Beyond Target Hypotheses: Generative Error Correction for Speech-to-Text Translation with Source Transcriptions

Anonymous ACL submission

Abstract

Generative error correction (GER) with large language models (LLMs) has shown strong promise for improving speech-to-text translation (ST) by reranking and rewriting N-best hypotheses. However, existing GER frameworks rely exclusively on ST hypotheses, leaving unexploited complementary acoustic-linguistic information encoded in in automatic speech recognition (ASR) transcriptions. We propose a unified GER approach that jointly conditions LLM-based ST correction on N-best ASR transcriptions and N-best ST hypotheses, bridging the source-language transcription and target-language translation streams within a single generative framework. Extensive experiments on the CoVoGER benchmark spanning 26 language pairs and BMELD dataset demonstrates that jointly conditioning on ASR and ST hypotheses consistently outperforms ST-only baselines in most settings, yielding modest and language-dependent gains.

1 Introduction

Speech-to-text translation (ST) has advanced considerably with large-scale end-to-end models such as Whisper (Radford et al., 2023) and SeamlessM4T (Barrault et al., 2023), which achieve strong performance across diverse languages and domains. Yet these systems still struggle in low-resource settings (Pakray et al., 2025) and with emotionally expressive speech (Lin et al., 2025). Conventional rescoring methods (Xu et al., 2022; Kang et al., 2024; Chen et al., 2023) rerank N-best hypotheses using external language models, but are inherently limited to selecting among existing candidates rather than producing improved translations.

Generative error correction (GER) offers a more flexible alternative, where large language models (LLMs) rewrite N-best hypotheses into more accurate outputs (Hu et al., 2024). Early GER work focused on ASR correction, demonstrating that LLMs can effectively leverage N-best ASR transcriptions to recover from recognition errors by exploiting their broad linguistic knowledge and contextual reasoning (Mu et al., 2024). Building on this, GER was extended to speech-to-text translation, conditioning the LLM on N-best ST hypotheses to rewrite translations (Yang et al., 2025b,a). Despite these advances, existing GER frameworks remain tightly coupled to a single hypothesis stream, leaving complementary cross-stream information largely unexplored. A natural source of such information lies in ASR transcriptions of the source speech, which capture the original spoken form directly, preserving lexical and structural details that may be lost or distorted during end-to-end translation. Multi-source approaches in neural machine translation have demonstrated the benefit of jointly conditioning on multiple input representations (Stoian et al., 2020), yet this insight has not been carried over to GER for ST. Incorporating both N-best ASR transcriptions and N-best ST hypotheses as joint input to the LLM could therefore provide a richer grounding signal, enabling more informed and accurate error correction.

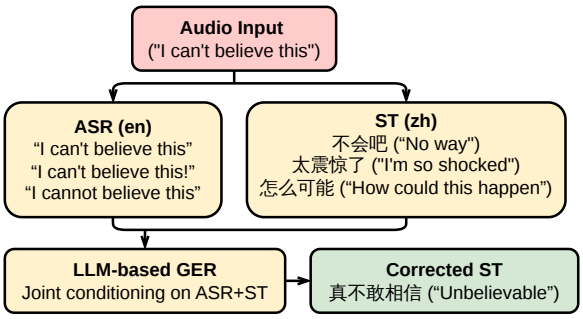

Figure 1: Overview of the proposed GER framework. N-best ASR hypotheses and N-best ST hypotheses are jointly conditioned to generate corrected speech translations.

