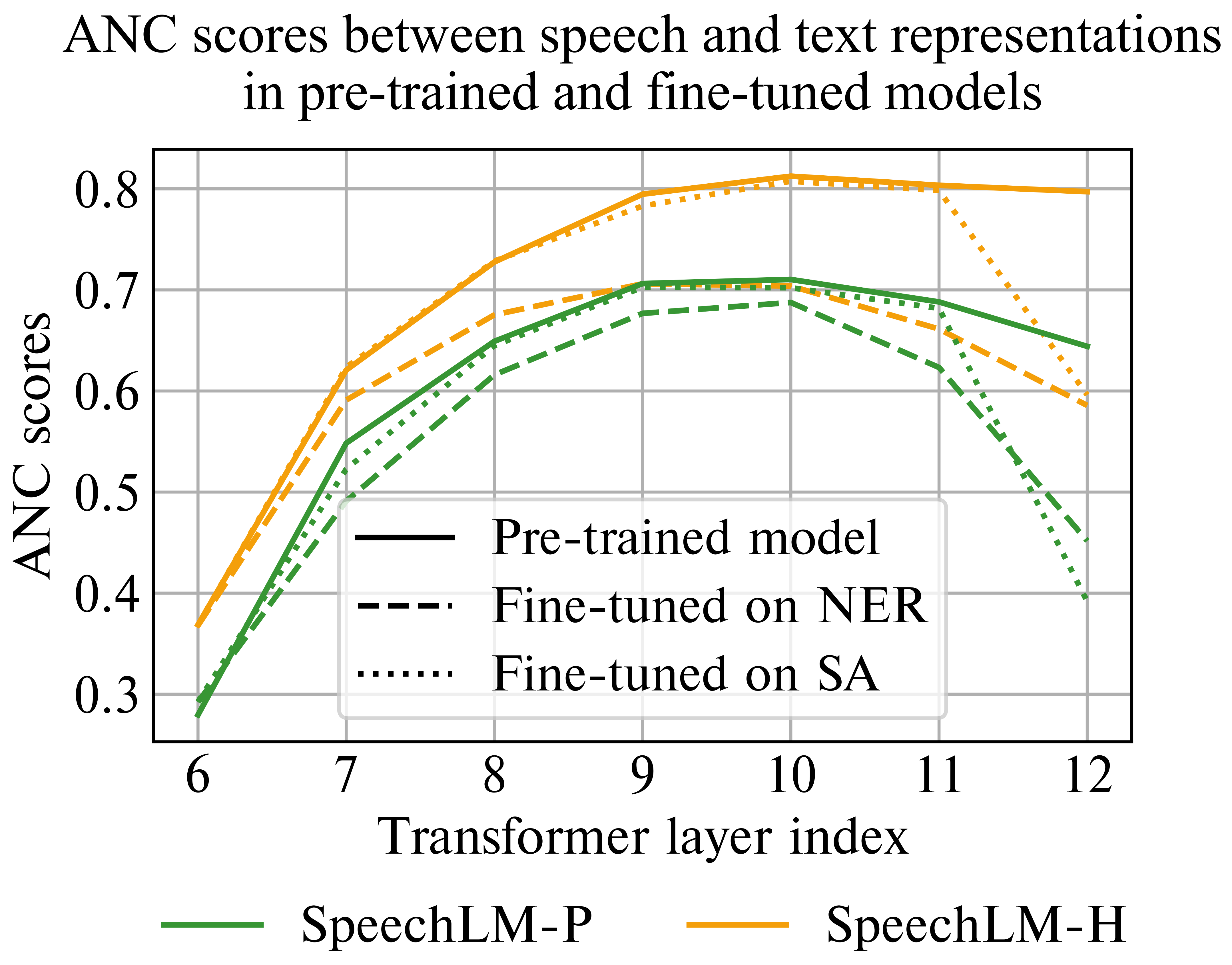
# Representation Alignment in Pre-Trained & Fine-Tuned Models

$corr(X_i, Y_i)$: whether speech & text representations are aligned.

