HydraQE: OSU's Submission for the IWSLT 2026 Speech Translation Metrics Shared Task

Kevin Krahn and Eric Fosler-Lussier
Dept. of Computer Science and Engineering
The Ohio State University
{krahn.6, fosler-lussier.1}@osu.edu

Abstract

We present HydraQE, our contribution to the IWSLT 2026 Speech Translation Metrics shared task. HydraQE is an end-to-end, reference-free quality estimation (QE) system for speech translation built on a Qwen3-ASR backbone, which accepts source audio and a translation hypothesis as joint input. Hidden states from all backbone layers are combined via a learnable sparsemax scalar mix, then re-encoded by a lightweight bidirectional Transformer to enable full cross-modal interaction prior to pooling into a shared embedding. Three independent prediction heads are trained on complementary supervision signals: human direct assessment (DA) annotations, MetricX-24 pseudo-labels, and xCOMET pseudo-labels. To address the scarcity of human-annotated data, we train on a combination of synthetically corrupted examples and silver pseudo-labeled machine translation outputs, using a curriculum that begins on synthetic and silver data and gradually shifts toward human-annotated examples. HydraQE outperforms cascaded text-based baselines and prior direct speech QE systems, demonstrating that end-to-end speech translation QE is competitive with cascaded approaches.

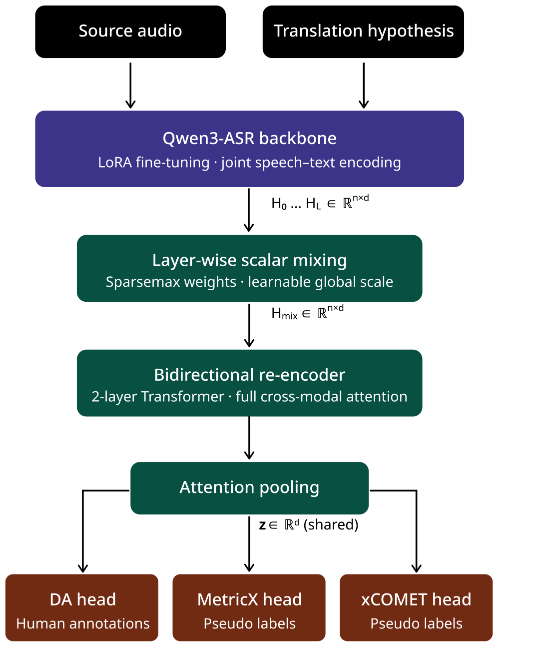

Figure 1: HydraQE architecture: Source audio and translation hypothesis are fed into the backbone model, followed by a scalar mixing module and bidirectional re-encoder. The token-wise embeddings are pooled and fed into multiple prediction heads.

1 Introduction

Quality estimation (QE) systems predict the quality of a machine translation given only the source and a candidate translation. While the ratings produced by recent QE systems correlate well with human judgments (Rei et al., 2023; Juraska et al., 2024), their emphasis has been on text translation, requiring a combination of automatic speech recognition (ASR) and text-QE for evaluating speech translations. Such a cascaded approach is inefficient and can pass ASR errors into the text-QE system, compounding errors and limiting performance.

In this paper, we present our submission to the IWSLT 2026 Speech Translation Metrics shared task (Adelani et al., 2026), focusing on an end-to-end speech QE system for reference-free evaluation of speech translation quality. Our design incorporates the Qwen3-ASR (Shi et al., 2026) model as our pretrained speech and text backbone along with multiple prediction heads trained on complementary supervision signals: (1) human direct assessment (DA) annotations from the official training data, (2) silver pseudo-labels generated by text metric models applied to machine-translated speech translation outputs, and (3) synthetically corrupted examples derived from speech translation datasets. Because human-annotated data is scarce, we employ a curriculum sampling strategy that begins training on synthetic and pseudo-labeled data and shifts toward human-labeled data.

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 323–331
July 3-4, 2026 ©2026 Association for Computational Linguistics

Our system outperforms cascaded text-based baselines and prior speech-based systems across language pairs, on both segment-level and system-level evaluations. Our results indicate that end-to-end speech translation QE systems can outperform cascaded approaches while being simpler and more efficient.

