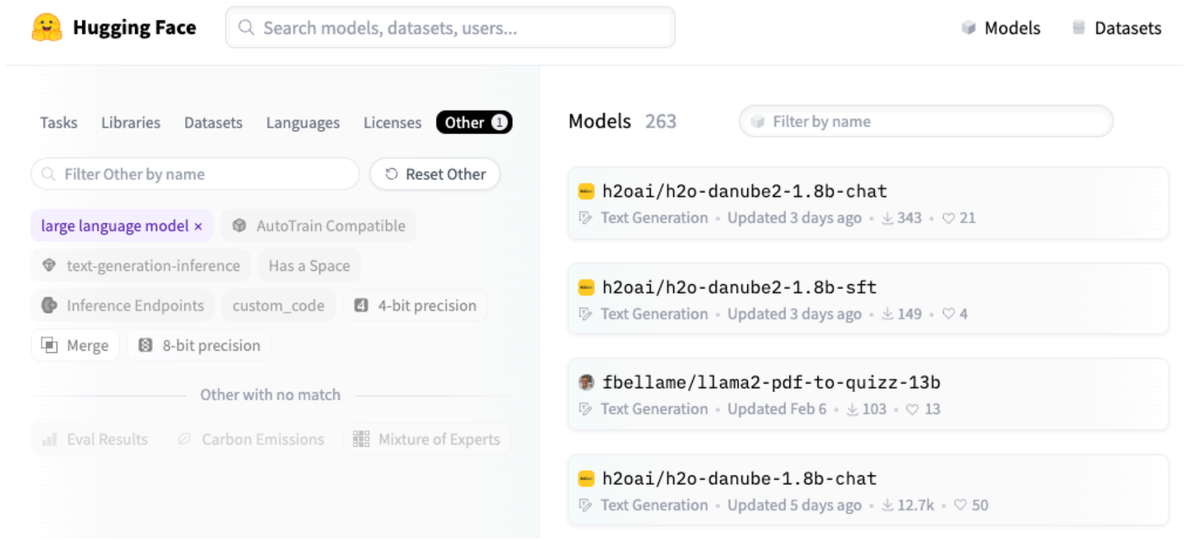
Thesis defense presentation

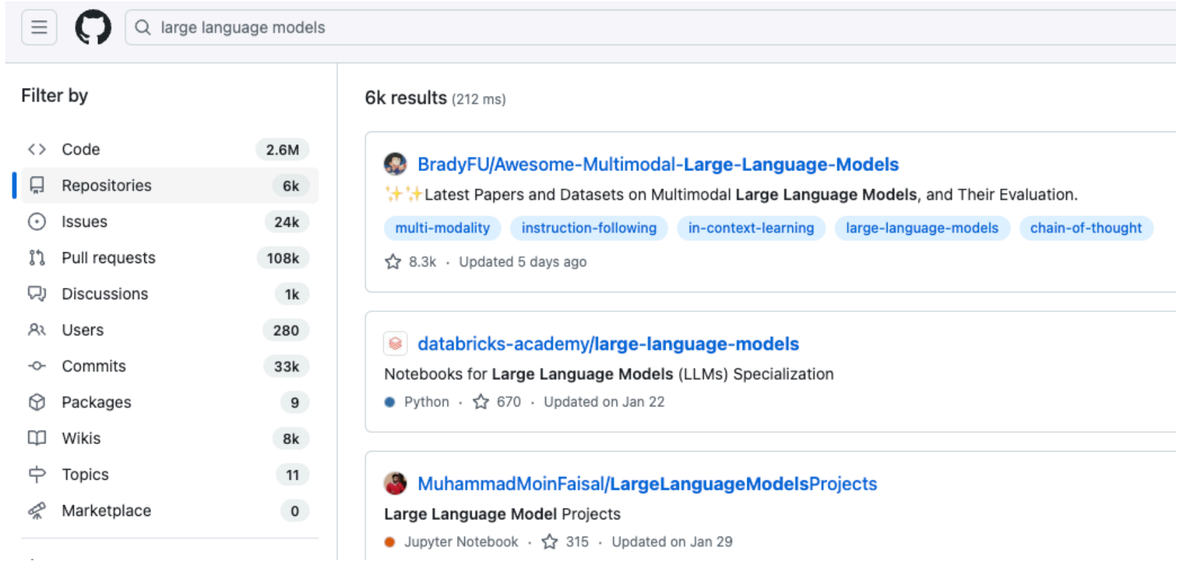
By Amir Djanibekov

MSc Student at NLP department

INTRODUCTION



https://huggingface.co/models?other=large+language+model&sort=trending

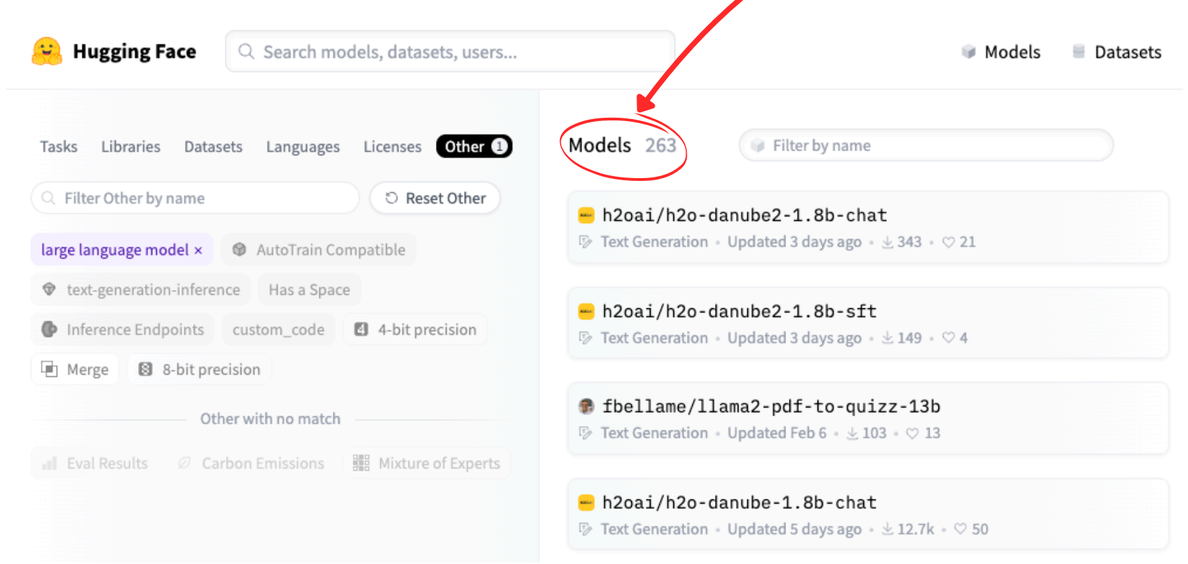


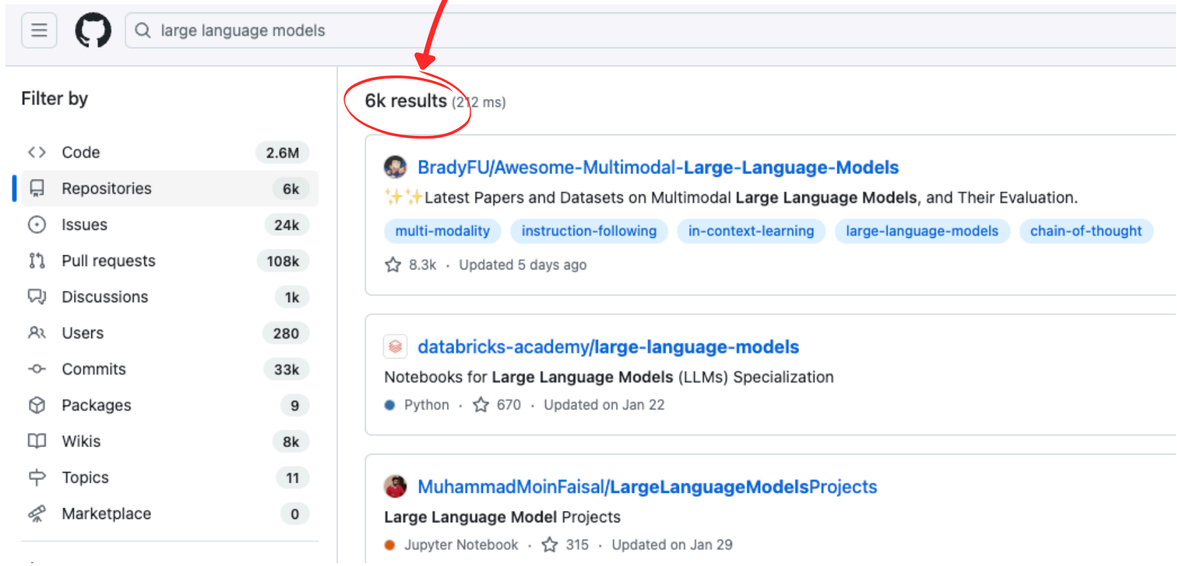
https://github.com/search?q=large+language+models&type=repositories

INTRODUCTION

....public models that you can experiment with

.... ongoing/completed works with large language models





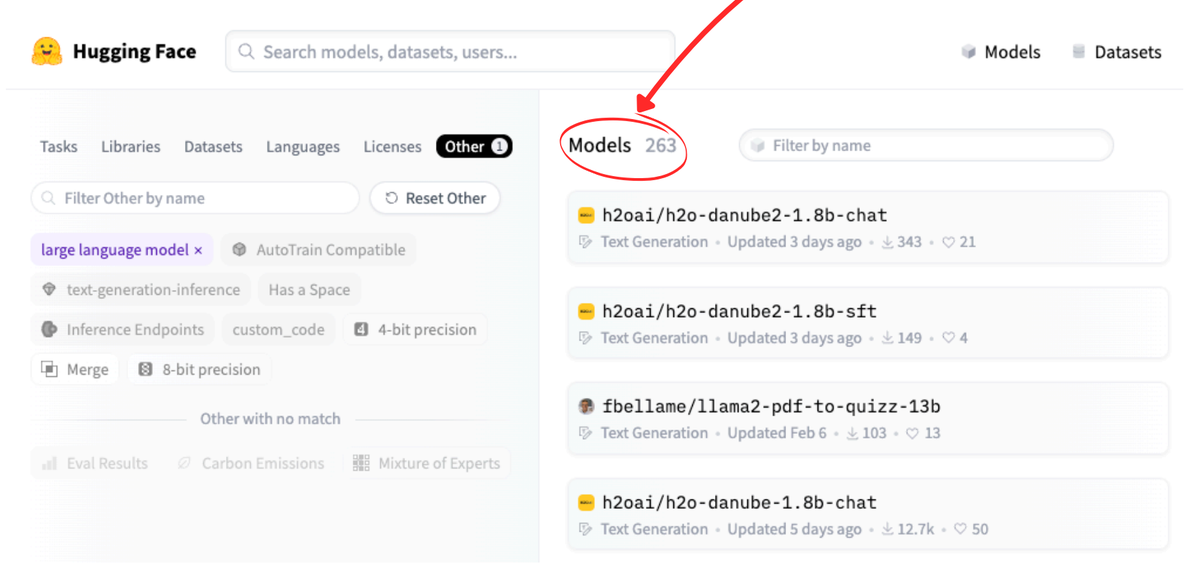
https://huggingface.co/models?other=large+language+model&sort=trending

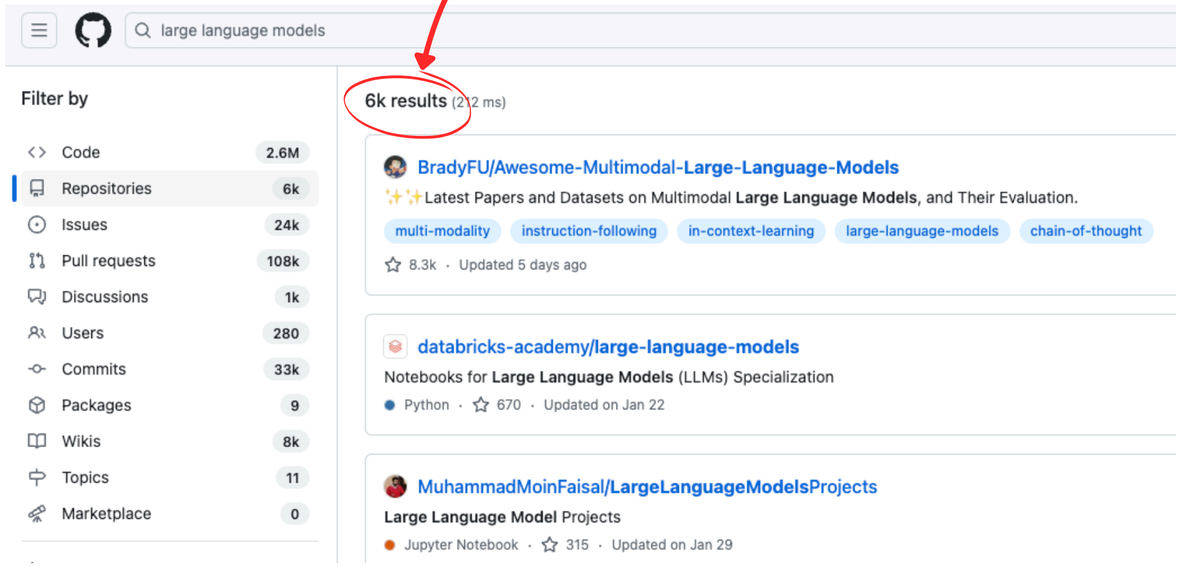
https://github.com/search?q=large+language+models&type=repositories

INTRODUCTION

....public models that you can experiment with

.... ongoing/completed works with large language models



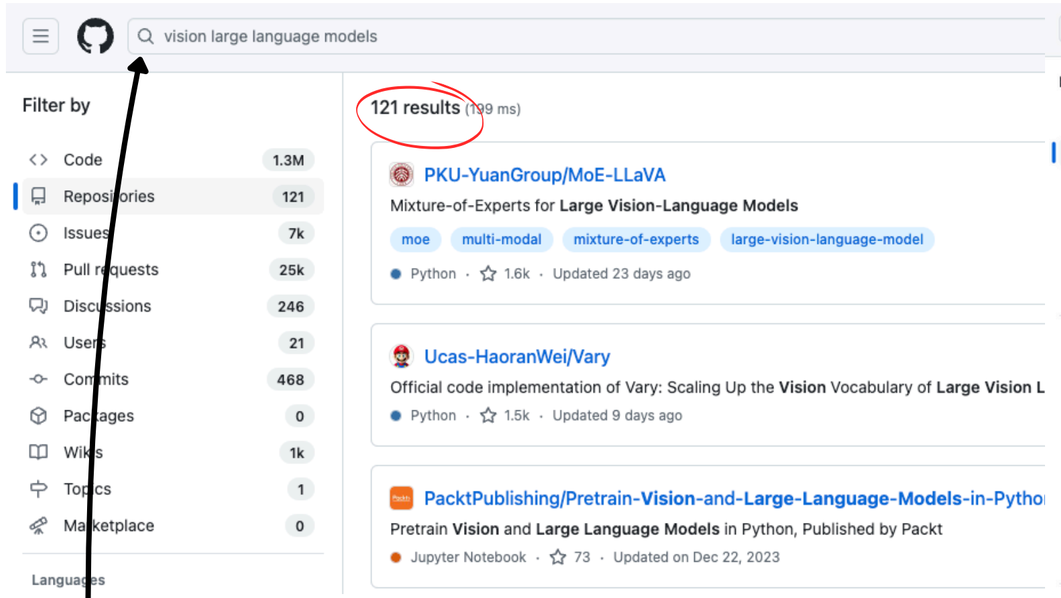


https://huggingface.co/models?other=large+language+model&sort=trending

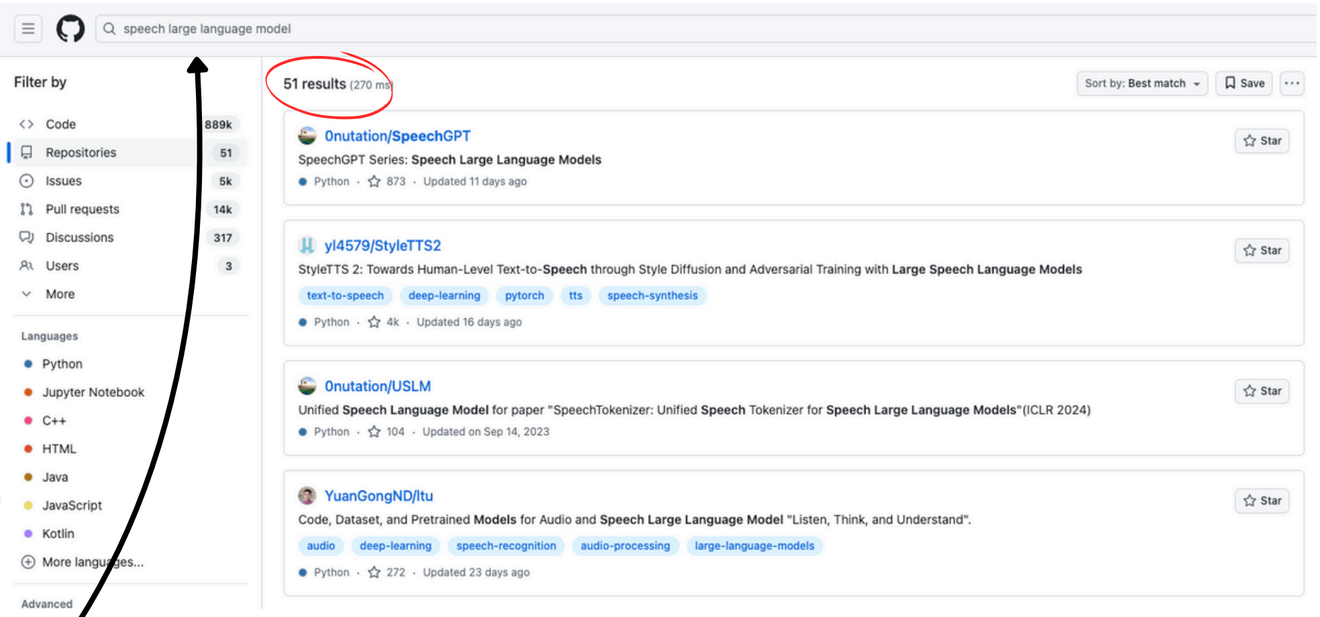
https://github.com/search?q=large+language+models&type=repositories

pre-train.... fine-tune.... instruction-tune....

