# Models for Trustworthy Speech Translation

*using Tower*

HyoJung Han

Computer Science
University of Maryland

DEPARTMENT OF
COMPUTER SCIENCE

IST-Unbabel Seminar, Nov 25th 2024

The rapid development of speech technology has expanded the use of speech translation (ST) applications in daily life.

The rapid development of speech technology has expanded the use of speech translation (ST) applications in daily life.

→ The needs to predict the reliability of their output is increasing.

# Models for Trustworthy Speech Translation

DEPARTMENT OF COMPUTER SCIENCE

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

## SpeechQE: Estimating the Quality of Direct Speech Translation

EMNLP2024

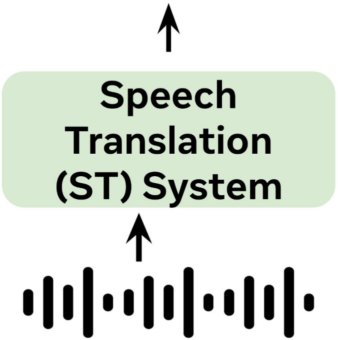

**HyoJung Han**
Computer Science
University of Maryland

**Kevin Duh**
HLTCOE
Johns Hopkins University

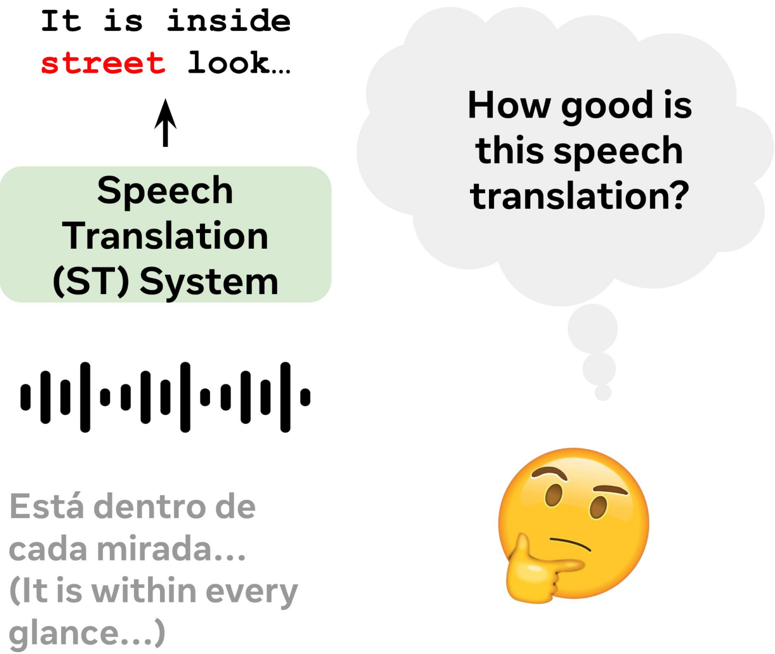
**Marine Carpuat**
Computer Science
University of Maryland

Está dentro de cada mirada…

It is inside
street look…




Está dentro de
cada mirada…
(It is within every
glance…)


It is inside street look…
Speech Translation (ST) System
Está dentro de cada mirada…
(It is within every glance…)
How good is this speech translation?

# Estimating Quality (QE) of Translation

Assessing the quality of translation is crucial as it help people rely on MT appropriately without reference.

Many works on QE*:

google/**metricx-23-qe-xxl-v2p0**

Unbabel/**wmt20-comet-qe-da**

**GEMBA**

Unbabel/**wmt22-cometkiwi-da**

Unbabel/**XCOMET-XL**

facebook/**blaser-2.0-qe**

**MaTESe**

**MS-Comet**

Unbabel/**wmt23-cometkiwi-da-xl**

google/**metricx-23-qe-xl-v2p0**

**UniTE**

Unbabel/**wmt21-comet-qe-mqm-marian**

**OpenKiwi**

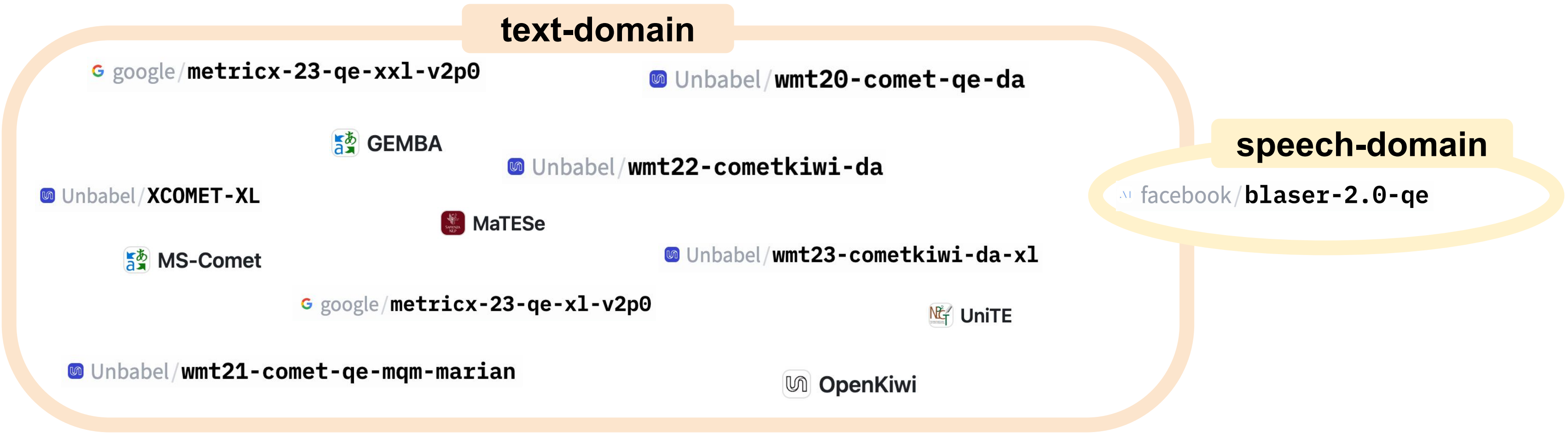


* Displayed is a subset of the QE works, which includes a runnable model.

# QE in Speech domain is underexplored

The rapid development of speech technology has expanded the use of speech translation (ST) applications in daily life,
thus <u>increasing the need to predict the reliability of their output</u>.

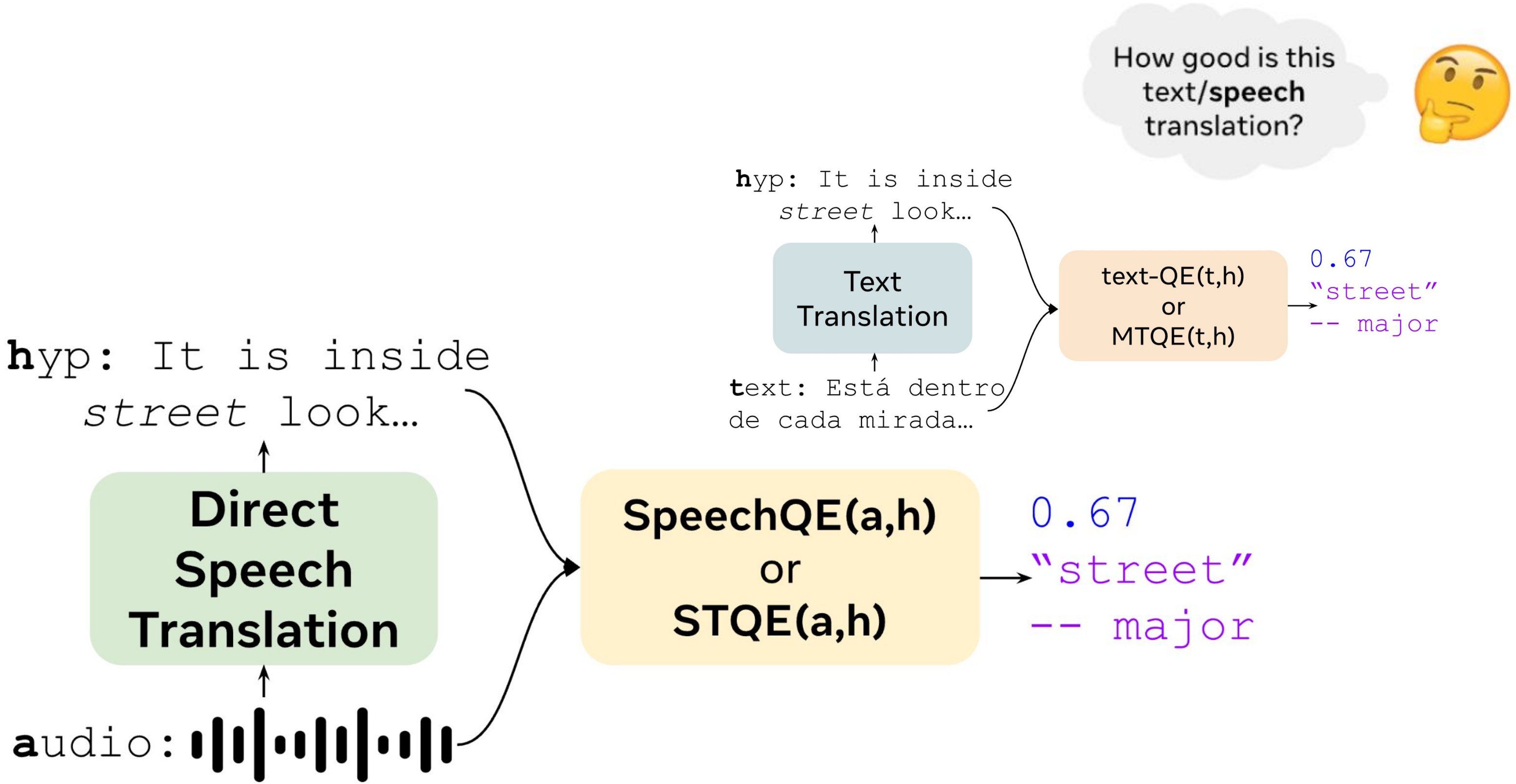
Many works on QE*:





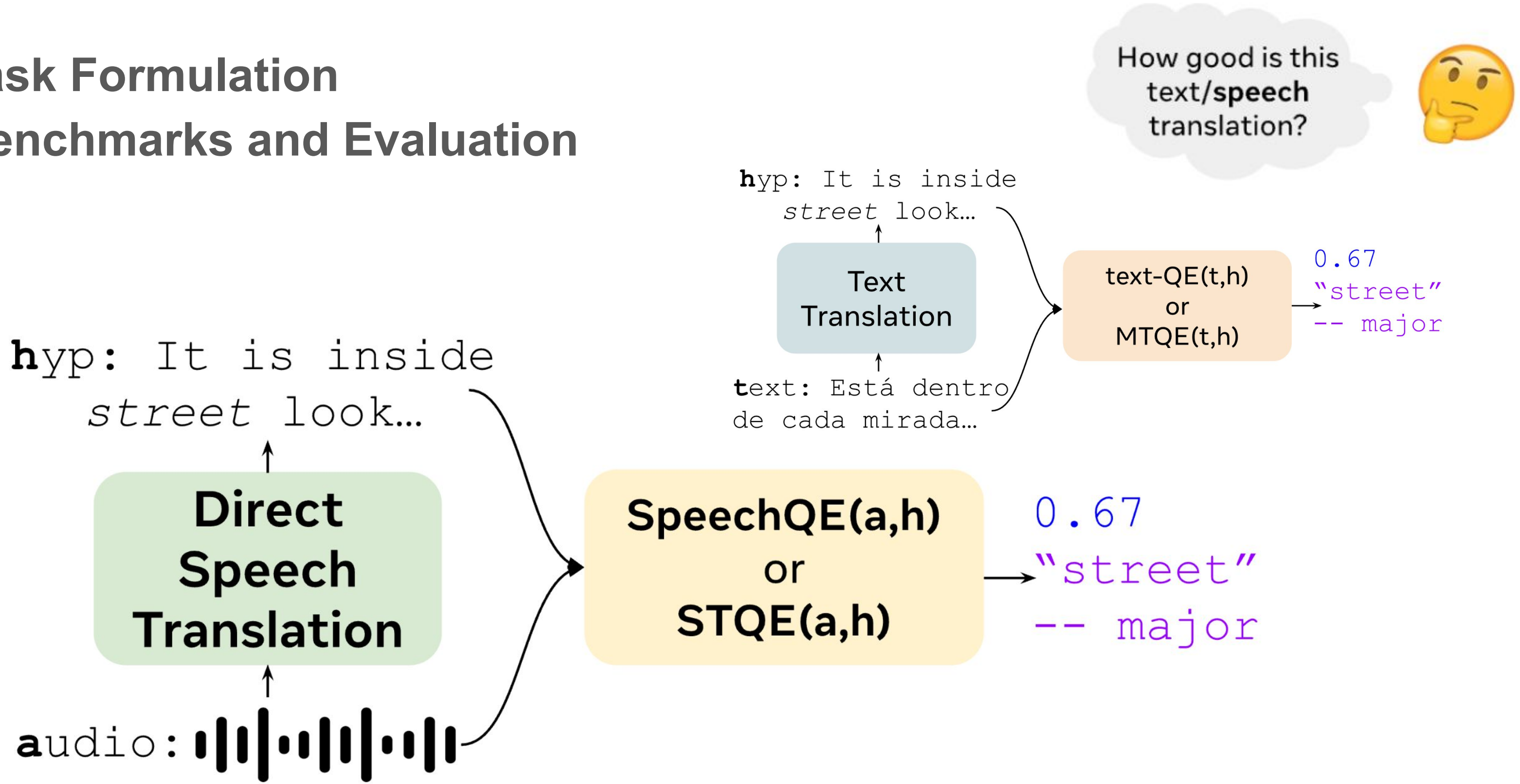

* Displayed is a subset of the QE works. The overall trends remain consistent.