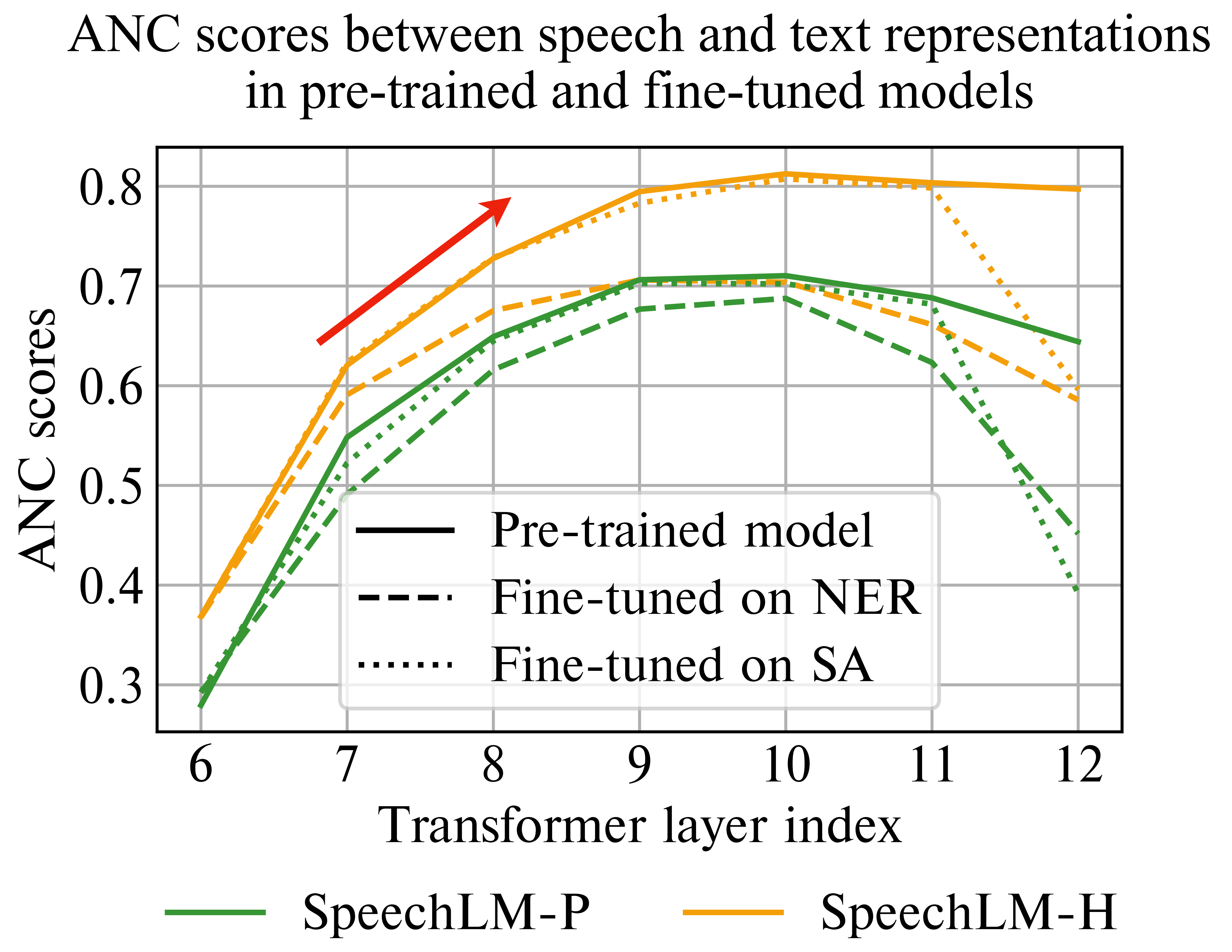


# Representation Alignment in Pre-Trained & Fine-Tuned Models

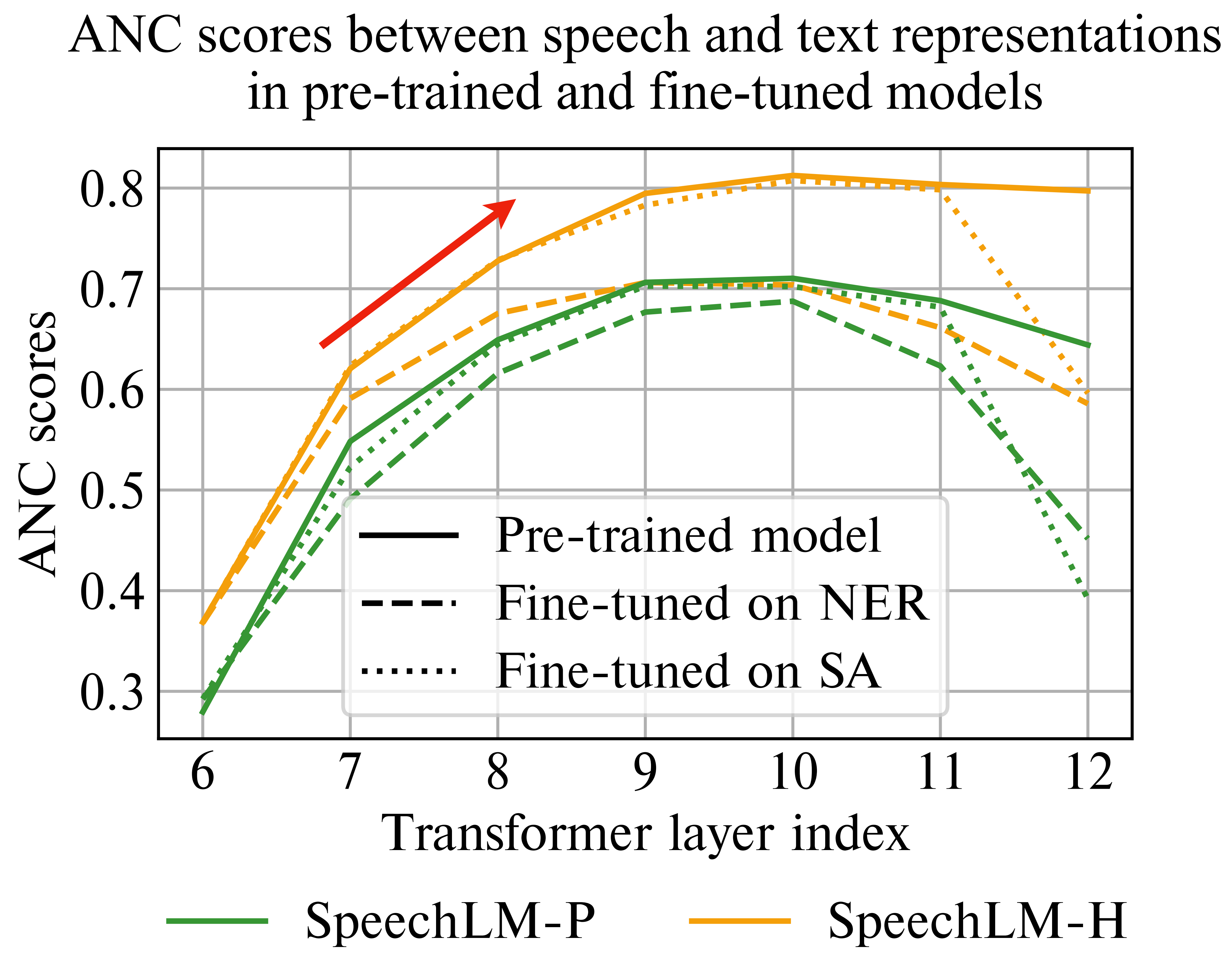


# Representation Alignment in Pre-Trained & Fine-Tuned Models



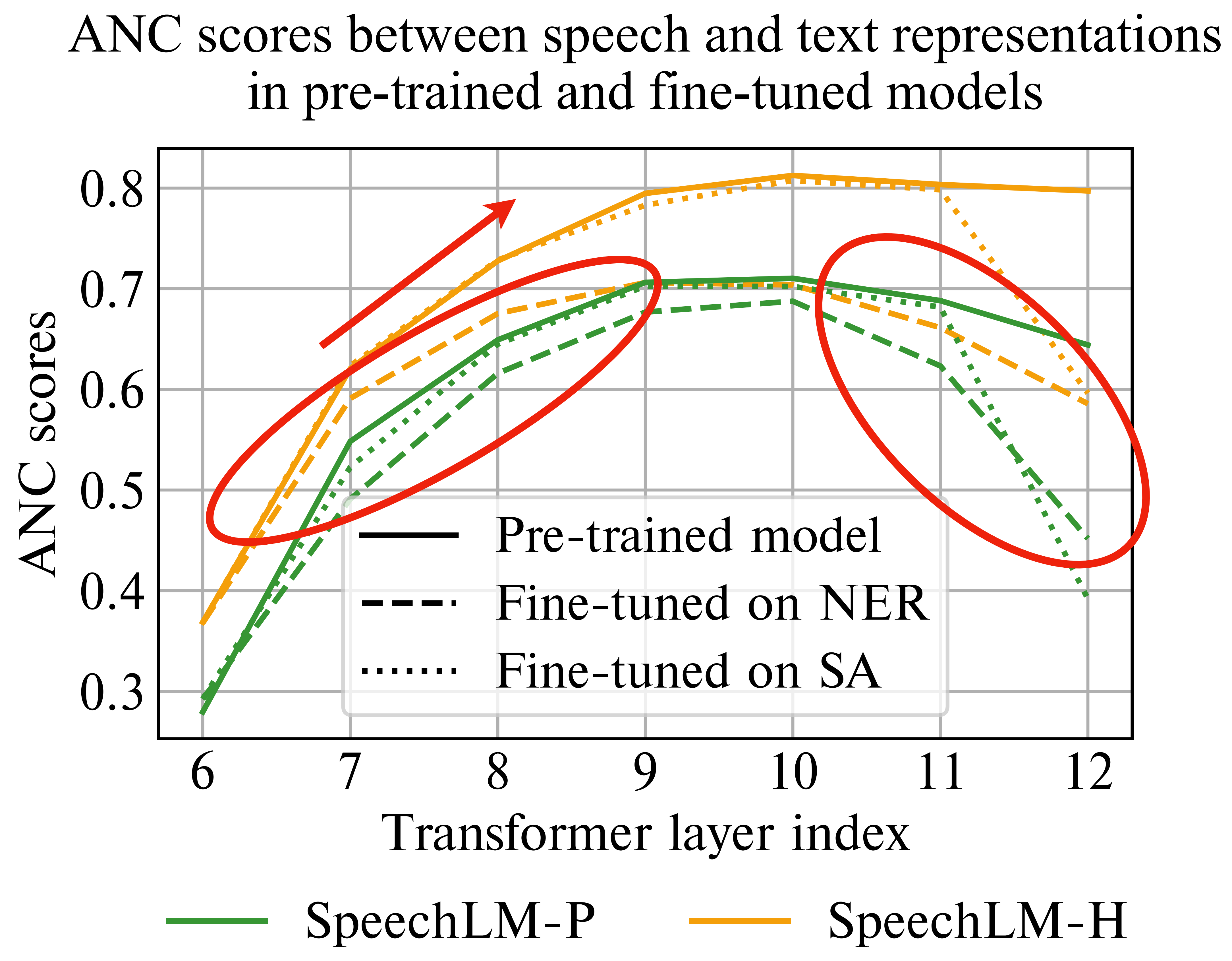
# Representation Alignment in Pre-Trained & Fine-Tuned Models

- Speech-text models learn aligned speech & text representations in bottom layers.