MOTIVATION



https://github.com/search?q=vision+large+language+model&type=repositories



https://github.com/search?q=speech+large+language+model&type=repositories

vision-language applications ≫ speech-language applications

.....

MOTIVATION

$$\underbrace{\texttt{speech-to-text}}_{ASR} \rightarrow \underbrace{\texttt{text-to-text}}_{LLM}$$

Speech recognition is a sequence classification problem

.... usually done by CTC training objective

$$\underbrace{B}_{\text{Batch size}} \times \underbrace{T}_{\text{Temporal dimension}} \times \underbrace{D}_{\text{Hidden dimension}} \rightarrow \underbrace{B}_{\text{Batch size}} \times \underbrace{T}_{\text{Temporal dimension}}$$

.... index from vocab

OVERALL

- LLMs approximate human-like understanding
  - integrating other modalities enriches LLMs contextual information
- Vision-language modeling already reached outstanding results in image captioning
- Discriminative Speech Models are good but Deep Discriminative Speech Models might better
  - Decoding tokens to text at decoder side
- (Motivation) Text Generative Speech Models another promising direction
  - LLM integration as vision-language

LIMITATIONS

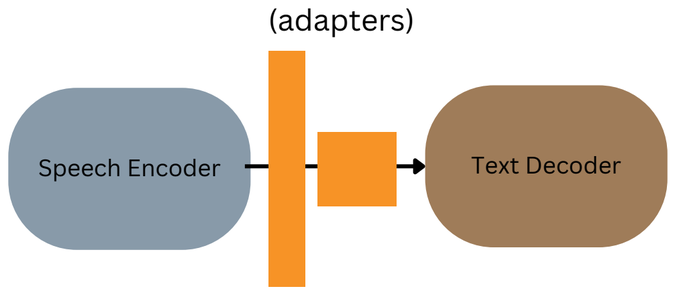
- Training LLM is compute exhaustive and expensive procedure
  - x2 pain collecting data
  - x3 pain verifying collected data
  - x4 pain training
  - x100 pain evaluating
- Speech-language modeling with LLM
  - next ‘x’ pain?

LIMITATIONS

- Training LLM is compute exhaustive and expensive procedure
  - x2 pain collecting data
  - x3 pain verifying collected data
  - x4 pain training
  - x100 pain evaluating
- Speech-language modeling with LLM
  - next ‘x’ pain?
  - we’ll try parameter efficient way by freezing LLM and speech encoder while training only modality connector/adapter

RELATED WORKS

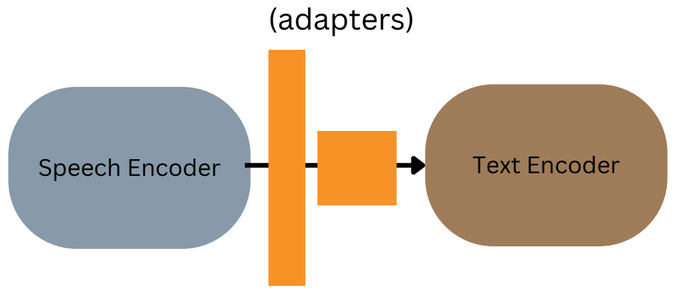
- End-to-End Speech Translation with Pre-trained Models and Adapters: UPC at IWSLT 2021



https://aclanthology.org/2021.iwslt-1.pdf#page=124

RELATED WORKS

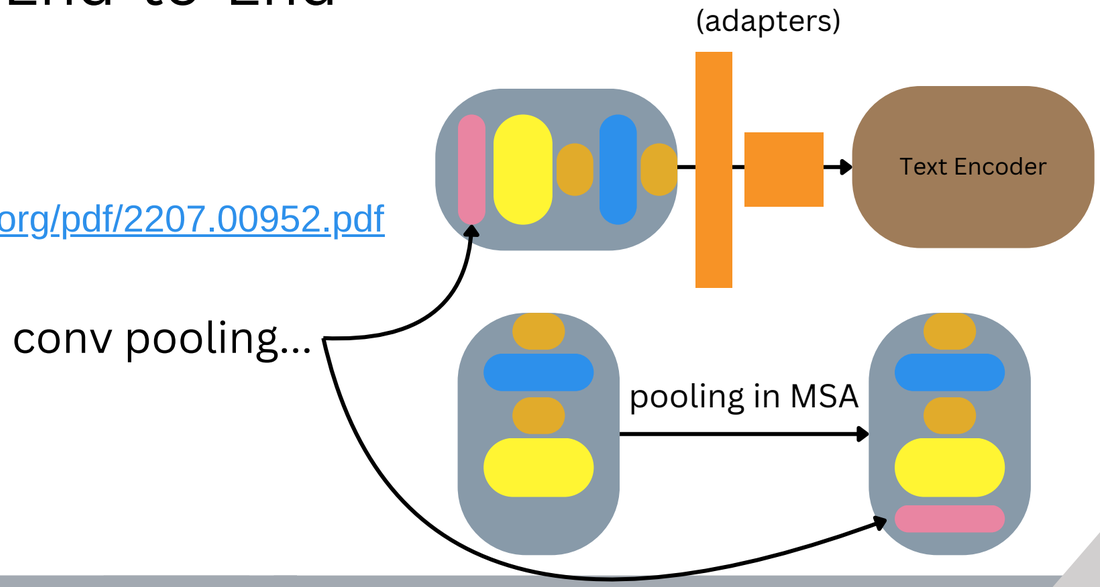
- End-to-End Speech Translation with Pre-trained Models and Adapters: UPC at IWSLT 2021



https://aclanthology.org/2021.iwslt-1.pdf#page=124

- M-Adapter: Modality Adaptation for End-to-End Speech-to-Text Translation

https://arxiv.org/pdf/2207.00952.pdf



RELATED WORKS

- SALM: Speech-augmented Language Model with In-context Learning for Speech Recognition and Translation

https://ieeexplore.ieee.org/abstract/document/10447553/

RELATED WORKS

- SALM: Speech-augmented Language Model with In-context Learning for Speech Recognition and Translation

https://ieeexplore.ieee.org/abstract/document/10447553/

- BLSP: Bootstrapping Language-Speech Pre-training via Behavior Alignment of Continuation Writing

https://arxiv.org/abs/2309.00916

RELATED WORKS

- SALM: Speech-augmented Language Model with In-context Learning for Speech Recognition and Translation

https://ieeexplore.ieee.org/abstract/document/10447553/

- BLSP: Bootstrapping Language-Speech Pre-training via Behavior Alignment of Continuation Writing

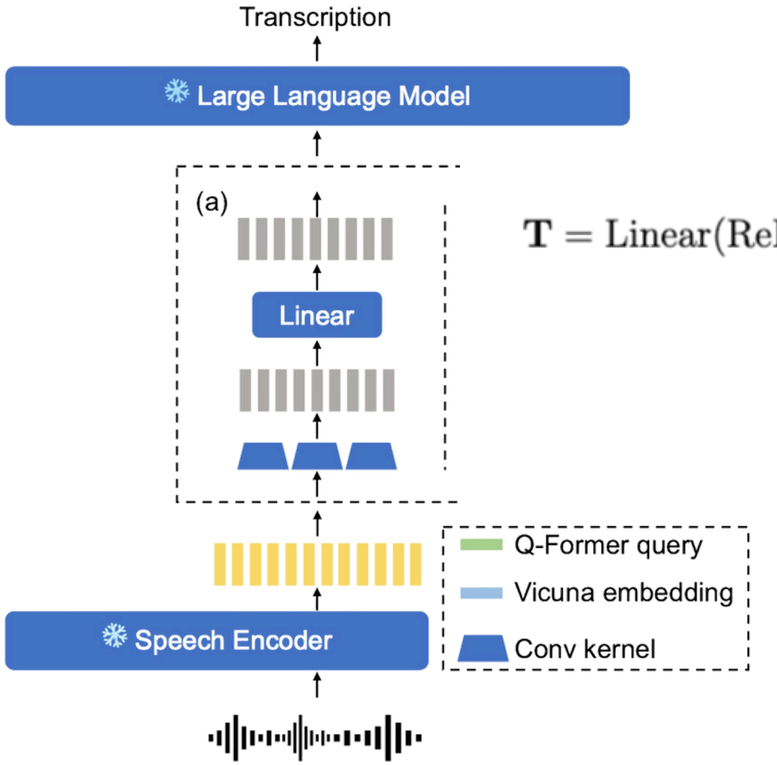
https://arxiv.org/abs/2309.00916

- SLM: Bridge the thin gap between Speech and Text foundation models

https://ieeexplore.ieee.org/abstract/document/10389703/

RELATED WORKS

- Connecting speech encoder and large language model for ASR



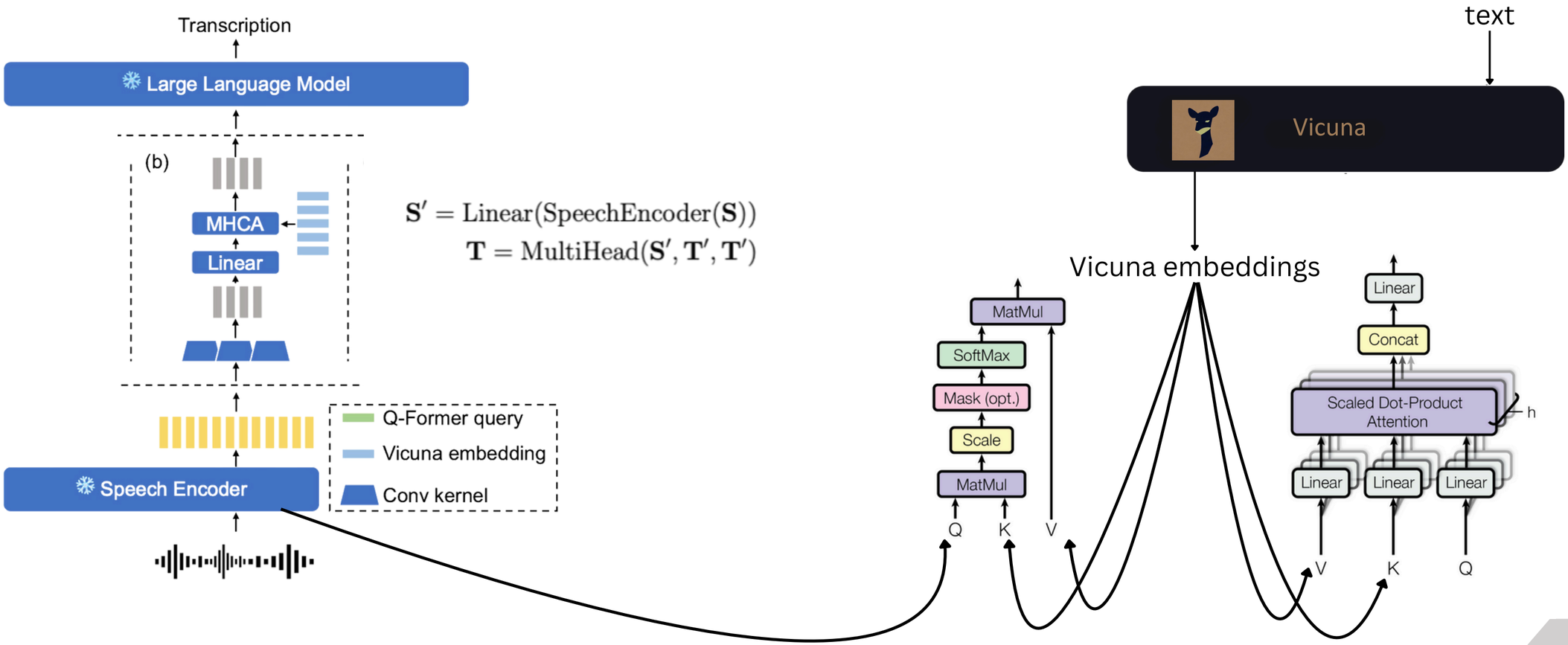
$$\mathbf{T} = \text{Linear}(\text{ReLU}(\text{Linear}(\text{SpeechEncoder}(\mathbf{S}))))$$

@InProceedings{pmlr-v97-houlsby19a,
title = {Parameter-Efficient Transfer Learning for {NLP}},
author = {Houlsby, Neil and Giurgiu, Andrei and Jastrzebski, Stanislaw and Morrone, Bruna and De Laroussilhe, Quentin and Gesmundo, Andrea and Attariyan, Mona and Gelly, Sylvain},
booktitle = {Proceedings of the 36th International Conference on Machine Learning},
pages = {2790--2799},
year = {2019},
editor = {Chaudhuri, Kamalika and Salakhutdinov, Ruslan},
volume = {97},
series = {Proceedings of Machine Learning Research},
month = {09--15 Jun},
publisher = {PMLR},
pdf = {http://proceedings.mlr.press/v97/houlsby19a/houlsby19a.pdf},
url = {https://proceedings.mlr.press/v97/houlsby19a.html},
abstract = {Fine-tuning large pretrained models is an effective transfer mechanism in NLP. However, in the presence of many downstream tasks, fine-tuning is parameter inefficient: an entire new model is required for every task. As an alternative, we propose transfer with adapter modules. Adapter modules yield a compact and extensible model; they add only a few trainable parameters per task, and new tasks can be added without revisiting previous ones. The parameters of the original network remain fixed, yielding a high degree of parameter sharing. To demonstrate adapter's effectiveness, we transfer the recently proposed BERT Transformer model to 26 diverse text classification tasks, including the GLUE benchmark. Adapters attain near state-of-the-art performance, whilst adding only a few parameters per task. On GLUE, we attain within $0.8%$ of the performance of full fine-tuning, adding only $3.6%$ parameters per task. By contrast, fine-tuning trains $100%$ of the parameters per task.}
}