# **SpeechQE**: Estimating Quality of Direct Speech Translation

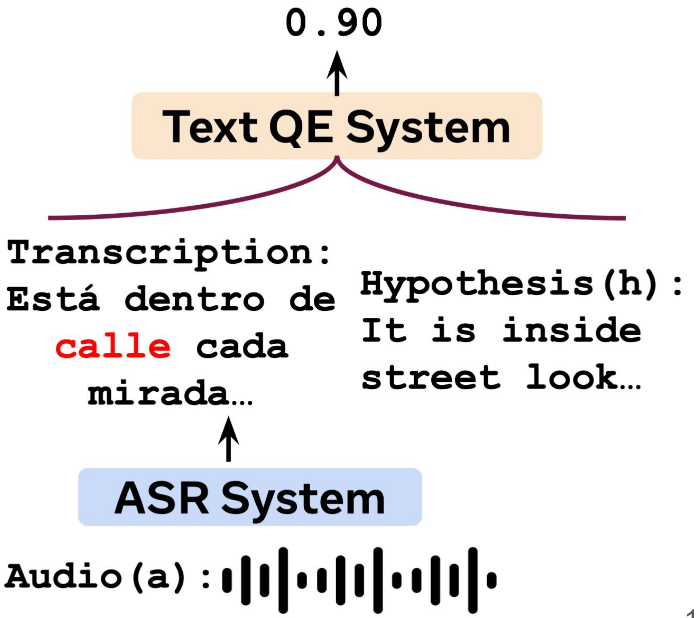


# **SpeechQE**: Estimating Quality of Direct Speech Translation

1. **Task Formulation**
2. **Benchmarks and Evaluation**



# Cascaded SpeechQE System

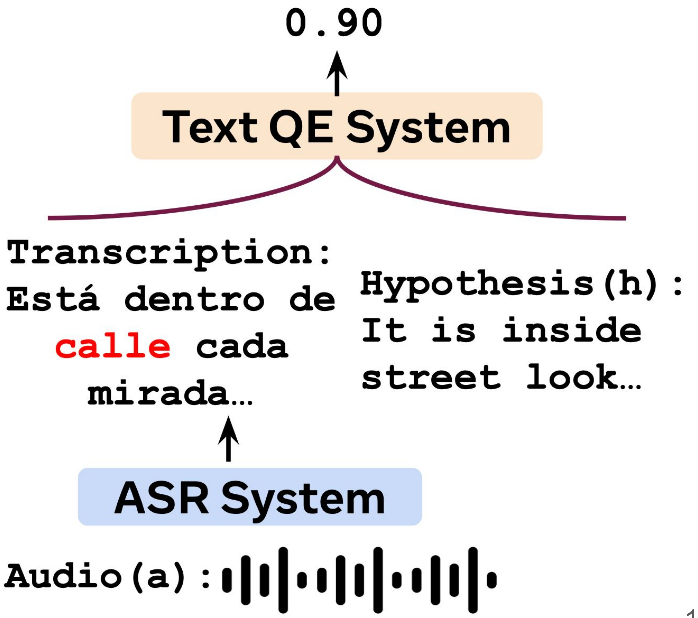


# Potential Issues of Cascaded SpeechQE System

1. **Efficiency**
   no (naturally occurring) intermediate ASR transcripts

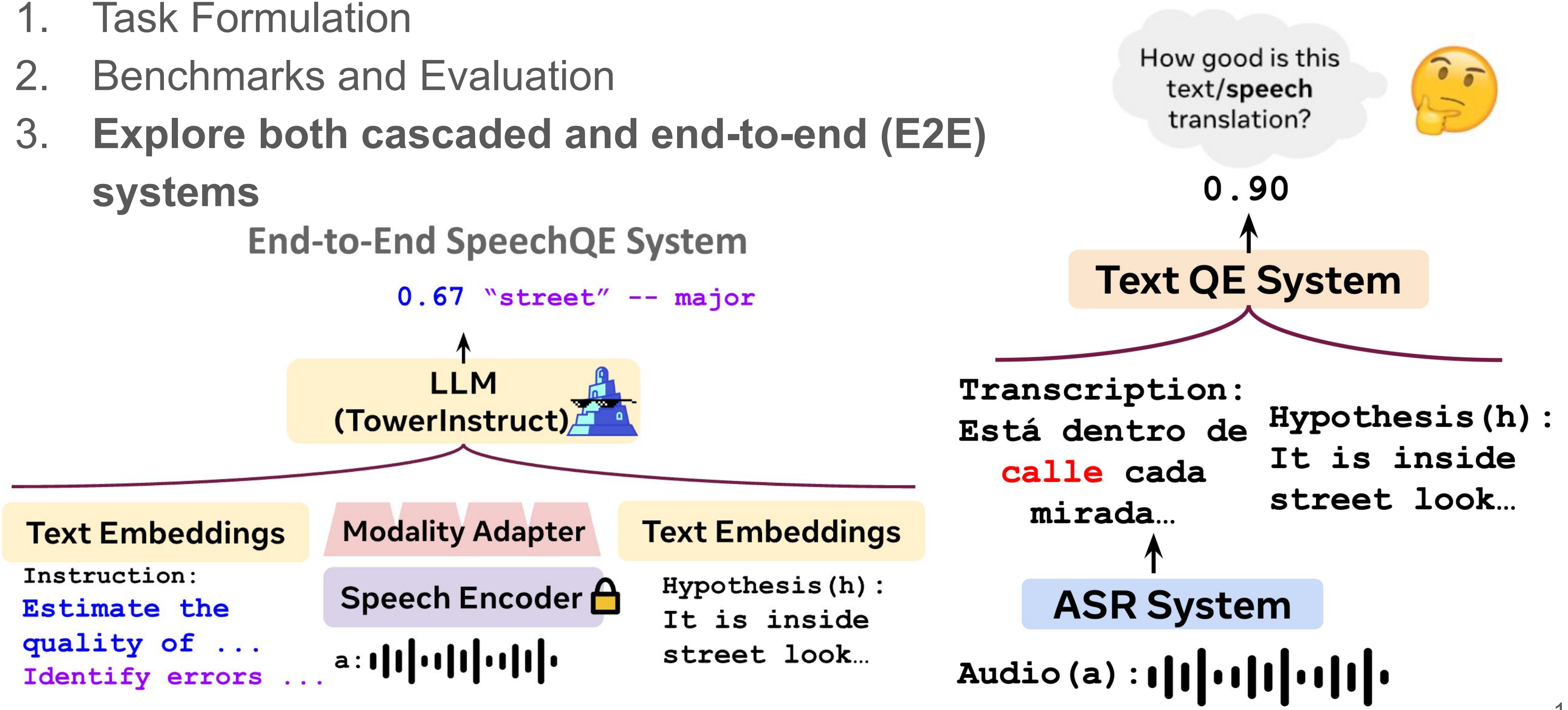
2. **Wrong speech representation** by ASR

3. **Modality mismatch**
   text-QE is not adapted to spoken language



# **SpeechQE**: Estimating Quality of Direct Speech Translation

1. Task Formulation
2. Benchmarks and Evaluation
3. **Explore both cascaded and end-to-end (E2E) systems**



# Preliminaries

h: hypothesis text
r: reference text
t: source text

"*Metric*": reference-based metric. e.g. BLEU, chrF, xCOMET, MetricX

score m: $m = metric(h, r)$ or $m = metric(t, h, r)$

# Preliminaries

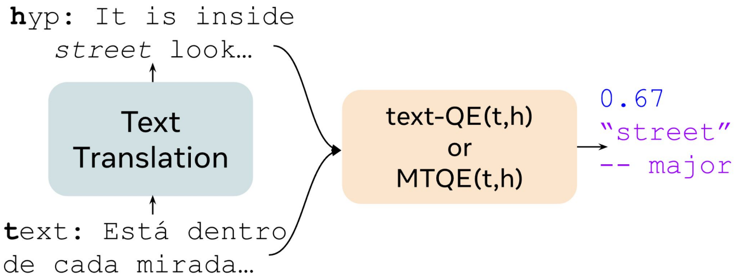
h: hypothesis text
r: reference text
t: source text

"*Metric*": reference-based metric. e.g. BLEU, chrF, xCOMET, MetricX

score m: $m = metric(h, r)$ or $m = metric(t, h, r)$

"text-QE": text quality estimation system. e.g. COMETkiwi, MetricX-QE

score q: $q = \text{text-QE}(t, h)$




* Among various ways in framing the QE task,
we mainly focus on sentence-level quality rating QE as Sentence-level quality rating instead of word-level.
We also experiment with Error Span Detection for a border QE scope in the paper.

# SpeechQE Task Formulation

h: hypothesis text

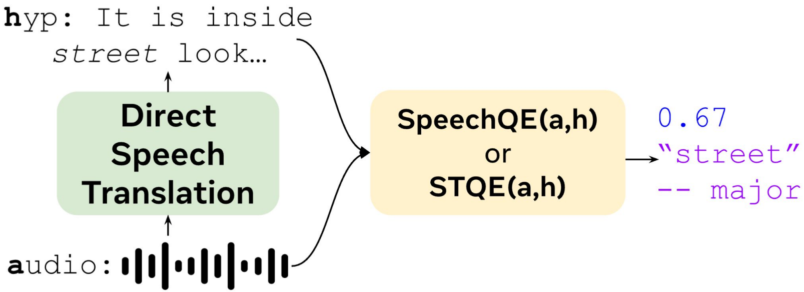
r: reference text

t: source text

a: source audio

“SpeechQE”: now the source input **a** is audio

score q: $q = SpeechQE(a, h)$

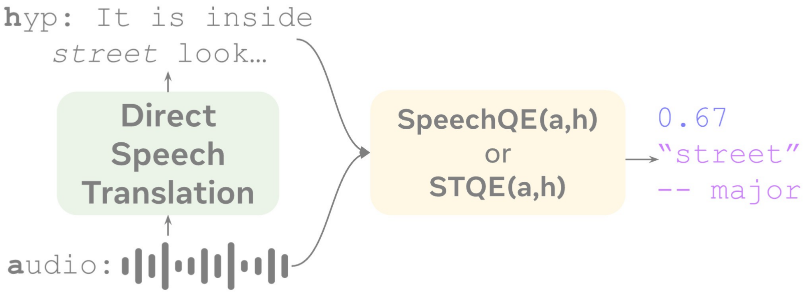


# SpeechQE Task Formulation

h: hypothesis text
r: reference text
t: source text
a: source audio

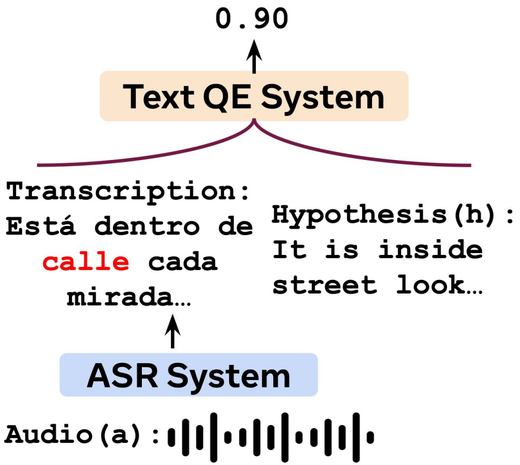
"SpeechQE": now the source input **a** is audio

score q: $q = SpeechQE(a, h)$



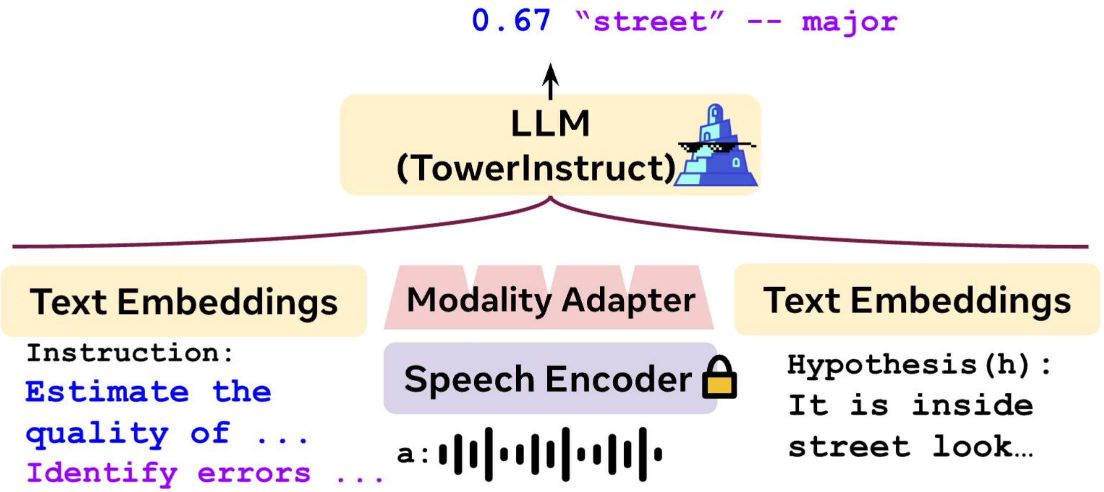

Cascaded SpeechQE System: outputs the score from a text-based QE system with the input of transcribed text ASR(a)

$$q_{cas} = \text{text-QE}(ASR(a), h)$$



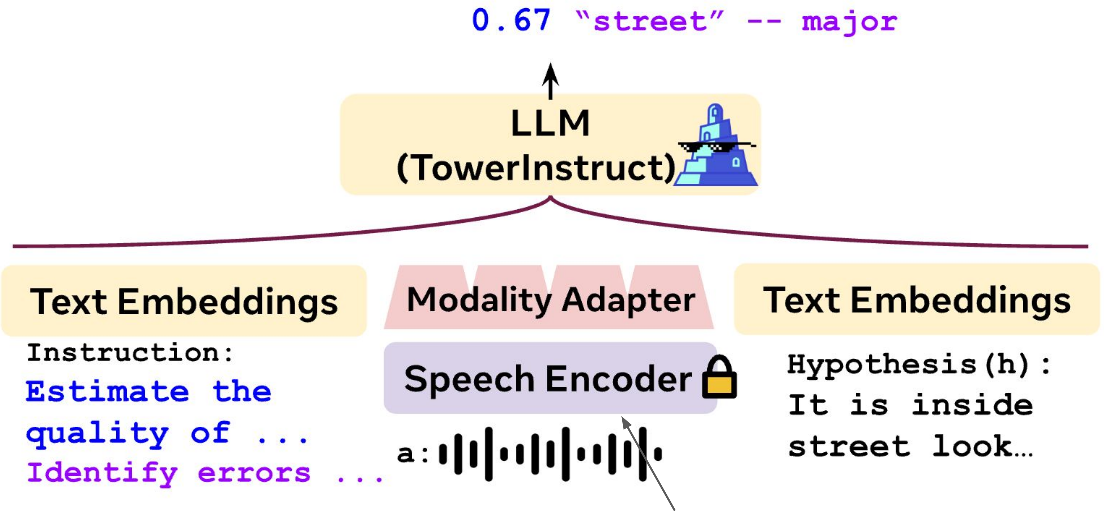
# Building End-to-End SpeechQE System

Incorporating a pre-trained speech encoder and a large language model (LLM)