2 Related Work

Reference-free quality estimation has advanced rapidly with the adoption of pre-trained cross-lingual encoders as backbones. COMET (Rei et al., 2020) established the paradigm of fine-tuning a multilingual encoder on human quality annotations, and subsequent work scaled this to larger models, larger datasets, and word-level annotations (Rei et al., 2022, 2023; Guerreiro et al., 2024). MetricX-24 (Juraska et al., 2024) demonstrated that large-scale synthetic data constructed by systematically corrupting reference translations provides a strong training signal for rare failure modes, and further introduced a unified hybrid architecture capable of reference-based and reference-free evaluation within a single model trained on a mixture of DA and MQM ratings. LLM-based approaches such as GEMBA-MQM (Kocmi and Federmann, 2023) have demonstrated strong zero-shot QE capabilities using generative models, especially at the system-level, but still fall short of trained metrics at the segment level (Lavie et al., 2025).

Evaluation of speech translation has historically followed MT evaluation practice by applying text metrics such as BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), and BLEURT (Sellam et al., 2020) to system output (Abdulmumin et al., 2025), but several recent systems support direct speech QE. The BLASER family of models operates in a shared cross-lingual and cross-modal speech–text embedding space, allowing for superior transfer across modalities and languages where labeled data is unavailable. The original BLASER (Chen et al., 2023) introduced a speech translation evaluation metric based on LASER (Jones et al., 2021) sentence embeddings, and BLASER 2.0 (Dale and Costa-jussà, 2024) improved upon this with multi-modal SONAR embeddings (Duquenne et al., 2023), using a multilingual speech encoder trained in a teacher–student fashion to make the speech encoder match the SONAR text encoder sentence embeddings. BLASER 3.0 (Omnilingual MT Team et al., 2026) adopts a multi-head architecture with separate prediction heads per training signal, which inspires our design, although it is not publicly available for testing. SpeechQE (Han et al., 2024) is a model designed specifically for speech translation QE, combining a pretrained speech encoder and text LLM trained on a QE task.

The effectiveness of direct speech QE compared to ASR-based approaches has been mixed; text metrics such as xCOMET and MetricX combined with ASR frequently achieve stronger correlation with human judgments than BLASER 2.0 (Han et al., 2024; Cettolo et al., 2026), while SpeechQE matches or slightly outperforms cascaded ASR approaches in some scenarios (Han et al., 2024).

Source	Language pairs	Segments
IWSLT 2023	en→de, en→ja, en→zh	12,480
WMT 2024	en→cs, en→es, en→is, en→ja, en→ru, en→uk, en→zh	10,193
WMT 2025	cs→de, cs→uk, en→cs, en→is, en→ja, en→ru, en→uk, en→zh	5,836
Total		**28,509**

Table 1: Official training data segments after filtering to the 8 most-frequent target languages (*zh, ja, de, uk, cs, ru, is, es*).

3 Methods

3.1 Dataset

Official shared task data We use the official IWSLT 2026 shared task dataset, which aggregates direct assessment (DA) annotations from IWSLT 2023 and WMT 2024–2025. We filter the dataset to the 8 most-frequent target languages for training (see Table 1). We use the provided development set, consisting of human annotations of IWSLT 2025 ACL Talks, for all model selection and hyperparameter tuning. We convert all training labels into the 0–1 range for consistency across sources.

Synthetic data To generate additional training samples, we apply several corruptions to the reference translations (see Table 2) in CoVoST 2(Wang et al., 2021) and FLEURS (Conneau et al., 2022), inspired by the synthetic data construction procedure of MetricX (Juraska et al., 2023, 2024). The `Unrelated`, `Gibberish`, and `Identity` samples have fixed scores, while the score targets for `Undertranslation` and `Corruption` are derived as the proportion of the character length left unmodified after corruption, then divided by two to

Corruption	Description	Score
`Undertranslation`	Randomly chosen sentence removed if segment is multi-sentence; otherwise, truncate 30–60% of words.	$[0.15, 0.5]$
`Corruption`	Replace a random 40–60% of tokens with random vocabulary tokens from the same language.	$[0.2, 0.3]$
`Unrelated`	Replace candidate translation with randomly chosen segment of similar length in the same language.	0.0
`Gibberish`	Sample random vocabulary tokens to produce a sequence with the same length as the candidate.	0.0
`Identity (Gold)`	Use the reference translation.	1.0

Table 2: Synthetic corruption approaches and score targets.

account for the disfluency incurred by the corruptions in addition to semantic changes. These synthetic examples primarily cover the low end of the score range for several reasons: (1) our QE datasets are biased toward high scores, so we attempt to partially balance the distribution; (2) it is difficult to procedurally generate useful scores for small corruptions, as changing one or two words could either drastically alter meaning or have little effect; (3) metric models struggle to detect undertranslations and fluent but unrelated translations (Juraska et al., 2024), so we target these failures modes directly.