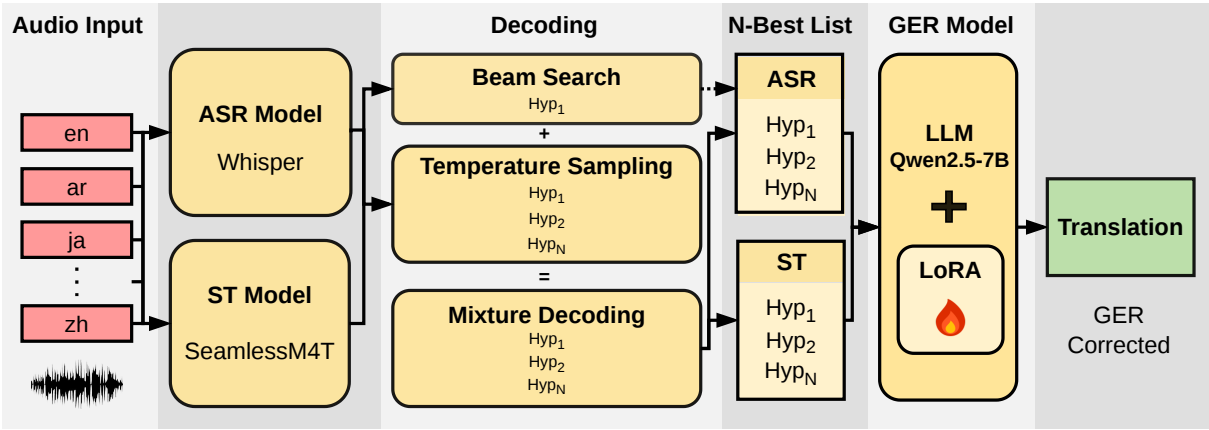

Figure 2: Overview of the proposed joint ASR+ST GER framework. Multilingual speech is processed in parallel by a Whisper ASR model and a SeamlessM4T ST model. Each model generates N-best hypotheses through a mixture decoding strategy that combines beam search and temperature sampling. The resulting ASR and ST N-best lists are jointly fed to a LoRA fine-tuned Qwen2.5-7B LLM to produce the final corrected translation.

In this paper, we propose a GER framework that jointly conditions LLM-based ST correction on N-best ASR transcriptions and N-best ST hypotheses, bridging the source-language transcription and target-language translation streams within a single generative framework. Experiments are conducted with Qwen2.5-7B (Hui et al., 2024) on CoVoST-2 (Wang et al., 2020; Ardila et al., 2020) and BMELD (Liang et al., 2021). We make three main contributions: (i) a joint ASR+ST GER framework that grounds LLM correction in both source transcriptions and translation hypotheses; (ii) evaluation across 26 CoVoST-2 language pairs and BMELD English-to-Chinese ST dataset; and (iii) a systematic comparison of 1-best vs. N-best ASR settings, revealing how transcription diversity influences correction quality in GER.

2 Methodology

As depicted in Figure 2, the framework comprises three core components: (1) a decoding strategy generating diverse N-best hypotheses from ASR and ST models, (2) a joint ASR-ST grounding mechanism that feeds both source-language transcriptions and target-language hypotheses into an LLM, and (3) a LoRA fine-tuning procedure that trains the LLM to produce corrected translations via a sequence-to-sequence objective.

2.1 Mixture Decoding

Multilingual speech is processed in parallel by a Whisper ASR model and a SeamlessM4T ST model. To generate N-best hypotheses, both models employ a mixture decoding strategy where beam search is used to obtain the 1-best hypothesis and temperature sampling is applied to generate diverse N-best hypotheses.

2.2 Joint ASR-ST Grounding

Given a speech segment in a source language, our goal is to generate an accurate translation in the target language by leveraging both N-best ASR transcriptions in the source language and N-best ST hypotheses in the target language as joint input to a large language model. Formally, let $\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$ denote the N-best ASR transcriptions and $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$ denote the N-best ST hypotheses, and let θ denote the parameters of the LoRA fine-tuned LLM. The corrected translation $\hat{y}$ is obtained by:

$$\hat{y} = \arg\max_{y} P_\theta(y \mid \mathcal{A}, \mathcal{H}) \tag{1}$$

2.3 Input Prompt Construction

The LLM input is a plain-text prompt concatenating the N-best ASR transcriptions $\mathcal{A}$ in the source language with the N-best ST hypotheses $\mathcal{H}$ in the target language. This lightweight design requires no architectural modifications and leverages the LLM's cross-lingual capabilities for implicit source-target alignment. For the BMELD dataset (Liang et al., 2021), the emotion label e and sentiment label s are prepended as explicit conditioning signals. We also examine a 1-best ASR setting, providing only a_1 to isolate the contribution of source-language grounding from transcription diversity. The instruction prompt format is detailed in Appendix C.