http://proceedings.mlr.press/v97/houlsby19a.html

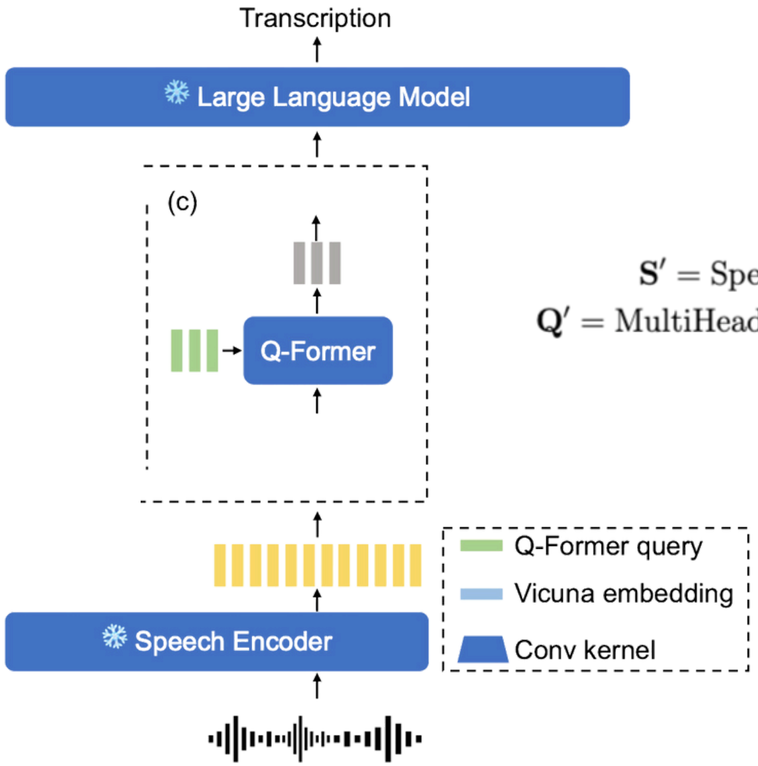
RELATED WORKS

- Connecting speech encoder and large language model for ASR

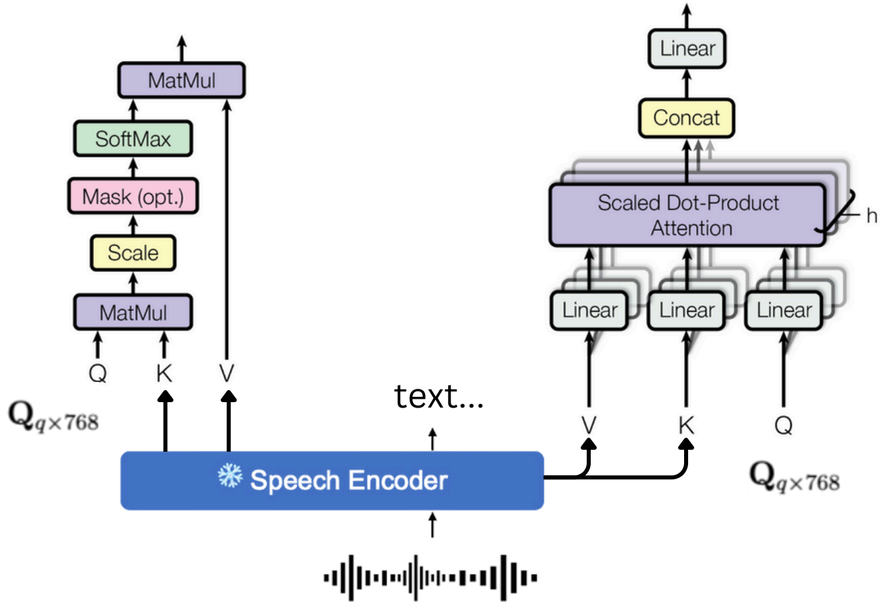


RELATED WORKS

- Connecting speech encoder and large language model for ASR

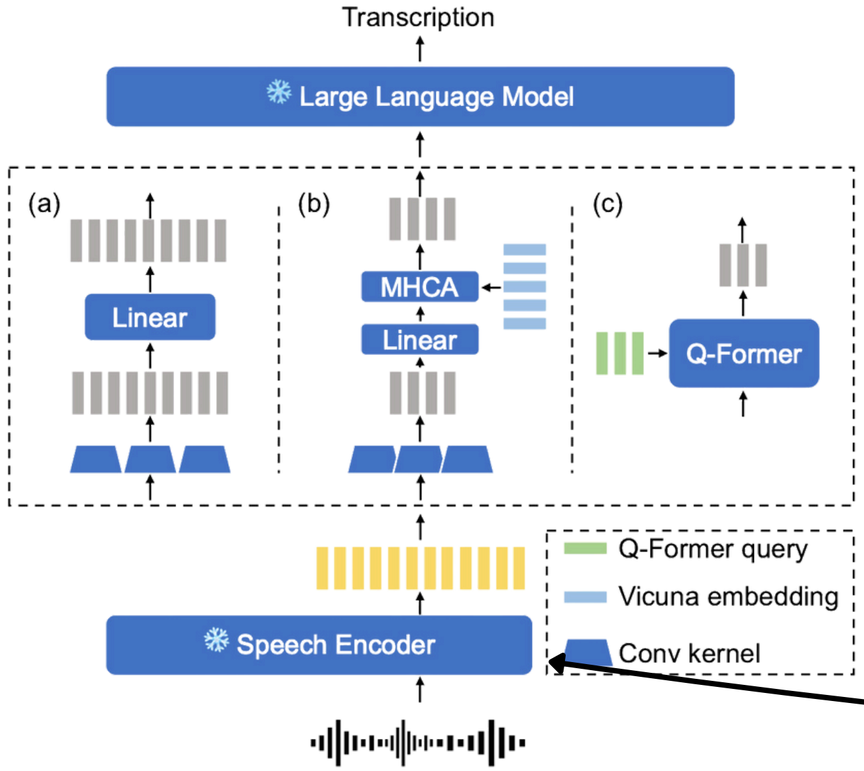


$$\mathbf{S}' = \text{SpeechEncoder}(\mathbf{S})$$
$$\mathbf{Q}' = \text{MultiHead}(\mathbf{Q}_{q\times 768}, \mathbf{S}', \mathbf{S}')$$



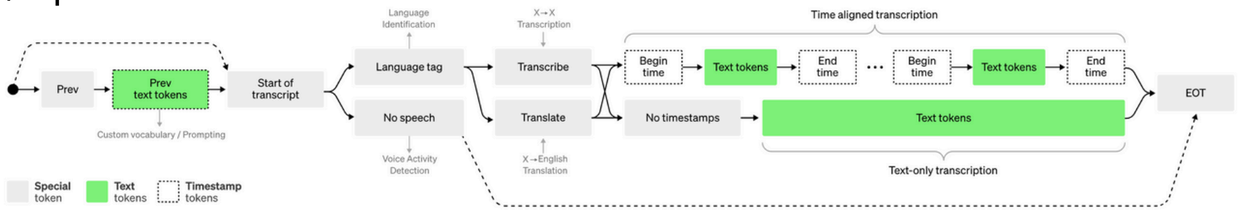
RELATED WORKS

- Connecting speech encoder and large language model for ASR



whisper speech encoder

(i) speech

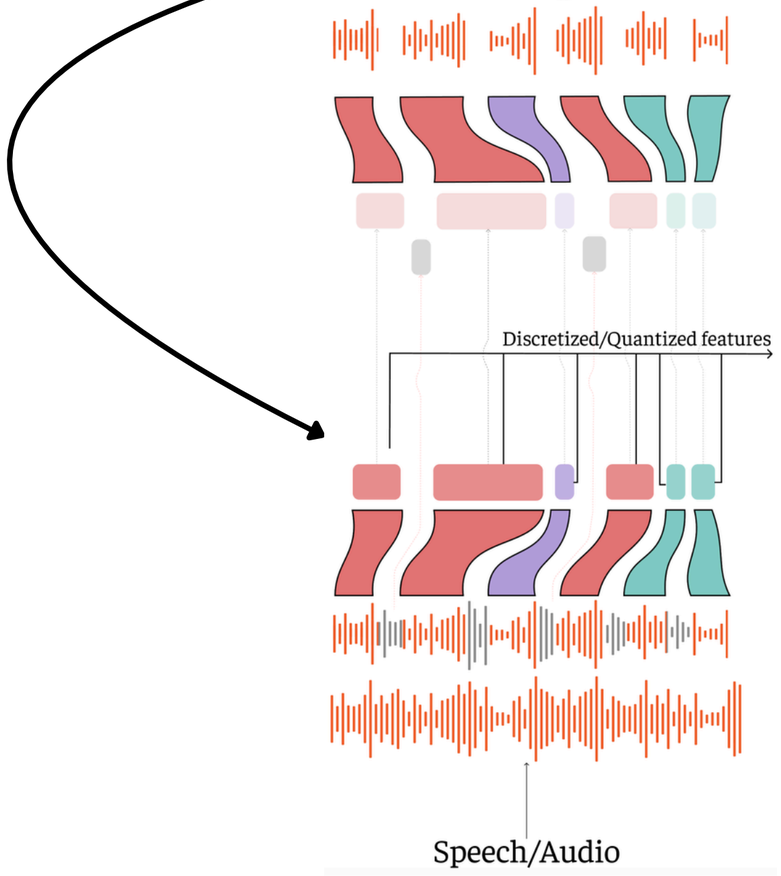


(o) text

BACKGROUND

- we’ll try parameter efficient way by freezing LLM and speech encoder while training only modality connector/adapter

...high level idea



speechcodes
or ssl features

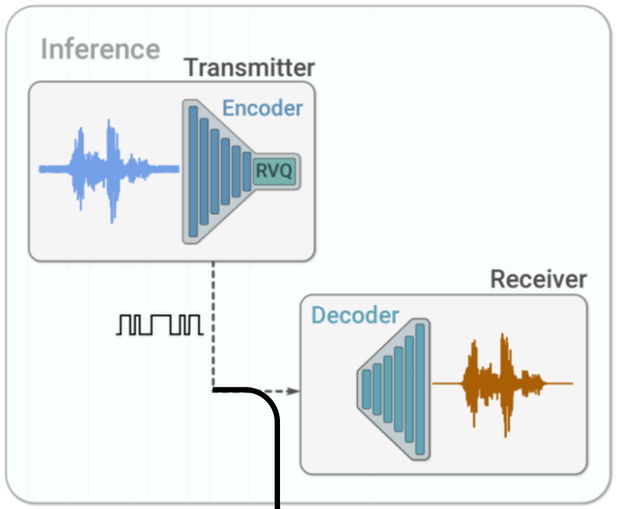
* ssl - self-supervised learning

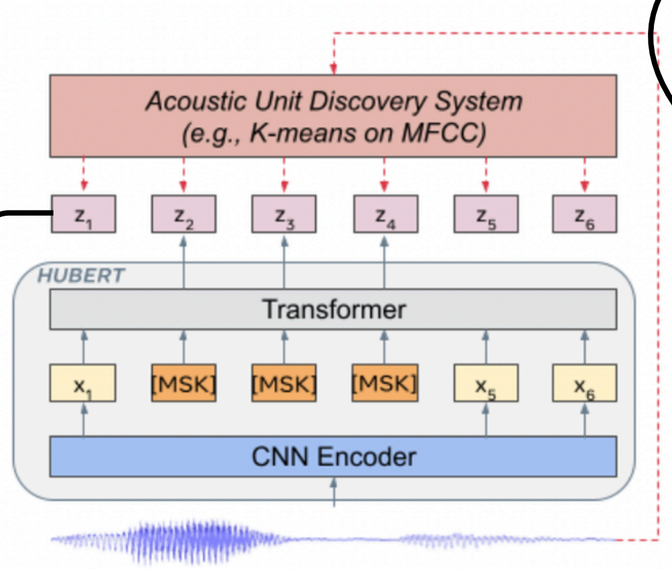
BACKGROUND

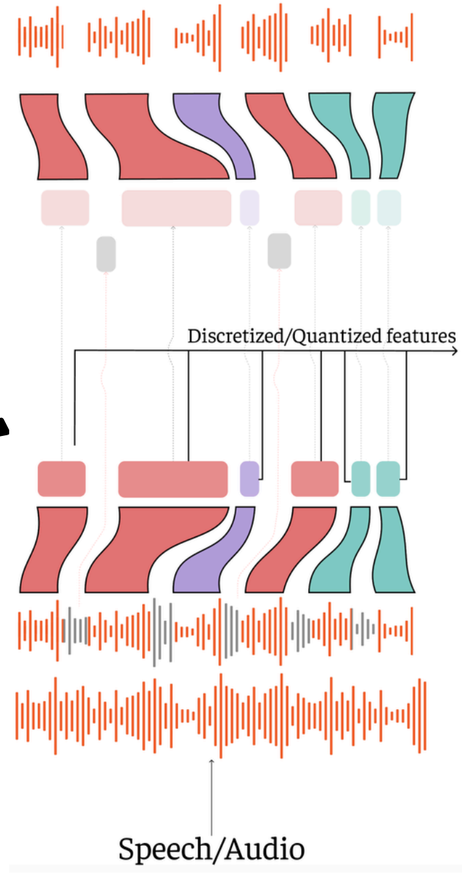
- we'll try parameter efficient way by freezing LLM and speech encoder while training only modality connector/adapter

...high level idea

...detailed







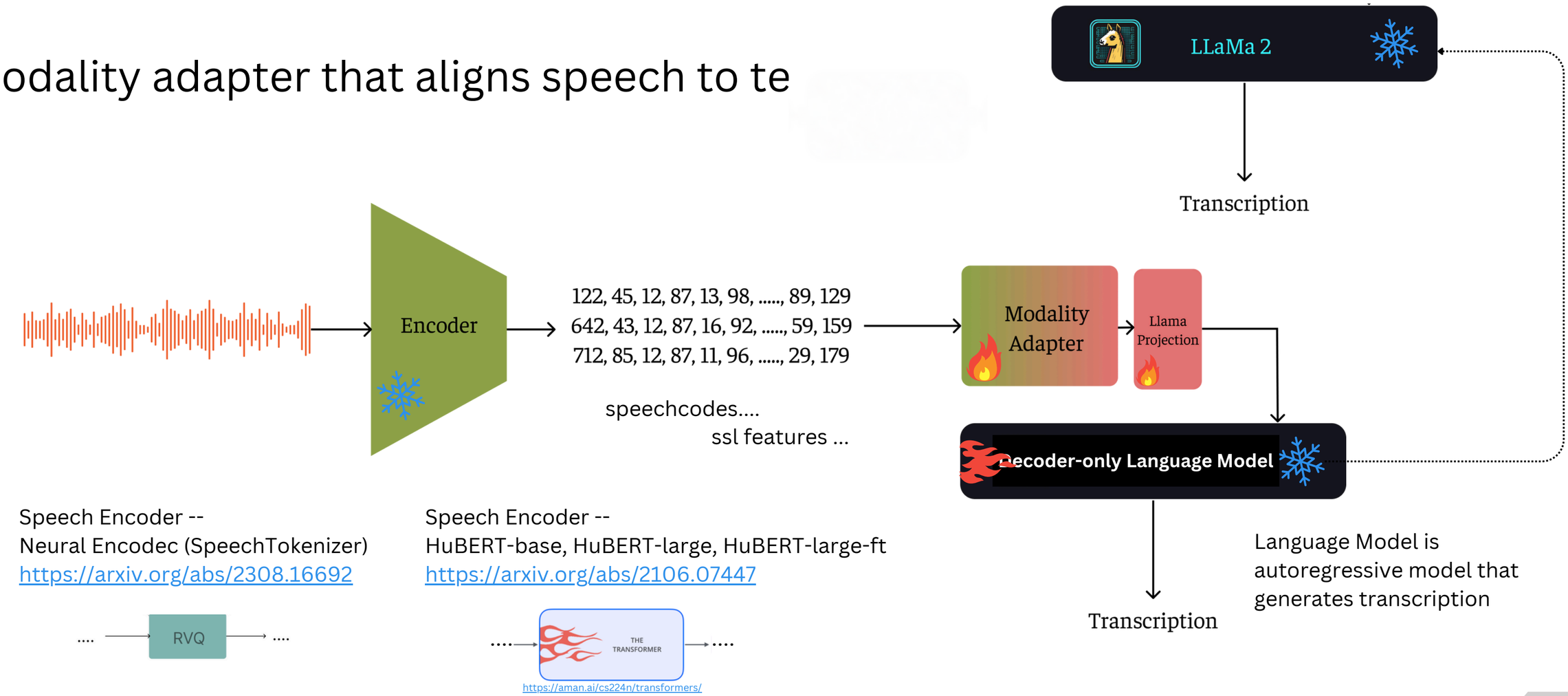
speechcodes
or ssl features

input to proposing connecter....