Silver data We also construct a large set of *silver* examples by running MT systems over CoVoST 2 and FLEURS English source segments to obtain machine-generated translation hypotheses for en→de and en→zh language pairs, then pseudo-labelling them with MetricX-24-XXL (Juraska et al., 2024) and xCOMET-XXL (Guerreiro et al., 2024) using the ground-truth reference translations as additional input. We transform the outputs of the MetricX-24 model into the $[0, 1]$ range: $(25 - \text{score})/25$. Training on these silver labels serves to distill these large text models into our speech QE model. We generate hypotheses using multiple MT systems: SeamlessM4T (Seamless Communication et al., 2023), which translates directly from the source audio; and NLLB (NLLB Team et al., 2022) and Hunyuan-MT-7B (Zheng et al., 2025), which translate from the gold transcripts. These systems vary widely in model size, training procedure, and architecture, thus diversifying hypothesis quality and style.

TTS-augmented text QE data Human-labeled data for speech translation quality is limited. To further increase training coverage, we incorporate human-labeled text data from the WMT 2022 Quality Estimation shared task (Freitag et al., 2022), which provides source–hypothesis pairs with human DA scores for the en→de and en→zh language pairs. We synthesize speech using Qwen3-TTS (Hu et al., 2026), allowing these examples to be used for speech-input QE training.

3.2 Architecture

The overall architecture for the system is shown in Figure 1; a breakdown of each component follows.

Backbone We use Qwen3-ASR-1.7B (Shi et al., 2026) as our backbone encoder. Qwen3-ASR supports speech and text inputs in 52 languages and dialects, allowing us to feed the source audio and the translation hypothesis as a single forced input and produce contextualized representations of both modalities jointly. We fine-tune the backbone using Low-Rank Adaptation (LoRA) (Hu et al., 2022), which inserts trainable low-rank matrices into the attention projections while keeping the original weights frozen, substantially reducing the number of trainable parameters and mitigating overfitting on the relatively small human-annotated training set. We additionally freeze the speech encoder parameters, as updating them degrades performance, with or without LoRA. Although this model was fine-tuned for ASR and not translation tasks, we find that using the translation hypothesis as forced input works well in practice, likely due to the rich pretrained capabilities of the base Qwen3 model.

Layer-wise scalar mixing Rather than using only the final hidden layer, we extract hidden states from all $L + 1$ layers of the backbone. For an input of n tokens, layer ℓ produces $H_\ell \in \mathbb{R}^{n\times d}$. Following Rei et al. (2023), we combine these with a scalar mix module that learns a task-specific weighted average across layers:

$$H_{\text{mix}} = \lambda \sum_{\ell=0}^{L} \beta_\ell H_\ell, \tag{1}$$

where $\lambda \in \mathbb{R}$ is a learnable global scale and $\boldsymbol{\beta} \in \Delta^L$ is a layer-importance distribution parameterised with SPARSEMAX (Martins and Astudillo, 2016). SPARSEMAX is preferred over SOFTMAX because it produces sparse weights, effectively selecting a small subset of layers rather than diffusely averaging all of them.

Bidirectional re-encoding Because Qwen3-ASR is a causal language model, each token in H_{mix} attends only to its left context. Consequently, early speech tokens have no direct access to the translation hypothesis and vice versa. To enable full cross-modal interaction, we pass H_{mix} through a lightweight two-layer bidirectional Transformer encoder. This re-encoding step is inexpensive relative to the backbone but improves the quality of the pooled representation by allowing every token to attend to the full source–hypothesis context. A final attention pooling layer then reduces the variable-length sequence into a single fixed-size embedding $\mathbf{z} \in \mathbb{R}^d$, which is shared across all prediction heads.

Multiple scoring heads Inspired by BLASER 3.0 (Omnilingual MT Team et al., 2026), we attach multiple independent prediction heads to $\mathbf{z}$, one per training signal. Unlike BLASER 3.0, which uses a single linear projection per head, each of our heads is a two-layer feed-forward network with a TANH activation, producing a scalar quality score. The additional capacity allows each head to model the distributional characteristics of its supervision source (human DA labels, MetricX pseudo-labels, and xCOMET pseudo-labels) while limiting interference with the shared representation. The pseudo-label heads effectively distill the reference-based metrics into a reference-free metric.