# Building End-to-End SpeechQE System

Incorporating a pre-trained speech encoder and a large language model (LLM)

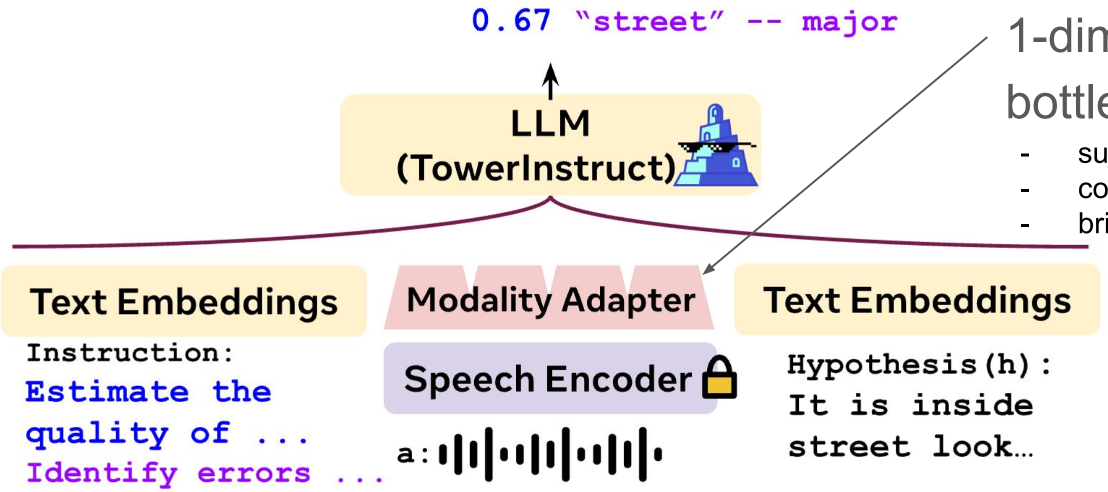



(Pre-trained) Speech Encoder: Whisper-large-v2

- extracts the audio feature from the raw audio.

# Building End-to-End SpeechQE System

Incorporating a pre-trained speech encoder and a large language model (LLM)




Modality Adapter: three 1-dim conv + 512-dim bottleneck layer

- subsamples the audio features
- compress the audio sequence
- bridges the speech representation

# Building End-to-End SpeechQE System

Incorporating a pre-trained speech encoder and a large language model (LLM)

(Pre-trained) text-LLM:
TowerInstruct-7B

- Input: text + audio embedding sequence.

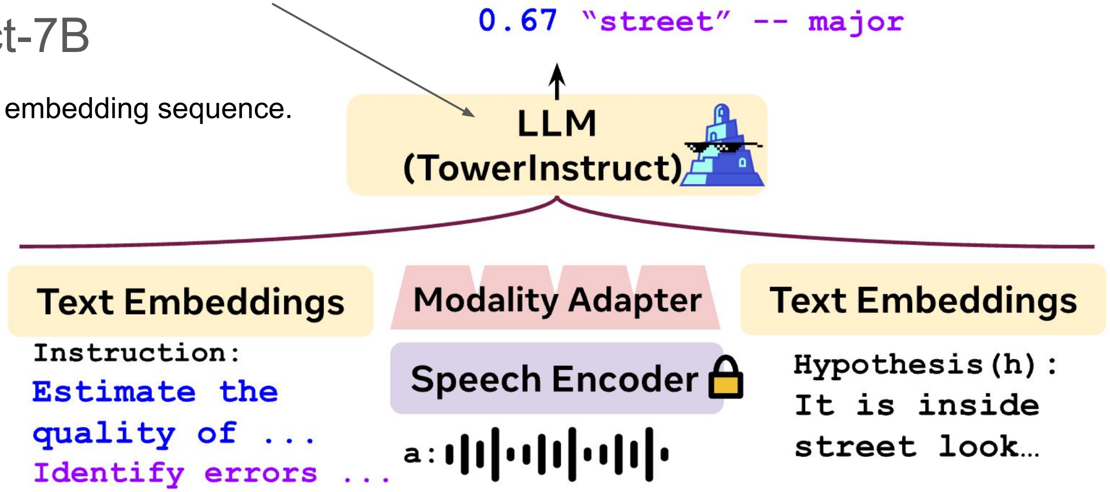


Table 10: Example of English-to-German speech translation and quality estimations of SpeechQE systems. Both cascaded and E2E SpeechQE systems could detect errors. However, the cascaded system estimates the severity lower than that of the metric labels partly due to ASR error while E2E could estimate the quality closely to labels.

```
# QE4ST task, training and testing

Given the German translation of the speech, estimate the quality of the translation as a
score between 0 to 1.

English: [[audio input]]

German translation: Wir modellieren den grasweisen, obstruktiven Summize-Ansatz mit zwei
verschiedenen Methoden.

# desired output in training or example output in testing

0.851
```

🔥 : Update weights

❄️ : Freeze weights

: LoRA Fine-tuning

# Training End-to-End SpeechQE System

SpeechQE {source audio (a), hypotheses (h), ratings (m)} + ASR + ST task

🔥 : Update weights

❄️ : Freeze weights

: LoRA Fine-tuning

# Training End-to-End SpeechQE System

SpeechQE {source audio (a), hypotheses (h), ratings (m)} + ASR + ST task

🔥 : Update weights

❄️ : Freeze weights

: LoRA Fine-tuning

# Training End-to-End SpeechQE System

SpeechQE {source audio (a), hypotheses (h), ratings (m)} + ASR + ST task

## Single-Phase Training

## Two-Phase Training

8 sample per batch, 1 GPU

🔥 : Updated weights

❄️ : Frozen weights

: LoRA Fine-tuned

# Training End-to-End SpeechQE System

SpeechQE {source audio (a), hypotheses (h), ratings (m)} + ASR + ST task

**Single-Phase Training**

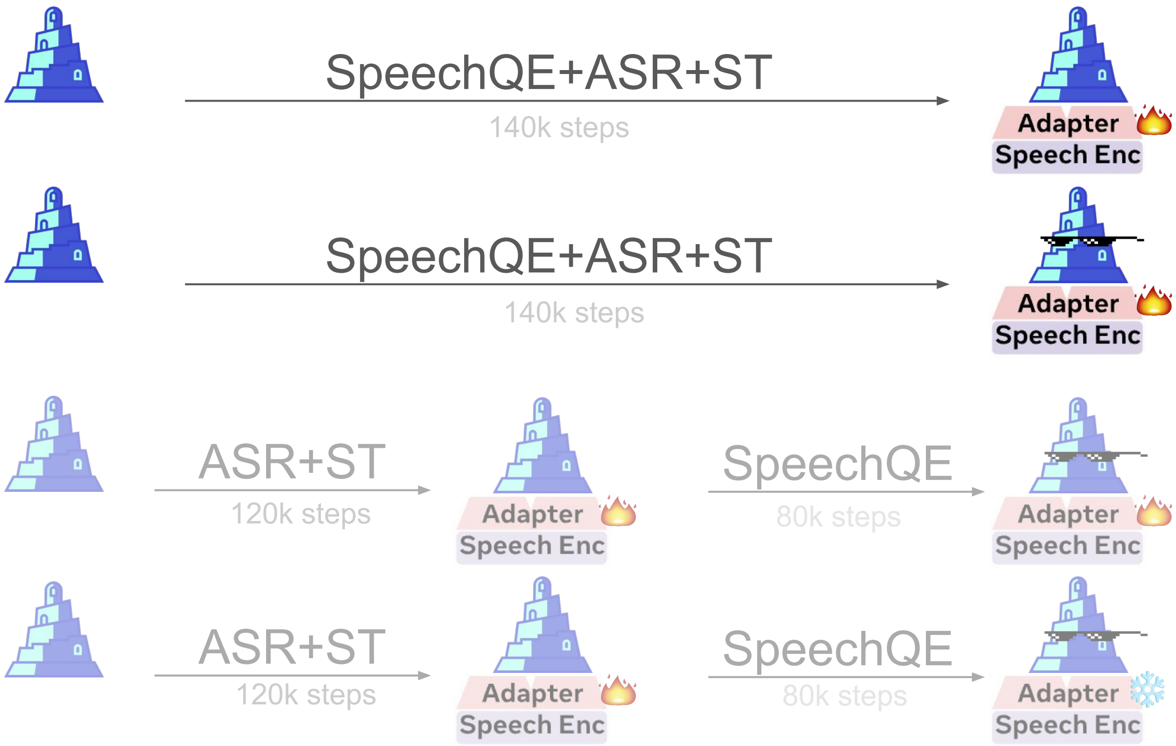
LLM-Fixed+Adapter

LLM-*LoRA*+Adapter

**Two-Phase Training**

LLM-*LoRA*
+Adapter-pre-train(pt)

LLM-*LoRA*
+Adapter-pt&Fixed




(Speech Encoder🔒 is always frozen)

8 sample per batch, 1 GPU

# Training End-to-End SpeechQE System

🔥 : Updated weights

❄️ : Frozen weights

: LoRA Fine-tuned

SpeechQE {source audio (a), hypotheses (h), ratings (m)} + ASR + ST task

**Single-Phase Training**

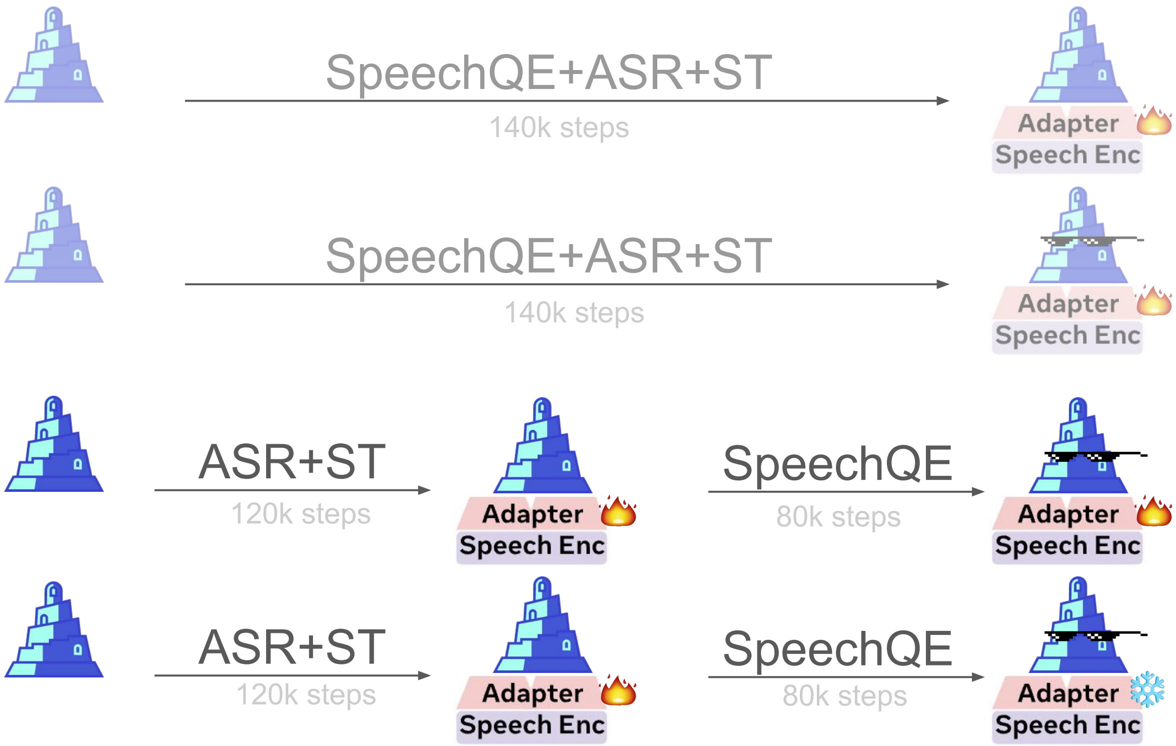
LLM-Fixed+Adapter

LLM-*LoRA*+Adapter

**Two-Phase Training**

LLM-*LoRA*
+Adapter-pre-train(pt)

LLM-*LoRA*
+Adapter-pt&Fixed




(Speech Encoder🔒 is always frozen)

8 sample per batch, 1 GPU

: Updated weights

: Frozen weights

: LoRA Fine-tuned

# Training End-to-End SpeechQE System

SpeechQE {source audio (a), hypotheses (h), ratings (m)} + ASR + ST task

**Single-Phase Training**

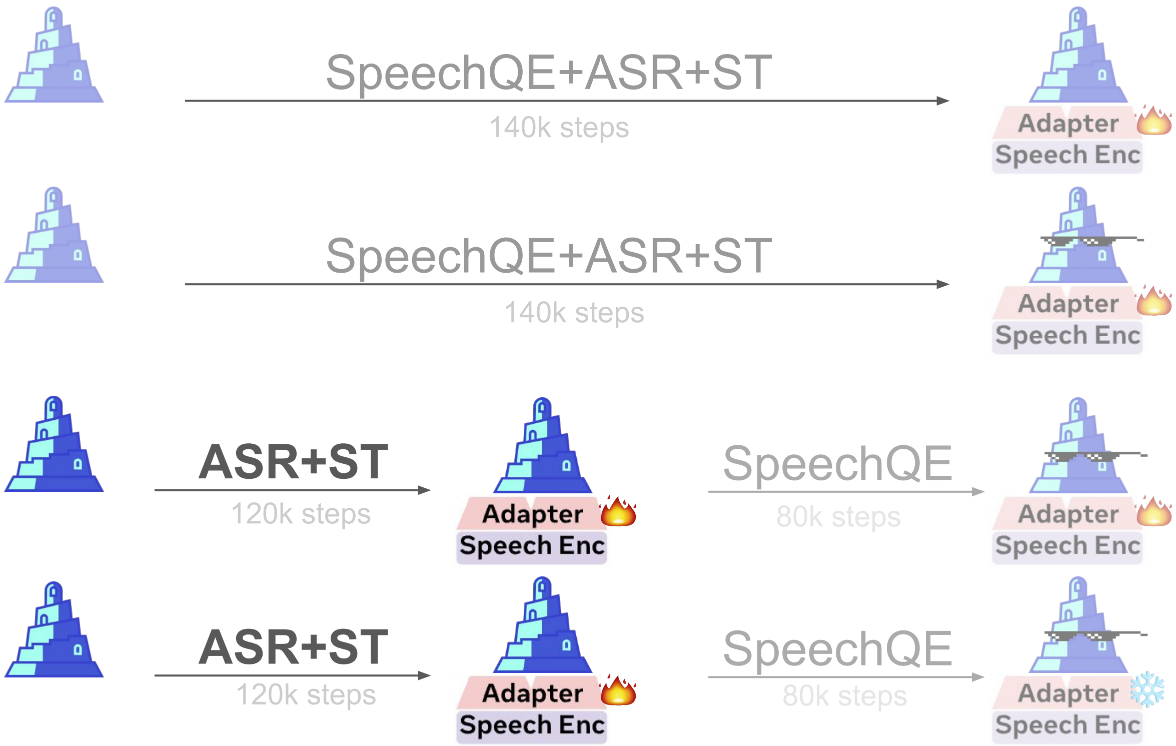
LLM-Fixed+Adapter

LLM-*LoRA*+Adapter

**Two-Phase Training**

LLM-*LoRA*
+Adapter-pre-train(pt)