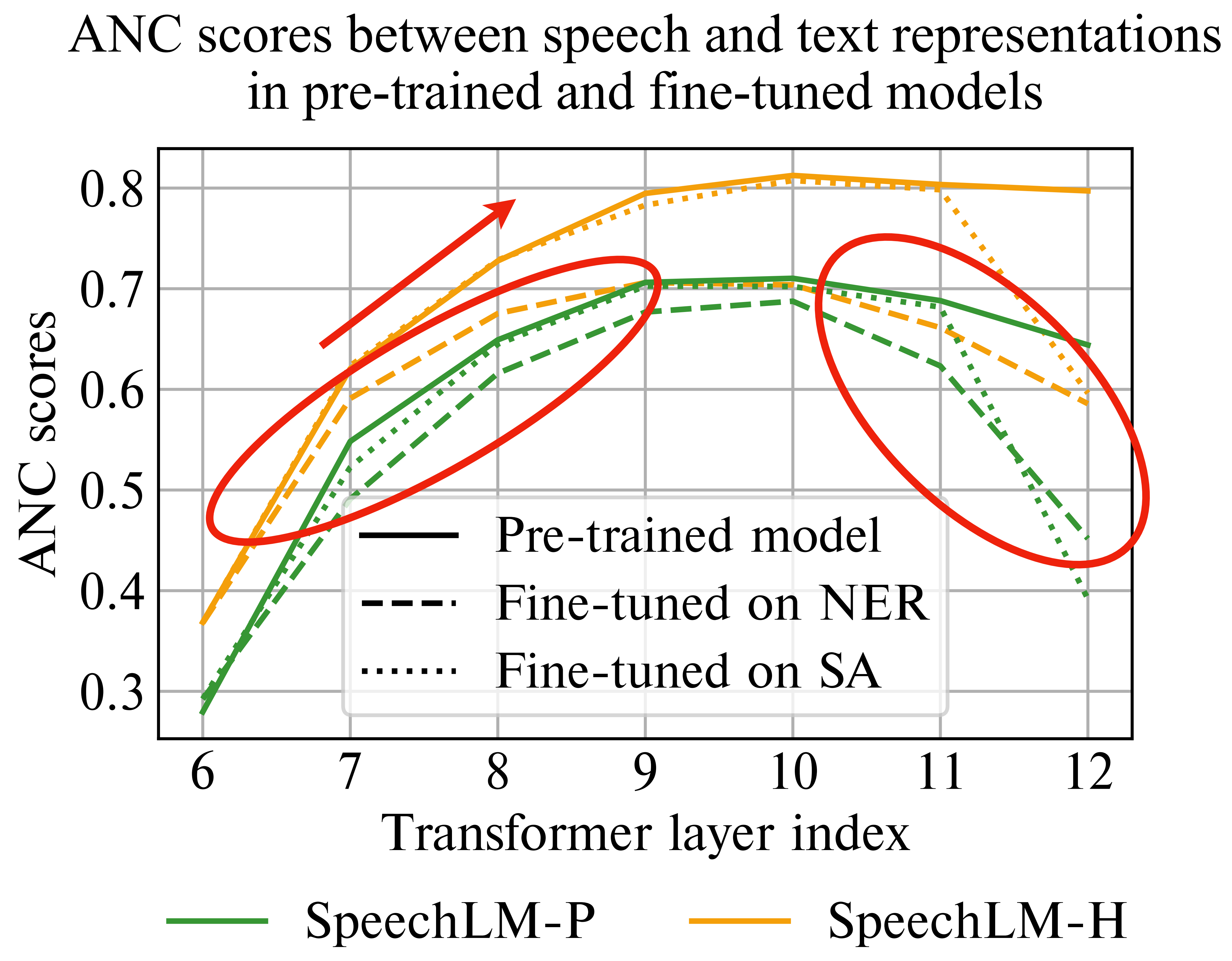
# Representation Alignment in Pre-Trained & Fine-Tuned Models

- Speech-text models learn aligned speech & text representations in bottom layers.



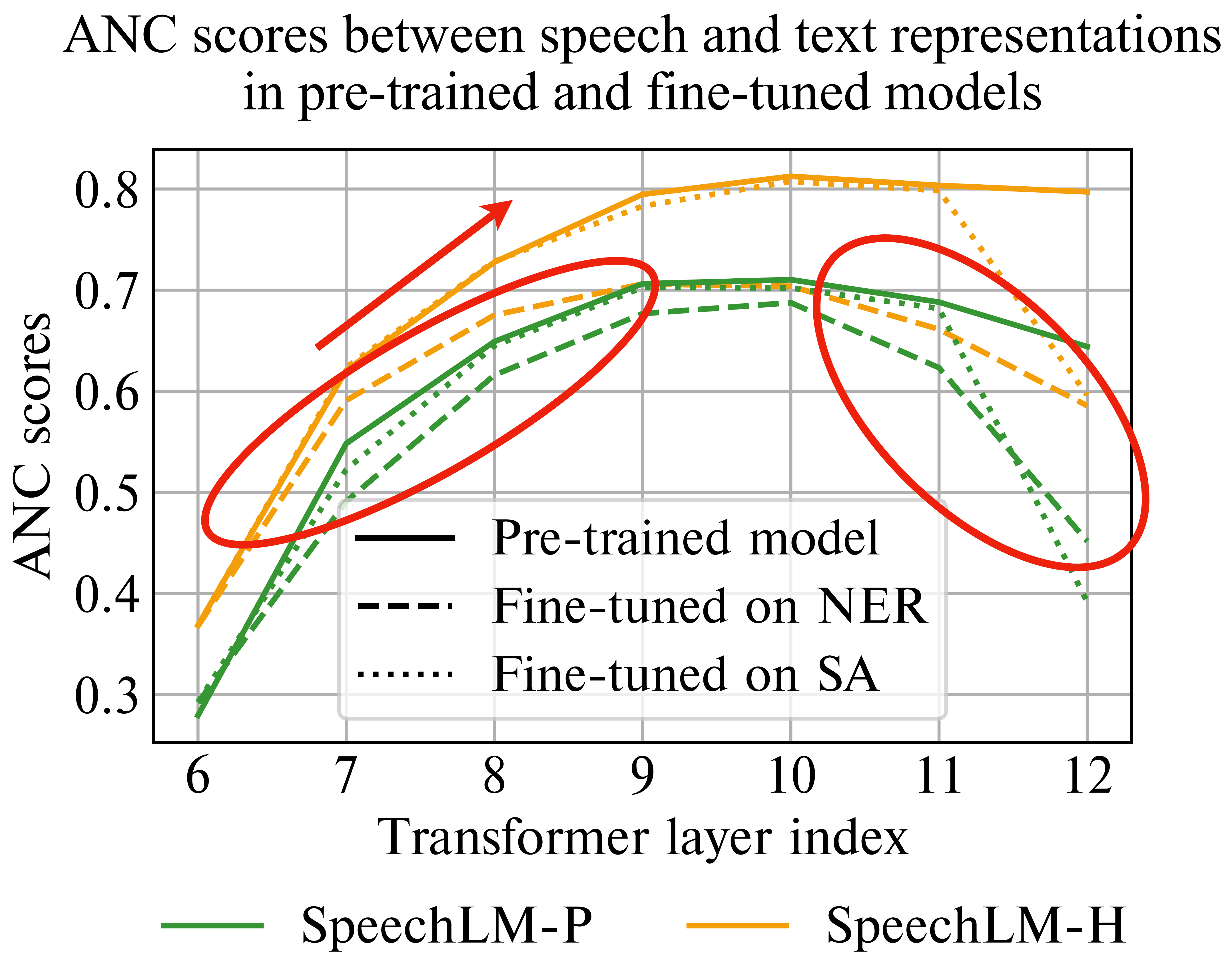
# Representation Alignment in Pre-Trained & Fine-Tuned Models

- Speech-text models learn aligned speech & text representations in bottom layers.
- Pre-trained & fine-tuned models are similar in bottom layers and differ more in top layers.