Setting	Ar-En	Ca-En	Cy-En	De-En	Et-En	Fa-En	Id-En	Ja-En	Lv-En	Sl-En	Sv-En	Tr-En	Zh-En	AVG
M-ST only	40.28	35.49	48.54	35.61	21.70	23.48	49.39	18.01	21.12	28.69	31.20	27.98	19.46	30.84
M-GER (ST only)	44.53	36.24	45.64	37.23	22.85	24.13	52.43	21.35	26.97	31.89	35.73	29.97	19.46	32.96
M-GER (1-ASR+ST)	$48.90^{\dagger\ddagger}$	$37.81^{\dagger\ddagger}$	$48.73^{\dagger\ddagger}$	$40.03^{\dagger\ddagger}$	$23.41^{\dagger\ddagger}$	$25.67^{\dagger\ddagger}$	$55.87^{\dagger\ddagger}$	$23.48^{\dagger\ddagger}$	$28.49^{\dagger\ddagger}$	$36.74^{\dagger\ddagger}$	$46.83^{\dagger\ddagger}$	$33.67^{\dagger\ddagger}$	$22.14^{\dagger\ddagger}$	$36.29^{\dagger\ddagger}$
M-GER (N-ASR+ST)	$49.17^{\dagger\ddagger}$	$37.94^{\dagger\ddagger}$	$46.61^{\dagger}$	$40.37^{\dagger\ddagger}$	$23.55^{\dagger\ddagger}$	$25.83^{\dagger\ddagger}$	$55.40^{\dagger\ddagger}$	$23.97^{\dagger\ddagger}$	$28.27^{\dagger\ddagger}$	$35.43^{\dagger\ddagger}$	$44.92^{\dagger\ddagger}$	$32.52^{\dagger\ddagger}$	$23.96^{\dagger\ddagger}$	$35.99^{\dagger\ddagger}$
L-ST only	45.25	38.36	**55.01**	38.88	26.43	25.60	51.26	21.89	26.57	37.50	38.23	31.34	20.82	35.16
L-GER (ST only)	47.73	38.41	52.60	38.96	27.10	26.13	53.75	23.96	31.35	39.57	41.41	32.50	21.04	36.50
L-GER (1-ASR+ST)	$50.45^{\dagger\ddagger}$	$\mathbf{39.16}^{\dagger\ddagger}$	$53.53^{\ddagger}$	$\mathbf{41.49}^{\dagger\ddagger}$	$27.03^{\ddagger}$	$\mathbf{27.01}^{\dagger\ddagger}$	$57.24^{\dagger\ddagger}$	$25.36^{\dagger\ddagger}$	$\mathbf{32.69}^{\dagger\ddagger}$	$40.33^{\dagger\ddagger}$	$48.43^{\dagger\ddagger}$	$34.49^{\dagger\ddagger}$	$\mathbf{24.77}^{\dagger\ddagger}$	$\mathbf{38.61}^{\dagger\ddagger}$
L-GER (N-ASR+ST)	$\mathbf{51.05}^{\dagger\ddagger}$	$38.08^{\ddagger}$	$52.04^{\ddagger}$	$40.38^{\dagger\ddagger}$	$\mathbf{27.31}^{\dagger\ddagger}$	$26.62^{\dagger\ddagger}$	$\mathbf{58.12}^{\dagger\ddagger}$	$\mathbf{25.55}^{\dagger\ddagger}$	$32.34^{\dagger\ddagger}$	$\mathbf{40.92}^{\dagger\ddagger}$	$\mathbf{48.54}^{\dagger\ddagger}$	$\mathbf{34.74}^{\dagger\ddagger}$	$25.12^{\dagger\ddagger}$	$38.52^{\dagger\ddagger}$

Table 1: BLEU results on CoVoST-2 for X-En across 13 language pairs using Qwen2.5-7B with LoRA fine-tuning. M and L denote Medium and Large model sizes for both the ASR (Whisper) and ST (SeamlessM4T) models respectively. Bold indicates the highest BLEU score per language pair. † and ‡ indicate statistically significant improvement over the same-size ST only and GER (ST only) baselines respectively at $p < 0.05$.