LLM-*LoRA*
+Adapter-pt&Fixed




(Speech Encoder is always frozen)

8 sample per batch, 1 GPU

# Training End-to-End SpeechQE System

🔥 : Updated weights

❄️ : Frozen weights

: LoRA Fine-tuned

SpeechQE {source audio (a), hypotheses (h), ratings (m)} + ASR + ST task


(Speech Encoder 🔒 is always frozen)

# Building SpeechQE Benchmark and Training Corpus

1. Subsampled about 80k segments from the training set and 500 from the dev and test of CoVoST2
2. Run ST models to get hypothesis.
3. Get automatic quality labels from (reference-based) metrics xCOMET-XL for train, adding MetricX-23-XL for test

| CoVoST2 | es2en | en2de | es2en | en2de |
|---|---|---|---|---|
| | *Train* | | *Dev / Test* | |
| ASR | 297k | 305k | | |
| ST | 79k | 290k | | |
| SpeechQE | 546k | 589k | 3.5k | 3.5k |

| Es2En diect ST systems | CoVoST2 BLEU |
|---|---|
| whisper-large-v3 | 39.05 |
| whisper-large-v2 | 39.53 |
| whisper-large | 38.11 |
| whisper-medium | 37.39 |
| whisper-small | 31.27 |
| whisper-base | 16.93 |
| whisper-tiny | 7.81 |
| **En2De direct ST systems** | **CoVoST2 BLEU** |
| seamless-m4t-v2-large | 43.12 |
| seamless-m4t-large | 40.55 |
| seamless-m4t-medium | 38.39 |
| s2t-wav2vec2-large-en-de | 26.98 |
| s2t-medium-mustc-multilingual-st | 8.08 |
| s2t-small-mustc-en-de-st | 7.82 |
| s2t-small-covost2-en-de-st | 14.19 |

Table 2: The list of seven direct ST models

# Building SpeechQE Benchmark and Training Corpus

1. **Subsampled** about **80k** segments from the **training** set and **500** from the **dev** and **test** of CoVoST2
2. Run ST models to get hypothesis.
3. Get automatic quality labels from (reference-based) metrics xCOMET-XL for train, adding MetricX-23-XL for test

| CoVoST2 | es2en | en2de | es2en | en2de |
|---|---|---|---|---|
| | *Train* | | *Dev / Test* | |
| ASR | 297k | 305k | | |
| ST | 79k | 290k | | |
| SpeechQE | 546k | 589k | 3.5k | 3.5k |

| Es2En diect ST systems | CoVoST2 BLEU |
|---|---|
| whisper-large-v3 | 39.05 |
| whisper-large-v2 | 39.53 |
| whisper-large | 38.11 |
| whisper-medium | 37.39 |
| whisper-small | 31.27 |
| whisper-base | 16.93 |
| whisper-tiny | 7.81 |
| **En2De direct ST systems** | **CoVoST2 BLEU** |
| seamless-m4t-v2-large | 43.12 |
| seamless-m4t-large | 40.55 |
| seamless-m4t-medium | 38.39 |
| s2t-wav2vec2-large-en-de | 26.98 |
| s2t-medium-mustc-multilingual-st | 8.08 |
| s2t-small-mustc-en-de-st | 7.82 |
| s2t-small-covost2-en-de-st | 14.19 |

Table 2: The list of seven direct ST models

# Building SpeechQE Benchmark and Training Corpus

1. Subsampled about 80k segments from the training set and 500 from the dev and test of CoVoST2
2. **Run ST models** to get hypothesis.
3. Get automatic quality labels from (reference-based) metrics xCOMET-XL for train, adding MetricX-23-XL for test

| CoVoST2 | es2en | en2de | es2en | en2de |
|---|---|---|---|---|
| | *Train* | | *Dev / Test* | |
| ASR | 297k | 305k | | |
| ST | 79k | 290k | | |
| SpeechQE | 546k | 589k | 3.5k | 3.5k |

| Es2En diect ST systems | CoVoST2 BLEU |
|---|---|
| whisper-large-v3 | 39.05 |
| whisper-large-v2 | 39.53 |
| whisper-large | 38.11 |
| whisper-medium | 37.39 |
| whisper-small | 31.27 |
| whisper-base | 16.93 |
| whisper-tiny | 7.81 |
| **En2De direct ST systems** | **CoVoST2 BLEU** |
| seamless-m4t-v2-large | 43.12 |
| seamless-m4t-large | 40.55 |
| seamless-m4t-medium | 38.39 |
| s2t-wav2vec2-large-en-de | 26.98 |
| s2t-medium-mustc-multilingual-st | 8.08 |
| s2t-small-mustc-en-de-st | 7.82 |
| s2t-small-covost2-en-de-st | 14.19 |

Table 2: The list of seven direct ST models

# Building SpeechQE Benchmark and Training Corpus

1. Subsampled about 80k segments from the training set and 500 from the dev and test of CoVoST2
2. Run ST models to get hypothesis.
3. **Get automatic quality labels** from (reference-based) **metrics xCOMET-XL for train**, adding MetricX-23-XL for test

| CoVoST2 | es2en | en2de | es2en | en2de |
|---|---|---|---|---|
| | *Train* | | *Dev / Test* | |
| ASR | 297k | 305k | | |
| ST | 79k | 290k | | |
| SpeechQE | 546k | 589k | 3.5k | 3.5k |

| Metric | | avg corr |
|---|---|---|
| XCOMET-Ensemble | **1** | **0.825** |
| XCOMET-QE-Ensemble* | 2 | 0.808 |
| MetricX-23 | 2 | 0.808 |
| MetricX-23-QE* | 2 | 0.800 |
| docWMT22CometDA | 4 | 0.768 |
| docWMT22CometKiwiDA* | 4 | 0.767 |
| Calibri-COMET22 | 4 | 0.767 |
| Calibri-COMET22-QE* | 4 | 0.755 |

Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent (Freitag et al., WMT 2023)

| metric | | avg corr |
|---|---|---|
| MetaMetrics-MT | **1** | **0.725** |
| MetricX-24-Hybrid | **1** | **0.721** |
| XCOMET | **1** | **0.719** |
| MetricX-24-Hybrid-QE* | 2 | 0.714 |
| gemba_esa* | 2 | 0.711 |
| XCOMET-QE* | 3 | 0.695 |

Are LLMs Breaking MT Metrics? Results of the WMT24 Metrics Shared Task (Freitag et al., WMT 2024)

# Building SpeechQE Benchmark and Training Corpus

1. Subsampled about 80k segments from the training set and 500 from the dev and test of CoVoST2
2. Run ST models to get hypothesis.
3. **Get automatic quality labels** from (reference-based) **metrics xCOMET-XL for train**, adding MetricX-23-XL for test

| CoVoST2 | es2en | en2de | es2en | en2de |
|---|---|---|---|---|
| | *Train* | | *Dev / Test* | |
| ASR | 297k | 305k | | |
| ST | 79k | 290k | | |
| SpeechQE | 546k | 589k | 3.5k | 3.5k |

$$m_{\text{xCOMET}} = \text{xCOMET}(\text{gold } t, h, r)$$

gold transcription text

generated hypothesis text

reference text

# Evaluation

We compute the correlations between SpeechQE scores (q) and :

1. m: Metric score of xCOMET-XL & MetricX-23-XL on En2En/En2De SpeechQE test (3.5k)
2. d: Human Direct Assessment score on IWSLT23-ACL En2De Speech Translation
   - source-based DA ratings of 416 hypotheses from each of the ten ST systems(4.1k)

| $\rho = corr(\mathbf{q}, \mathbf{m} \text{ or } \mathbf{d})$ | CoVoST2 Es2En | | CoVoST2 En2De | | IWSLT23 |
|---|---|---|---|---|---|
| | $\mathrm{m}_{\mathrm{xCOMET}}$ | $\mathrm{m}_{\mathrm{MetricX}}$ | $\mathrm{m}_{\mathrm{xCOMET}}$ | $\mathrm{m}_{\mathrm{MetricX}}$ | En2De $d$ |
| ***Cascaded SpeechQE Systems Correlations*** | | | | | |
| $\mathrm{q}_{cas} = \text{xCOMET-qe}(\text{ASR}(a), h)$ | | | | | |
| $\mathrm{q}_{cas} = \text{MetricX-qe}(\text{ASR}(a), h)$ | | | | | |
| $\mathrm{q}_{cas} = \text{text-BLASER2.0-qe}(\text{ASR}(a), h)$ | | | | | |
| ***End-to-End SpeechQE Systems Correlations*** | | | | | |
| $\mathrm{q}_{e2e} = \text{BLASER2.0-qe}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-Fixed+Adapter}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt-Fixed}(a, h)$ | | | | | |

$$\boxed{\rho = corr(\mathbf{q}, \mathbf{m})}$$

$$\mathrm{m}_{\mathrm{xCOMET}} = \text{xCOMET}(\text{gold } t, h, r)$$

$$\mathrm{m}_{\mathrm{MetricX}} = \text{MetricX}(h, r)$$

$$\boxed{\rho = corr(\mathbf{q}, \mathbf{d})}$$

Human Direct Assessment score



Evaluating the IWSLT2023 Speech Translation Tasks: Human Annotations, Automatic Metrics, and Segmentation (Sperber et al., LREC-COLING 2024)

# Evaluation

We compute the correlations between SpeechQE scores (q) and :

1. m: Metric score of xCOMET-XL & MetricX-23-XL on En2En/En2De SpeechQE test (3.5k)
2. d: Human Direct Assessment score on IWSLT23-ACL En2De Speech Translation
   - source-based DA ratings of 416 hypotheses from each of the ten ST systems(4.1k)

| $\rho = corr(\mathbf{q}, \mathbf{m} \text{ or } \mathbf{d})$ | CoVoST2 Es2En | | CoVoST2 En2De | | IWSLT23 |
|---|---|---|---|---|---|
| | $\mathrm{m}_{\text{xCOMET}}$ | $\mathrm{m}_{\text{MetricX}}$ | $\mathrm{m}_{\text{xCOMET}}$ | $\mathrm{m}_{\text{MetricX}}$ | En2De $d$ |
| ***Cascaded SpeechQE Systems Correlations*** | | | | | |
| $\mathrm{q}_{cas} = \text{xCOMET-qe}(\text{ASR}(a), h)$ | | | | | |
| $\mathrm{q}_{cas} = \text{MetricX-qe}(\text{ASR}(a), h)$ | | | | | |
| $\mathrm{q}_{cas} = \text{text-BLASER2.0-qe}(\text{ASR}(a), h)$ | | | | | |
| ***End-to-End SpeechQE Systems Correlations*** | | | | | |
| $\mathrm{q}_{e2e} = \text{BLASER2.0-qe}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-Fixed+Adapter}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt-Fixed}(a, h)$ | | | | | |

$$\rho = corr(\mathbf{q}, \mathbf{m})$$

$$\mathrm{m}_{\text{xCOMET}} = \text{xCOMET}(\text{gold } t, h, r)$$
$$\mathrm{m}_{\text{MetricX}} = \text{MetricX}(h, r)$$

$$\rho = corr(\mathbf{q}, \mathbf{d})$$

Human Direct Assessment score



Evaluating the IWSLT2023 Speech Translation Tasks: Human Annotations, Automatic Metrics, and Segmentation (Sperber et al., LREC-COLING 2024)

# Evaluation

We compute the correlations between SpeechQE scores (q) and :

1. **m: Metric score of xCOMET-XL & MetricX-23-XL** on En2En/En2De SpeechQE test (3.5k)
2. d: Human Direct Assessment score on IWSLT23-ACL En2De Speech Translation
   - source-based DA ratings of 416 hypotheses from each of the ten ST systems(4.1k)

| $\rho = corr(\mathbf{q}, \mathbf{m} \text{ or } \mathbf{d})$ | CoVoST2 Es2En $\mathrm{m_{xCOMET}}$ | CoVoST2 Es2En $\mathrm{m_{MetricX}}$ | CoVoST2 En2De $\mathrm{m_{xCOMET}}$ | CoVoST2 En2De $\mathrm{m_{MetricX}}$ | IWSLT23 En2De $d$ |
|---|---|---|---|---|---|
| ***Cascaded SpeechQE Systems Correlations*** | | | | | |
| $\mathrm{q}_{cas} = \text{xCOMET-qe}(\text{ASR}(a), h)$ | | | | | |
| $\mathrm{q}_{cas} = \text{MetricX-qe}(\text{ASR}(a), h)$ | | | | | |
| $\mathrm{q}_{cas} = \text{text-BLASER2.0-qe}(\text{ASR}(a), h)$ | | | | | |
| ***End-to-End SpeechQE Systems Correlations*** | | | | | |
| $\mathrm{q}_{e2e} = \text{BLASER2.0-qe}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-Fixed+Adapter}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt-Fixed}(a, h)$ | | | | | |

$$\rho = corr(\mathbf{q}, \mathbf{m})$$

$$\mathrm{m_{xCOMET}} = \text{xCOMET}(\text{gold } t, h, r)$$
$$\mathrm{m_{MetricX}} = \text{MetricX}(h, r)$$

$$\rho = corr(\mathbf{q}, \mathbf{d})$$

Human Direct Assessment score



Evaluating the IWSLT2023 Speech Translation Tasks: Human Annotations, Automatic Metrics, and Segmentation (Sperber et al., LREC-COLING 2024)

