CoVoGER: A Multilingual Multitask Benchmark for Speech-to-text Generative Error Correction with Large Language Models

Zhengdong Yang[1], **Zhen Wan**[1], **Sheng Li**[2], **Chao-Han Huck Yang**[3], **Chenhui Chu**[1]
[1]Kyoto University [2]Institute of Science Tokyo [3]NVIDIA Research
`{zd-yang,zhenwan}@nlp.ist.i.kyoto-u.ac.jp` `li.s.az@m.titech.ac.jp`
`hucky@nvidia.com` `chu@i.kyoto-u.ac.jp`

Abstract

Large language models (LLMs) can rewrite the N-best hypotheses from a speech-to-text model, often fixing recognition or translation errors that traditional rescoring cannot. Yet research on generative error correction (GER) has been focusing on monolingual automatic speech recognition (ASR), leaving its multilingual and multitask potential underexplored. We introduce CoVoGER, a benchmark for GER that covers both ASR and speech-to-text translation (ST) across 15 languages and 28 language pairs. CoVoGER is constructed by decoding Common Voice 20.0 and CoVoST-2 with Whisper of three model sizes and SeamlessM4T of two model sizes, providing 5-best lists obtained via a mixture of beam search and temperature sampling. We evaluated various instruction-tuned LLMs, including commercial models in zero-shot mode and open-sourced models with LoRA fine-tuning, and found that the mixture decoding strategy yields the best GER performance in most settings. CoVoGER will be released to promote research on reliable language-universal speech-to-text GER. The code and data for the benchmark are available at `https://github.com/N-Orien/CoVoGER`.

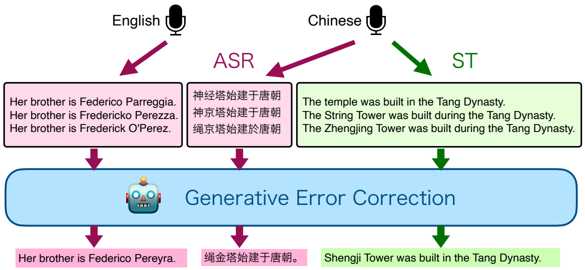
English
Chinese
ASR
ST
Her brother is Federico Parreggia.
Her brother is Fredericko Perezza.
Her brother is Frederick O'Perez.
神经塔始建于唐朝
神京塔始建于唐朝
绳京塔始建於唐朝
The temple was built in the Tang Dynasty.
The String Tower was built during the Tang Dynasty.
The Zhengjing Tower was built during the Tang Dynasty.
Generative Error Correction
Her brother is Federico Pereyra.
绳金塔始建于唐朝。
Shengji Tower was built in the Tang Dynasty.

Figure 1: An example of a multilingual multitask GER system. Boxes above the GER model represent N-best lists generated by first-pass models, and boxes below the GER model represent the model's predictions of the corrected sentences.

1 Introduction

Automatic speech recognition (ASR) and speech-to-text translation (ST) systems (Zue, 1985; Ney, 1999) are increasingly deployed in real-world applications, from voice assistants and captioning services to cross-lingual communication tools. However, even state-of-the-art models can produce *transcription errors*, especially under noisy conditions or with accented speech. These errors often lead to miscommunication, which leads to a growing need for methods to correct such ASR/ST errors on the fly. Recent advances in large language models (LLMs) (Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023; Bai et al., 2023) offer a promising new pathway to tackle this challenge: leveraging LLMs to revise or repair the initial output of speech-to-text systems, thereby enhancing both the accuracy and the cognitive readability of the transcribed content.

Generative error correction (GER) (Chen et al., 2023a; Yang et al., 2023) has emerged as a new paradigm to leverage LLMs for refining speech outputs. Unlike traditional rescoring (Xu et al., 2022; Udagawa et al., 2022; Chen et al., 2023b), which merely re-rank the existing hypotheses in an N-best list, GER approaches utilize LLM to generate an improved final transcription. This approach enables the LLM to aggregate evidence from multiple hypotheses and leverage its linguistic knowledge and contextual reasoning to correct errors, marking a transition toward active, generative correction within a multi-pass voice-agentic[1] system (*i.e.*, an ASR/ST agent followed by an LLM agent).