Setting	En-Ar	En-Ca	En-Cy	En-De	En-Et	En-Fa	En-Id	En-Ja	En-Lv	En-Sl	En-Sv	En-Tr	En-Zh	AVG
M-ST only	22.60	39.61	31.82	33.03	25.01	18.46	37.00	24.00	21.07	30.92	39.53	21.40	32.04	28.96
M-GER (ST only)	22.81	38.69	30.87	33.83	24.28	17.91	37.26	30.92	20.30	29.87	39.23	20.04	44.35	30.03
M-GER (1-ASR+ST)	$23.10^{\dagger\ddagger}$	38.50	$32.59^{\dagger\ddagger}$	$34.61^{\dagger\ddagger}$	24.54	$20.26^{\dagger\ddagger}$	$37.61^{\dagger\ddagger}$	$32.75^{\dagger\ddagger}$	$21.31^{\dagger\ddagger}$	30.39	39.42	20.63	$47.70^{\dagger\ddagger}$	$31.03^{\dagger\ddagger}$
M-GER (N-ASR+ST)	$23.19^{\dagger\ddagger}$	38.83	$32.78^{\dagger\ddagger}$	$34.93^{\dagger\ddagger}$	24.44	$20.19^{\dagger\ddagger}$	$37.73^{\dagger\ddagger}$	$26.12^{\dagger}$	$21.27^{\dagger\ddagger}$	$30.93^{\dagger}$	39.70	20.36	$48.00^{\dagger\ddagger}$	$30.65^{\dagger\ddagger}$
L-ST only	25.13	**42.09**	34.18	36.23	**28.90**	19.78	39.41	25.59	24.23	**35.42**	**42.93**	**24.25**	35.83	31.84
L-GER (ST only)	25.73	40.02	33.26	35.75	27.57	18.21	38.84	32.47	21.47	34.05	41.64	22.29	46.90	32.17
L-GER (1-ASR+ST)	$25.85^{\dagger\ddagger}$	$40.95^{\ddagger}$	$\mathbf{35.25}^{\dagger\ddagger}$	$36.76^{\dagger\ddagger}$	$28.44^{\ddagger}$	$21.71^{\dagger\ddagger}$	$\mathbf{39.52}^{\dagger\ddagger}$	$33.49^{\dagger\ddagger}$	$24.09^{\ddagger}$	$34.90^{\ddagger}$	$42.49^{\ddagger}$	$22.91^{\ddagger}$	$49.52^{\dagger\ddagger}$	$33.53^{\dagger\ddagger}$
L-GER (N-ASR+ST)	$\mathbf{25.93}^{\dagger\ddagger}$	$40.90^{\ddagger}$	$35.23^{\dagger\ddagger}$	$\mathbf{36.96}^{\dagger\ddagger}$	$28.84^{\ddagger}$	$\mathbf{21.83}^{\dagger\ddagger}$	$39.40^{\dagger\ddagger}$	$\mathbf{33.53}^{\dagger\ddagger}$	$\mathbf{24.28}^{\dagger\ddagger}$	$35.19^{\ddagger}$	$42.85^{\ddagger}$	$23.12^{\ddagger}$	$\mathbf{49.93}^{\dagger\ddagger}$	$\mathbf{33.69}^{\dagger\ddagger}$

Table 2: BLEU results on CoVoST-2 for En-X across 13 language pairs using Qwen2.5-7B with LoRA fine-tuning. M and L denote Medium and Large model sizes for both the ASR (Whisper) and ST (SeamlessM4T) models respectively. Bold indicates the highest BLEU score per language pair. † and ‡ indicate statistically significant improvement over the same-size ST only and GER (ST only) baselines respectively at $p < 0.05$.

2.4 Model Training

Following the GER fine-tuning paradigm of Yang et al. (2025b), we train the LLM to generate the corrected translation $\hat{y}$ from the joint ASR and ST hypothesis inputs using a sequence-to-sequence approach with cross-entropy loss:

$$\mathcal{L}_{CE} = \sum_{l=1}^{L} -\log P_\theta(t_l^* \mid t_{l-1}^*, \ldots, t_1^*; \mathcal{A}, \mathcal{H}, \mathcal{I}) \tag{2}$$

where t_l^* is the l-th target token, L is the sequence length, $\mathcal{I}$ is the instruction prompt, and θ denotes the learnable parameters.