# Evaluation

We compute the correlations between SpeechQE scores (q) and :

1. m: Metric score of xCOMET-XL & MetricX-23-XL on En2En/En2De SpeechQE test (3.5k)
2. **d: Human Direct Assessment score** on **IWSLT23-ACL** En2De **Speech Translation**
   - source-based DA ratings of 416 hypotheses from each of the ten ST systems (4.1k)

| $\rho = corr(\mathbf{q}, \mathbf{m} \text{ or } \mathbf{d})$ | CoVoST2 Es2En | | CoVoST2 En2De | | IWSLT23 |
|---|---|---|---|---|---|
| | $\mathrm{m}_{\mathrm{xCOMET}}$ | $\mathrm{m}_{\mathrm{MetricX}}$ | $\mathrm{m}_{\mathrm{xCOMET}}$ | $\mathrm{m}_{\mathrm{MetricX}}$ | En2De $d$ |
| ***Cascaded SpeechQE Systems Correlations*** | | | | | |
| $\mathrm{q}_{cas} = \text{xCOMET-qe}(\text{ASR}(a), h)$ | | | | | |
| $\mathrm{q}_{cas} = \text{MetricX-qe}(\text{ASR}(a), h)$ | | | | | |
| $\mathrm{q}_{cas} = \text{text-BLASER2.0-qe}(\text{ASR}(a), h)$ | | | | | |
| ***End-to-End SpeechQE Systems Correlations*** | | | | | |
| $\mathrm{q}_{e2e} = \text{BLASER2.0-qe}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-Fixed+Adapter}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt}(a, h)$ | | | | | |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt-Fixed}(a, h)$ | | | | | |

$$\rho = corr(\mathbf{q}, \mathbf{m})$$

$$\mathrm{m}_{\mathrm{xCOMET}} = \text{xCOMET}(\text{gold } t, h, r)$$

$$\mathrm{m}_{\mathrm{MetricX}} = \text{MetricX}(h, r)$$

$$\rho = corr(\mathbf{q}, \mathbf{d})$$

Human Direct Assessment score



Evaluating the IWSLT2023 Speech Translation Tasks: Human Annotations, Automatic Metrics, and Segmentation (Sperber et al., LREC-COLING 2024)

# Correlation with Reference-based Metrics

| $\rho = corr(\mathbf{q}, \mathbf{m})$ | $\mathrm{m}_{\mathrm{xCOMET}} = \mathrm{xCOMET}(\text{gold } t, h, r)$ | Es2En | | En2De | |
|---|---|---|---|---|---|
| | $\mathrm{m}_{\mathrm{MetricX}} = \mathrm{MetricX}(h, r)$ | $\mathrm{m}_{\mathrm{xCOMET}}$ | $\mathrm{m}_{\mathrm{MetricX}}$ | $\mathrm{m}_{\mathrm{xCOMET}}$ | $\mathrm{m}_{\mathrm{MetricX}}$ |
| ***Cascaded SpeechQE Systems Correlations*** $\rho_{cas} = corr(\mathbf{q}_{cas}, \mathbf{m})$ | | | | | |
| $\mathrm{q}_{cas} = \mathrm{xCOMET\text{-}qe}(\mathrm{ASR}(a), h)$ | | $\overline{0.892}$ | 0.782 | $\overline{0.910}$ | 0.821 |
| $\mathrm{q}_{cas} = \mathrm{MetricX\text{-}qe}(\mathrm{ASR}(a), h)$ | | 0.803 | 0.803 | 0.854 | $\overline{0.871}$ |
| $\mathrm{q}_{cas} = \mathrm{text\text{-}BLASER2.0\text{-}qe}(\mathrm{ASR}(a), h)$ | | 0.776 | 0.711 | 0.813 | 0.771 |
| ***End-to-End SpeechQE Systems Correlations*** $\rho_{e2e} = corr(\mathbf{q}_{e2e}, \mathbf{m})$ | | | | | |
| $\mathrm{q}_{e2e} = \mathrm{BLASER2.0\text{-}qe}(a, h)$ | | 0.780 | 0.712 | 0.856 | 0.819 |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-Fixed+Adapter}(a, h)$ | | 0.862 | 0.797 | 0.882 | 0.848 |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter}(a, h)$ | | 0.882 | 0.818 | 0.914 | 0.867 |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt}(a, h)$ | | 0.890 | 0.833 | 0.922 | 0.872 |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt-Fixed}(a, h)$ | | **0.895** | **0.834** | **0.925** | **0.873** |



ASR = Whisper-large-v3

# Correlation with Metrics - Cross Comparison

| $\rho = corr(\mathbf{q}, \mathbf{m})$ | $\mathrm{m}_{\text{xCOMET}} = \text{xCOMET}(\text{gold } t, h, r)$<br>$\mathrm{m}_{\text{MetricX}} = \text{MetricX}(h, r)$ | Es2En $\boxed{\mathrm{m}_{\text{xCOMET}}}$ | | Es2En $\boxed{\mathrm{m}_{\text{MetricX}}}$ | En2De $\boxed{\mathrm{m}_{\text{xCOMET}}}$ | | En2De $\boxed{\mathrm{m}_{\text{MetricX}}}$ |
|---|---|---|---|---|---|---|---|
| ***Cascaded SpeechQE Systems Correlations*** $\rho_{cas} = corr(\mathbf{q}_{cas}, \mathbf{m})$ | | | | | | | |
| $\mathrm{q}_{cas} = \boxed{\text{xCOMET-qe}}(\text{ASR}(a), h)$ | | $\overline{0.892}$ | > | 0.782 | $\overline{0.910}$ | | 0.821 |
| $\mathrm{q}_{cas} = \boxed{\text{MetricX-qe}}(\text{ASR}(a), h)$ | | 0.803 | | 0.803 | 0.854 | < | $\overline{0.871}$ |
| $\mathrm{q}_{cas} = \text{text-BLASER2.0-qe}(\text{ASR}(a), h)$ | | 0.776 | | 0.711 | 0.813 | | 0.771 |

A single xCOMET and MetricX model for *metric* and *QE* settings

→ matching QE and metric model could favor the output from the model similar to its own.



ASR = Whisper-large-v3

# Correlation with Metrics - Cross Comparison

| $\rho = corr(\mathbf{q}, \mathbf{m})$ $\quad$ $\mathrm{m}_{\text{xCOMET}} = \text{xCOMET}(\text{gold } t, h, r)$ $\quad$ $\mathrm{m}_{\text{MetricX}} = \text{MetricX}(h, r)$ | Es2En | | En2De | |
|---|---|---|---|---|
| | $\boxed{\mathrm{m}_{\text{xCOMET}}}$ | $\boxed{\mathrm{m}_{\text{MetricX}}}$ | $\boxed{\mathrm{m}_{\text{xCOMET}}}$ | $\boxed{\mathrm{m}_{\text{MetricX}}}$ |
| ***Cascaded SpeechQE Systems Correlations*** $\rho_{cas} = corr(\mathbf{q}_{cas}, \mathbf{m})$ | | | | |
| $\mathrm{q}_{cas} = \boxed{\text{xCOMET-qe}}(\text{ASR}(a), h)$ | $\overline{0.892}$ > | 0.782 | $\overline{0.910}$ | 0.821 |
| $\mathrm{q}_{cas} = \boxed{\text{MetricX-qe}}(\text{ASR}(a), h)$ | 0.803 | 0.803 | 0.854 < | $\overline{0.871}$ |
| $\mathrm{q}_{cas} = \text{text-BLASER2.0-qe}(\text{ASR}(a), h)$ | 0.776 | 0.711 | 0.813 | 0.771 |

**A single xCOMET and MetricX model for *metric* and *QE* settings**

→ matching QE and metric model could favor the output from the model similar to its own.



ASR = Whisper-large-v3

# Correlation with Metrics - Cross Comparison

| $\rho = corr(\mathbf{q}, \mathbf{m})$ | $\mathrm{m}_{\text{xCOMET}} = \text{xCOMET}(\text{gold } t, h, r)$<br>$\mathrm{m}_{\text{MetricX}} = \text{MetricX}(h, r)$ | Es2En<br>$\mathrm{m}_{\text{xCOMET}}$ | | Es2En<br>$\mathrm{m}_{\text{MetricX}}$ | En2De<br>$\mathrm{m}_{\text{xCOMET}}$ | | En2De<br>$\mathrm{m}_{\text{MetricX}}$ |
|---|---|---|---|---|---|---|---|
| ***Cascaded SpeechQE Systems Correlations*** $\rho_{cas} = corr(\mathbf{q}_{cas}, \mathbf{m})$ | | | | | | | |
| $\mathrm{q}_{cas} = \text{xCOMET-qe}(\text{ASR}(a), h)$ | | $\overline{0.892}$ | > | 0.782 | $\overline{0.910}$ | | 0.821 |
| $\mathrm{q}_{cas} = \text{MetricX-qe}(\text{ASR}(a), h)$ | | 0.803 | | 0.803 | 0.854 | < | $\overline{0.871}$ |
| $\mathrm{q}_{cas} = \text{text-BLASER2.0-qe}(\text{ASR}(a), h)$ | | 0.776 | | 0.711 | 0.813 | | 0.771 |

A single xCOMET and MetricX model for *metric* and *QE* settings

→ **matching QE and metric model could favor the output from the model similar to its own.**

ASR = Whisper-large-v3 LLM Evaluators Recognize and Favor Their Own Generations (Panickssery et al., NeurIPS 2024)

# Correlation with Reference-based Metrics

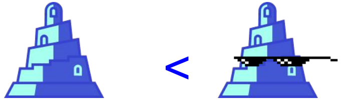
| $\rho = corr(\mathbf{q}, \mathbf{m})$ $\quad \mathrm{m}_{\text{xCOMET}} = \text{xCOMET}(\text{gold } t, h, r)$ | Es2En | | En2De | |
|---|---|---|---|---|
| $\mathrm{m}_{\text{MetricX}} = \text{MetricX}(h, r)$ | $\mathrm{m}_{\text{xCOMET}}$ | $\mathrm{m}_{\text{MetricX}}$ | $\mathrm{m}_{\text{xCOMET}}$ | $\mathrm{m}_{\text{MetricX}}$ |
| ***Cascaded SpeechQE Systems Correlations*** $\rho_{cas} = corr(\mathbf{q}_{cas}, \mathbf{m})$ | | | | |
| $\mathrm{q}_{cas} = \text{xCOMET-qe}(\text{ASR}(a), h)$ | $\overline{0.892}$ | 0.782 | $\overline{0.910}$ | 0.821 |
| $\mathrm{q}_{cas} = \text{MetricX-qe}(\text{ASR}(a), h)$ | 0.803 | 0.803 | 0.854 | $\overline{0.871}$ |
| $\mathrm{q}_{cas} = \text{text-BLASER2.0-qe}(\text{ASR}(a), h)$ | 0.776 | 0.711 | 0.813 | 0.771 |
| ***End-to-End SpeechQE Systems Correlations*** $\rho_{e2e} = corr(\mathbf{q}_{e2e}, \mathbf{m})$ | | | | |
| $\mathrm{q}_{e2e} = \text{BLASER2.0-qe}(a, h)$ | 0.780 | 0.712 | 0.856 | 0.819 |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-Fixed+Adapter}(a, h)$ | 0.862 | 0.797 | 0.882 | 0.848 |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter}(a, h)$ | 0.882 | 0.818 | 0.914 | 0.867 |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt}(a, h)$ | 0.890 | 0.833 | 0.922 | 0.872 |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt-Fixed}(a, h)$ | **0.895** | **0.834** | **0.925** | **0.873** |



ASR = Whisper-large-v3

# Correlation with Reference-based Metrics

| $\rho = corr(\mathbf{q}, \mathbf{m})$ | $\mathrm{m}_{\text{xCOMET}} = \text{xCOMET}(\text{gold } t, h, r)$ <br> $\mathrm{m}_{\text{MetricX}} = \text{MetricX}(h, r)$ | Es2En | | En2De | |
|---|---|---|---|---|---|
| | | $\mathrm{m}_{\text{xCOMET}}$ | $\mathrm{m}_{\text{MetricX}}$ | $\mathrm{m}_{\text{xCOMET}}$ | $\mathrm{m}_{\text{MetricX}}$ |
| ***Cascaded SpeechQE Systems Correlations*** $\rho_{cas} = corr(\mathbf{q}_{cas}, \mathbf{m})$ | | | | | |
| $\mathrm{q}_{cas} = \text{xCOMET-qe}(\text{ASR}(a), h)$ | | 0.892 | 0.782 | 0.910 | 0.821 |
| $\mathrm{q}_{cas} = \text{MetricX-qe}(\text{ASR}(a), h)$ | | 0.803 | 0.803 | 0.854 | 0.871 |
| $\mathrm{q}_{cas} = \text{text-BLASER2.0-qe}(\text{ASR}(a), h)$ | | 0.776 | 0.711 | 0.813 | 0.771 |
| ***End-to-End SpeechQE Systems Correlations*** $\rho_{e2e} = corr(\mathbf{q}_{e2e}, \mathbf{m})$ | | | | | |
| $\mathrm{q}_{e2e} = \text{BLASER2.0-qe}(a, h)$ | | 0.780 | 0.712 | 0.856 | 0.819 |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-Fixed+Adapter}(a, h)$ | | 0.862 | 0.797 | 0.882 | 0.848 |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter}(a, h)$ | | 0.882 | 0.818 | 0.914 | 0.867 |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt}(a, h)$ | | 0.890 | 0.833 | 0.922 | 0.872 |
| $\mathrm{q}_{e2e} = \textit{TowerInstruct-LoRA+Adapter-pt-Fixed}(a, h)$ | | **0.895** | **0.834** | **0.925** | **0.873** |