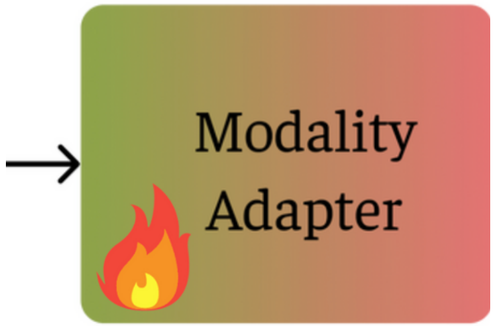
* ssl - self-supervised learning

PROPOSED METHODS

modality adapter that aligns speech to te

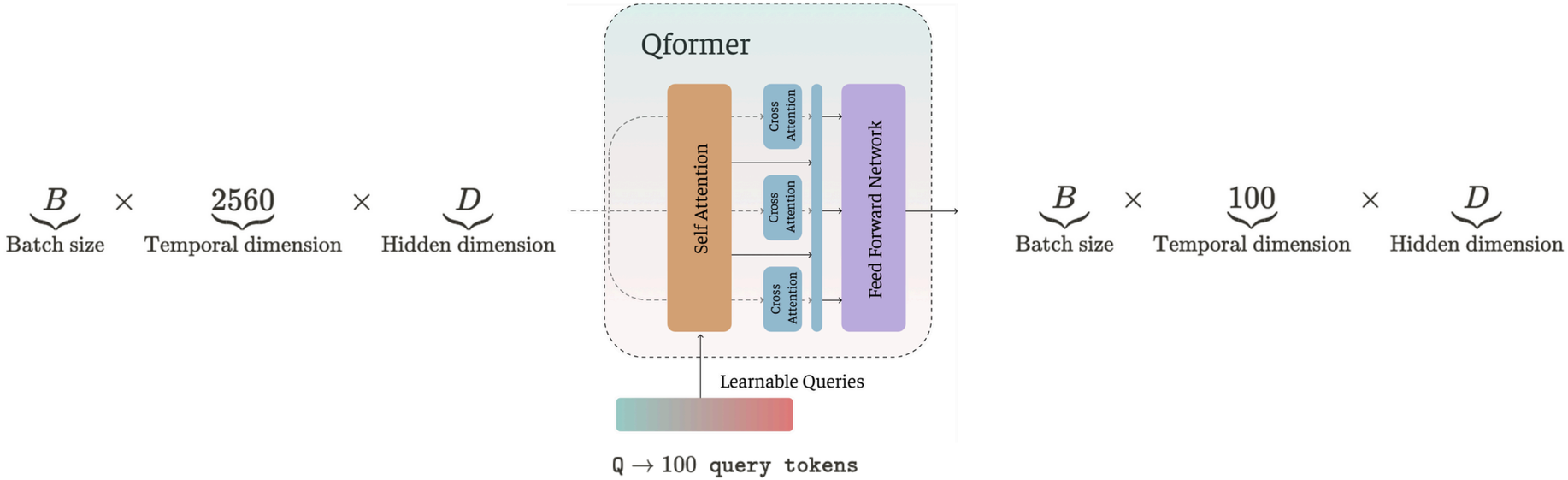


CURRENT WORKS

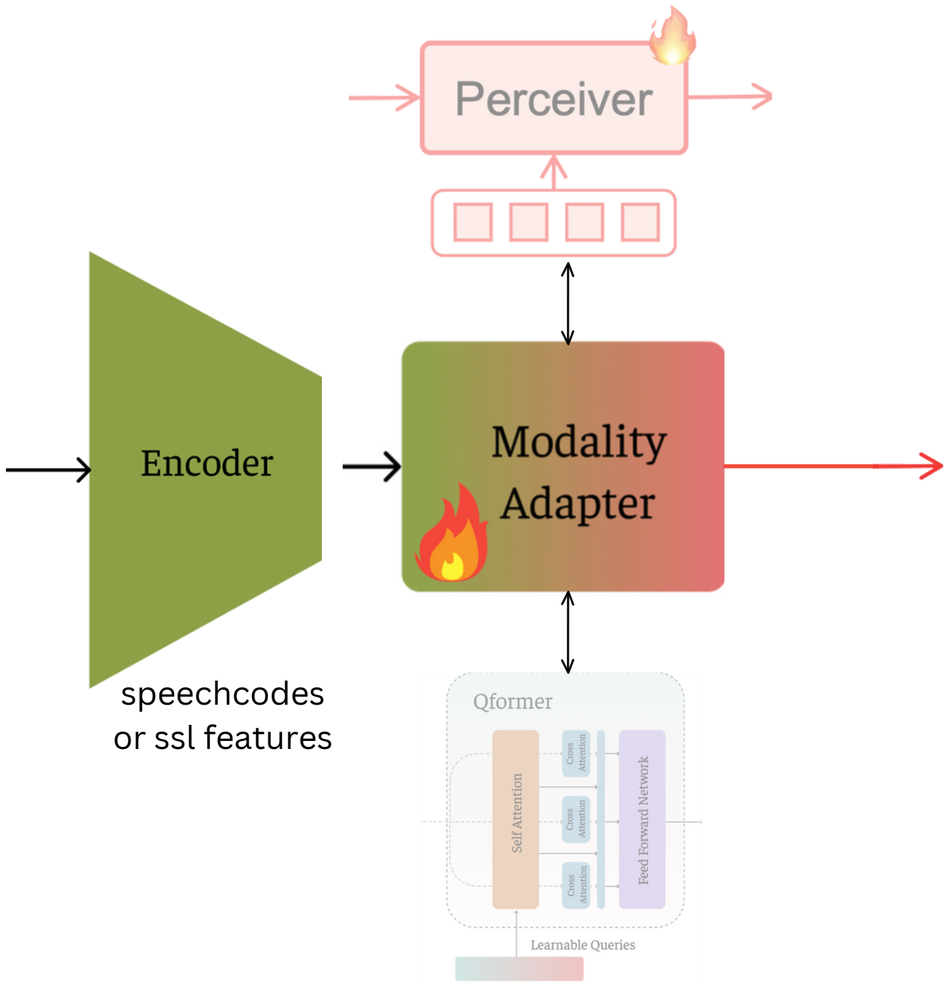


BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models
https://arxiv.org/abs/2301.12597

QFormer - Querying Transformer

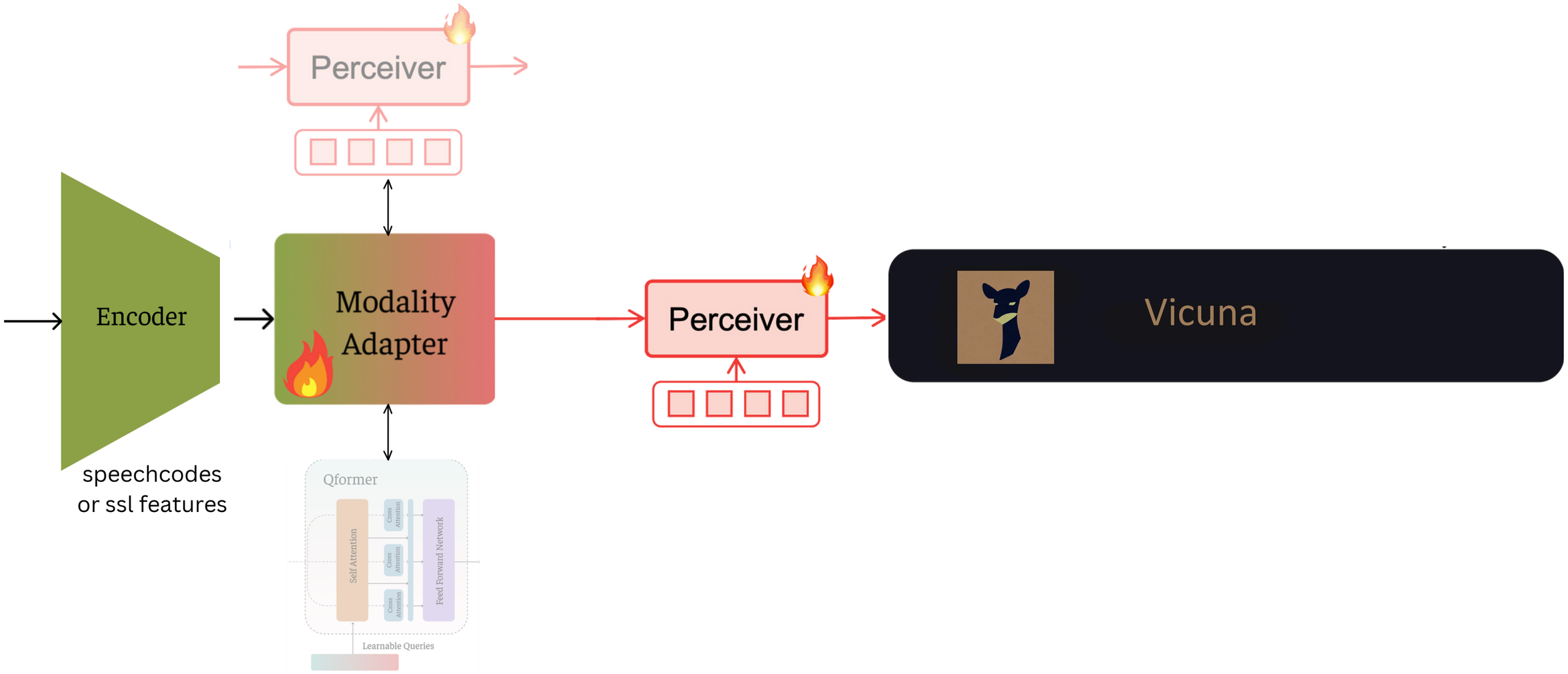


PROPOSED METHOD



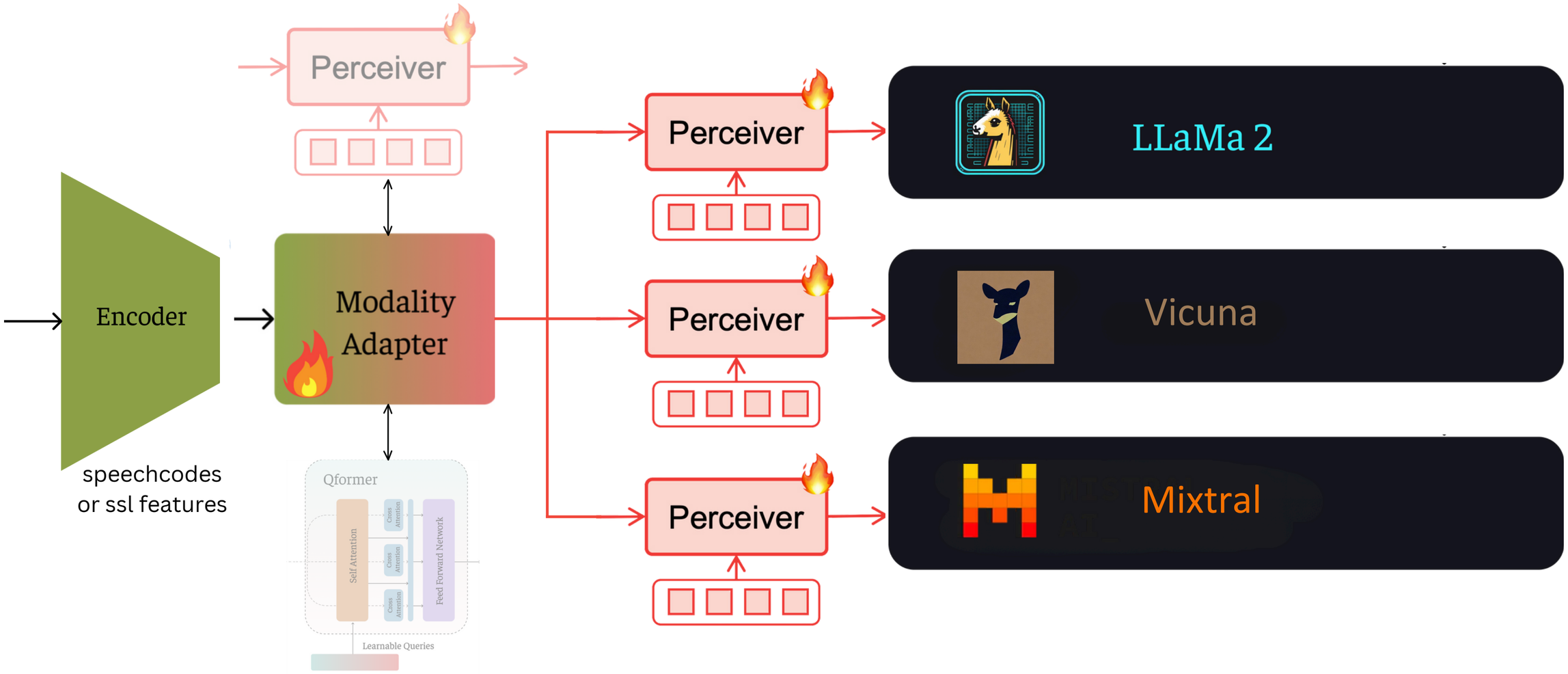
* Perceiver is a Modality adapter, predefined set of query tokens serve as adapter tokens

PROPOSED METHOD



* Perceiver is a Modality adapter, predefined set of query tokens serve as adapter tokens

PROPOSED METHOD



* Perceiver is a Modality adapter, predefined set of query tokens serve as adapter tokens

DATA PREPROCESSING

Table 4.1: Table Example: Comparative analysis of target transcriptions versus predictions after processing through a neural EnCodec.

№	Target Transcription	Predicted Transcription
(1)	and now the squires despatched to make the prize came up bringing with them two gentlemen on horseback two pilgrims on foot and a coach full of women with some six servants on foot and on horseback in attendance on them and a couple of muleteers whom the gentlemen had with them	and now the squires dispatched to make the prize came up, bringing with them two gentlemen on horseback two pilgrims on foot, and a coachful of women with some six servants on foot and on horseback in attendance on them and a coachful of women with some six servants on foot and on horseback in attendance on them and a coachful of women with some six servants on foot and on horseback in attendance on them and a coachful
(2)	otto winked at me	i don't win to me
(3)	sonya smiled no	saw your smiles no
(4)	daphne exclaimed tenderly what more is needed	to have me explain tenderly what morrisley did
(5)	they look too funny for anything the opossums are the only marsupials in this country continued old mother nature	..... (non-English symbols)
(6)	our prostrate brother homo ventrambulans razor	a prostrate brother a moe then trampolines razor
(7)	hog n	pog now
(8)	repugnant alike to reason and conscience resigned to growing infirmities resist a common adversary	they don't end up alike, they're reasoned and conscious. redesigned, they're cruelly made in their medias. they're reasoned, they come in adversity
(9)	for a sound that would mean the near approach of danger hush sh sh came again as a gentle murmur from below and the something that moved and breathed in the darkness seemed to draw nearer to yvonne	for a sound that would maim the nearer put to danger hush shh shh came again as a gentle mammoth for below and the something that moved and breathed in the darkness seemed to draw nearer to yvonne

Assess the reconstruction capabilities of the speech encoder.

Whisper Large, one of the most precise commercial options for speech recognition, resulted in 22.5%

Further, text processing on predicted transcription that involves punctuation removal, normalizing letters reduced WER to 10.5% and CER to 5.6%

DATA PREPROCESSING

Table 4.1: Table Example: Comparative analysis of target transcriptions versus predictions after processing through a neural EnCodec.

№	Target Transcription	Predicted Transcription
(1)	and now the squires despatched to make the prize came up bringing with them two gentlemen on horseback two pilgrims on foot and a coach full of women with some six servants on foot and on horseback in attendance on them and a couple of muleteers whom the gentlemen had with them	and now the squires dispatched to make the prize came up, bringing with them two gentlemen on horseback two pilgrims on foot, and a coachful of women with some six servants on foot and on horseback in attendance on them and a coachful of women with some six servants on foot and on horseback in attendance on them and a coachful of women with some six servants on foot and on horseback in attendance on them and a coachful
(2)	otto winked at me	i don't win to me
(3)	sonya smiled no	saw your smiles no
(4)	daphne exclaimed tenderly what more is needed	to have me explain tenderly what morrisley did
(5)	they look too funny for anything the opossums are the only marsupials in this country continued old mother nature	..... (non-English symbols)
(6)	our prostrate brother homo ventrambulans razor	a prostrate brother a moe then trampolines razor
(7)	hog n	pog now
(8)	repugnant alike to reason and conscience resigned to growing infirmities resist a common adversary	they don't end up alike, they're reasoned and conscious. redesigned, they're cruelly made in their medias. they're reasoned, they come in adversity
(9)	for a sound that would mean the near approach of danger hush sh sh came again as a gentle murmur from below and the something that moved and breathed in the darkness seemed to draw nearer to yvonne	for a sound that would maim the nearer put to danger hush shh shh came again as a gentle mammoth for below and the something that moved and breathed in the darkness seemed to draw nearer to yvonne

Some predicted transcription examples may not closely resemble the actual transcription:

- (2), (3), and (8) examples

There was cases where ASR model Failed to recognize

- (5) example

.....why?!

no paralinguistic features

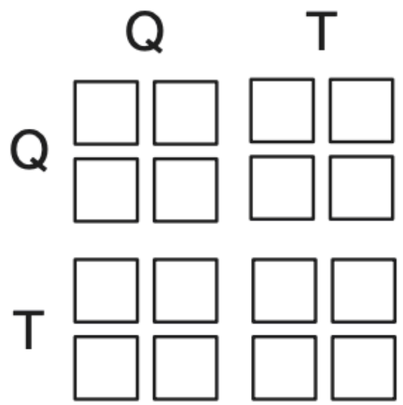
- (7) example. However,
  - this was not always the case,
- (9) example, which retained its paralinguistic elements.

TRAINING OBJECTIVE

Pre-Train Speech-Text Qformer ....