However, most GER studies concentrate exclusively on English (Yang et al., 2023; Hu et al., 2024a; Ghosh et al., 2024). Non-English (Udagawa et al., 2024; Robatian et al., 2025) and multilingual (Hu et al., 2024b) variants are emerging, but coverage remains fragmentary and lacks a unified

[1] We refer to the recent chained voice agents setup: `https://platform.openai.com/docs/guides/voice-agents`.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6302–6314
November 4-9, 2025 ©2025 Association for Computational Linguistics

evaluation framework. Furthermore, most studies address ASR and ST tasks in isolation, overlooking their well-known synergies. It therefore remains an open question whether GER can benefit from joint training across both speech-to-text tasks.

Meanwhile, the first-pass decoding setup remains underexplored. GER performance hinges on the quality of the N-best lists, which depend on both generation methods and upstream models. Yet most prior work relies solely on beam search with a single first-pass model to produce these lists (Chen et al., 2023a; Hu et al., 2024b), while alternatives like temperature sampling and a systematic analysis of decoding choices are largely absent.

To address these research gaps in speech-to-text research, we make the following contributions:

- We propose **CoVoGER**, the first benchmark for GER that spans multiple languages and multiple speech-to-text tasks (ASR and ST), evaluating GER models' capabilities shown in Figure 1.
- A systematic investigation of first-pass decoding setups has been introduced, which includes decoding strategies and model sizes. Our results uncover the impact on GER performance and motivate a new approach blending beam search with temperature sampling.
- We conduct extensive experiments with various LLMs in both zero-shot and parameter-efficient fine-tuning (PEFT) settings for the benchmark to highlight potential trade-offs.
- Public release of the reproducible CoVoGER benchmark and dataset will foster further research and development of multilingual speech-to-text GER methods.

2 Related Work

Yang et al. (2023) first introduced this generative modeling idea in GER-based ASR, directly rewriting an N-best list rather than selecting a single hypothesis, which also prompted LLMs with instructions and demonstrated that minimal fine-tuning closes most of the gap to oracle WER. Chen et al. (2023a) later formalized HYPORADISE, showing that prompting GPT-style models with N=5 hypotheses can significantly reduce English WER by discovering up to N=20 beam size.

Recent multilingual work by Li et al. (2024) tackles over 100 languages through fine-tuning a single LLM. Their model corrects grammar and spelling and even hallucinates missing words via cross-script transfer. However, the input is limited to only a single hypothesis, leaving the richer N-best setting and its potential diversity untouched.

The GER paradigm has been initialized to speech translation (ST) or agentic setups (Cheng et al., 2024). For instance, (Hu et al., 2024b) focuses solely on multilingual ST. To our knowledge, no existing dataset simultaneously covers both ASR and ST across multiple languages. **CoVoGER** bridges this gap by supporting 15 ASR languages and 28 source–target ST directions, yielding 40M N-best lists for a compact multilingual evaluation.

For the investigation on generating N-best lists, ProGRes (Tur et al., 2024) prompts an LLM to produce additional transcription hypotheses based on the ASR's N-best outputs, but leaves the first-pass decoding setups unchanged. As a study close to ours, Ma et al. (2025) varies the size of the ASR model for GER, yet it still relies on the 1-best input, without analysis of different decoding strategies. Although there are studies in the text-generation community (Shen et al., 2022) that investigate decoding strategies such as beam search and temperature sampling, no similar exploration has been conducted for GER. **CoVoGER** fills this gap by generating 5-best lists using three Whisper model sizes and two SeamlessM4T model sizes, comparing beam, sampling, and a mixture of both, and quantifying their impact on GER.

3 Generative Error Correction

3.1 Task Formulation

Given an utterance's N-best list $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$ produced by a *first-pass* speech–to–text model (either ASR or ST), GER seeks a mapping $f : \mathcal{H} \to \hat{\mathbf{y}}$ such that the generated sequence $\hat{\mathbf{y}} = f(h_1, \ldots, h_N)$ is closer to the reference transcription/translation $\mathbf{y}$ than any $h_i \in \mathcal{H}$.