ASR = Whisper-large-v3

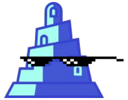
# Correlation with Reference-based Metrics

| $\rho = corr(\mathbf{q}, \mathbf{m})$ $\quad$ $\mathrm{m}_{\text{xCOMET}} = \text{xCOMET}(\text{gold } t, h, r)$, $\mathrm{m}_{\text{MetricX}} = \text{MetricX}(h, r)$ | Es2En $\mathrm{m}_{\text{xCOMET}}$ | Es2En $\mathrm{m}_{\text{MetricX}}$ | En2De $\mathrm{m}_{\text{xCOMET}}$ | En2De $\mathrm{m}_{\text{MetricX}}$ |
|---|---|---|---|---|
| ***Cascaded SpeechQE Systems Correlations*** $\rho_{cas} = corr(\mathbf{q}_{cas}, \mathbf{m})$ | | | | |
| $\mathrm{q}_{cas}$ = xCOMET-qe(ASR($a$), $h$) | $\overline{0.892}$ | 0.782 | $\overline{0.910}$ | 0.821 |
| $\mathrm{q}_{cas}$ = MetricX-qe(ASR($a$), $h$) | 0.803 | 0.803 | 0.854 | $\overline{0.871}$ |
| $\mathrm{q}_{cas}$ = text-BLASER2.0-qe(ASR($a$), $h$) | 0.776 | 0.711 | 0.813 | 0.771 |
| ***End-to-End SpeechQE Systems Correlations*** $\rho_{e2e} = corr(\mathbf{q}_{e2e}, \mathbf{m})$ | | | | |
| $\mathrm{q}_{e2e}$ = BLASER2.0-qe($a$, $h$) | 0.780 | 0.712 | 0.856 | 0.819 |
| $\mathrm{q}_{e2e}$ = *TowerInstruct-Fixed+Adapter*($a$, $h$) | 0.862 | 0.797 | 0.882 | 0.848 |
| $\mathrm{q}_{e2e}$ = *TowerInstruct-LoRA+Adapter*($a$, $h$) | 0.882 | 0.818 | 0.914 | 0.867 |
| $\mathrm{q}_{e2e}$ = *TowerInstruct-LoRA+Adapter-pt*($a$, $h$) | 0.890 | 0.833 | 0.922 | 0.872 |
| $\mathrm{q}_{e2e}$ = *TowerInstruct-LoRA+Adapter-pt-Fixed*($a$, $h$) | **0.895** | **0.834** | **0.925** | **0.873** |

**Weight updates at least partially are necessary when a text-LLM is not fine-tuned with QE tasks**

No QE rating capability

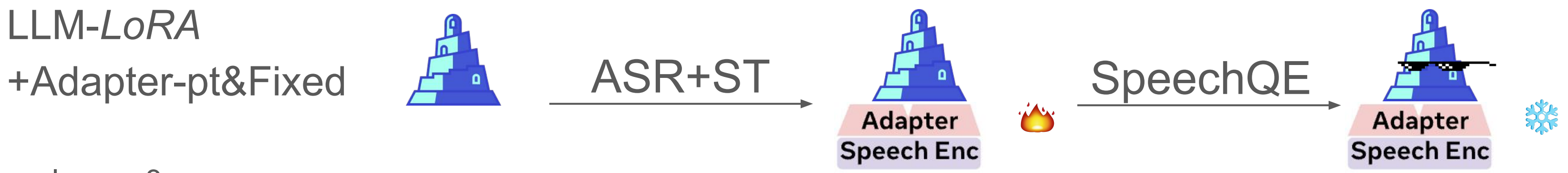
<



ASR = Whisper-large-v3

# Correlation with Reference-based Metrics

| $\rho = corr(\mathbf{q}, \mathbf{m})$ | $\mathrm{m_{xCOMET}} = \mathrm{xCOMET}(\text{gold } t, h, r)$ <br> $\mathrm{m_{MetricX}} = \mathrm{MetricX}(h, r)$ | Es2En $\mathrm{m_{xCOMET}}$ | Es2En $\mathrm{m_{MetricX}}$ | En2De $\mathrm{m_{xCOMET}}$ | En2De $\mathrm{m_{MetricX}}$ |
|---|---|---|---|---|---|
| ***Cascaded SpeechQE Systems Correlations*** $\rho_{cas} = corr(\mathbf{q}_{cas}, \mathbf{m})$ | | | | | |
| $\mathrm{q}_{cas}$ | = xCOMET-qe(ASR($a$), $h$) | 0.892 | 0.782 | 0.910 | 0.821 |
| $\mathrm{q}_{cas}$ | = MetricX-qe(ASR($a$), $h$) | 0.803 | 0.803 | 0.854 | 0.871 |
| $\mathrm{q}_{cas}$ | = text-BLASER2.0-qe(ASR($a$), $h$) | 0.776 | 0.711 | 0.813 | 0.771 |
| ***End-to-End SpeechQE Systems Correlations*** $\rho_{e2e} = corr(\mathbf{q}_{e2e}, \mathbf{m})$ | | | | | |
| $\mathrm{q}_{e2e}$ | = BLASER2.0-qe($a$, $h$) | 0.780 | 0.712 | 0.856 | 0.819 |
| $\mathrm{q}_{e2e}$ | = *TowerInstruct-Fixed+Adapter*($a$, $h$) | 0.862 | 0.797 | 0.882 | 0.848 |
| $\mathrm{q}_{e2e}$ | = *TowerInstruct-LoRA+Adapter*($a$, $h$) | 0.882 | 0.818 | 0.914 | 0.867 |
| $\mathrm{q}_{e2e}$ | = *TowerInstruct-LoRA+Adapter-pt*($a$, $h$) | 0.890 | 0.833 | 0.922 | 0.872 |
| $\mathrm{q}_{e2e}$ | = *TowerInstruct-LoRA+Adapter-pt-Fixed*($a$, $h$) | **0.895** | **0.834** | **0.925** | **0.873** |



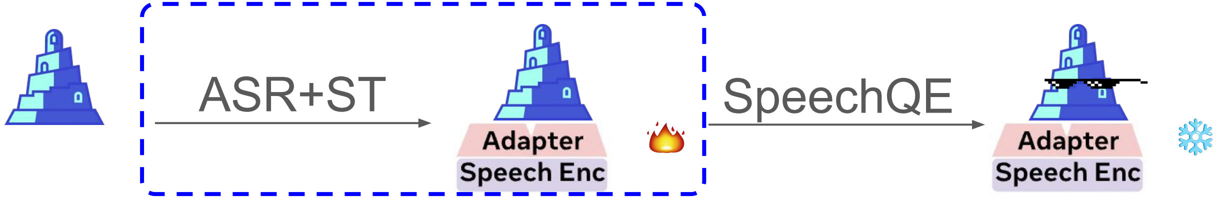

ASR = Whisper-large-v3

# Correlation with Reference-based Metrics

| $\rho = corr(\mathbf{q}, \mathbf{m})$ $\quad$ $\mathrm{m_{xCOMET}} = \mathrm{xCOMET}(\text{gold } t, h, r)$ $\quad$ $\mathrm{m_{MetricX}} = \mathrm{MetricX}(h, r)$ | Es2En $\mathrm{m_{xCOMET}}$ | Es2En $\mathrm{m_{MetricX}}$ | En2De $\mathrm{m_{xCOMET}}$ | En2De $\mathrm{m_{MetricX}}$ |
|---|---|---|---|---|
| ***Cascaded SpeechQE Systems Correlations*** $\rho_{cas} = corr(\mathbf{q}_{cas}, \mathbf{m})$ | | | | |
| $\mathrm{q}_{cas} = \mathrm{xCOMET\text{-}qe}(\mathrm{ASR}(a), h)$ | 0.892 | 0.782 | 0.910 | 0.821 |
| $\mathrm{q}_{cas} = \mathrm{MetricX\text{-}qe}(\mathrm{ASR}(a), h)$ | 0.803 | 0.803 | 0.854 | 0.871 |
| $\mathrm{q}_{cas} = \mathrm{text\text{-}BLASER2.0\text{-}qe}(\mathrm{ASR}(a), h)$ | 0.776 | 0.711 | 0.813 | 0.771 |
| ***End-to-End SpeechQE Systems Correlations*** $\rho_{e2e} = corr(\mathbf{q}_{e2e}, \mathbf{m})$ | | | | |
| $\mathrm{q}_{e2e} = \mathrm{BLASER2.0\text{-}qe}(a, h)$ | 0.780 | 0.712 | 0.856 | 0.819 |
| $\mathrm{q}_{e2e} = \mathit{TowerInstruct\text{-}Fixed{+}Adapter}(a, h)$ | 0.862 | 0.797 | 0.882 | 0.848 |
| $\mathrm{q}_{e2e} = \mathit{TowerInstruct\text{-}LoRA{+}Adapter}(a, h)$ | 0.882 | 0.818 | 0.914 | 0.867 |
| $\mathrm{q}_{e2e} = \mathit{TowerInstruct\text{-}LoRA{+}Adapter\text{-}pt}(a, h)$ | 0.890 | 0.833 | 0.922 | 0.872 |
| $\mathrm{q}_{e2e} = \mathit{TowerInstruct\text{-}LoRA{+}Adapter\text{-}pt\text{-}Fixed}(a, h)$ | **0.895** | **0.834** | **0.925** | **0.873** |

**a separate training phase for mapping speech-to-text perception is critical.**

LLM-*LoRA*
+Adapter-pt&Fixed






ASR = Whisper-large-v3

# Correlation with Human Score

| IWSLT23-ACL En2De Test set $\rho = corr(\mathbf{q}, \mathbf{d})$ | Human DA score $d$ |
|---|---|
| ***Cascaded SpeechQE*** *and Human DA* $\rho = corr(\mathbf{q}_{cas}, \mathbf{d})$ | |
| $q$ = xCOMET-qe(ASR($a$), $h$) | 0.485 |
| $q$ = MetricX-qe(ASR($a$), $h$) | 0.495 |
| $q$ = wmt23-cometkiwi-da-xl(ASR($a$), $h$) | **0.503** |
| $q$ = wmt22-cometkiwi-da(ASR($a$), $h$) | 0.486 |
| $q$ = text-BLASER2.0-qe(ASR($a$), $h$) | 0.428 |
| ***E2E SpeechQE*** *& Human DA correlation* $\rho = corr(\mathbf{q}_{e2e}, \mathbf{d})$ | |
| $q$ = BLASER2.0-qe($a$, $h$) | 0.420 |
| $q$ = *TowerInst-LoRA+Adapter-pt*($a$, $h$) | 0.492 |
| $q$ = *TowerInst-LoRA+Adapter-pt-Fixed*($a$, $h$) | **0.509** |

ASR = Azure API, output provided by

Evaluating Multilingual Speech Translation under Realistic Conditions with Resegmentation and Terminology (Salesky et al., IWSLT 2023)

# Correlation with Human Score

| IWSLT23-ACL En2De Test set $\rho = corr(\mathbf{q}, \mathbf{d})$ | Human DA score $d$ |
|---|---|
| ***Cascaded SpeechQE*** *and Human DA* $\rho = corr(\mathbf{q}_{cas}, \mathbf{d})$ | |
| $q$ = xCOMET-qe(ASR$(a)$, $h$) | 0.485 |
| $q$ = MetricX-qe(ASR$(a)$, $h$) | 0.495 |
| $q$ = wmt23-cometkiwi-da-xl(ASR$(a)$, $h$) | **0.503** |
| $q$ = wmt22-cometkiwi-da(ASR$(a)$, $h$) | 0.486 |
| $q$ = text-BLASER2.0-qe(ASR$(a)$, $h$) | 0.428 |
| ***E2E SpeechQE*** *& Human DA correlation* $\rho = corr(\mathbf{q}_{e2e}, \mathbf{d})$ | |
| $q$ = BLASER2.0-qe$(a, h)$ | 0.420 |
| $q$ = *TowerInst-LoRA+Adapter-pt*$(a, h)$ | 0.492 |
| $q$ = *TowerInst-LoRA+Adapter-pt-Fixed*$(a, h)$ | **0.509** |

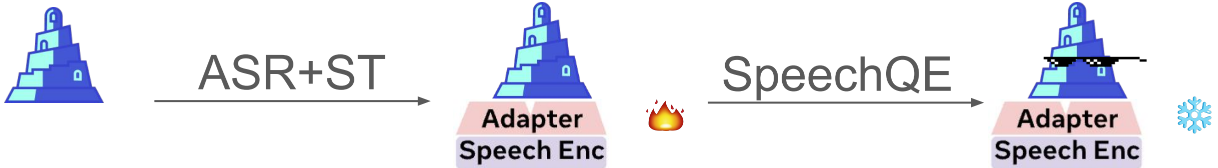
ASR = Azure API, output provided by

Evaluating Multilingual Speech Translation under Realistic Conditions with Resegmentation and Terminology (Salesky et al., IWSLT 2023)

# Correlation with Human Score

| IWSLT23-ACL En2De Test set<br>$\rho = corr(\mathbf{q}, \mathbf{d})$ | Human DA<br>score $d$ |
|---|---|
| ***Cascaded SpeechQE*** *and Human DA* $\rho = corr(\mathbf{q}_{cas}, \mathbf{d})$ | |
| $q$ = xCOMET-qe(ASR($a$), $h$) | 0.485 |
| $q$ = MetricX-qe(ASR($a$), $h$) | 0.495 |
| $q$ = wmt23-cometkiwi-da-xl(ASR($a$), $h$) | **0.503** |
| $q$ = wmt22-cometkiwi-da(ASR($a$), $h$) | 0.486 |
| $q$ = text-BLASER2.0-qe(ASR($a$), $h$) | 0.428 |
| ***E2E SpeechQE*** *& Human DA correlation* $\rho = corr(\mathbf{q}_{e2e}, \mathbf{d})$ | |
| $q$ = BLASER2.0-qe($a$, $h$) | 0.420 |
| $q$ = *TowerInst-LoRA+Adapter-pt*($a$, $h$) | 0.492 |
| $q$ = *TowerInst-LoRA+Adapter-pt-Fixed*($a$, $h$) | **0.509** |

LLM-*LoRA*
+Adapter-pt&Fixed



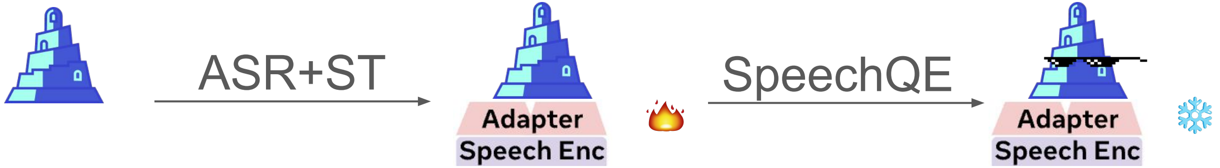

ASR = Azure API, output provided by
Evaluating Multilingual Speech Translation under Realistic Conditions with Resegmentation and Terminology (Salesky et al., IWSLT 2023)