3 Experiments

3.1 Datasets

We evaluate on CoVoST-2 (Wang et al., 2020) and BMELD (Liang et al., 2021). CoVoST-2 is a large-scale multilingual ST corpus, from which we use 26 language pairs, utilizing the N-best ST hypotheses provided in the CoVoGER benchmark (Yang et al., 2025b) and decoding N-best ASR transcriptions from the same audio using Whisper MEDIUM and LARGE in ASR mode. BMELD is an emotion-aware English-Chinese ST dataset with emotion and sentiment labels, which are incorporated into the LLM instructions during fine-tuning. For all datasets, N-best ASR transcriptions are decoded using the same mixture decoding strategy. Dataset statistics are provided in Appendix A.

3.2 Experimental Settings

3.2.1 N-best Hypothesis Generation

ASR N-best Transcriptions. We decode N-best ASR transcriptions using Whisper (Radford et al., 2023) in ASR mode under the same mixture decoding strategy as CoVoGER (Yang et al., 2025b): the 1-best hypothesis via beam search and four additional hypotheses via temperature sampling, yielding a 5-best list per segment. We use MEDIUM and LARGE Whisper variants throughout.

ST N-best Hypotheses. The N-best ST hypotheses are taken directly from the CoVoGER benchmark (Yang et al., 2025b), decoded from SeamlessM4T (Barrault et al., 2023) using mixture decoding strategy across both MEDIUM and LARGE model sizes. We design four conditions to assess the contribution of ASR transcriptions to GER quality:

(i) **ST-only (no GER)**: Direct 1-best ST output, serving as the lower bound.

(ii) **GER w/ N-best ST**: LLM conditioned solely on N-best ST hypotheses (Yang et al., 2025b,a), serving as our primary baseline.

(iii) **GER w/ 1-best ASR + N-best ST**: LLM jointly conditioned on a_1 and $\mathcal{H}$, isolating

source language grounding without transcription diversity.

(iv) **GER w/ N-best ASR + N-best ST**: LLM jointly conditioned on $\mathcal{A}$ and $\mathcal{H}$, assessing whether ASR diversity adds further benefit.

Comparing (ii)–(iii) isolates source grounding; comparing (iii)–(iv) isolates ASR diversity. Qwen2.5-7B (Hui et al., 2024) is fine-tuned with LoRA (Hu et al., 2022) via LitGPT[1] using rank r=8, scaling α=16, and dropout 0.05, following Yang et al. (2025b) for fair comparison. Training runs for 25,000 iterations with an effective batch size of 64 on a single H100 GPU. Models are fine-tuned separately on the training splits of CoVoST-2 and BMELD, with reference translations as target outputs. Translation quality is measured with SacreBLEU (Post, 2018).

3.3 Results

We evaluate our proposed framework across two benchmarks: CoVoST-2 for multilingual speech translation and BMELD for emotion-aware speech translation. Tables 1 and 2 report BLEU scores across 13 language pairs in both X-En and En-X directions. The ASR quality of the Whisper Medium and Large models used for source-language grounding in Appendix D. In the X-En direction, GER with ASR grounding consistently outperforms both the ST-only baseline and the transcription-free GER (ST only) setting across model sizes. The benefit is most pronounced for morphologically rich or typologically distant languages such as Estonian, Latvian, and Turkish, where surface-form grounding from the source transcript appears to compensate for higher ambiguity in the acoustic signal. Gains are comparatively modest for language pairs already well-served by the ST backbone, such as Welsh (Cy-En) and Indonesian (Id-En). Using N-best ASR hypotheses instead of a single-best transcript yields marginal further improvement on average, suggesting that source-language grounding itself is the primary driver of performance rather than transcription diversity.