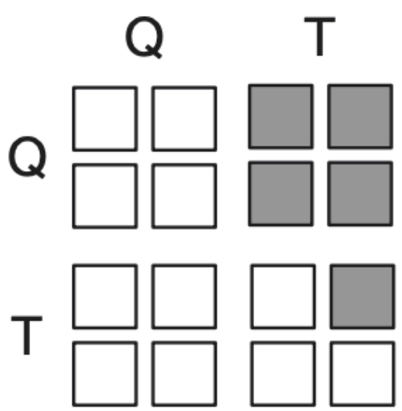
Q: query token positions; **T**: text token positions.

masked unmasked



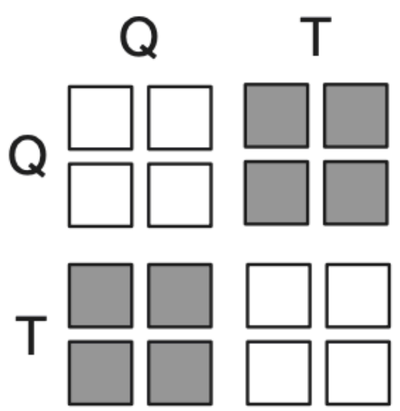
Bi-directional Self-Attention Mask

Speech-Text Matching



Multi-modal Causal Self-Attention Mask

Speech-Text Generation



Uni-modal Self-Attention Mask

Speech-Text Contrastive Learning

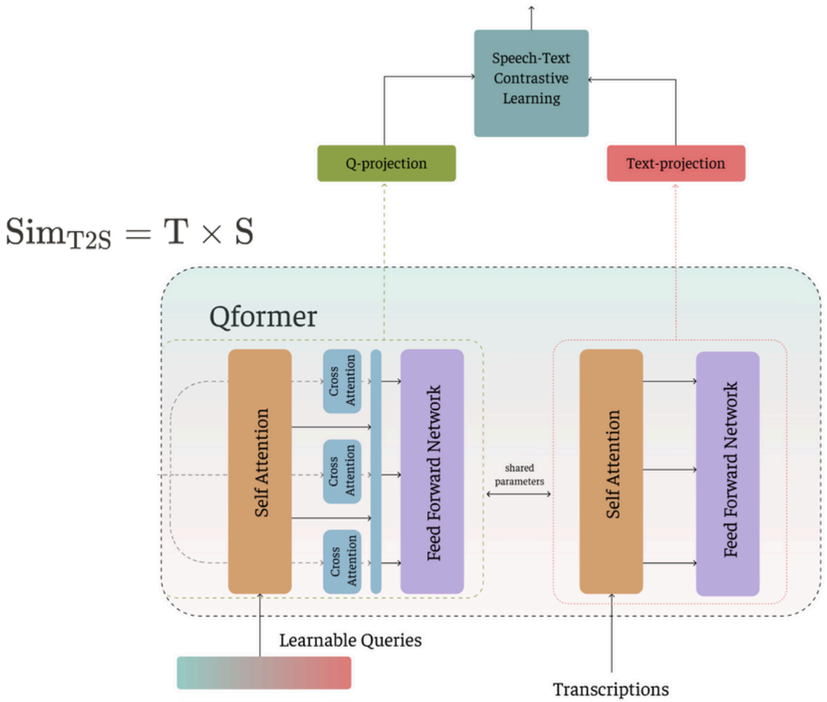
3 traning losses: **STM**, **SGT**, **STCL**

$$\sum$$

STM - Speech Text Matching
SGT - Speech-Text Generation
STCL - Speech-Text Contrastive Learning

TRAINING OBJECTIVE

“Speech-Text Contrastive Training”

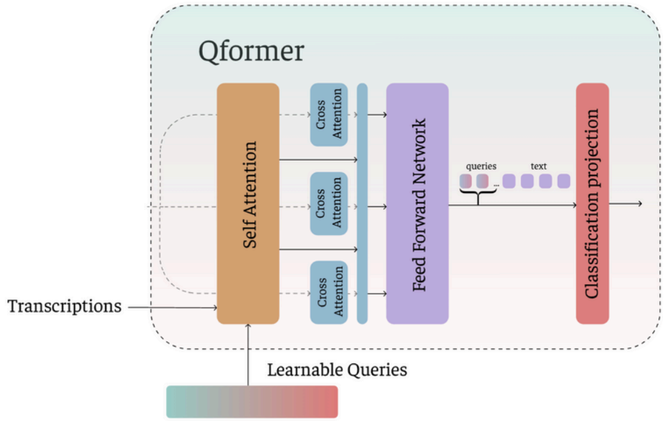


$$\mathrm{Sim}_{\mathrm{T2S}} = \mathrm{T} \times \mathrm{S}$$

$$\sum_{i=0}^{N} \underbrace{\frac{H(\mathrm{Sim}_{\mathrm{t2s}}, T_i) + H(\mathrm{Sim}_{\mathrm{s2t}}, T_i)}{2}}_{\text{Speech-Text contrastive loss}}$$

TRAINING OBJECTIVE

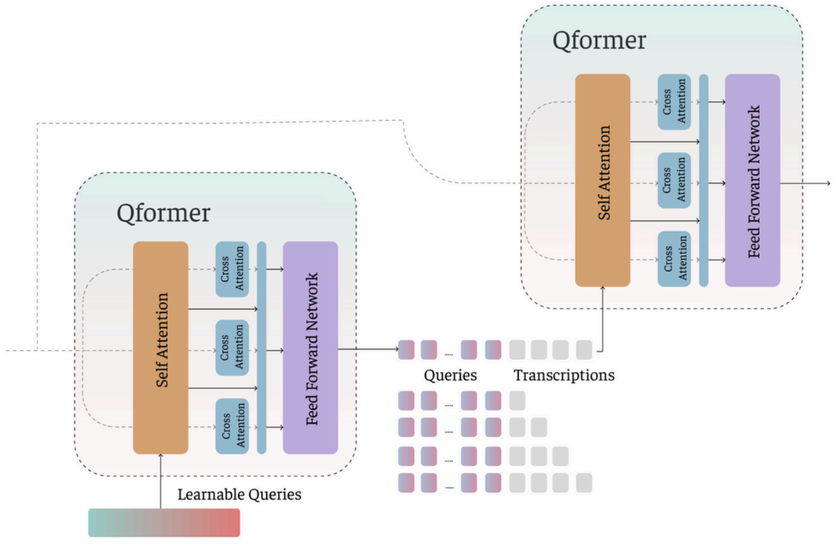
“Speech-Text Matching Training”



$$\sum_{i=0}^{N} \underbrace{(-(y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p)))}_{\text{Speech-Text Matching loss}}$$

TRAINING OBJECTIVE

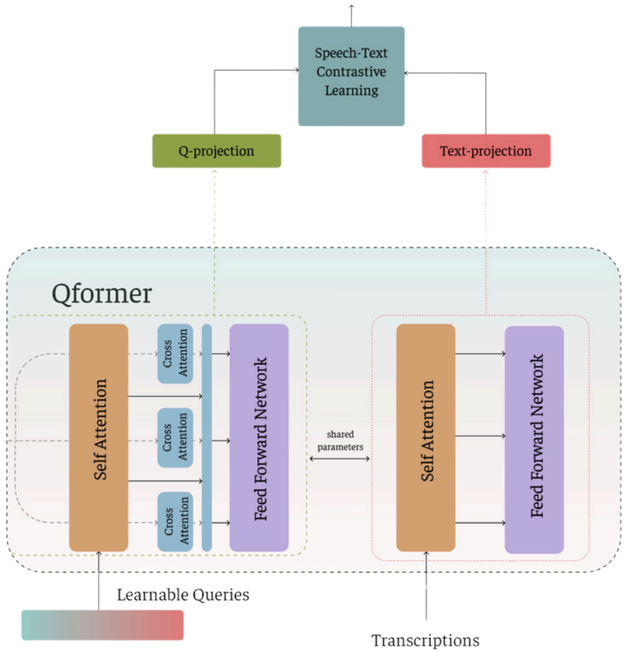
“Speech-Text Generation Training”

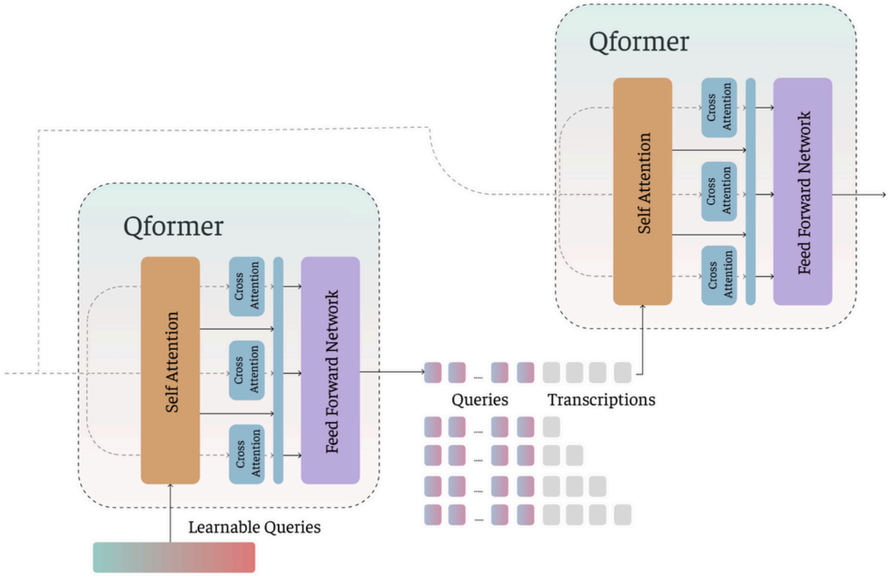


$$\sum_{i=1}^{N} \log\left(\mathbb{P}(u_i|u_{i-k},\ldots,u_{i-1},\Theta)\right)$$

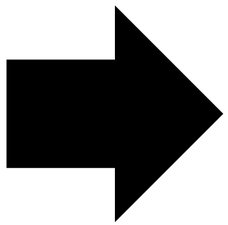
TRAINING OBJECTIVE

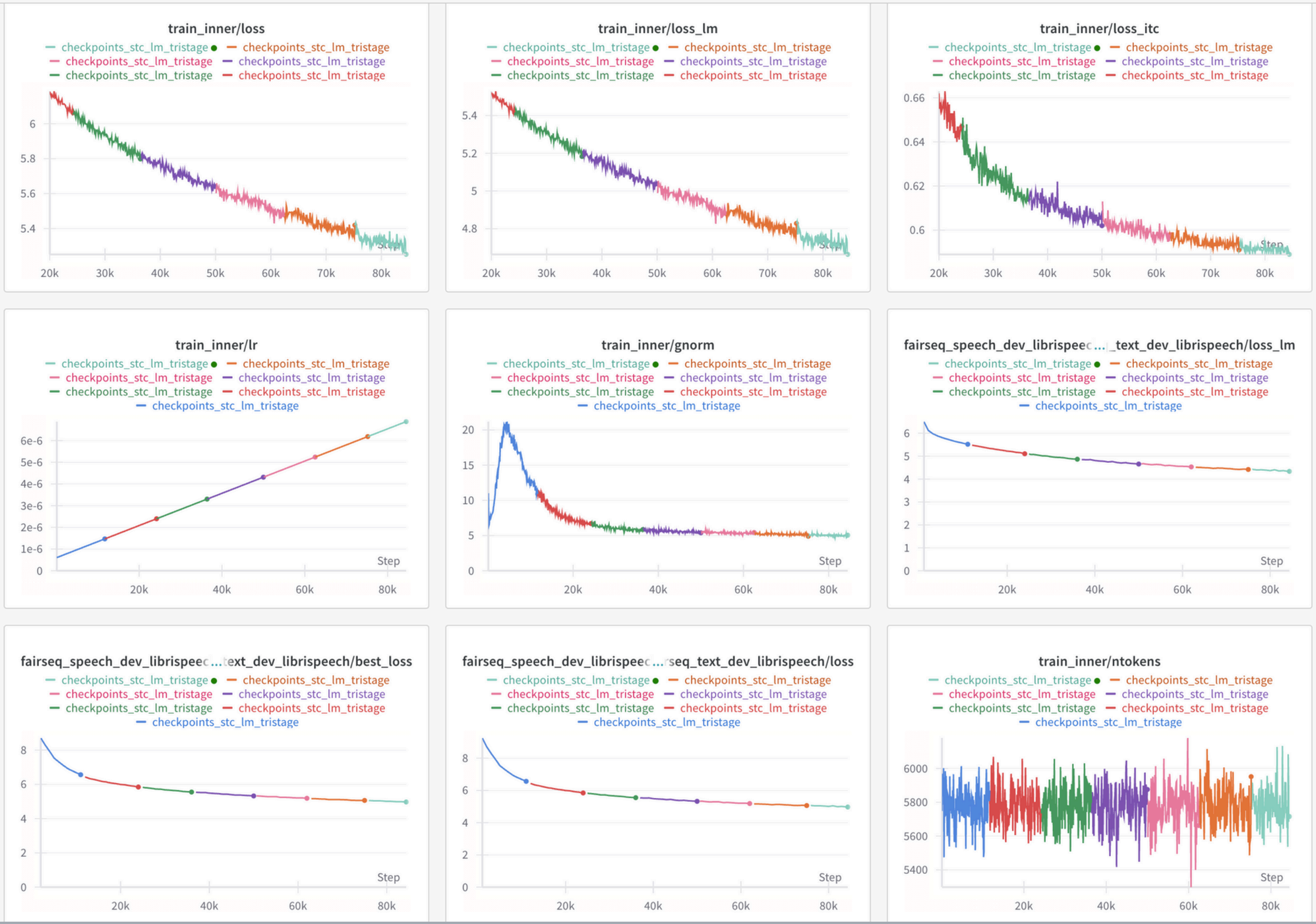
“Joint Speech-Text Contrasting & Speech-Text Generation Training”





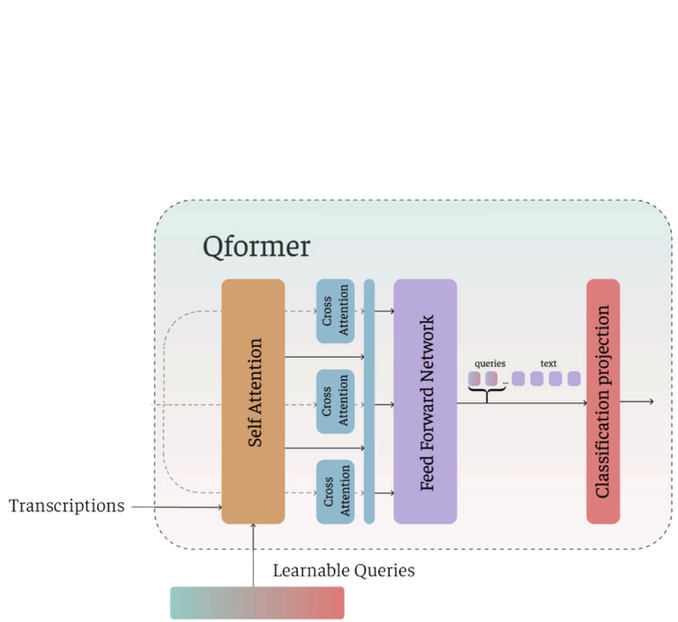
$$\underbrace{\sum_{i=0}^{N} \frac{H(\text{Sim}_{\text{t2s}}, T_i) + H(\text{Sim}_{\text{s2t}}, T_i)}{2}}_{\text{Speech-Text contrastive loss}} + \sum_{i=1}^{N} \log\left(\mathbb{P}(u_i | u_{i-k}, \ldots, u_{i-1}, \Theta)\right)$$

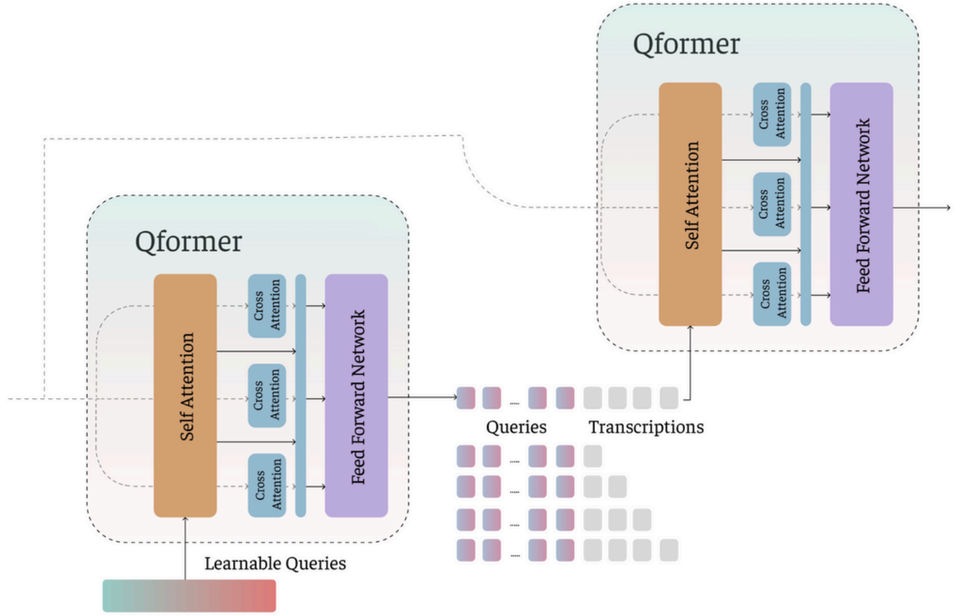



train_inner/loss
checkpoints_stc_lm_tristage
6
5.8
5.6
5.4
20k 30k 40k 50k 60k 70k 80k
train_inner/loss_lm
checkpoints_stc_lm_tristage
5.4
5.2
5
4.8
20k 30k 40k 50k 60k 70k 80k
train_inner/loss_itc
checkpoints_stc_lm_tristage
0.66
0.64
0.62
0.6
Step
20k 30k 40k 50k 60k 70k 80k
train_inner/lr
checkpoints_stc_lm_tristage
6e-6
5e-6
4e-6
3e-6
2e-6
1e-6
0
Step
20k 40k 60k 80k
train_inner/gnorm
checkpoints_stc_lm_tristage
20
15
10
5
0
Step
20k 40k 60k 80k
fairseq_speech_dev_librispeec ... _text_dev_librispeech/loss_lm
checkpoints_stc_lm_tristage
6
5
4
3
2
1
0
Step
20k 40k 60k 80k
fairseq_speech_dev_librispeec ...ext_dev_librispeech/best_loss
checkpoints_stc_lm_tristage
8
6
4
2
0
Step
20k 40k 60k 80k
fairseq_speech_dev_librispeec ...seq_text_dev_librispeech/loss
checkpoints_stc_lm_tristage
8
6
4
2
0
Step
20k 40k 60k 80k
train_inner/ntokens
checkpoints_stc_lm_tristage
6000
5800
5600
5400
Step
20k 40k 60k 80k