3.3 Training Procedure

Multi-head loss Each training example is routed to its corresponding prediction head based on data source: human DA annotations are passed to the DA head, and MetricX and xCOMET pseudo-labeled examples are passed to their respective silver heads. The total loss is the sum of the mean squared error (MSE) between the predicted and target scores, with the DA head weighted separately:

$$L_{total} = L_{\text{MetricX}} + L_{\text{xCOMET}} + \lambda_{DA} \cdot L_{\text{DA}} \quad (2)$$

We find that $\lambda_{DA} = 1.5$ works well in practice; we tried values in $\{1.0, 1.5, 2.0\}$ and observed only marginal differences, with 1.5 performing slightly better on the development set. Gradients flow through the active head and the shared backbone, so the backbone is updated by all data sources while each head specializes to its own distribution.

Curriculum sampling Because human DA annotations are scarce relative to the synthetic and silver data, we train with a curriculum that gradually shifts the sampling distribution toward official data. We control the DA/non-DA split with a mixing coefficient α, which remains at 0 for the first 10,000 steps (a warm-up phase on synthetic and silver data), then increases linearly to 0.09 over the next 20,000 steps. At each step, we sample from the official DA pool with probability α and from the synthetic–silver pool with probability $1 - \alpha$.

Within the non-DA pool, a second coefficient β controls the synthetic/silver balance: synthetic examples are drawn with probability β and silver examples with probability $1 - \beta$. β decreases linearly from 1.0 to 0.005 over the first 5,000 steps. The synthetic data thus contributes heavily at first but quickly tapers off into a small but consistent signal. All schedule values (α, β, and the step counts) were tuned on the development set.

To avoid the model overfitting to common sequence lengths and score ranges, silver and synthetic examples are drawn via stratified sampling: we maintain five equal-mass bins along the sequence-length axis and, for silver data, five equal-mass bins along the pseudo-label score axis, sampling uniformly across bins before sampling within them. Additionally, we balance languages within each pool so each language pair contributes equally.

Synthetic examples are routed to the silver heads but are deliberately withheld from the DA head. Routing synthetic examples to the DA head slightly degrades performance on human DA annotations, likely because the deterministic corruption-based scores used for synthetic data do not match the distributional characteristics of human judgments.

Checkpoint selection We train the model for 70,000 steps with a batch size of 8, using the AdamW optimizer. We evaluate performance on the dev set every 1,000 steps. Through experimentation with the augmented WMT 2022 QE en→de and en→zh TTS data, we find that even when sampled with low probability the inclusion of this data slightly hurt performance, and therefore exclude it from our final training run.

Based on dev scores, a tradeoff exists in the choice of prediction head: the DA head achieves

System	Dev						Test (scored by organizers)					
	Segment-level (τ)			System-level			Segment-level (τ)			System-level		
	de	zh	Avg	de	zh	Avg	de	zh	Avg	de	zh	Avg
COMET (partial)†	11.3	12.0	11.6	44.4	68.7	56.6	13.1	14.8	14.0	85.7	78.2	81.9
COMET†	32.6	36.5	34.6	86.2	92.6	89.4	-	-	-	-	-	-
SpeechQE†	26.6	31.8	29.2	78.6	93.4	86.0	22.1	23.3	22.7	89.8	96.1	92.9
BLASER 2.0 QE†	22.0	26.8	24.4	86.0	67.7	76.9	18.5	20.3	19.4	88.8	88.2	88.5
CometKiwi†*	31.4	35.9	33.7	91.6	34.9	63.3	23.7	33.3	28.5	93.1	97.2	95.2
MetricX-24-XXL QE	29.7	31.2	30.5	99.6	**91.3**	**95.5**	-	-	-	-	-	-
xCOMET-XXL QE	30.4	35.5	33.0	87.7	45.8	66.8	-	-	-	-	-	-
HydraQE (DA head)	32.5	**39.9**	**36.2**	**99.7**	69.2	84.5	25.3	**33.6**	29.5	94.7	98.1	96.4
HydraQE (MetricX head)	31.8	39.3	35.6	98.8	90.1	94.5	24.3	32.9	28.6	94.2	97.1	95.7
HydraQE (xCOMET head)	**32.7**	37.1	34.9	96.9	34.6	65.7	**26.7**	32.9	**29.8**	94.8	**98.7**	96.7
HydraQE (all heads avg.)	32.0	38.7	35.4	99.5	56.0	77.7	25.3	33.3	29.3	**95.1**	98.4	**96.8**
HydraQE (primary)	32.1	39.8	36.0	99.4	79.3	89.4	24.8	33.5	29.1	94.5	97.9	96.2