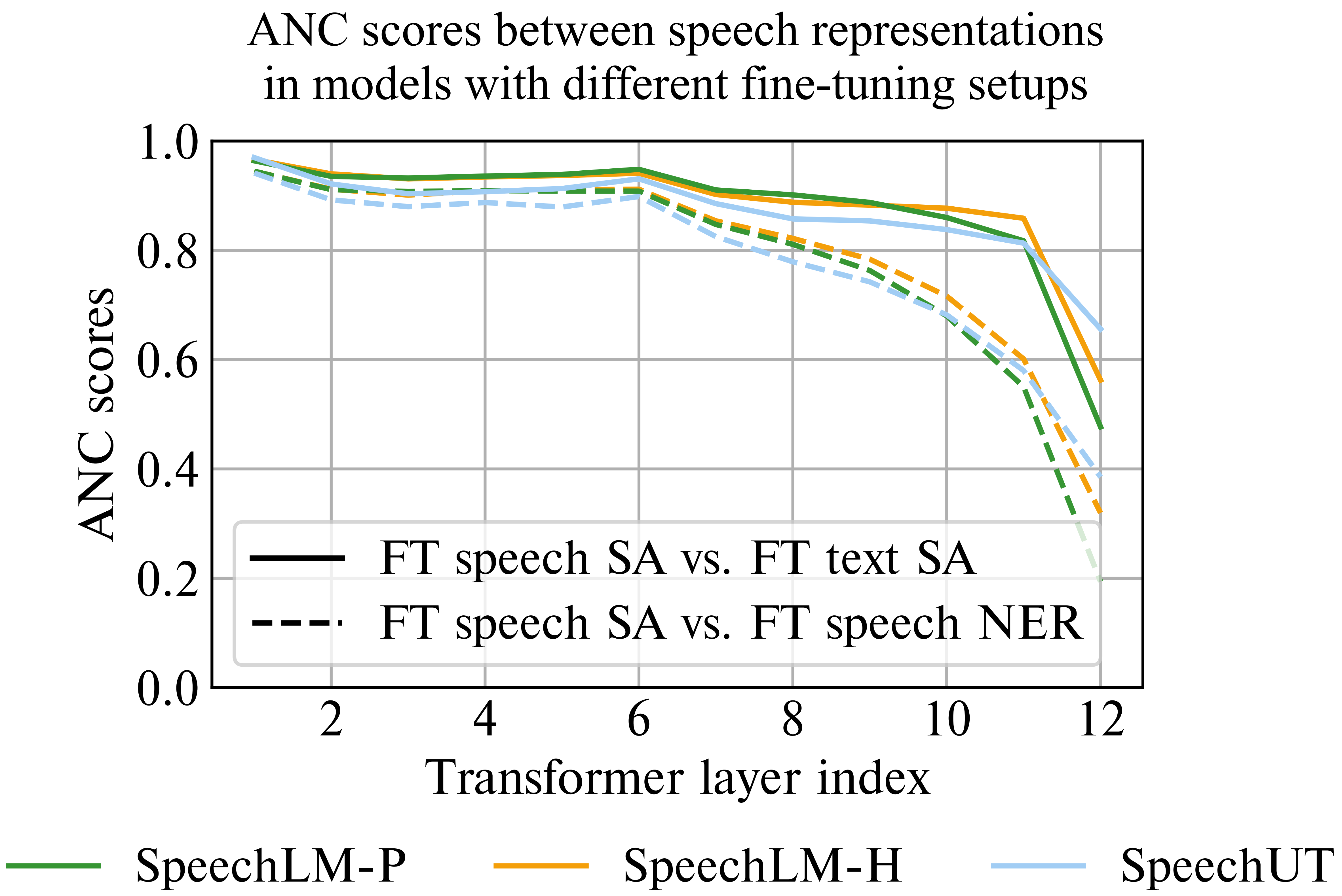


# Representation Alignment in Pre-Trained & Fine-Tuned Models

- Speech-text models learn aligned speech & text representations in bottom layers.
- Pre-trained & fine-tuned models are similar in bottom layers and differ more in top layers.
  - Fine-tuning affects top layers more.