A similar pattern holds in the En-X direction, where both GER variants with ASR conditioning outperform the transcription-free GER baseline. The advantage is especially visible for morphologically complex target languages (Estonian, Latvian) and logographic scripts (Japanese, Chinese), consistent with the hypothesis that explicit source conditioning helps the LLM produce more faithful target-side surface forms. Notably, Medium-size GER with ASR grounding already approaches or exceeds the Large GER (ST only) baseline in several language pairs, pointing to the efficiency of source-language conditioning as a substitute for raw model capacity.

Size	E/S	Settings	BLEU
M	–	ST only	11.50
M	✓	GER (ST only)	13.09
M	✓	GER (1-ASR+ST)	$20.42^{\dagger\ddagger}$
M	✓	GER (N-ASR+ST)	$19.01^{\dagger\ddagger}$
L	–	ST only	11.87
L	✓	GER (ST only)	15.88
L	✓	GER (1-ASR+ST)	$\mathbf{20.99}^{\dagger\ddagger}$
L	✓	GER (N-ASR+ST)	$19.46^{\dagger\ddagger}$

Table 3: BLEU results on BMELD (English→Chinese). M and L denote Medium and Large model sizes for both the ASR (Whisper) and ST (SeamlessM4T) models, respectively. The GER model used is Qwen2.5-7B. Bold indicates the highest BLEU score. † and ‡ indicate statistically significant improvement over the same-size ST only and GER (ST only) baselines respectively at $p < 0.05$.

Table 3 reports results on English→Chinese emotion-aware ST. For Qwen2.5-7B, the GER (1-ASR+ST) pipeline achieves SacreBLEU scores of 20.99 and 20.42 using Medium and Large ASR models, respectively. As illustrated in Table 8 (see Appendix E for full qualitative analysis), incorporating N-best hypotheses from both ASR and ST enables GER to recover finer-grained details such as numeral form and natural expression that are lost when relying on a single hypothesis. Following Yang et al. (2025a), we additionally evaluate Llama-2-7B on BMELD to maintain comparability with prior work, results are reported in Appendix B.

4 Conclusion

In this work, we introduced a GER framework for LLM-based speech translation correction that jointly leverages N-best ASR transcriptions in the source language alongside N-best ST hypotheses, extending existing approaches that condition solely on target-side hypotheses. Through systematic experimentation across 1-best and N-best ASR settings, we examined the role of transcription diversity in correction quality. Evaluated on CoVoST-2 and BMELD, our framework consistently improves translation quality over ST-only baselines across diverse language pairs in most settings.

[1] https://github.com/Lightning-AI/litgpt

Limitations

While our proposed framework demonstrates consistent improvements over ST-only baselines, several limitations should be noted. First, we decode ASR transcriptions using Whisper Medium and Whisper Large only, and the impact of other ASR systems on GER quality remains unexplored. Second, emotion and sentiment labels are incorporated as plain text in the prompt, and more sophisticated integration strategies such as learned embeddings or cross-attention mechanisms may yield further improvements. Finally, due to computational constraints, we do not evaluate larger model variants for GER, which may show different trends from the fine-tuned results reported here.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: Massively multilingual and multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Tongzhou Chen, Cyril Allauzen, Yinghui Huang, Daniel Park, David Rybach, W Ronny Huang, Rodrigo Cabrera, Kartik Audhkhasi, Bhuvana Ramabhadran, Pedro J Moreno, and 1 others. 2023. Large-scale language model rescoring on long-form data. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and EnSiong Chng. 2024. Large language models are efficient learners of noise-robust speech recognition. *arXiv preprint arXiv:2401.10446*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Iwen E Kang, Christophe Van Gysel, and Man-Hung Siu. 2024. Transformer-based model for asr n-best rescoring and rewriting. *arXiv preprint arXiv:2406.08207*.

Yunlong Liang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Modeling bilingual conversational characteristics for neural chat translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5711–5724.

Hsi-Che Lin, Yi-Cheng Lin, Huang-Cheng Chou, and Hung-yi Lee. 2025. Improving speech emotion recognition in under-resourced languages via speech-to-speech translation with bootstrapping data selection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Bingshen Mu, Xucheng Wan, Naijun Zheng, Huan Zhou, and Lei Xie. 2024. Mmger: Multi-modal and multi-granularity generative error correction with llm for joint accent and speech recognition. *IEEE Signal Processing Letters*, 31:1940–1944.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the third conference on machine translation: Research papers*, pages 186–191.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing asr pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7909–7913. IEEE.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*.