Table 3: Segment-level Kendall-Tau (τ)(↑) and System-level Soft Pairwise Accuracy (↑) of HydraQE and baselines on the IWSLT 2026 Metrics Shared Task development and test sets, with *en* as the source language and *de* and *zh* as the target languages. The primary submission is a weighted average of the DA and MetricX heads. † denotes official shared task baselines; text-based baselines use gold transcripts provided with the dataset. * For CometKiwi, we report dev results using the XXL variant; on the test set we report the organizers' results.

the highest segment-level scores, while the MetricX head achieves the highest system-level scores. To balance between these tasks for our primary model, we independently choose the checkpoint with the best segment-level scores per head and combine their predictions using the following sum:

$$\text{Score} = \frac{3}{4}\text{Head}_{\text{DA}} + \frac{1}{4}\text{Head}_{\text{MetricX}}. \tag{3}$$

The higher weight for the DA head is chosen to prioritize segment-level over system-level scores.

4 Evaluation

4.1 Main Results

Table 3 reports segment-level Kendall-τ and system-level Soft Pairwise Accuracy for all HydraQE configurations and baselines. We report our results on the development set and the IWSLT organizers' official results on the test set.

On segment-level evaluation, HydraQE achieves the strongest performance across both the dev and test sets, reaching $\tau = 33.6$ on en→zh on the test set using the DA head, and 29.8 averaged across language pairs using the xCOMET head, outperforming all baselines including the strongest text-based baseline, CometKiwi (28.5), which has access to gold transcripts. The fact that an end-to-end speech QE system exceeds a text-based cascade that uses gold transcripts suggests that the joint audio–hypothesis encoding captures quality-relevant information beyond what is available in transcripts alone. The speech-native baselines, SpeechQE and BLASER 2.0 QE, trail substantially at 22.7 and 19.4 respectively, confirming that prior direct speech QE systems leave considerable room for improvement. The poor performance of BLASER 2.0 is possibly due to it being trained on XSTS annotations rather than DA annotations.

System-level results present a different picture. On the dev set, MetricX-24-XXL achieves the highest system-level average (95.5), driven by a near-perfect score on en→de (99.6). The MetricX head of HydraQE closely tracks this behavior (94.5), while the DA head, despite its strong segment-level performance, drops to 84.5 due to weaker performance on en→zh (69.2). The xCOMET head performs poorly on en→zh (34.6), lower than the teacher metric xCOMET (45.8). These dev set results motivated us to blend the DA and MetricX heads for our primary submission, trading a small amount of segment-level performance relative to the DA head alone in exchange for a substantially improved system-level average. However, on the test set all heads achieve excellent average system scores (> 95). In fact, the worst HydraQE head on the dev set, xCOMET, is the best individual head on the test set. We also note that the system-level dev scores oscillated dramatically during training, making these scores somewhat untrustworthy. The ensemble of all heads yielded the highest average system test score (96.8), surpassing the xCOMET head (96.7) by a narrow margin.

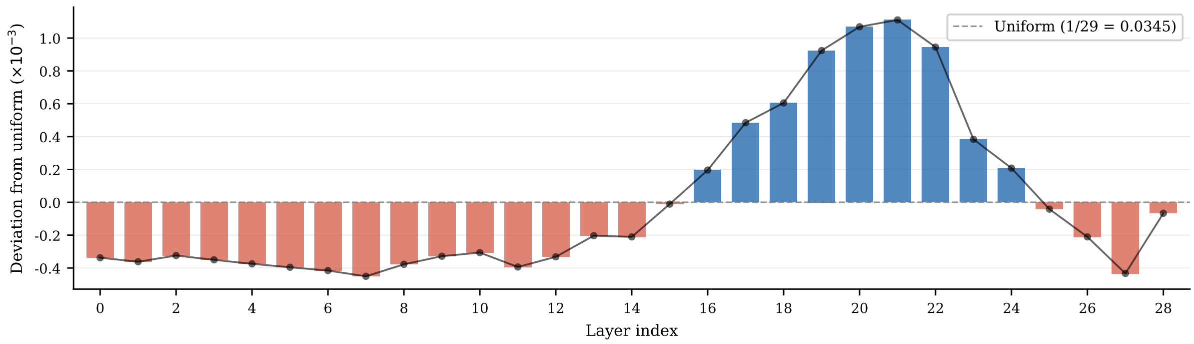

Figure 2: Learned scalar-mix weights of layer-wise mixing module, shown as deviation from uniform weighting ($1/29 \approx 0.034$). The model up-weights upper-middle layers (16–24), consistent with those layers carrying richer semantic representations, while suppressing the earliest and topmost layers.