3.2 Learning the Mapping f

During training, we minimize a sequence–to–sequence loss over the reference data:

$$\mathcal{L} = -\sum_{t=1}^{|\mathbf{y}|} \log p_\phi(y_t \mid y_{<t}, \mathcal{H}), \qquad (1)$$

where p_ϕ is the conditional distribution parameterized by an LLM with parameters ϕ. In practice, we

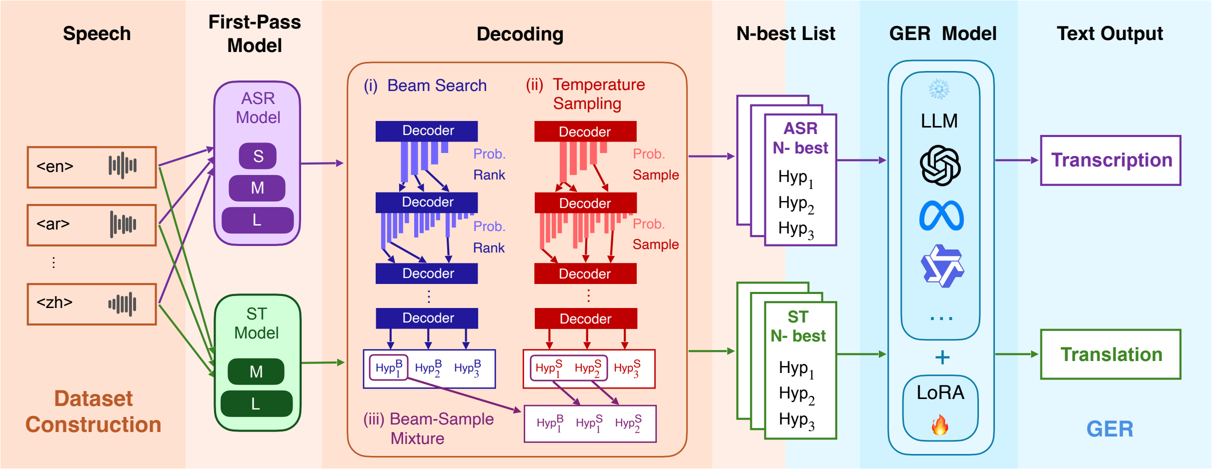

Figure 2: Overview of the CoVOGER benchmerk. To construct the dataset, speech from multiple languages is transcribed or translated by several first-pass models with different model sizes. The process can be conducted with various decoding strategies, including beam search, temperature sampling, or a mixture of both. The resulting ASR and ST 5-best lists (the figure only shows 3-best for presentation) are used as inputs to train and evaluate GER models in our benchmark.

adopt token-level targets with the standard cross-entropy loss.

3.3 Low-Rank Adaptation

To adapt large LLMs without updating all weights, we employ the PEFT method Low-Rank Adaptation (LoRA) (Hu et al., 2022). LoRA freezes the pretrained weights and injects rank-r update matrices $\mathbf{A}, \mathbf{B}$ into each attention projection:

$$\begin{aligned} \mathbf{W}_{\text{Q}} &\leftarrow \mathbf{W}_{\text{Q}} + \alpha\, \mathbf{B}_{\text{Q}} \mathbf{A}_{\text{Q}}, \\ \mathbf{B}_{\text{Q}} \in \mathbb{R}^{d \times r}, &\qquad \mathbf{A}_{\text{Q}} \in \mathbb{R}^{r \times d}. \end{aligned} \tag{2}$$

Only $\mathbf{A}, \mathbf{B}$ (and layer-norm biases) are trained, greatly reducing the number of updated parameters while preserving the forward pass latency of the base LLM. The LoRA-augmented model is optimized with the same loss as Eq. (1).

4 CoVoGER

In this section, we present **CoVoGER**, with an overview illustrated in Figure 2.

4.1 Source Speech Datasets

To construct the CoVoGER benchmark, we decode speech from two large-scale public datasets, Common Voice 20.0[2] for ASR and CoVoST-2 (Wang et al., 2021) for ST. Tables 1 and 2 summarize the amount of N-best lists decoded from these two datasets.

CoVoST-2 A multilingual ST corpus derived from Common Voice. We select 14 non-English source languages and pair each of them bidirectionally with English, yielding 28 ST directions.[3] For each direction, we keep the official train/validation/test splits.