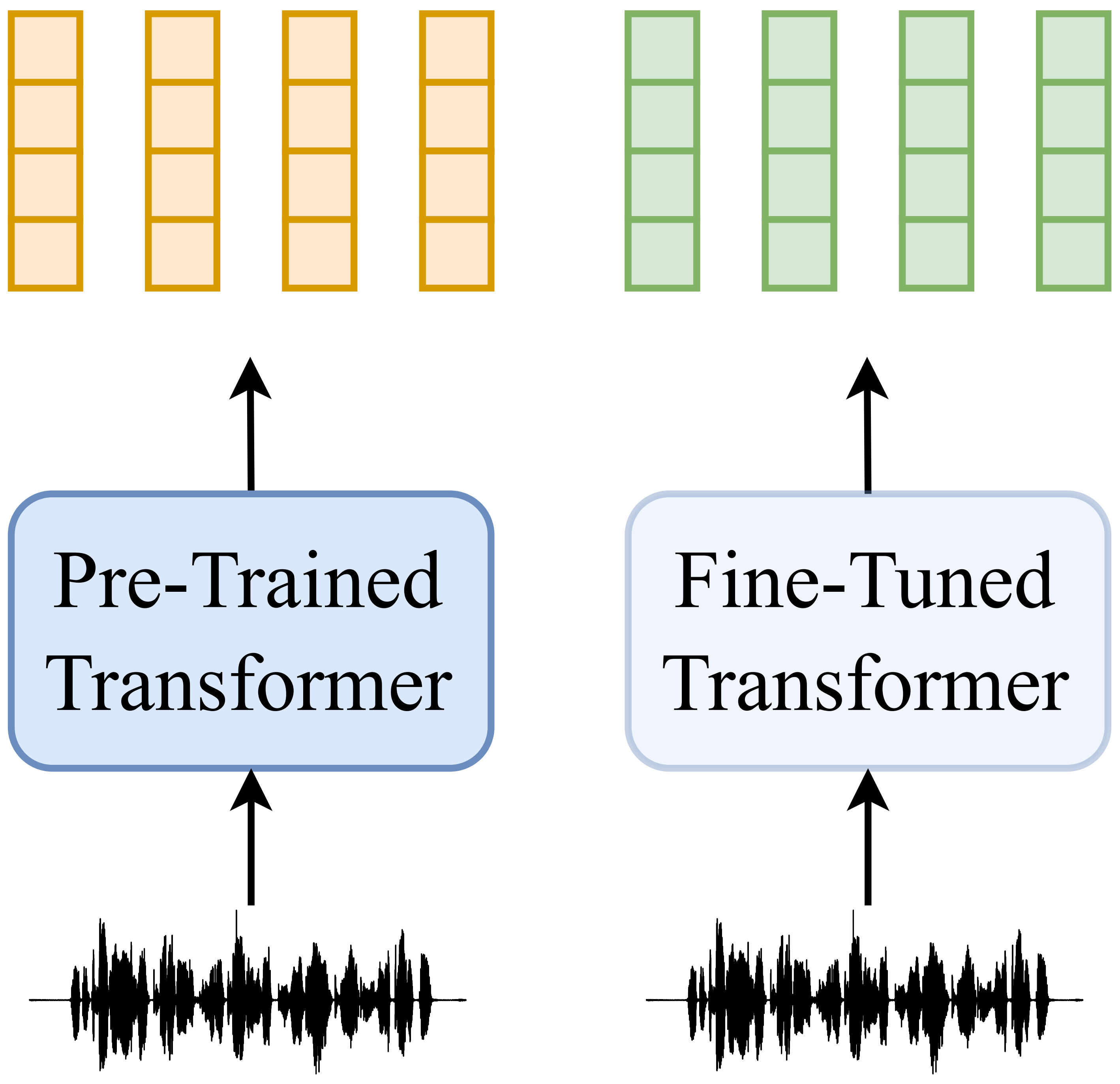


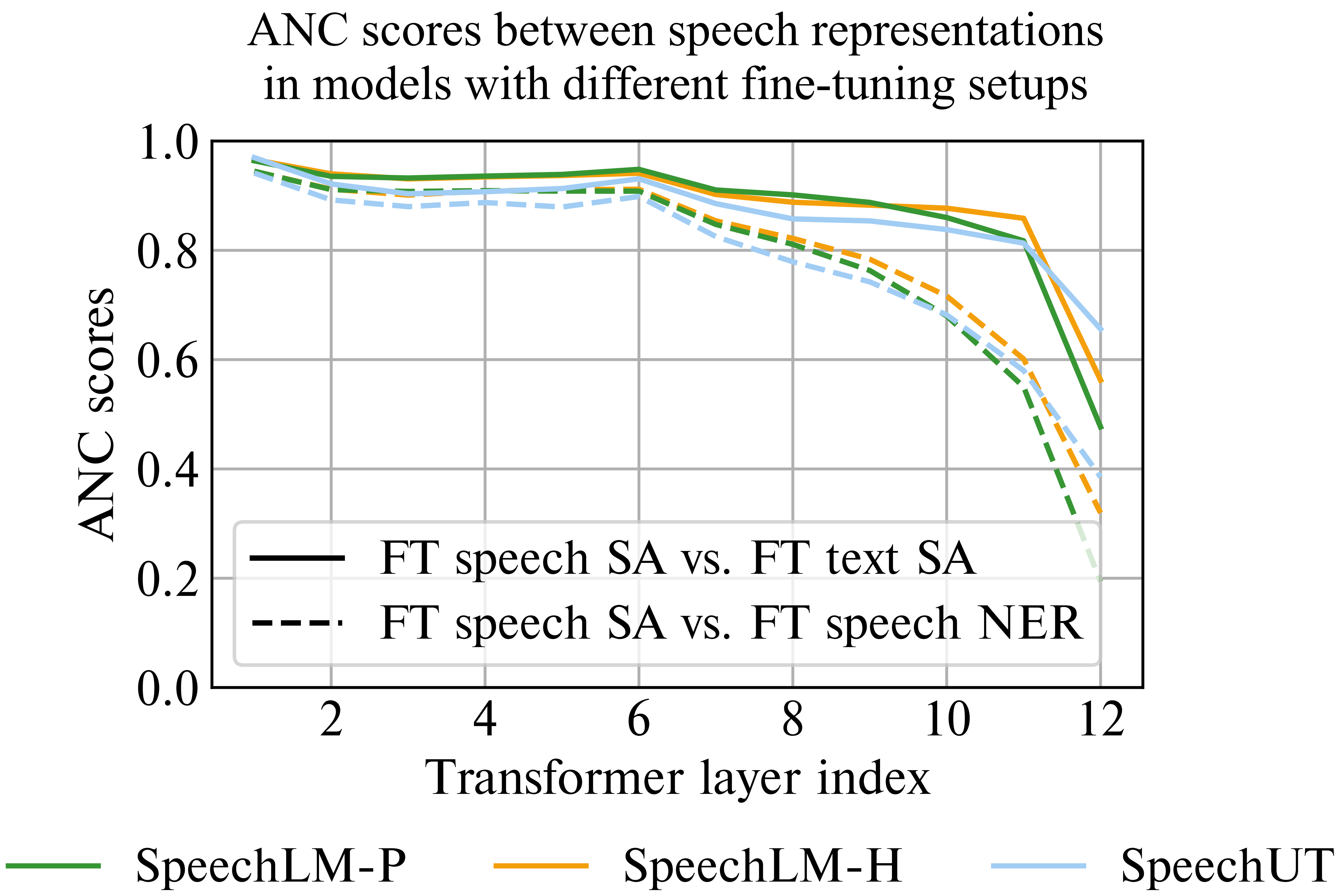
# Top Layers Are Task Specific



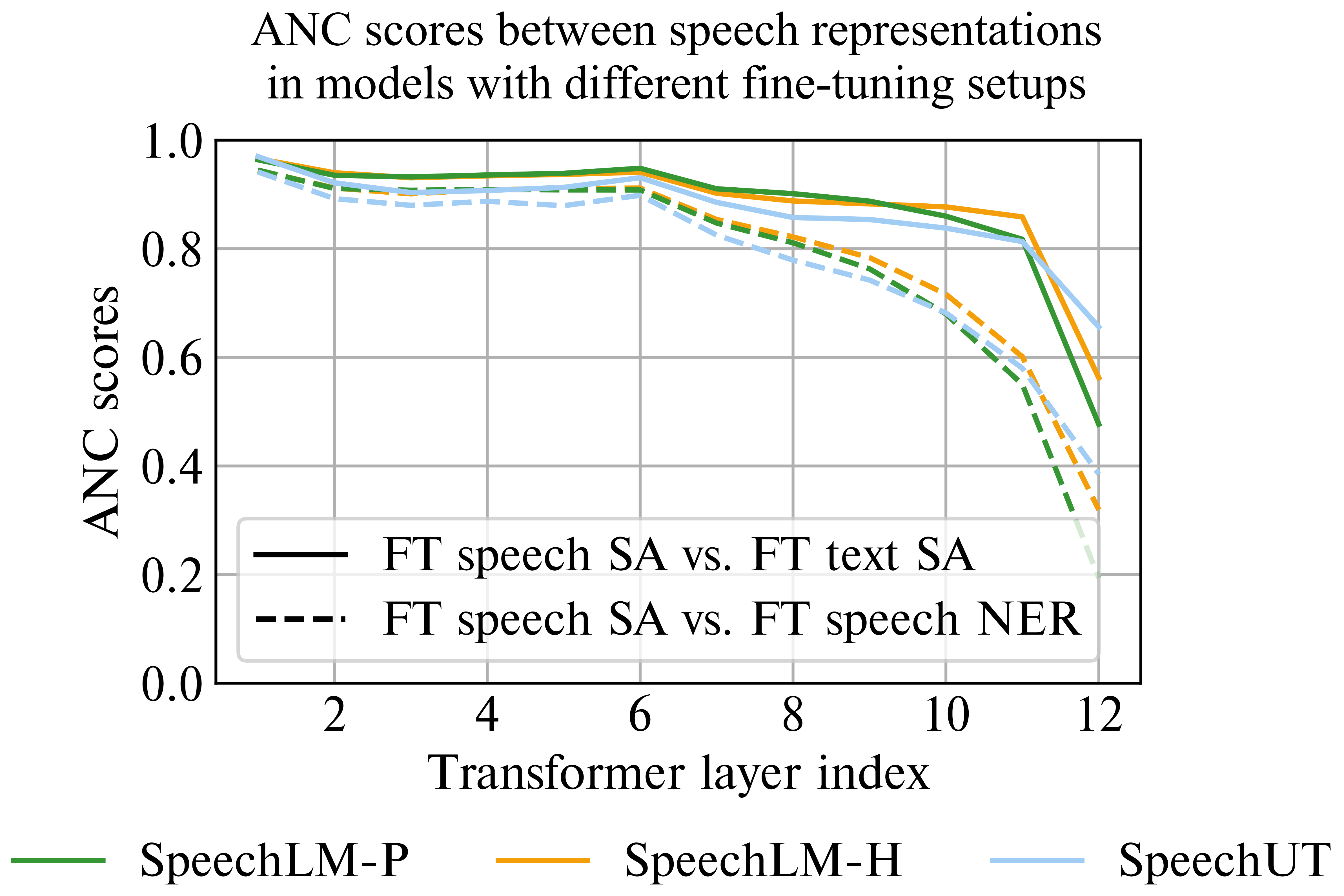
# Top Layers Are Task Specific

$corr(X_i, Y_i)$: how much pre-trained and fine-tuned models differ.