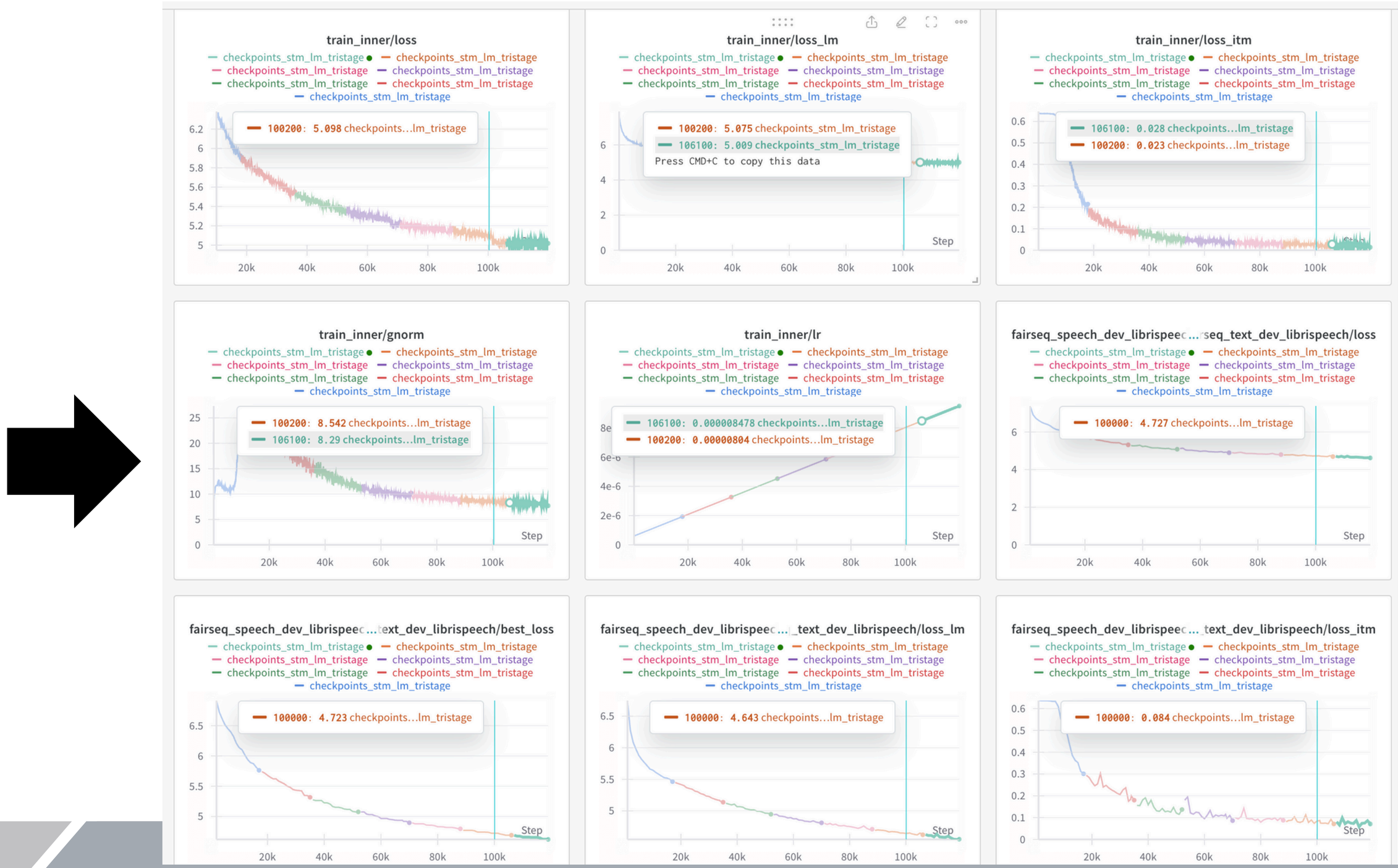
TRAINING OBJECTIVE

“Joint Speech-Text Matching & Speech-Text Generation Training”



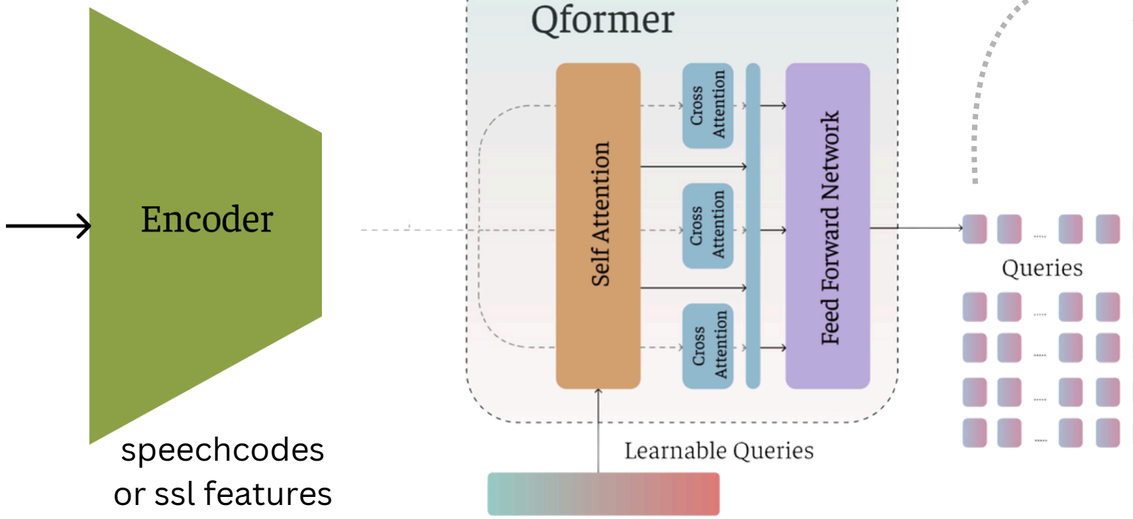


$$\sum_{i=0}^{N} \underbrace{(-(y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p)))}_{\text{Speech-Text Matching loss}} + \sum_{i=1}^{N} \log\left(\mathbb{P}(u_i | u_{i-k}, \ldots, u_{i-1}, \Theta)\right)$$

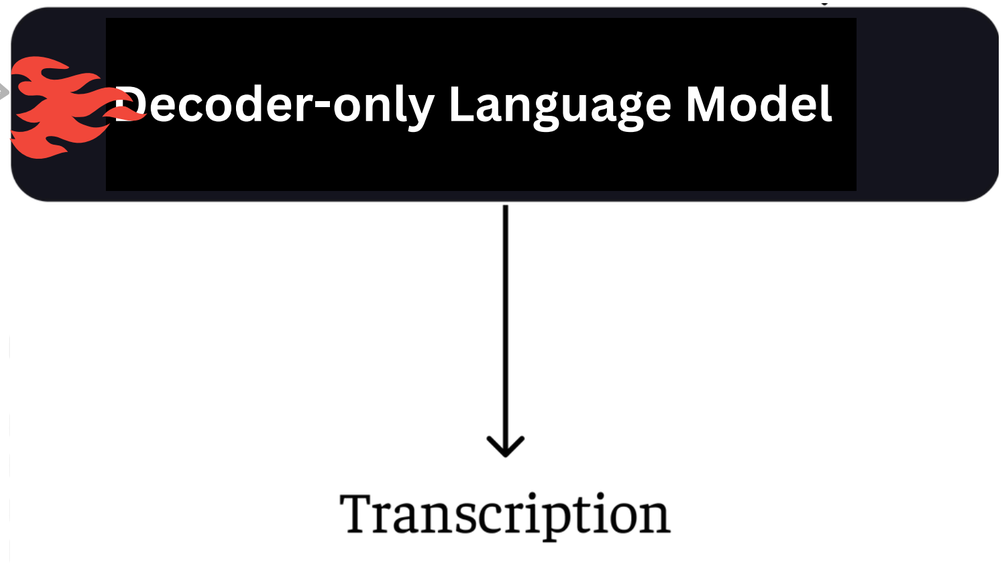

train_inner/loss
checkpoints_stm_lm_tristage
100200: 5.098 checkpoints...lm_tristage
train_inner/loss_lm
100200: 5.075 checkpoints_stm_lm_tristage
106100: 5.009 checkpoints_stm_lm_tristage
Press CMD+C to copy this data
Step
train_inner/loss_itm
106100: 0.028 checkpoints...lm_tristage
100200: 0.023 checkpoints...lm_tristage
train_inner/gnorm
100200: 8.542 checkpoints...lm_tristage
106100: 8.29 checkpoints...lm_tristage
train_inner/lr
106100: 0.000008478 checkpoints...lm_tristage
100200: 0.00000804 checkpoints...lm_tristage
fairseq_speech_dev_librispeec...seq_text_dev_librispeech/loss
100000: 4.727 checkpoints...lm_tristage
fairseq_speech_dev_librispeec...text_dev_librispeech/best_loss
100000: 4.723 checkpoints...lm_tristage
fairseq_speech_dev_librispeec..._text_dev_librispeech/loss_lm
100000: 4.643 checkpoints...lm_tristage
fairseq_speech_dev_librispeec...text_dev_librispeech/loss_itm
100000: 0.084 checkpoints...lm_tristage

PROPOSED METHOD

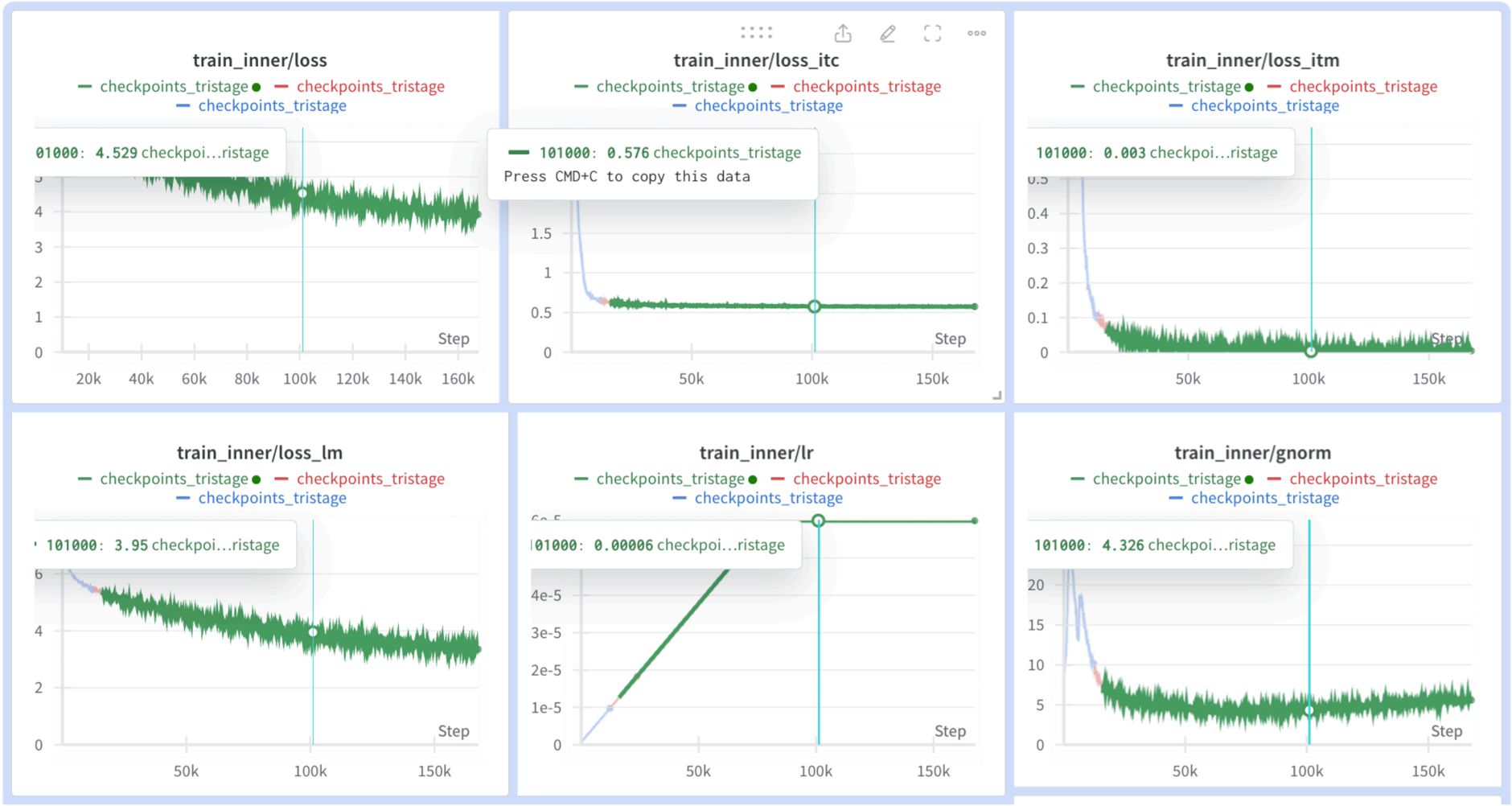
overall...



detailed....



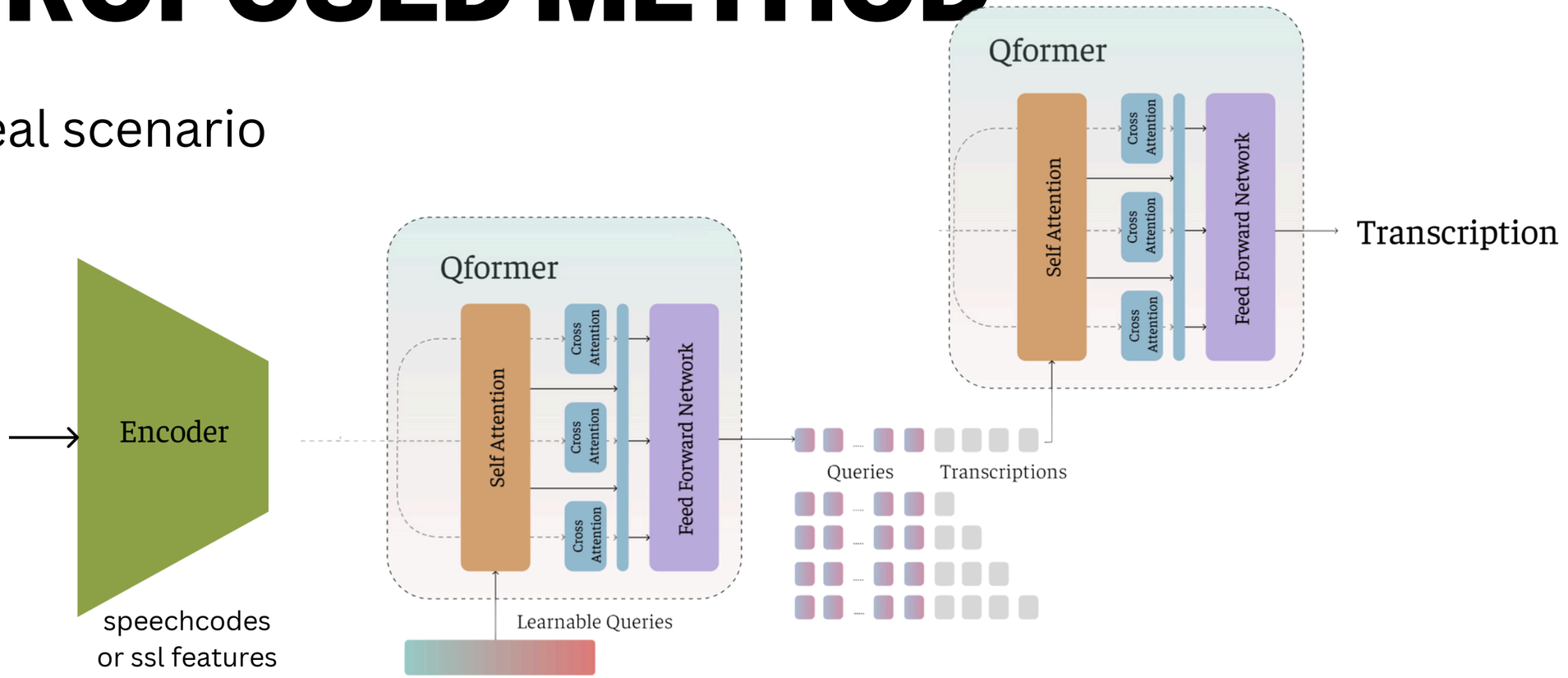
PROPOSED METHOD (TRAIN-STATS)



PROPOSED METHOD (DEV-STATS)

PROPOSED METHOD

real scenario



LIMITATIONS

The proposed system’s connectivity was assessed based on its capability to generate corresponding text pairs not taking into account NLU specific datasets.