# Correlation with Human Score

**The best-practice E2E system is more effective in aligning with human judgments.**

| IWSLT23-ACL En2De Test set<br>$\rho = corr(\mathbf{q}, \mathbf{d})$ | Human DA<br>score $d$ |
|---|---|
| ***Cascaded SpeechQE*** *and Human DA* $\rho = corr(\mathbf{q}_{cas}, \mathbf{d})$ | |
| $q$ = xCOMET-qe(ASR($a$), $h$) | 0.485 |
| $q$ = MetricX-qe(ASR($a$), $h$) | 0.495 |
| $q$ = wmt23-cometkiwi-da-xl(ASR($a$), $h$) | **0.503** |
| $q$ = wmt22-cometkiwi-da(ASR($a$), $h$) | 0.486 |
| $q$ = text-BLASER2.0-qe(ASR($a$), $h$) | 0.428 |
| ***E2E SpeechQE*** *& Human DA correlation* $\rho = corr(\mathbf{q}_{e2e}, \mathbf{d})$ | |
| $q$ = BLASER2.0-qe($a$, $h$) | 0.420 |
| $q$ = *TowerInst-LoRA+Adapter-pt*($a$, $h$) | 0.492 |
| $q$ = *TowerInst-LoRA+Adapter-pt-Fixed*($a$, $h$) | **0.509** |

LLM-*LoRA*
+Adapter-pt&Fixed




ASR = Azure API, output provided by
Evaluating Multilingual Speech Translation under Realistic Conditions with Resegmentation and Terminology (Salesky et al., IWSLT 2023)

# Cascaded Model Size and Architecture

Is the dominance of E2E over cascaded models due to the E2E parameter size or architecture rather than its end-to-end nature?

| $\rho = corr(\mathbf{q}, \mathbf{m} \text{ or } \mathbf{d})$ | CoVoST2 Es2En Test | | | | IWSLT23 |
|---|---|---|---|---|---|
| | $m_{\text{xCOMET-XL}}$ | $m_{\text{xCOMET-XXL}}$ | $m_{\text{MetricX-XL}}$ | $m_{\text{MetricX-XXL}}$ | En2De $d$ |
| ***Cascaded Model with XXL Size vs E2E speech-LLM*** | | | | | |
| $q_{cas}$ = ASR (1.5B) → xCOMET-XL-qe (3.5B) | 0.892 | 0.800 | 0.782 | 0.788 | 0.485 |
| $q_{cas}$ = ASR (1.5B) → xCOMET-XXL-qe (10.7B) | 0.787 | **0.873** | 0.708 | 0.734 | 0.486 |
| $q_{cas}$ = ASR (1.5B) → MetricX-XL-qe (3.7B) | 0.803 | 0.758 | 0.803 | 0.766 | 0.495 |
| $q_{cas}$ = ASR (1.5B) → MetricX-XXL-qe (13B) | 0.700 | 0.677 | 0.652 | 0.694 | 0.502 |
| ***Cascaded text-LLM vs E2E speech-LLM*** | | | | | |
| $q_{cas}$ = ASR (1.5B) → *text-TowerInstruct-LoRA* (7B) | 0.852 | 0.816 | 0.780 | 0.785 | _ |
| $q_{e2e}$ = *TowerInstruct-LoRA+Adapter-pt-Fixed* (7.5B) | **0.895** | 0.827 | **0.834** | **0.834** | **0.509** |



ASR = Whisper-large-v3(1.5B)

# Cascaded Model with Larger Size

| $\rho = corr(\mathbf{q}, \mathbf{m} \text{ or } \mathbf{d})$ | CoVoST2 Es2En Test | | | | IWSLT23 |
|---|---|---|---|---|---|
| | $\mathrm{m}_{\text{xCOMET-XL}}$ | $\mathrm{m}_{\text{xCOMET-XXL}}$ | $\mathrm{m}_{\text{MetricX-XL}}$ | $\mathrm{m}_{\text{MetricX-XXL}}$ | En2De $d$ |
| ***Cascaded Model with XXL Size vs E2E speech-LLM*** | | | | | |
| $\mathrm{q}_{cas}$ = ASR (1.5B) → xCOMET-XL-qe (3.5B) | 0.892 | 0.800 | 0.782 | 0.788 | 0.485 |
| $\mathrm{q}_{cas}$ = ASR (1.5B) → xCOMET-XXL-qe (10.7B) | 0.787 | **0.873** | 0.708 | 0.734 | 0.486 |
| $\mathrm{q}_{cas}$ = ASR (1.5B) → MetricX-XL-qe (3.7B) | 0.803 | 0.758 | 0.803 | 0.766 | 0.495 |
| $\mathrm{q}_{cas}$ = ASR (1.5B) → MetricX-XXL-qe (13B) | 0.700 | 0.677 | 0.652 | 0.694 | 0.502 |
| ***Cascaded text-LLM vs E2E speech-LLM*** | | | | | |
| $\mathrm{q}_{cas}$ = ASR (1.5B) → *text-TowerInstruct-LoRA* (7B) | 0.852 | 0.816 | 0.780 | 0.785 | – |
| $\mathrm{q}_{e2e}$ = *TowerInstruct-LoRA+Adapter-pt-Fixed* (7.5B) | **0.895** | 0.827 | **0.834** | **0.834** | **0.509** |



ASR = Whisper-large-v3(1.5B)

# Cascaded Model with Similar Architecture

| $\rho = corr(\mathbf{q}, \mathbf{m} \text{ or } \mathbf{d})$ | CoVoST2 Es2En Test | | | | IWSLT23 |
|---|---|---|---|---|---|
| | $m_{\text{xCOMET-XL}}$ | $m_{\text{xCOMET-XXL}}$ | $m_{\text{MetricX-XL}}$ | $m_{\text{MetricX-XXL}}$ | En2De $d$ |
| ***Cascaded Model with XXL Size vs E2E speech-LLM*** | | | | | |
| $q_{cas}$ = ASR (1.5B) → xCOMET-XL-qe (3.5B) | 0.892 | 0.800 | 0.782 | 0.788 | 0.485 |
| $q_{cas}$ = ASR (1.5B) → xCOMET-XXL-qe (10.7B) | 0.787 | **0.873** | 0.708 | 0.734 | 0.486 |
| $q_{cas}$ = ASR (1.5B) → MetricX-XL-qe (3.7B) | 0.803 | 0.758 | 0.803 | 0.766 | 0.495 |
| $q_{cas}$ = ASR (1.5B) → MetricX-XXL-qe (13B) | 0.700 | 0.677 | 0.652 | 0.694 | 0.502 |
| ***Cascaded text-LLM vs E2E speech-LLM*** | | | | | |
| $q_{cas}$ = ASR (1.5B) → *text-TowerInstruct-LoRA* (7B) | 0.852 | 0.816 | 0.780 | 0.785 | _ |
| $q_{e2e}$ = *TowerInstruct-LoRA+Adapter-pt-Fixed* (7.5B) | **0.895** | 0.827 | **0.834** | **0.834** | **0.509** |



ASR = Whisper-large-v3(1.5B)

# Cascaded Model with Similar Architecture

**The improvements are coming from the E2E nature of the approach rather than the LLM-based solution or larger parameters.**

| $\rho = corr(\mathbf{q}, \mathbf{m} \text{ or } \mathbf{d})$ | CoVoST2 Es2En Test | | | | IWSLT23 |
|---|---|---|---|---|---|
| | $m_{\text{xCOMET-XL}}$ | $m_{\text{xCOMET-XXL}}$ | $m_{\text{MetricX-XL}}$ | $m_{\text{MetricX-XXL}}$ | En2De $d$ |
| ***Cascaded Model with XXL Size vs E2E speech-LLM*** | | | | | |
| $q_{cas}$ = ASR (1.5B) → xCOMET-XL-qe (3.5B) | 0.892 | 0.800 | 0.782 | 0.788 | 0.485 |
| $q_{cas}$ = ASR (1.5B) → xCOMET-XXL-qe (10.7B) | 0.787 | **0.873** | 0.708 | 0.734 | 0.486 |
| $q_{cas}$ = ASR (1.5B) → MetricX-XL-qe (3.7B) | 0.803 | 0.758 | 0.803 | 0.766 | 0.495 |
| $q_{cas}$ = ASR (1.5B) → MetricX-XXL-qe (13B) | 0.700 | 0.677 | 0.652 | 0.694 | 0.502 |
| ***Cascaded text-LLM vs E2E speech-LLM*** | | | | | |
| $q_{cas}$ = ASR (1.5B) → *text-TowerInstruct-LoRA* (7B) | 0.852 | 0.816 | 0.780 | 0.785 | _ |
| $q_{e2e}$ = *TowerInstruct-LoRA+Adapter-pt-Fixed* (7.5B) | **0.895** | 0.827 | **0.834** | **0.834** | **0.509** |

→ E2E system is better suited for SpeechQE task than the cascaded



ASR = Whisper-large-v3(1.5B)

# Cascaded Model with Similar Architecture

The improvements are coming from the E2E nature of the approach rather than the LLM-based solution or larger parameters.

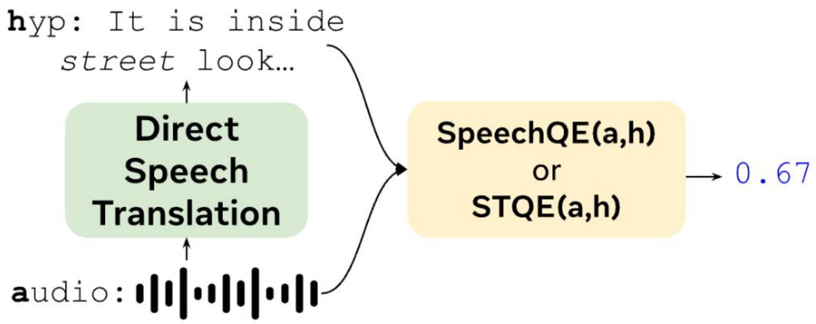
| $\rho = corr(\mathbf{q}, \mathbf{m} \text{ or } \mathbf{d})$ | CoVoST2 Es2En Test | | | | IWSLT23 |
|---|---|---|---|---|---|
| | $\mathrm{m}_{\text{xCOMET-XL}}$ | $\mathrm{m}_{\text{xCOMET-XXL}}$ | $\mathrm{m}_{\text{MetricX-XL}}$ | $\mathrm{m}_{\text{MetricX-XXL}}$ | En2De $d$ |
| ***Cascaded Model with XXL Size vs E2E speech-LLM*** | | | | | |
| $\mathrm{q}_{cas}$ = ASR (1.5B) → xCOMET-XL-qe (3.5B) | 0.892 | 0.800 | 0.782 | 0.788 | 0.485 |
| $\mathrm{q}_{cas}$ = ASR (1.5B) → xCOMET-XXL-qe (10.7B) | 0.787 | **0.873** | 0.708 | 0.734 | 0.486 |
| $\mathrm{q}_{cas}$ = ASR (1.5B) → MetricX-XL-qe (3.7B) | 0.803 | 0.758 | 0.803 | 0.766 | 0.495 |
| $\mathrm{q}_{cas}$ = ASR (1.5B) → MetricX-XXL-qe (13B) | 0.700 | 0.677 | 0.652 | 0.694 | 0.502 |
| ***Cascaded text-LLM vs E2E speech-LLM*** | | | | | |
| $\mathrm{q}_{cas}$ = ASR (1.5B) → *text-TowerInstruct-LoRA* (7B) | 0.852 | 0.816 | 0.780 | 0.785 | – |
| $\mathrm{q}_{e2e}$ = *TowerInstruct-LoRA+Adapter-pt-Fixed* (7.5B) | **0.895** | 0.827 | **0.834** | **0.834** | **0.509** |

→ **E2E system is better suited for SpeechQE task than the cascaded**



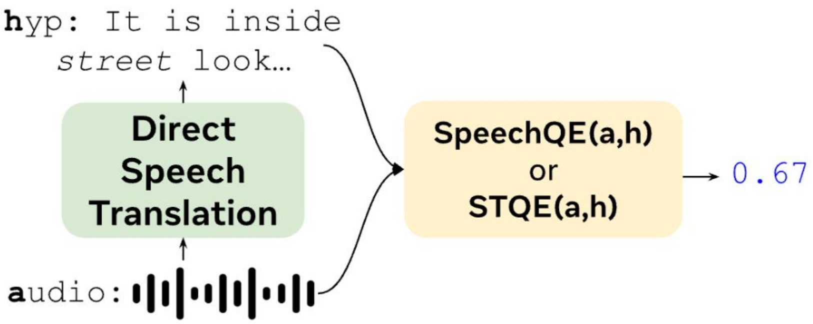
ASR = Whisper-large-v3(1.5B)

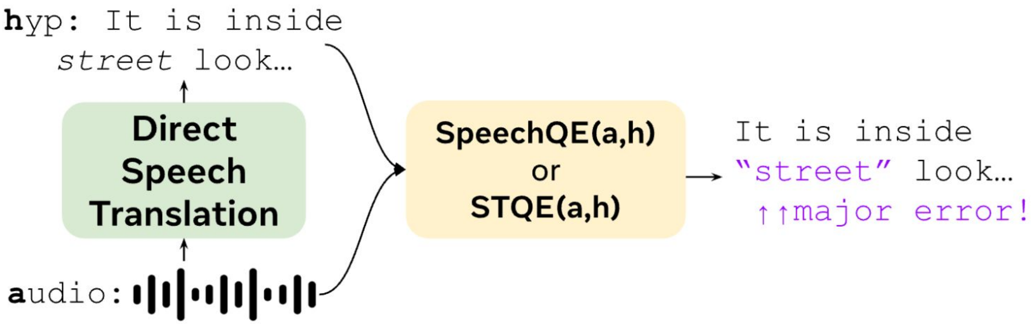
SpeechQE
(numerical rating)



# Error Span Detection for ST

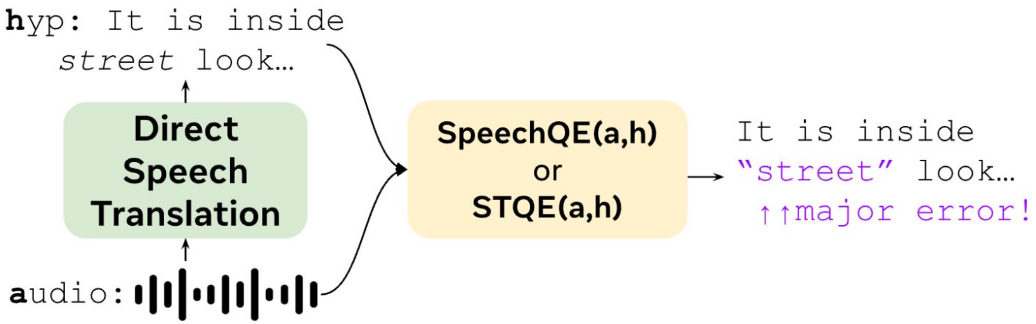
SpeechQE
(numerical rating)