Liyan Xu, Yile Gu, Jari Kolehmainen, Haidar Khan, Ankur Gandhe, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko. 2022. Rescorebert: Discriminative speech recognition rescoring with bert. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 6117–6121. IEEE.

Zhengdong Yang, Sheng Li, and Chenhui Chu. 2025a. Generative error correction for emotion-aware speech-to-text translation. In *Findings of*

the Association for Computational Linguistics: ACL 2025, pages 20413–20421.

Zhengdong Yang, Zhen Wan, Sheng Li, Chao-Han Huck Yang, and Chenhui Chu. 2025b. Covoger: A multilingual multitask benchmark for speech-to-text generative error correction with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6313–6325.

A Dataset Statistics

We evaluate on two datasets (Tables 5 and 4). BMELD is a bilingual English–Chinese extension of MELD (*Friends* TV series), annotated with emotion and sentiment labels, comprising 9,987 training, 1,084 validation, and 2,601 test samples. CoVoST-2 is a large-scale multilingual ST benchmark covering 21 translation directions, yielding over 4 million training instances across all language pairs.

Table 4: Number of N-best lists decoded from CoVoST 2. "En-X" aggregates all 13 English→X directions.

Language Pair	Train	Validation	Test
En-X	$13 \times 289{,}392$	$13 \times 15{,}520$	$13 \times 15{,}526$
Ar-En	1,832	1,587	1,695
Ca-En	95,854	12,730	12,730
Cy-En	937	184	690
De-En	127,824	13,511	13,511
Et-En	1,782	1,576	1,571
Fa-En	51,423	782	3,445
Id-En	928	792	844
Ja-En	1,119	635	684
Lv-En	2,337	1,125	1,629
Sl-En	1,843	509	360
Sv-En	2,157	1,349	1,595
Tr-En	3,494	731	1,629
Zh-En	7,085	4,843	4,898
Total	**4,060,711**	**247,119**	**263,431**

Table 5: BMELD (en-zh) statistics per emotion and sentiment label.

Label	Train	Dev	Test
Emotion			
Neutral	4,709	460	1,251
Joy	1,743	162	400
Sadness	682	109	208
Fear	268	40	50
Anger	1,109	146	345
Surprise	1,205	146	279
Disgust	271	21	68
Sentiment			
Neutral	4,709	460	1,251
Positive	2,334	231	518
Negative	2,944	393	832
Total	**9,987**	**1,084**	**2,601**

B Llama-2-7B Results on BMELD

Table 3 presents results on BMELD (English→Chinese) using Llama-2-7B as the GER model. Incorporating ASR transcriptions alongside ST hypotheses yields consistent improvements over the ST-only baseline, with further gains observed when expanding to N-best hypotheses. The Large model variants outperform their Medium counterparts in the ST-only setting, though this gap narrows once ASR hypotheses are incorporated, suggesting that GER effectively compensates for weaker acoustic model capacity.

Table 6: Results on BMELD (English→Chinese). M and L denote Medium and Large model sizes for Whisper (ASR) and SeamlessM4T (ST). The GER model is Llama-2-7B.

Size	Settings	E/S	BLEU
M	GER (ST only)	✓	13.08
M	GER (1-ASR+ST)	✓	17.18
M	GER (N-ASR+ST)	✓	17.63
L	GER (ST only)	✓	16.28
L	GER (1-ASR+ST)	✓	17.42
L	GER (N-ASR+ST)	✓	17.86

C Instruction Prompts

The following are the instruction prompts used for GER fine-tuning in our experiments. We use three prompt variants: **GER (ST only)**, which uses only speech translation hypotheses; **GER (1-ASR+ST)**, which incorporates a single ASR transcription with translation hypotheses; and **GER (N-ASR+ST)**, which incorporates N-best ASR transcriptions together with translation hypotheses.

GER (N-ASR+ST)

Instruction:
You are given N-best Chinese translations for the same English speech segment. Your task is to output one corrected translation that best matches the intended speech.