- **TLDR**, no NLU assesment

While an examination of attention heads and layers offered insights into the system’s internal representations, we did not conduct any subsequent studies to delve deeper into these aspects.

- **TLDR**, no XAI, interpretations of attention

No ablation study examining training strategy

RESULTS

Experimental setup		WER↓		CER↓	
ASR-Qformer-∗	Queries	dev-clean	dev-other	dev-clean	dev-other
SpeechTokenizer	50	103.56%	107.79%	79.15%	82.01%
	100	103.55%	109.16%	79.50%	83.54%
HuBERT-base	50	77.56%	77.56%	62.32%	62.32%
	100	74.80%	79.96%	60.70%	64.66%
HuBERT-large	50	52.61%	53.86%	44.02%	45.16%
	100	34.47%	39.61%	31.15%	32.84%

WER results are poor!! ....
BUT

RESULTS

Experimental setup ASR-Qformer-∗	Queries	WER↓ dev-clean	WER↓ dev-other	CER↓ dev-clean	CER↓ dev-other
SpeechTokenizer	50	103.56%	107.79%	79.15%	82.01%
	100	103.55%	109.16%	79.50%	83.54%
HuBERT-base	50	77.56%	77.56%	62.32%	62.32%
	100	74.80%	79.96%	60.70%	64.66%
HuBERT-large	50	52.61%	53.86%	44.02%	45.16%
	100	34.47%	39.61%	31.15%	32.84%

WER results are poor!! ....
BUT

Experimental setup ASR-Qformer-∗	Queries	Precision↑ dev-clean	Precision↑ dev-other	Recall↑ dev-clean	Recall↑ dev-other	F1↑ dev-clean	F1↑ dev-other
SpeechTokenizer	50	84.74%	81.18%	84.83%	84.41%	84.78%	84.29%
	100	84.79%	84.06%	85.62%	84.58%	84.92%	84.31%
HuBERT-base	50	88.18%	87.10%	88.59%	87.90%	88.37%	87.49%
	100	88.46%	87.33%	88.87%	88.15%	88.65%	87.73%
HuBERT-large	50	92.11%	91.49%	92.32%	91.99%	91.73%	92.06%
	100	94.49%	93.56%	94.60%	93.83%	94.60%	93.83%

Semantic Similarity is better!!!

RESULTS

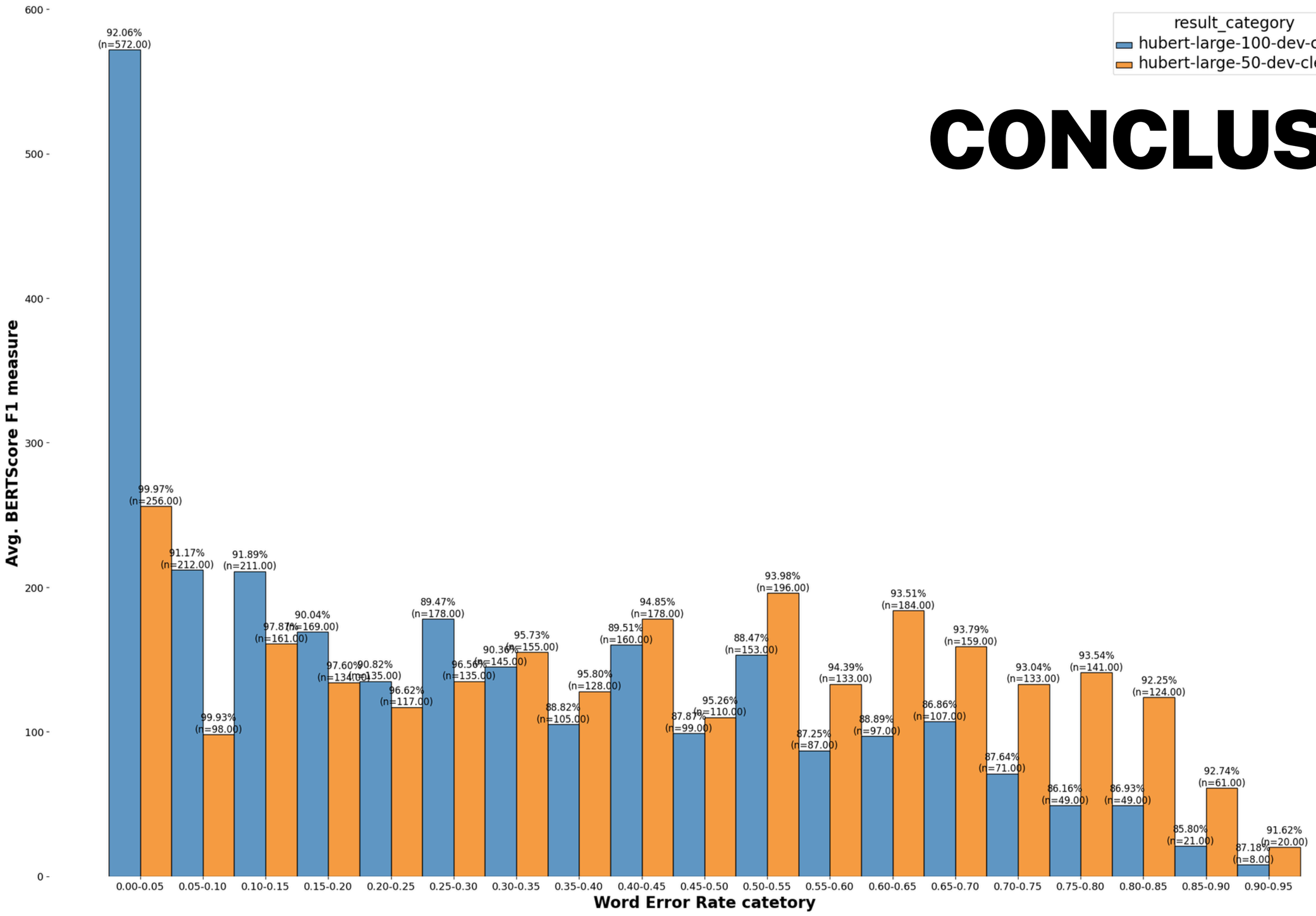
Experimental setup ASR-Qformer-∗	Target Transcription	Predicted Transcription	WER
HuBERT-base-50	there is a seat in the garden at the side of the house again she hesitated	there is a seat in the garden at the side of the house again she hesitated	0.0%
	never had any act seemed so impossible	never had any accident seemed so impossible	14.28%
	no she shut her eyes	no she shut her eyes and i	40.00%
	then she turned towards the quarter indicated and disappeared round the laurel bushes	then she turned toward the quarter below town and ordered lord chelford the coroner's court	61.53%
	look here he said this is all nonsense you know you are tired out and there's something wrong what is it	look here he said this is all you know and there is something wrong here you are tired and tired	57.14%
HuBERT-large-50	i shall lock up all the doors and windows in the house and then i shall give you my latch key and you can let yourself in and stay the night here there is no one in the house	i shall lock up all the doors in the house and then i shall lock up your door and let you in there is no one in the house and you can stay all the night here myself and the lazaretto	61.53%
	i will catch the night train and bring my mother up to morrow then we will see what can be done	i will catch up the night train and bring my mother to morrow then we will see what can be done	9.52%
	when she opened the door of it the bright fire which beenie undesired had kindled there startled her the room looked unnatural uncanny because it was cheerful	when she opened the door of it the bright fire sparkled there unlike the room which linley had described the room was unmoved because it looked untenanted	51.85%

Experimental setup ASR-Qformer-∗	Queries	WER↓ dev-clean	WER↓ dev-other	CER↓ dev-clean	CER↓ dev-other
SpeechTokenizer	50	103.56%	107.79%	79.15%	82.01%
	100	103.55%	109.16%	79.50%	83.54%
HuBERT-base	50	77.56%	77.56%	62.32%	62.32%
	100	74.80%	79.96%	60.70%	64.66%
HuBERT-large	50	52.61%	53.86%	44.02%	45.16%
	100	34.47%	39.61%	31.15%	32.84%

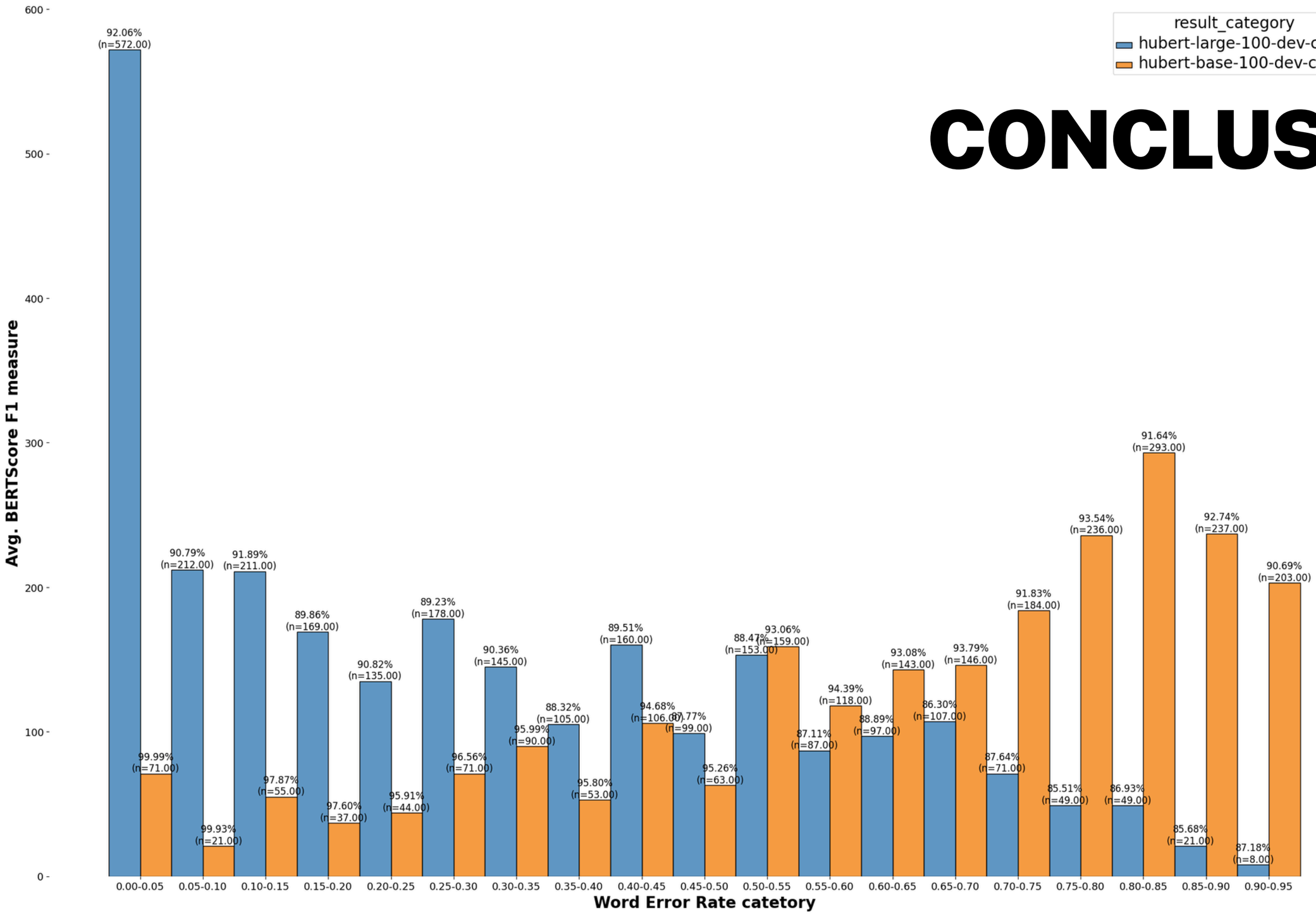
Experimental setup	WER↓ dev-clean	WER↓ dev-other	CER↓ dev-clean	CER↓ dev-other
CTC/AED FBank [48]	2.5%	6.3%	—	—
CTC/AED dicrete tokens [48]	2.2%	4.5%	—	—
ASR-Qformer-HuBERT-large-100	34.47%	39.61%	31.15%	32.84%

Table 5.2: WER Comparison on LibriSpeech from existing methods

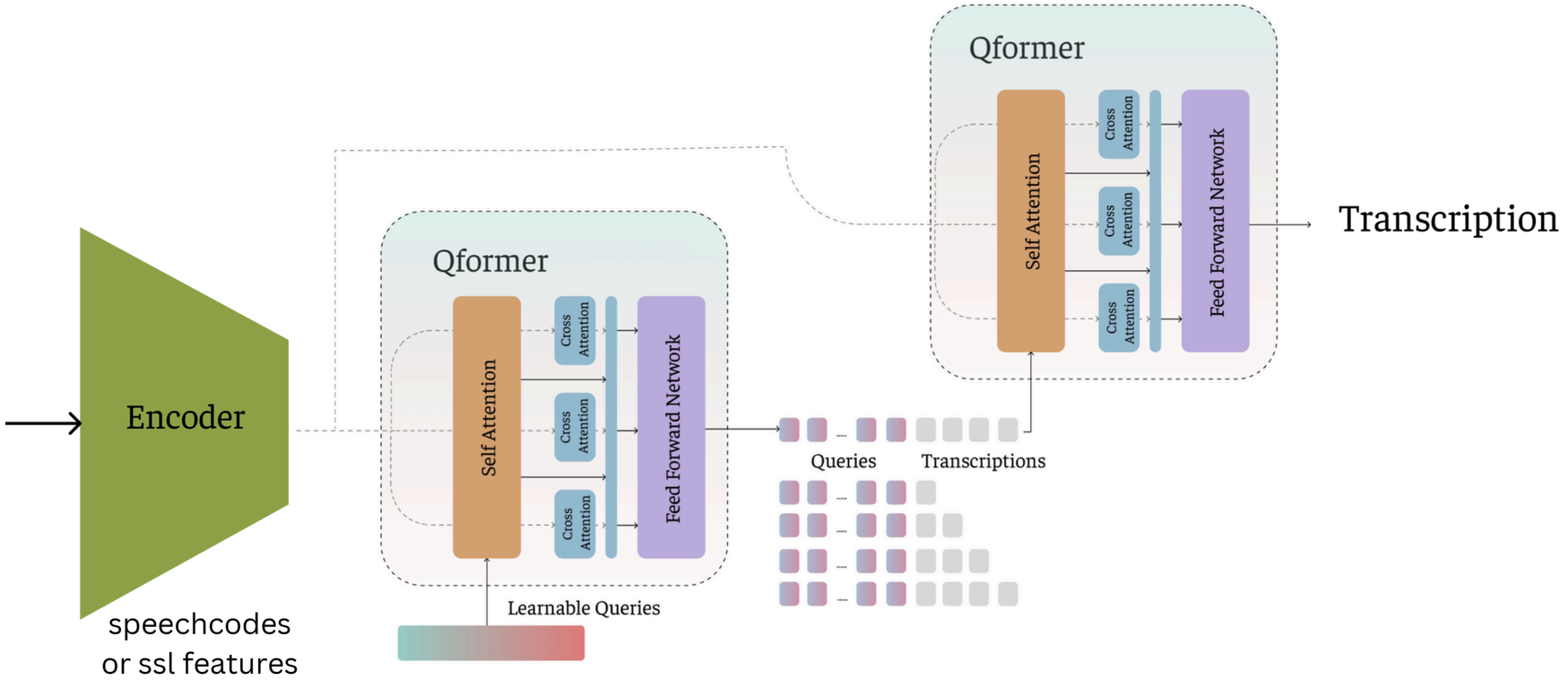
CONCLUSSION



CONCLUSSION



PROPOSED METHOD FOR AST



....same architecture as for ASR

* AST - Automatic Speech Translation

RESULTS

System	BERTScore	
	En→De	En→Fr
STRONGBASELINE* [51]	77.44%	81.75%
WEAKBASELINE* [51]	74.86%	77.28%
APPTEK-Constrained [51]	77.32%	–
HW-TSC-Constrained [51]	–	76.11%
HW-TSC-Unconstrained [51]	75.79%	–
APV-Unconstrained [51]	73.68%	77.77%
HW-TSC-Constrained [51]	74.07%	76.11%
AST-Qformer-HuBERT-large-50	71.24%	65.39