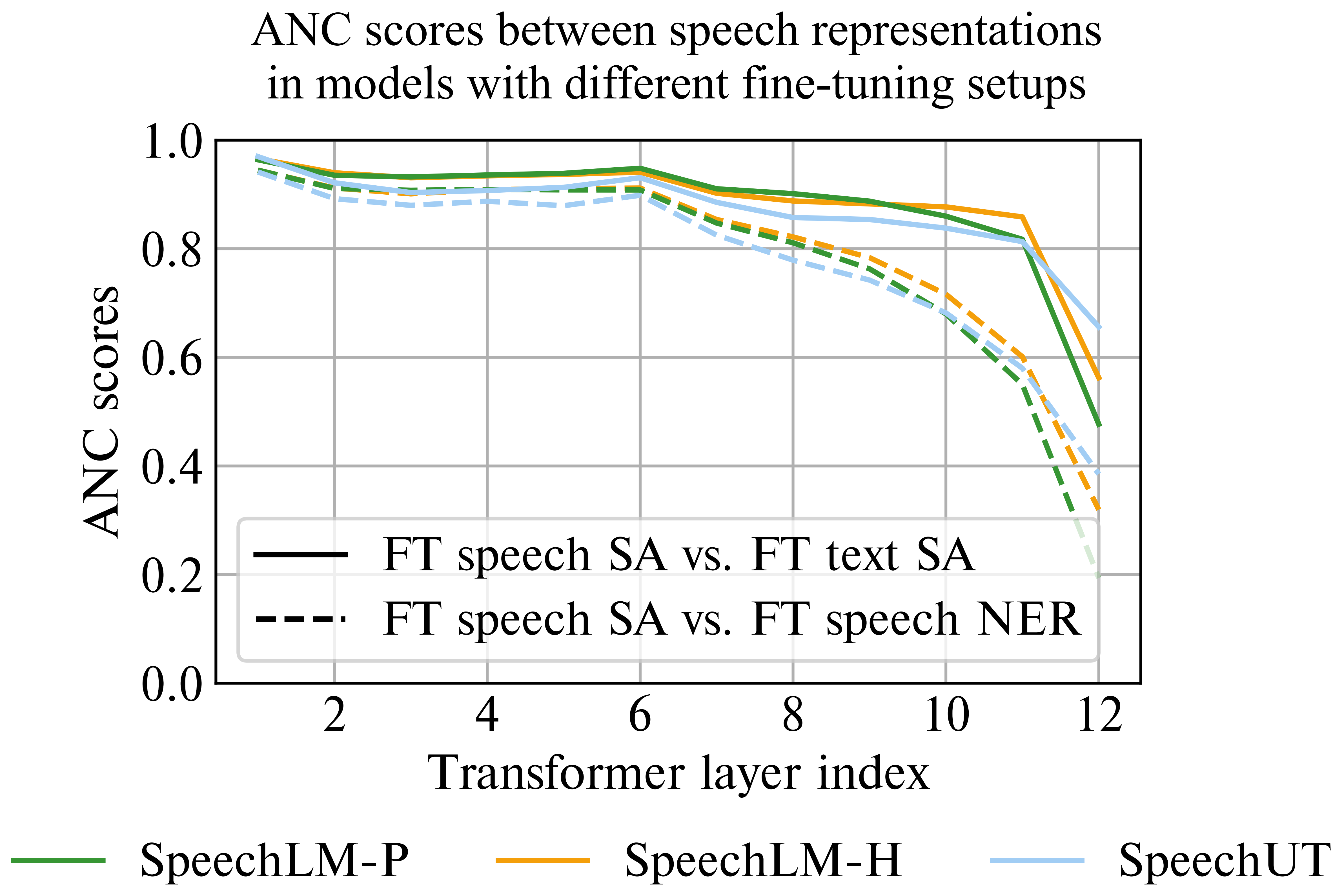



# Top Layers Are Task Specific



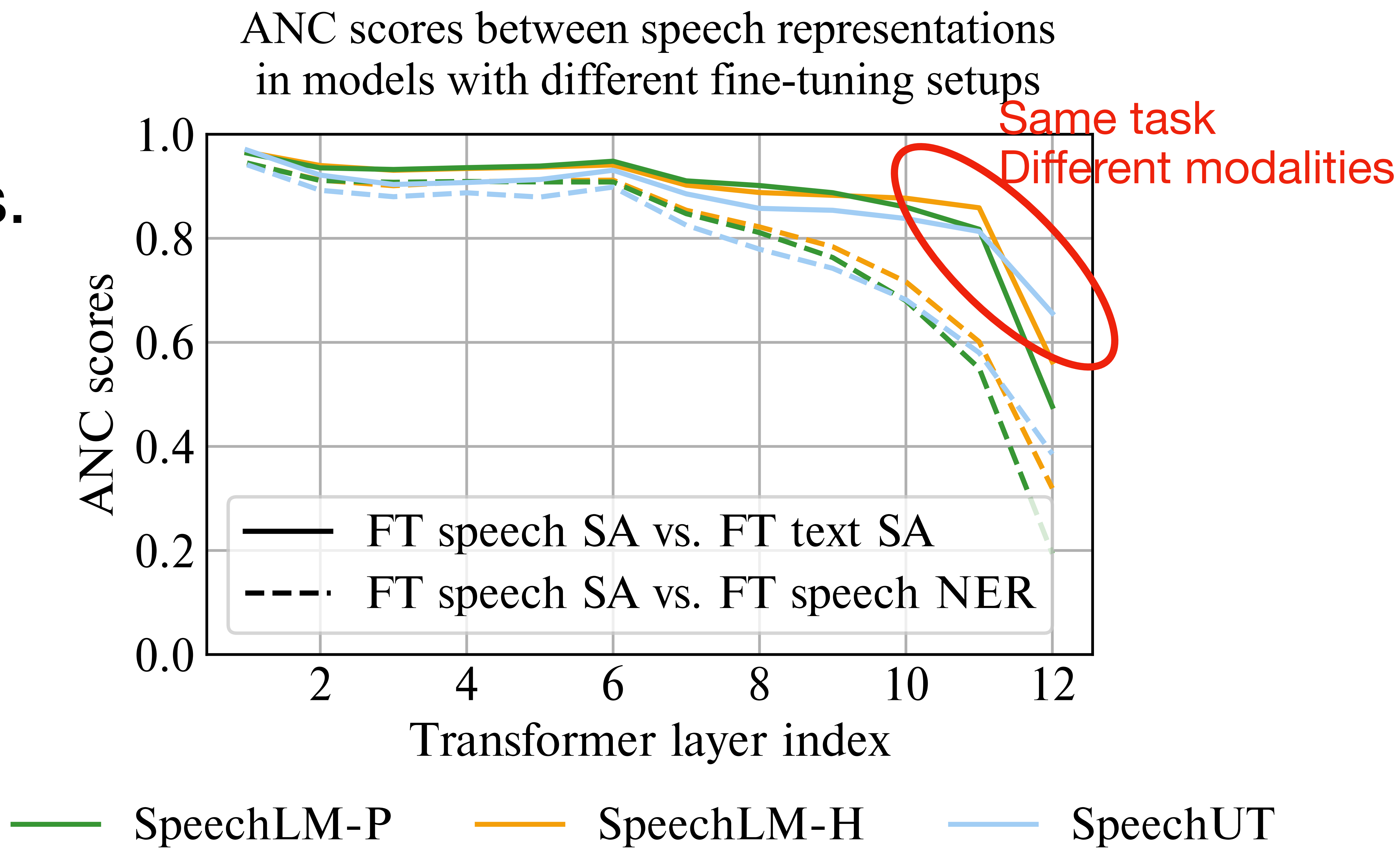
# Top Layers Are Task Specific

- We compare



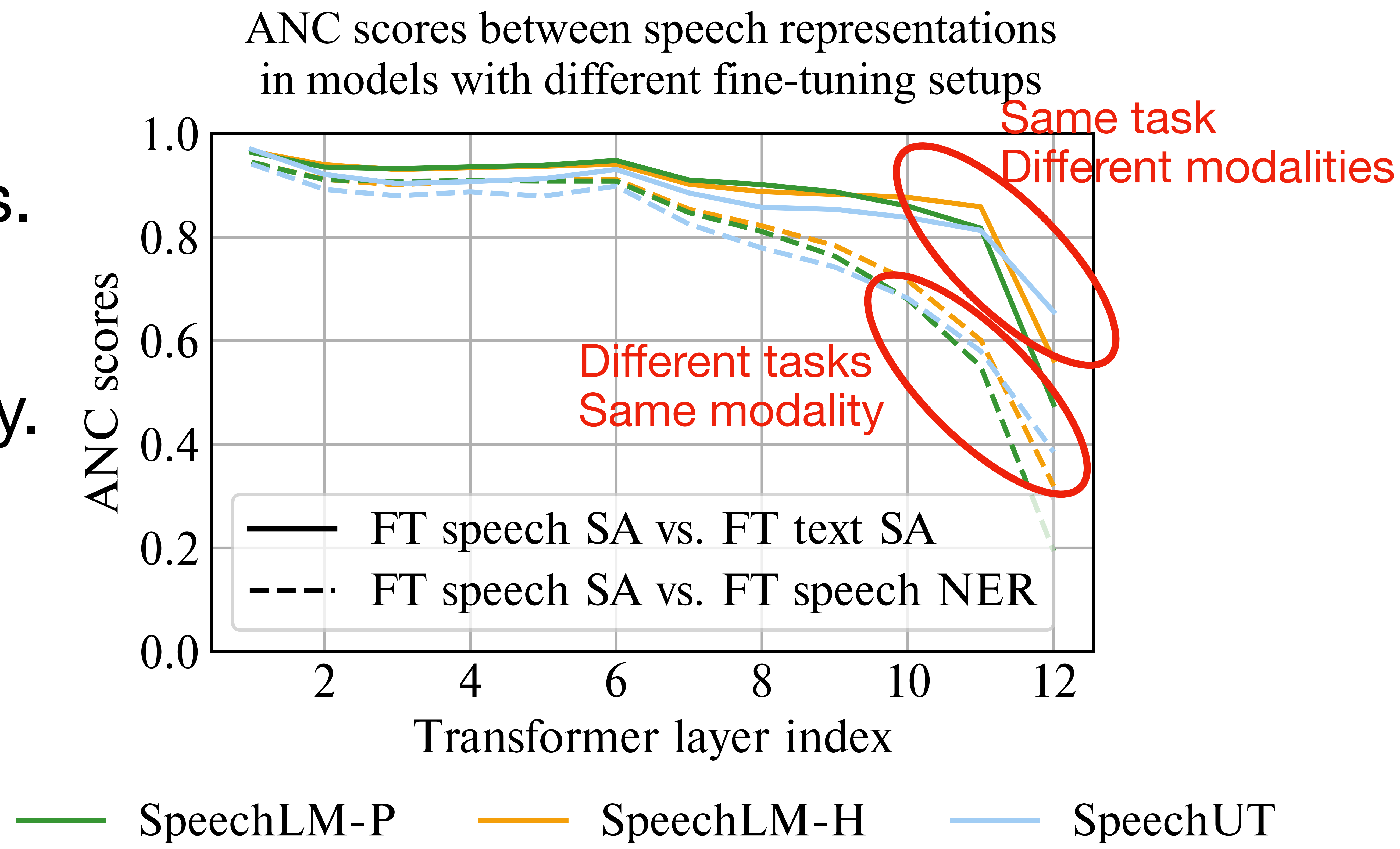
# Top Layers Are Task Specific

- We compare
  - Models fine-tuned on the same task with different input modalities.