# SpeechESD: Error Span Detection for ST

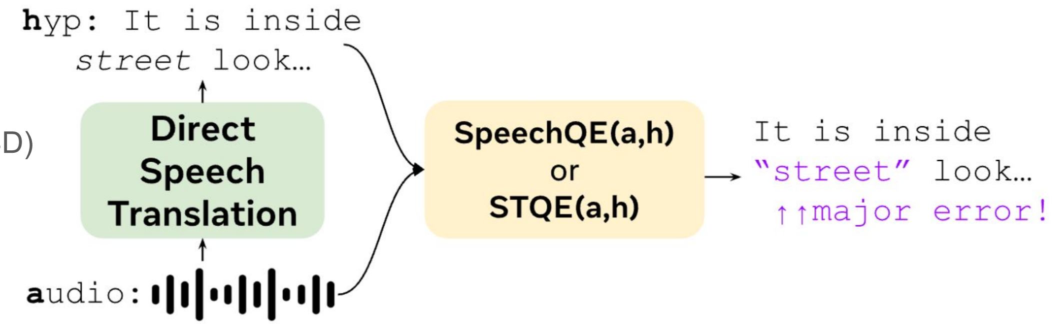
Speech
Error Span
Detection
(SpeechESD)



# Zero-Shot Error Span Detection for ST

No QE rating capability
But **has error span detection ability**

Error Span Detection (ESD)

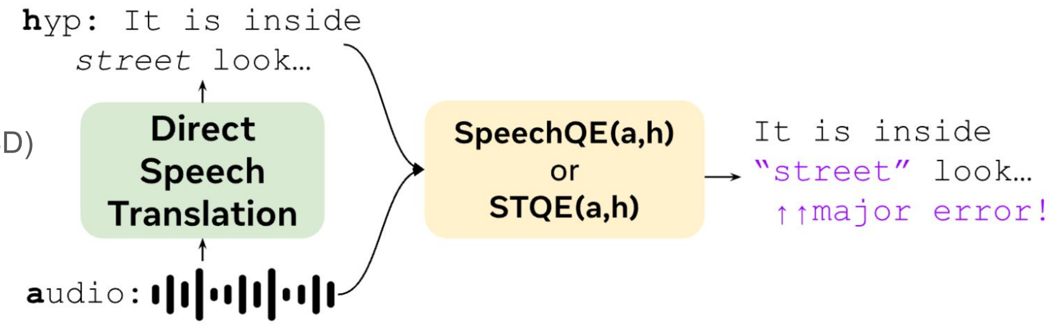


# Zero-Shot Error Span Detection for ST

How effectively the method of injecting speech modality generalizes the capability of text-LLM to speech LLM without explicitly training the target speech task?

No QE rating capability
But **has error span detection ability**

Error Span Detection (ESD)



# SpeechESD Experiment Setting

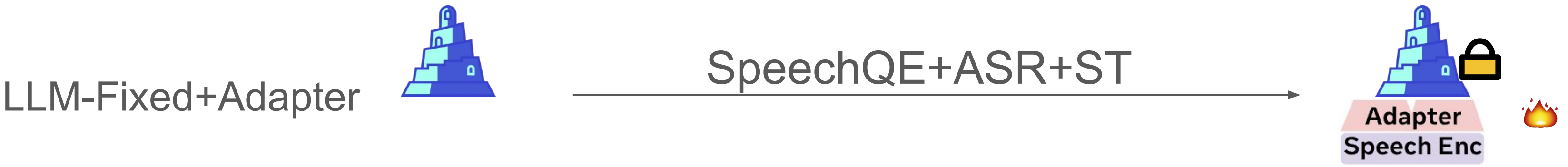
**Reference-based error span labels from Error span output of the xCOMET metric function**

- xCOMET outputs Error Spans.

Compare:

1. the E2E SpeechESD based on TowerInstruct in zero-shot way
2. cascaded system where TowerInstruct is text-ESD model

Fixed-LLM for E2E: lose ESD capabilities after fine-tuned with non-ESD tasks.



# SpeechESD Experiment Setting

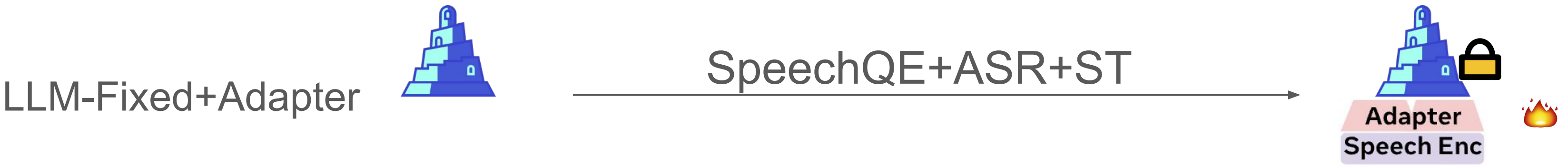
Reference-based error span labels from Error span output of the xCOMET metric function

- xCOMET has ESD capabilities.

**Compare:**

1. **E2E SpeechESD** based on **TowerInstruct** in **zero-shot way**
2. **cascaded system** where **TowerInstruct** is **text-ESD** model

Fixed-LLM for E2E: lose ESD capabilities after fine-tuned with non-ESD tasks.



# SpeechESD Experiment Setting

Reference-based error span labels from Error span output of the xCOMET metric function

- xCOMET has ESD capabilities.

Compare:

1. the E2E SpeechESD based on TowerInstruct in zero-shot way
2. cascaded system where TowerInstruct is text-ESD model

**Fixed-LLM for E2E: lose ESD capabilities after fine-tuned with non-ESD tasks.**

LLM-Fixed+Adapter

# Zero-Shot SpeechESD

| **ESD for ST** | Precision | Recall | F1 Score |
|---|---|---|---|
| ***Cascaded Systems*** | | | |
| txt-ESD(gold $t, h$) | 0.438 | 0.591 | 0.503 |
| txt-ESD(whisper-large-v2$(a), h$) | 0.434 | 0.550 | 0.485 |
| txt-ESD(whisper-medium$(a), h$) | 0.429 | 0.540 | 0.478 |
| txt-ESD(whisper-small$(a), h$) | 0.413 | 0.535 | 0.466 |
| txt-ESD(whisper-base$(a), h$) | 0.385 | 0.550 | 0.453 |
| ***End-to-End Systems*** | | | |
| *TowerInst-Fixed+Adt*$(a, h)$ | 0.411 | 0.542 | 0.467 |

(Example analysis is in the paper)

# Zero-Shot SpeechESD

Cascaded systems remain the preferred choice for achieving the highest performance when we do not have speech training data for the target task.

| **ESD for ST** | Precision | Recall | F1 Score |
|---|---|---|---|
| ***Cascaded Systems*** | | | |
| txt-ESD(gold $t$, $h$) | 0.438 | 0.591 | 0.503 |
| txt-ESD(whisper-large-v2($a$), $h$) | 0.434 | 0.550 | 0.485 |
| txt-ESD(whisper-medium($a$), $h$) | 0.429 | $\overline{0.540}$ | 0.478 |
| txt-ESD(whisper-small($a$), $h$) | $\underline{0.413}$ | 0.535 | $\overline{0.466}$ |
| txt-ESD(whisper-base($a$), $h$) | 0.385 | 0.550 | 0.453 |
| ***End-to-End Systems*** | | | |
| *TowerInst-Fixed+Adt*($a$, $h$) | 0.411 | 0.542 | 0.467 |



(Example analysis is in the paper)

# Zero-Shot SpeechESD

Cascaded systems remain the preferred choice for achieving the highest performance when we do not have speech training data for the target task.

Still, the E2E model performs decently in zero-shot

| **ESD for ST** | Precision | Recall | F1 Score |
|---|---|---|---|
| ***Cascaded Systems*** | | | |
| txt-ESD(gold $t, h$) | 0.438 | 0.591 | 0.503 |
| txt-ESD(whisper-large-v2($a$), $h$) | 0.434 | 0.550 | 0.485 |
| txt-ESD(whisper-medium($a$), $h$) | 0.429 | $\overline{0.540}$ | 0.478 |
| txt-ESD(whisper-small($a$), $h$) | $\underline{0.413}$ | 0.535 | $\overline{0.466}$ |
| txt-ESD(whisper-base($a$), $h$) | 0.385 | 0.550 | 0.453 |
| ***End-to-End Systems*** | | | |
| *TowerInst-Fixed+Adt*($a, h$) | 0.411 | 0.542 | 0.467 |



(Example analysis is in the paper)

# Zero-Shot SpeechESD

Cascaded systems remain the preferred choice for achieving the highest performance when we do not have speech training data for the target task.

Still, the E2E model performs decently in zero-shot

| ESD for ST | Precision | Recall | F1 Score |
|---|---|---|---|
| ***Cascaded Systems*** | | | |
| txt-ESD(gold $t, h$) | 0.438 | 0.591 | 0.503 |
| txt-ESD(whisper-large-v2($a$), $h$) | 0.434 | 0.550 | 0.485 |
| txt-ESD(whisper-medium($a$), $h$) | 0.429 | 0.540 | 0.478 |
| txt-ESD(whisper-small($a$), $h$) | 0.413 | 0.535 | 0.466 |
| txt-ESD(whisper-base($a$), $h$) | 0.385 | 0.550 | 0.453 |
| ***End-to-End Systems*** | | | |
| *TowerInst-Fixed+Adt*($a, h$) | 0.411 | 0.542 | 0.467 |

→ **text-LLM ability is transferable to speech LLM in a zero-shot manner.**



(Example analysis is in the paper)

# SpeechQE: Estimating the Quality of Direct Speech Translation

**HyoJung Han** 
Computer Science
University of Maryland
hjhan@cs.umd.edu

**Kevin Duh** 
HLTCOE
Johns Hopkins University
kevinduh@cs.jhu.edu

**Marine Carpuat** 
Computer Science
University of Maryland
marine@cs.umd.edu

github.com/h-j-han/SpeechQE    Huggingface Hub

- **SpeechQE task**: formulation, benchmarks, evaluation of cascaded and E2E architectures
- **E2E SpeechQE model**: methods for corpus creation, training strategies, and architectural design
- **E2E systems are generally better** suited to estimate the quality of direct speech translation
- **SpeechQE need more attention!** It deserves dedicated attention as a separate problem from text-QE quality. Releasing our data and models to guide further research in this space.

# SpeechQE: Estimating the Quality of Direct Speech Translation

**HyoJung Han**

Computer Science
University of Maryland
hjhan@cs.umd.edu

**Kevin Duh**

HLTCOE
Johns Hopkins University
kevinduh@cs.jhu.edu

**Marine Carpuat**

Computer Science
University of Maryland
marine@cs.umd.edu

github.com/h-j-han/SpeechQE        Huggingface Hub

- **SpeechQE task**: formulation, benchmarks, evaluation of cascaded and E2E architectures
- **E2E SpeechQE model**: methods for corpus creation, training strategies, and architectural design
- **E2E systems are generally better** suited to estimate the quality of direct speech translation
- **SpeechQE need more attention!** It deserves dedicated attention as a separate problem from text-QE quality. Releasing our data and models to guide further research in this space.

# SpeechQE: Estimating the Quality of Direct Speech Translation

**HyoJung Han**
Computer Science
University of Maryland
hjhan@cs.umd.edu

**Kevin Duh**
HLTCOE
Johns Hopkins University
kevinduh@cs.jhu.edu

**Marine Carpuat**
Computer Science
University of Maryland
marine@cs.umd.edu

github.com/h-j-han/SpeechQE    Huggingface Hub

- **SpeechQE task**: formulation, benchmarks, evaluation of cascaded and E2E architectures
- **E2E SpeechQE model**: methods for corpus creation, training strategies, and architectural design
- **E2E systems are generally better** suited to estimate the quality of direct speech translation
- **SpeechQE need more attention!** It deserves dedicated attention as a separate problem from text-QE quality. Releasing our data and models to guide further research in this space.

# SpeechQE: Estimating the Quality of Direct Speech Translation

**HyoJung Han**

Computer Science
University of Maryland
hjhan@cs.umd.edu

**Kevin Duh**

HLTCOE
Johns Hopkins University
kevinduh@cs.jhu.edu

**Marine Carpuat**

Computer Science
University of Maryland
marine@cs.umd.edu

github.com/h-j-han/SpeechQE    Huggingface Hub

- **SpeechQE  task**: formulation, benchmarks, evaluation of cascaded and E2E architectures
- **E2E SpeechQE model**:  methods for corpus creation, training strategies, and architectural design
- **E2E systems are generally better** suited to estimate the quality of direct speech translation
- **SpeechQE need more attention!** It deserves dedicated attention as a separate problem from text-QE quality. Releasing our data and models to guide further research in this space.

# SpeechQE: Estimating the Quality of Direct Speech Translation

**HyoJung Han** 
Computer Science
University of Maryland
hjhan@cs.umd.edu

**Kevin Duh** 
HLTCOE
Johns Hopkins University
kevinduh@cs.jhu.edu

**Marine Carpuat** 
Computer Science
University of Maryland
marine@cs.umd.edu

**github.com/h-j-han/SpeechQE** **Huggingface Hub**

- **SpeechQE task**: formulation, benchmarks, evaluation of cascaded and E2E architectures
- **E2E SpeechQE model**: methods for corpus creation, training strategies, and architectural design
- **E2E systems are generally better** suited to estimate the quality of direct speech translation
- **SpeechQE need more attention!** It deserves dedicated attention as a separate problem from text-QE quality. Releasing our data and models to guide further research in this space.

# Models for Trustworthy Speech Translation

## Trustworthiness in Speech Translation

Email: hjhan@umd.edu X: @h__j___han

DEPARTMENT OF COMPUTER SCIENCE

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

# SpeechQE: Estimating the Quality of Direct Speech Translation

**EMNLP2024**

**HyoJung Han**
Computer Science
University of Maryland

**Kevin Duh**
HLTCOE
Johns Hopkins University

**Marine Carpuat**
Computer Science
University of Maryland