Input:

N-best Translations:
1. 我的职责吗? (笑声)
2. 我的职责? (笑声)
3. 我的职责吗？好吧
4. 我的职责? (笑)
5. 我的职责

Output:
我的职责？好吧。

GER (1-ASR+ST)

Instruction:
You are given an ASR transcription and N-best Chinese translations for the same English speech segment. Your task is to output one corrected translation that best matches the intended speech. Use the ASR transcription to resolve ambiguity and prefer information consistent across hypotheses.

Input:

ASR Transcription (source speech):
My duties? All right.

N-best Translations:
1. 我的职责吗? (笑声)
2. 我的职责? (笑声)
3. 我的职责吗？好吧
4. 我的职责? (笑)
5. 我的职责

Output:
我的职责？好吧。

GER (N-ASR+ST)

Instruction:
You are given N-best transcriptions and N-best Chinese translations for the same English speech segment. Your task is to output one corrected translation that best matches the intended speech. Use the ASR transcriptions to resolve ambiguity and prefer information consistent across hypotheses.

Input:

ASR Transcription (source speech):
1. My duties? All right.
2. My duty? All right?
3. My duties all right.
4. My duties all rite!
5. My duties all riot

N-best Translations:
1. 我的职责吗? (笑声)
2. 我的职责? (笑声)
3. 我的职责吗？好吧
4. 我的职责? (笑)
5. 我的职责

Output:
我的职责？好吧。

D ASR Transcription Quality

Table 7 reports the Word Error Rate (WER) of Whisper Medium and Large across all source languages in the CoVoST-2 dataset.

E Qualitative Analysis

Table 8 shows a BMELD (En→Zh) example illustrating the benefit of joint ASR-ST grounding. Without ASR input, GER (ST only) mistranslates the number, producing a verbose written-form equivalent of "twenty-seven", while GER (N-ASR+ST) correctly recovers the compact numeric form "No. 27", closely matching the reference. This confirms that ASR grounding helps the LLM resolve numeric and stylistic ambiguities that ST hypotheses alone maynot.

Table 7: WER (%) of Whisper Medium (M) and Large (L) across source languages in CoVoST-2. Lower is better. *CER is reported for character-level languages (ja, zh).

Language	**M**	**L**
Arabic (ar)	27.08	16.44
Catalan (ca)	13.60	08.62
Welsh (cy)	40.30	33.10
German (de)	10.44	06.28
English (en)	13.47	11.41
Estonian (et)	38.45	28.24
Persian (fa)	55.40	42.56
Indonesian (id)	08.88	05.11
Japanese (ja)*	24.25	17.47
Latvian (lv)	31.40	19.60
Slovenian (sl)	20.65	11.96
Swedish (sv)	10.80	06.52
Turkish (tr)	19.05	14.87
Chinese (zh)*	22.88	13.83

*CER for character-level languages (ja, zh).

Emotion: Anger | Sentiment: Negative *Src: "That way if one of them is missing, Monica can be like: Where's No. 27?"*	
ASR (N-best)	1. That way if any of them goes missing, Monica could be like, where is twenty-seven? 4. That way if one of them is missing, Monica can say, where is No. 27?
ST (N-best)	1. 这样的话，如果其中一个不见了，莫妮卡可能会问：第二十七号去哪了? (*That way, if one of them is missing, Monica might ask, where did number twenty-seven go?*) 4. 这样的话，如果其中一个丢失了，莫妮卡会说：27号在哪里? (*That way, if one gets lost, Monica would say, where is No. 27?*)
GER (ST only)	这样的话，如果其中一个不见了，莫妮卡可能会问：第二十七号去哪了? (*That way, if one is missing, Monica might ask: where did number twenty-seven go?*)
GER (N-ASR+ST)	这样的话，如果他们中有一个不见了，莫妮卡会说："**27**号在哪里？" (*That way, if one of them is missing, Monica can be like: "Where is No. 27?"*)

Table 8: Qualitative example from BMELD (En→Zh). ASR hypotheses are in English (source language); ST hypotheses and GER outputs are in Chinese (target language).