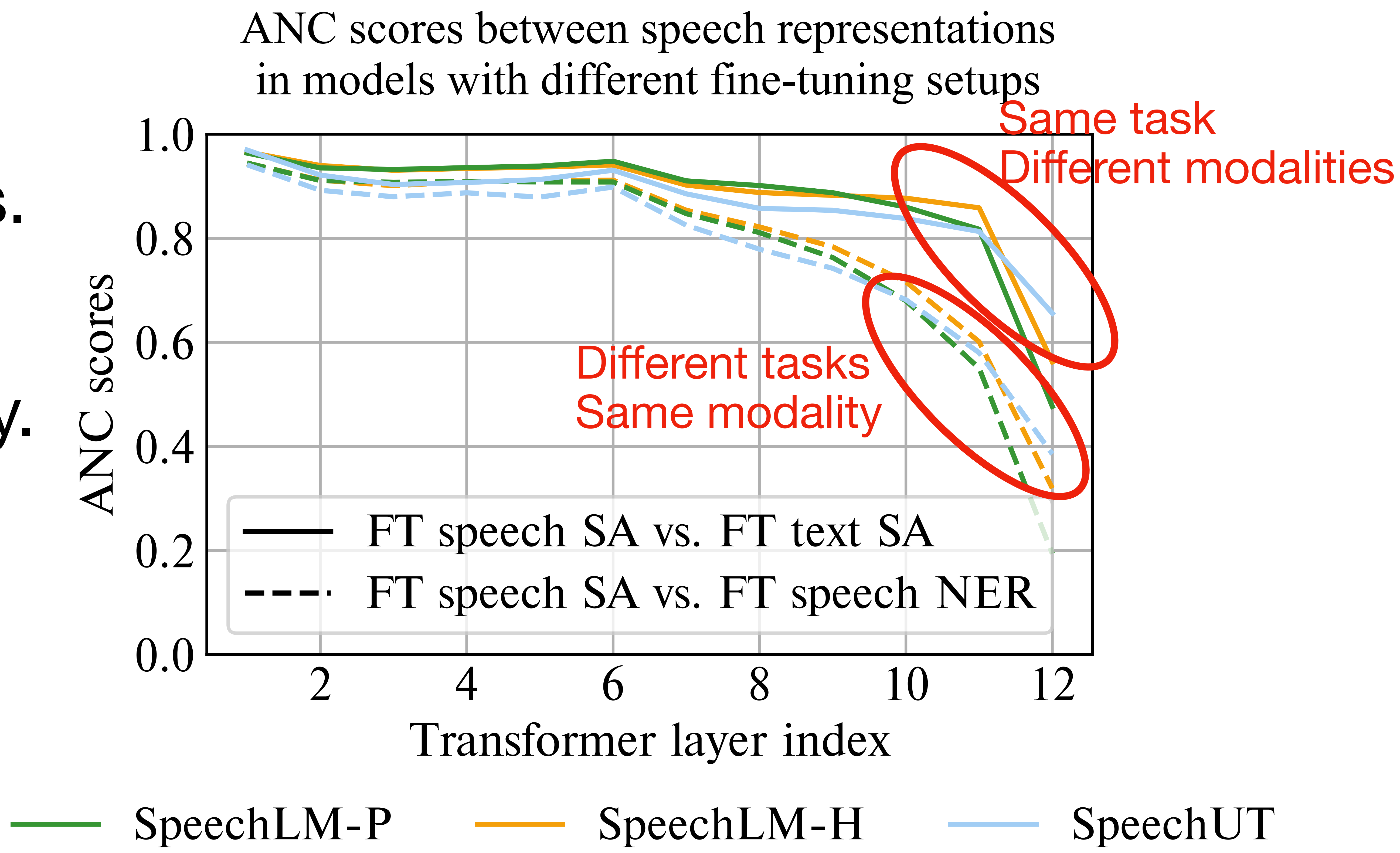
# Top Layers Are Task Specific

- We compare
  - Models fine-tuned on the same task with different input modalities.
  - Models fine-tuned on different tasks with the same input modality.



# Top Layers Are Task Specific

- We compare
  - Models fine-tuned on the same task with different input modalities.
  - Models fine-tuned on different tasks with the same input modality.
- During fine-tuning, the task makes a larger difference than the input mo-daily to top layers.



# Inspired By the Analysis…

# Inspired By the Analysis…

- Bottom layers align speech & text representations.

# Inspired By the Analysis…

- Bottom layers align speech & text representations.
  - Should not be affected by fine-tuning.

# Inspired By the Analysis…

- Bottom layers align speech & text representations.
  - Should not be affected by fine-tuning.
- Top layers are task specific.

# Inspired By the Analysis…

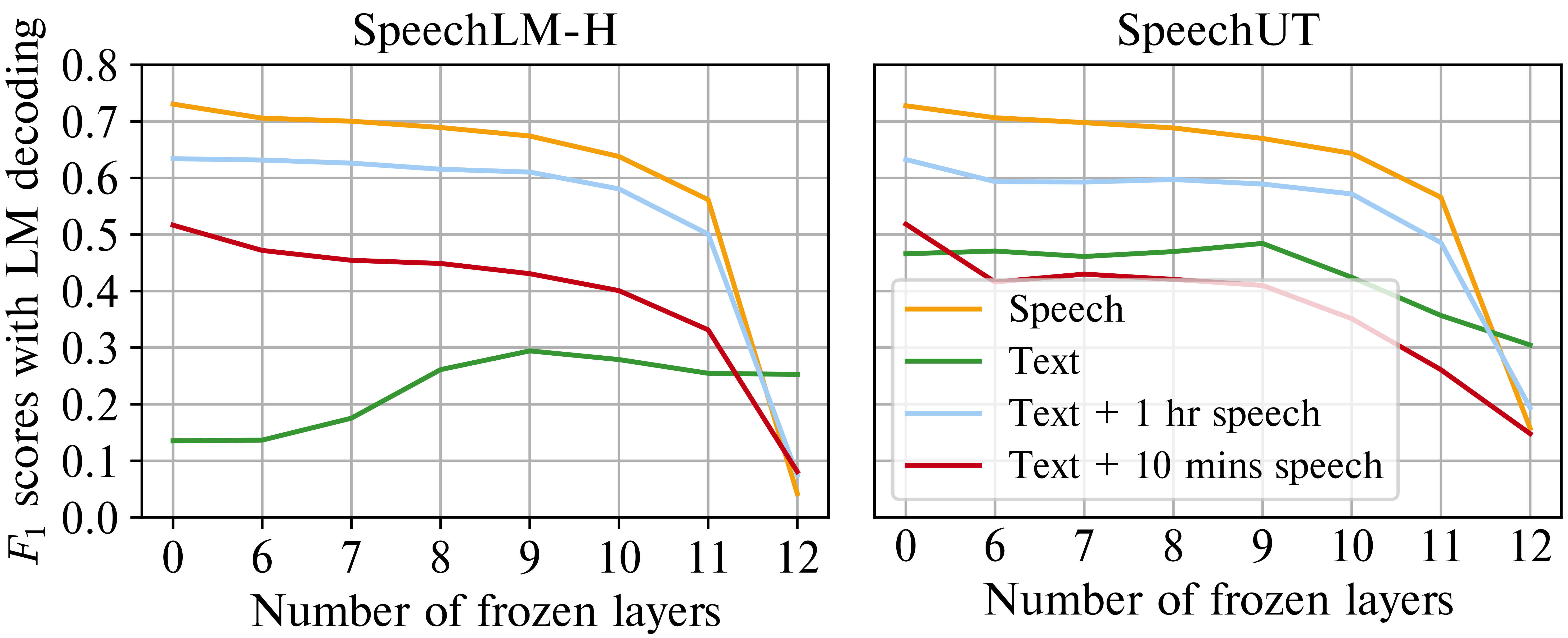
- Bottom layers align speech & text representations.
  - Should not be affected by fine-tuning.
- Top layers are task specific.
  - Should be fine-tuned.