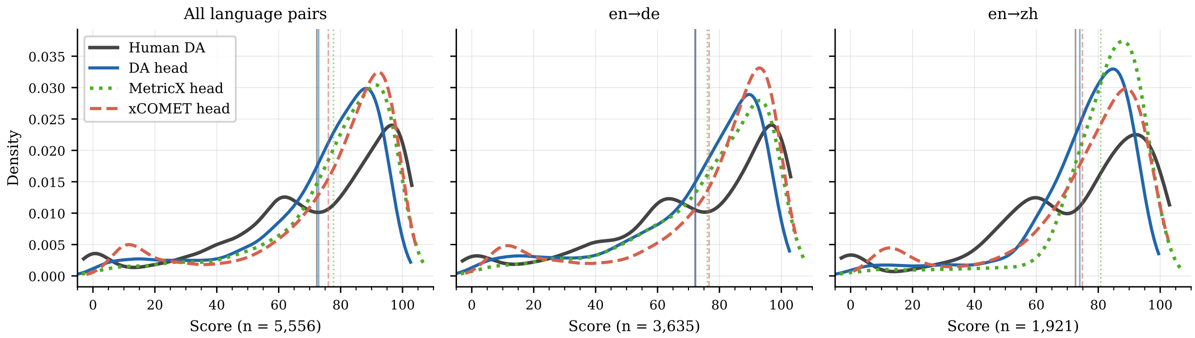

Figure 3: Distributions of human scores and predicted scores from each HydraQE head on the IWSLT 2026 dev set.

The primary submission combines the DA and MetricX heads, based on dev set evidence. The test results show this to be a reasonable, but not optimal choice, not beating an individual head in any category but also outperforming the worst head in all categories. For comparison with other submissions, we refer the reader to the official evaluation results for the IWSLT metrics shared task (Adelani et al., 2026).

4.2 Analysis

Layer-wise scalar mixing Figure 2 shows that the learned scalar mix concentrates weight in the upper-middle layers (16–24) of the backbone, with near-zero weight assigned to the lowest and highest layers. This is consistent with probing studies of transformer language models, which find that the topmost layers tend to specialize on token prediction rather than general-purpose meaning (Jawahar et al., 2019; Tenney et al., 2019; Liu et al., 2024). The pattern suggests that the scalar mix is exploiting the backbone as a semantic encoder rather than relying on its ASR-tuned output layer, which is encouraging given that the model was not originally trained for translation quality assessment.

Score distributions by head Figure 3 reveals qualitatively different output distributions across the three prediction heads. The human DA labels are concentrated overall toward the high-end, but show distinct, smaller concentrations in the middle and low end of the score range, reflecting annotators' tendency to cluster translations scores rather than using the entire score range. The DA head trained on human labels produces a smoother, unimodal distribution concentrated toward the high end. The two pseudo-label heads similarly concentrate toward the high end, consistent with prior work showing automatic metrics to be overly optimistic and generous with scores for low quality translations (Zouhar et al., 2024). The inclusion of synthetic data covering the extreme low end was unsuccessful at mitigating this issue.

The two pseudo-label heads also exhibit language-specific behavior: the MetricX head concentrates more density at the high end for en→de, while the xCOMET head does so for en→zh. This asymmetry likely reflects differing calibration of the two teacher metrics across languages and may partly explain the complementarity between the DA and MetricX heads observed in the dev results.

5 Conclusion

In this work we presented HydraQE, an end-to-end reference-free quality estimation system for speech translation built on a Qwen3-ASR backbone. By jointly encoding source audio and translation hypothesis and training multiple prediction heads on complementary supervision signals via curriculum sampling, HydraQE achieves stronger segment-level correlation with human judgments on the IWSLT 2026 Metrics Shared Task test set compared to both cascaded text-based baselines and prior direct speech QE systems.