Table 5.6: Combined En→De and En→Fr translation evaluation for BERTScore

Experimental setup	Precision↑		Recall↑		F1↑	
ASR-Qformer-∗	dev	tst-COMMON	dev	tst-COMMON	dev	tst-COMMON
HuBERT-large En→De	71.35%	71.48%	71.20%	71.24%	71.25%	71.33%
HuBERT-large En→Fr	69.37%	70.54%	69.47%	70.27%	69.39%	70.37%

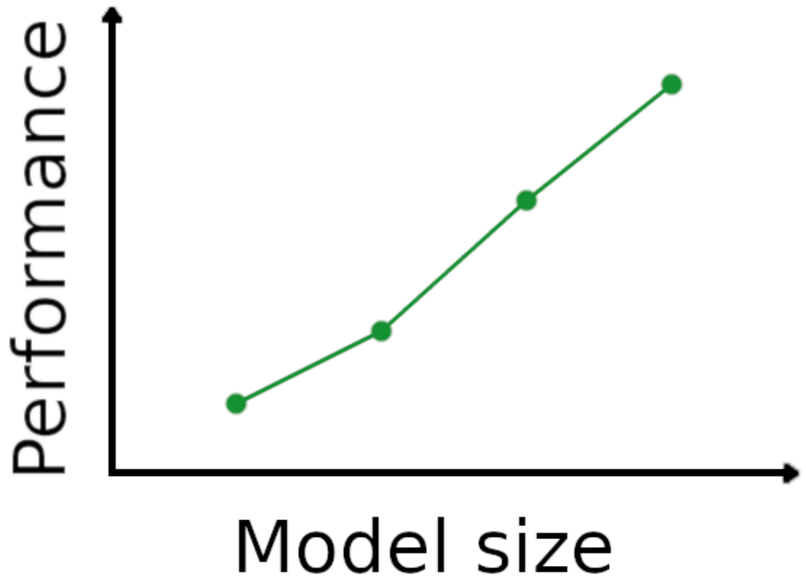
Table 5.7: BERTScore Comparison on LibriSpeech development and test splits

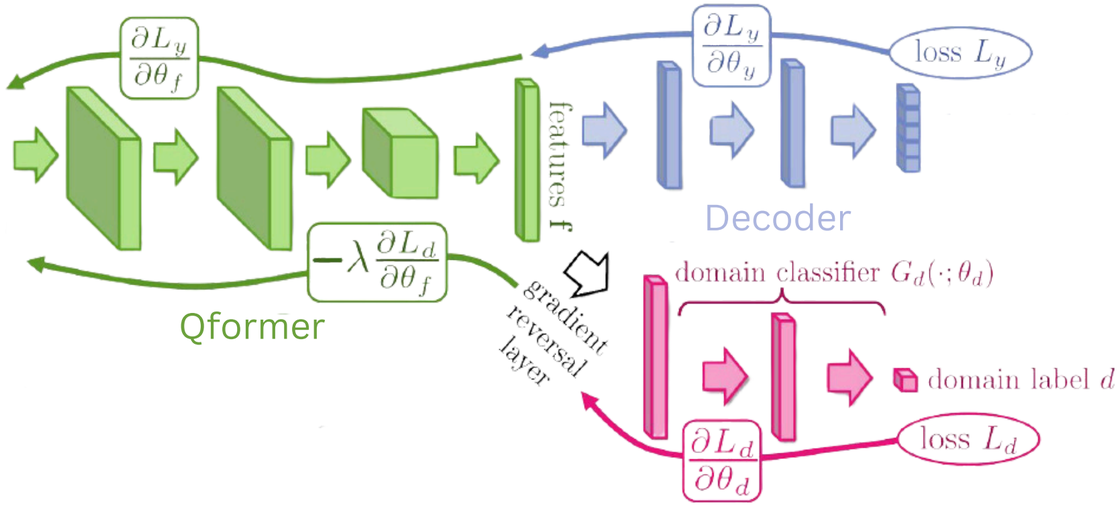
BLEU exhibits same problem as WER & CER discussed previously

... BERTScore

DISCUSSION

Insights from the drawn figures and tables emphasize the importance of query size and model depth as a critical factor in the design and optimization of speech recognition systems, highlighting its impact on the system's ability to process and interpret speech data effectively. Results from test-clean, test-other split from LibriSpeech indicate that the proposed method suffers distributional shifts raising the need for a domain adaptation strategy method. We believe that there is plenty of potential iterative work in the direction of ASR-Qformer in the provided research.



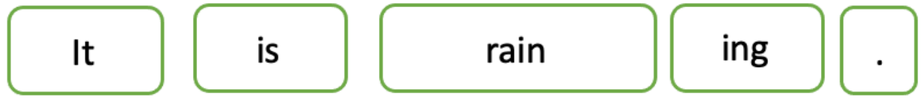


DISCUSSION

Normally, current ASR models operate with vocabulary on character level with classification objective, however, Qformer uses BERT's weights and vocabulary. Thus in our experiments we align speech representation with BERT's vocabulary that operates on word pieces rather than usual character level. This obviously adds complexity to the system.

Character level tokenization

Sub-word level tokenization



SpeechTokenizer uses CNN feature encoder/extractor and RVQ with 8 vector quantization block each with 1024 sized codebook

HuBERT used CNN feature encoder/extractor

example: 21seconds audio with sampling rate 16K is 336K frames

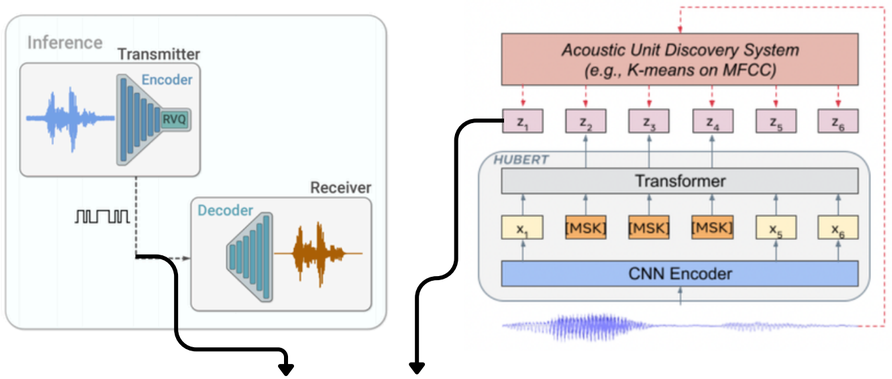
- SpeechTokenizer input tensor is **[1, 1050, 1024]**
- HuBERT input tensor is **[1, 1049, 768]**
- Tokenized text tensor is **[1, 75, 768]**

←--------due to semantic knowledge distillation

DISCUSSION

The outcomes derived from the translation tasks further emphasize the potential for advancement within our architectural framework. The comparative analysis with benchmarks from the IWSLT 2022 Isometric Spoken Language Translation challenge illuminates the disparities in performance, particularly in semantic understanding and translation accuracy as reflected by our lower BERTScore. This revelation not only highlights the commendable capabilities of competing networks but also serves as a motivation for the refinement of our architecture.

Overall, there is still a gap between speech and text modality, despite the main outcomes suggesting that the proposed training method is feasible and experiments representing semantic similarity to the target transcription. By that, we mean systems that connect LLM with speech encoders that were not created with speech-to-text objective. Moreover, positive insights are not revealed by the neural encoder's speech code results. The latent representation we transferred from a quantized codebook may be the cause of the subpar performance. Proposed hypothesis posited that the encoder's representations with hidden dimension were optimally suited as inputs for the Qformer, might be wrong, yielding the need in introducing embedding layer.

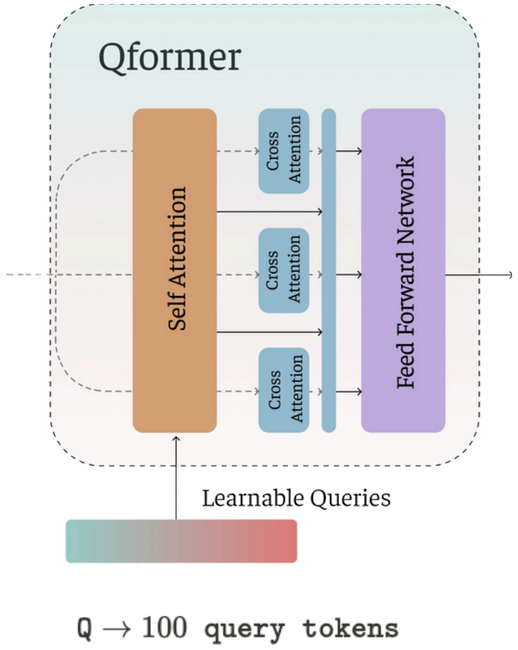


input to proposing connecter....

Suggestions...

1. Clustering HuBERT's intermediate layer representation
2. Introducing emebdding layer for speechcodes from SpeechTokenizer rather than using codebook's latent representation

CONCLUSSION



TLDR; “We introduce an alignment strategy inspired by the vision-language model BLIP2, adapting it by replacing speech representations. This approach does not iterate upon BLIP's original methodology but rather expands the application of discrete speech representation. Our study investigates the effectiveness of using speech units as inputs across various speech processing tasks, such as Automatic Speech Recognition and Automatic Speech Translation. To assess the versatility of discrete units, we utilized two distinct types of speech encoders: one focused on reconstruction and the other on a masking objective.

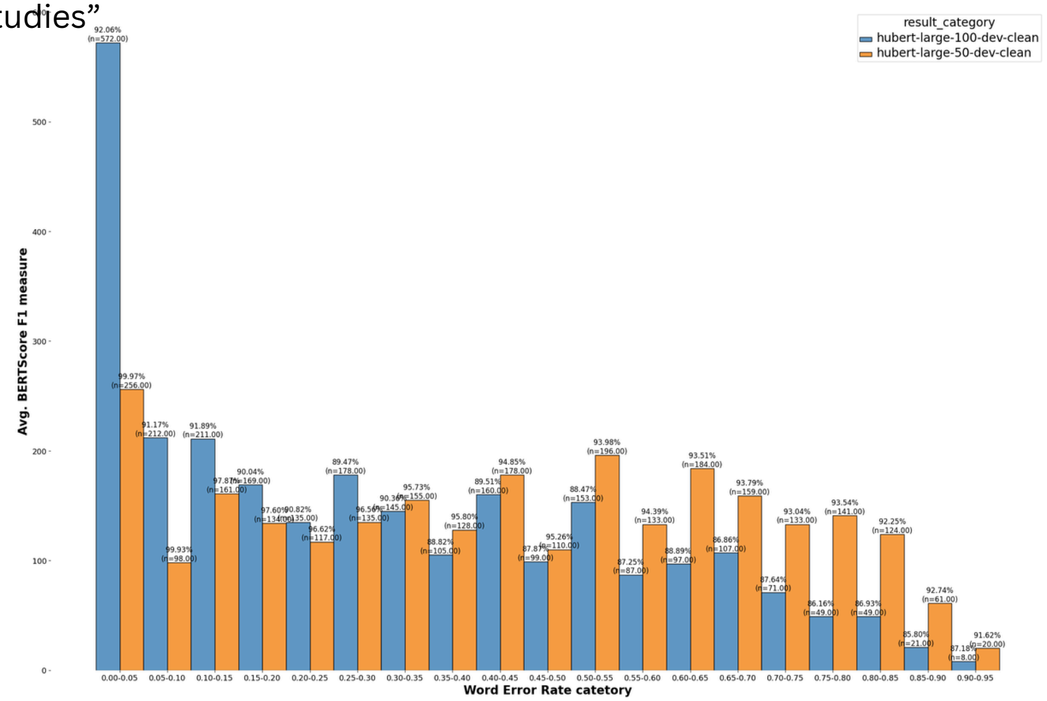
Our experiments conducted on the LibriSpeech dataset demonstrate the efficacy of our method, though it does present challenges in precisely matching the target text. By comparing our model's performance to existing benchmarks, we found a interrelationship between the outputs and the semantic representation of the gold standard labels. However, the neural encoder designed for speech reconstruction, which we employed, did not produce favorable results based on our assumptions, thereby suggesting that features derived from the masking model offer a more suitable context for extracting queries. The research questions that we outlined in the beginning of the paper raise two main problems that were addressed in this study.”

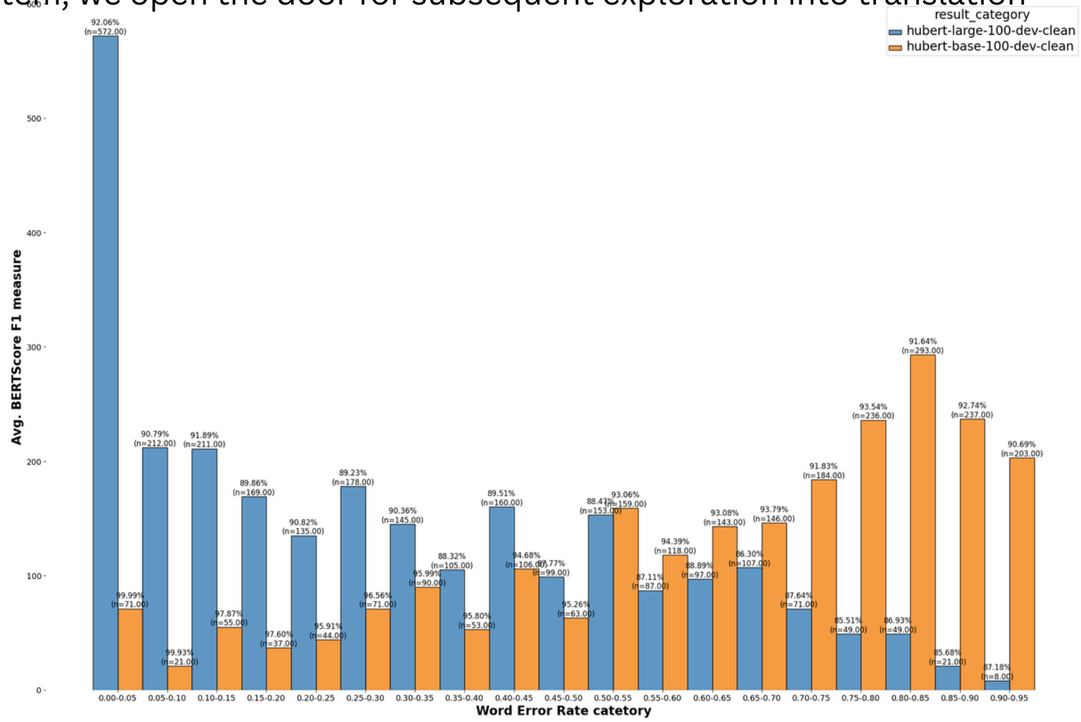
NLPR; (not long pleaser read)

- Introduced alignment strategy tested on ASR and AST
- Speechcodes and SSL representation is tested as input
- LibriSpeech results suggest that exact word matching did not succeed but semantic similarity is preserved

CONCLUSSION

TLDR; “Firstly, our investigation into whether speech-encoded representations, derived either from compression algorithms or speech self-supervised learning methods, can bridge the modality gap between text and speech concludes positively. Evidence from our ASR-HuBERT experiments supports the efficacy of larger encoders with an increased number of query tokens. A more detailed analysis provided in our evaluation section's Figure\ref{fig:hubert_base_vs_large_100} depicts that, across various error categories represented on the x-axis, semantic meaning is preserved with approximately 90\% similarity. This underscores the ability of self-supervised learning features to align closely with text. The resutls of our findings are particularly promising for future research focused on low-resource languages with substantial amounts of untranscribed speech resources. By demonstrating that training a speech model without actual transcriptions and training a text model with a modest quantity of speech-text pairs is sufficient to construct an ASR system, we open the door for subsequent exploration into translation studies”





CONCLUSSION

TLDR; “Finally, this study contributes a novel reduction strategy through the use of a fixed number of query tokens to extract information from the speech modality. This approach addresses a common challenge within the speech research community: the search for compressing strategy in high-dimensional speech representations into more manageable, lower-dimensional forms to reduce computational resource and storage requirements. Our methodology effectively condenses speech representations into a fixed number of tokens that approximate the dimensionality of text tokenized representations, offering a practical solution to these challenges.”

NLPR

- Another advantage of the system is fixed-sized reduced speech representation that might used for autoregressive decoding