# Inspired By the Analysis…

- Bottom layers align speech & text representations.
  - Should not be affected by fine-tuning.
- Top layers are task specific.
  - Should be fine-tuned.
- How about fine-tuning only top layers and keeping bottom layers frozen?

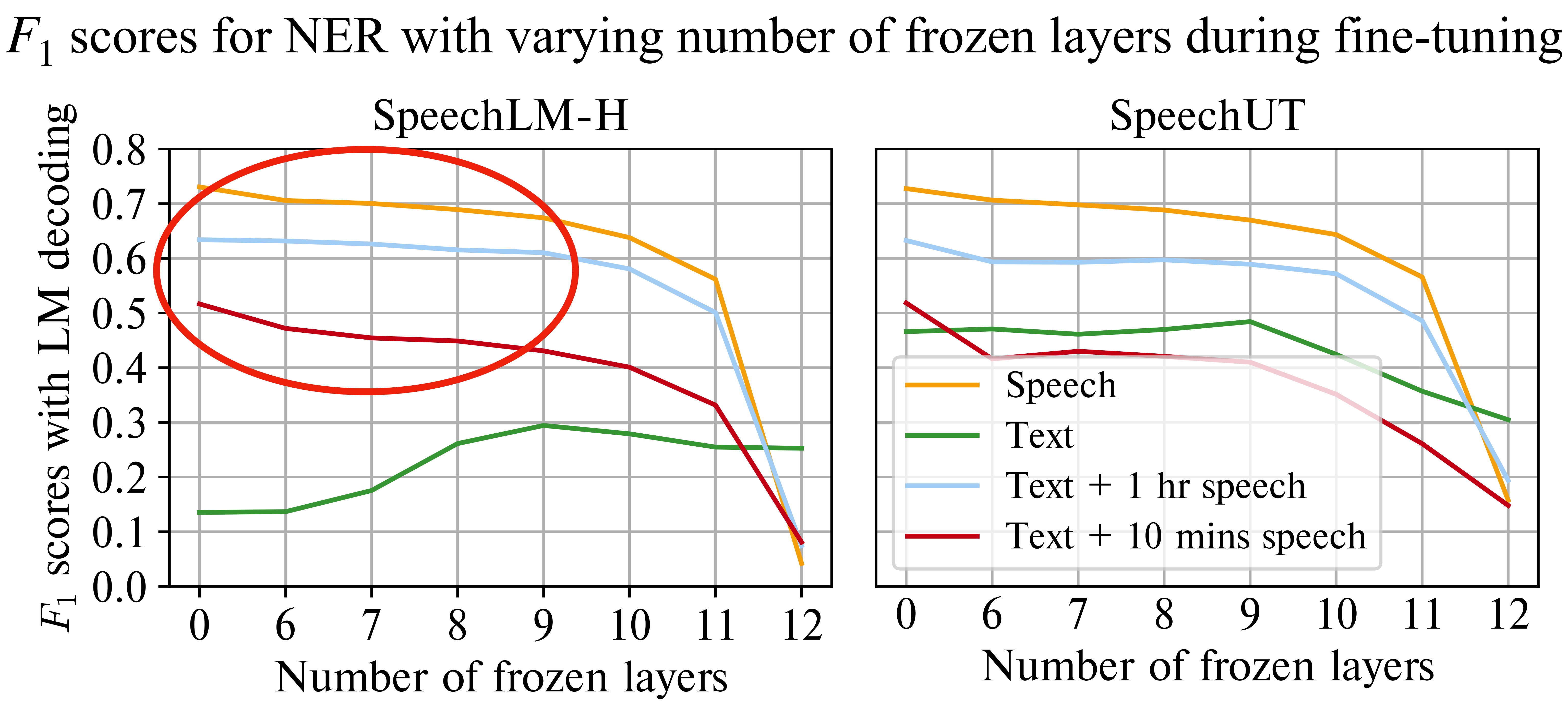
# Fine-Tuning with Bottom Layers Frozen

$F_1$ scores for NER with varying number of frozen layers during fine-tuning



# Fine-Tuning with Bottom Layers Frozen

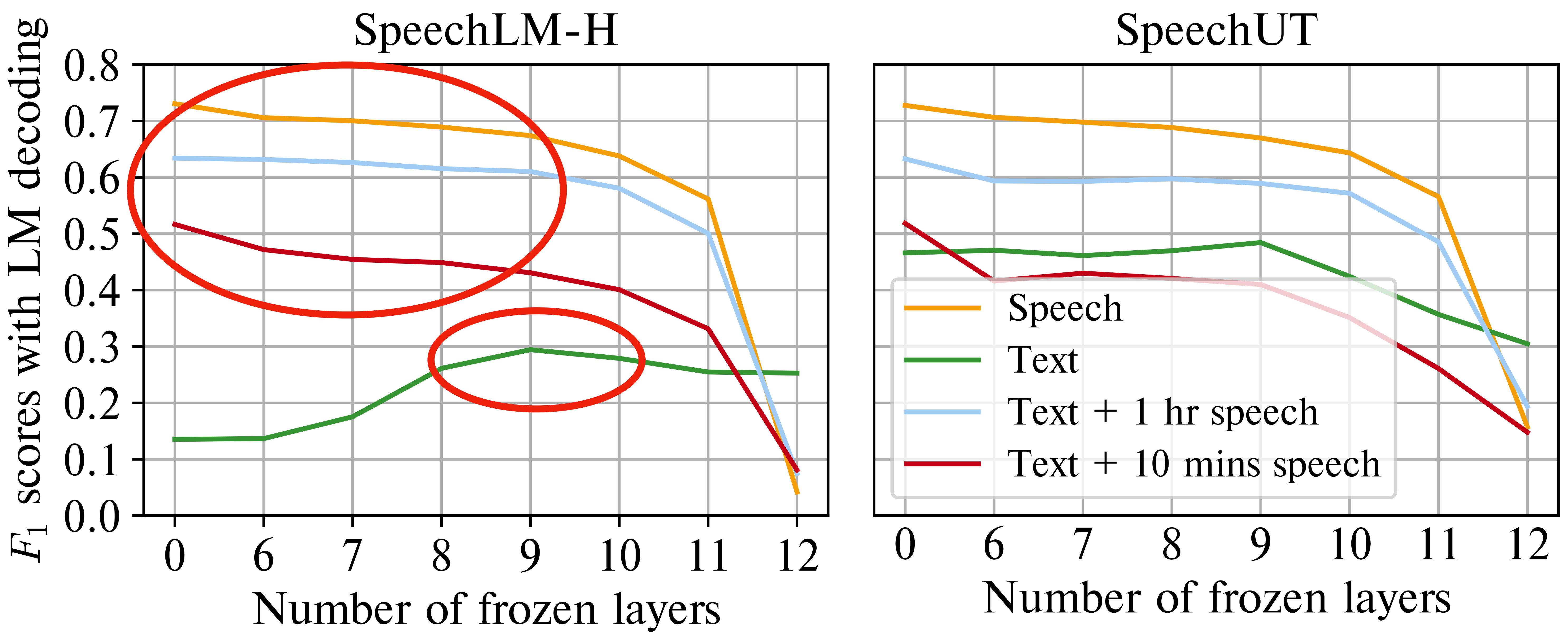
- All-speech & few-shot: slight performance reduction.



# Fine-Tuning with Bottom Layers Frozen

- All-speech & few-shot: slight performance reduction.
- Zero-shot: significant improvements in text-to-speech transferability.

$F_1$ scores for NER with varying number of frozen layers during fine-tuning



# Conclusion

- Speech-text models for few-shot SLU.
  - Speech-text models exhibit zero-shot transferability from text to speech.
  - Few-shot performance matches previous work trained with only 20% of speech data.
- Analysis of speech-text models.
  - Bottom layers are task-agnostic and top layers are task-specific.
  - Freezing bottom layers enhances zero-shot performance.