A key finding is that the choice of prediction head involves a meaningful tradeoff in per-language performance as well as segment vs system-level performance. No head dominates all metrics simultaneously. Our primary submission attempts to balance segment and system-level performance by combining two heads based on dev performance, but our post-hoc test set analysis showed no benefit over the single best head or an average over all heads.

Several directions remain open for future work. The curriculum sampling schedule and head weighting were tuned on a single development set covering only two language pairs; evaluation across a broader set of source and target languages would strengthen confidence in these design choices. The exclusion of TTS-augmented text QE data, which marginally hurt development performance, warrants further investigation, as synthesized speech may introduce domain mismatch that targeted data augmentation could address.

Acknowledgments

This work was carried out in partnership with the ETEN Innovation Lab. We are grateful to Joel Mathew for his guidance and support throughout this project, and to the members of the OSU SLaTe Lab for many helpful discussions.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 33 others. 2025. Findings of the IWSLT 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vienna, Austria (in-person and online). Association for Computational Linguistics.

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, and 33 others. 2026. Speech translation and metrics in 2026: Findings of the iwslt campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Mauro Cettolo, Marco Gaido, Matteo Negri, Sara Papi, and Luisa Bentivogli. 2026. How to evaluate speech translation with source-aware neural mt metrics. *Preprint*, arXiv:2511.03295.

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2023. BLASER: A text-free speech-to-speech translation evaluation metric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9064–9079, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. FLEURS: few-shot learning evaluation of universal representations of speech. In *IEEE Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023*, pages 798–805. IEEE.

David Dale and Marta R. Costa-jussà. 2024. BLASER 2.0: a metric for evaluation and quality estimation of massively multilingual speech and text translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16075–16085, Miami, Florida, USA. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *Preprint*, arXiv:2308.11466.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection.

Transactions of the Association for Computational Linguistics, 12:979–995.

HyoJung Han, Kevin Duh, and Marine Carpuat. 2024. SpeechQE: Estimating the quality of direct speech translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21852–21867, Miami, Florida, USA. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-tts technical report. *Preprint*, arXiv:2601.15621.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Alexander Jones, William Yang Wang, and Kyle Mahowald. 2021. A massively multilingual analysis of cross-linguality in shared embedding space. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5833–5847, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. 2024. Fantastic semantics and where to find them: Investigating which layers of generative LLMs reflect lexical semantics. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14551–14558, Bangkok, Thailand. Association for Computational Linguistics.

André F. T. Martins and Ramón Fernandez Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, JMLR Workshop and Conference Proceedings, pages 1614–1623. JMLR.org.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Omnilingual MT Team, Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-Eric Saint-James, Ioannis Tsiamas, Chierh Cheng, Joe Chuang, Paul-Ambroise Duquenne, Mark Duppenthaler, Nate Ekberg, Cynthia Gao, Pere Lluís Huguet Cabot, João Maria Janeiro, Jean Maillard, and 12 others. 2026. Omnilingual mt: Machine translation for 1,600 languages. *Preprint*, arXiv:2603.16309.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference* *on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023. Seamlessm4t: Massively multilingual & multimodal machine translation. *Preprint*, arXiv:2308.11596.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-asr technical report. *Preprint*, arXiv:2601.21337.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. Covost 2 and massively multilingual speech translation. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 2247–2251. ISCA.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report. *Preprint*, arXiv:2509.05209.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

A Hyperparameters

Hyperparameter	**Value**
Optimization	
Optimizer	AdamW
AdamW β_1, β_2	0.9, 0.999
Learning rate (backbone)	5e-6
Learning rate (new params)	2e-5
Weight decay	0.05
Warmup steps	5,000
Batch size	8
Training steps	70,000
Evaluation interval	1,000 steps
Dropout	0.1
LoRA Rank	32
LoRA Alpha	64
LoRA Dropout	0.1
Re-encoder	
Layers	2
Hidden size	1024
Attention heads	8
Activation	SwiGLU
Prediction Heads	
Hidden sizes	3072, 1024
Activation	Tanh
Curriculum sampling	
DA warm-up duration	10,000 steps
DA ramp duration	20,000 steps
DA mixing coefficient α (final)	0.09
Synthetic ramp duration	5,000 steps
Synthetic coefficient β (start)	1.0
Synthetic coefficient β (final)	0.005

Table 4: Hyperparameters used for the final HydraQE model.