Common Voice 20.0 To provide substantially larger ASR training data in the same language set, we extract utterances of the 15 languages (the 14 above plus English) from Mozilla Common Voice 20.0. We also adopt the dataset's original train/validation/test splits.

Because some speech segments appear in both datasets, we filter the data to prevent leakage: any utterance in one dataset's validation or test split is removed from the other dataset's training split, and any utterance in one's test split is deleted from the other's validation split.

4.2 First–Pass ASR and ST Models

The N-best lists used in CoVoGER are generated with two state-of-the-art, publicly available foundation models: **Whisper** (Radford et al., 2023) for ASR and **SeamlessM4T** (Barrault et al., 2023) for ST. For each model family, we select various model sizes to study how first-pass model performance and hypothesis diversity influence downstream GER. As GER may compensate for weaker ASR/ST models, comparing different first-pass

[2] https://commonvoice.mozilla.org/en/datasets

[3] We exclude Mongolian because the first-pass model (Whisper) exhibits an error rate exceeding 100%, making N-best generation unreliable.

	Train	Validation	Test
Ar	28,524	10,405	10,497
Ca	1,172,032	15,148	16,412
Cy	8,000	5,392	5,399
De	583,678	11,061	16,191
En	1,108,326	9,871	16,398
Et	3,128	2,421	2,807
Fa	29,422	10,625	10,629
Id	4,973	3,210	3,690
Ja	14,477	7,766	7,786
Lv	13,870	7,536	7,578
Sl	1,448	1,216	1,328
Sv	7,419	4,744	5,345
Ta	46,095	12,067	12,203
Tr	38,992	11,645	11,660
Zh	25,231	8,478	10,630
Total	3,085,615	121,585	138,553

Table 1: Number of N-best lists decoded from ASR dataset Common Voice 20.0 with one decoding setup.

	Train	Validation	Test
En-X	14 × 289,392	14 × 15,520	14 × 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
Ta–En	815	273	786
Tr–En	3,494	731	1,629
Zh–En	7,085	4,843	4,898
Total	4,350,918	257,907	263,431

Table 2: Number of N-best lists decoded from ST dataset CoVoST 2 with one decoding setup. The "En-X" row aggregates the 14 English → X directions.

model sizes quantifies how much baseline accuracy remains necessary after correction. In addition, smaller models may yield more diverse hypotheses, potentially benefiting GER even if their 1-best accuracy is lower.

Whisper A multilingual encoder–decoder model trained on 680k hours of web-scale speech. We adopt three released models[4]—SMALL, MEDIUM, and LARGE to decode Common Voice 20.0 (Table 1).

SeamlessM4T A massively multilingual model that unifies ASR, S2T, T2T, and S2S in a single architecture. We use the MEDIUM and LARGE models[5] to decode the 28 CoVoST-2 directions (Table 2).

Task	Model	Parameters
ASR	Whisper$_{\text{small}}$	244 M
ASR	Whisper$_{\text{medium}}$	769 M
ASR	Whisper$_{\text{large}}$	1.55 B
ST	SeamlessM4T$_{\text{medium}}$	1.2 B
ST	SeamlessM4T$_{\text{large}}$	2.3 B

Table 3: First-pass speech models used to generate N-best lists for GER.

Table 3 summarizes the models and parameter counts used throughout our experiments.

4.3 First-Pass Decoding Strategies

A first-pass model $p_\theta(\mathbf{y} \mid \mathbf{x})$ generates an N-best list $H = \{h_1, \ldots, h_N\}$ that is later fed to the GER model. We examine two complementary decoding schemes, *beam search* and *temperature sampling*, and finally combine them to obtain a diverse yet accurate hypothesis set.

Beam search. Beam search heuristically approximates the maximum a posteriori sequence

$$\begin{aligned} h^\star &= \arg\max_{\mathbf{y}} p_\theta(\mathbf{y} \mid \mathbf{x}) \\ &= \arg\max_{\mathbf{y}} \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, \mathbf{x}) \,, \end{aligned} \tag{3}$$

by expanding the B highest-scoring partial candidates at each time-step t, and retaining only the top B of their continuations. After termination, we collect the N highest-scoring finished hypotheses $H^{\text{beam}} = \{h_1^{\text{beam}}, \ldots, h_N^{\text{beam}}\}$ ranked by length-normalized log-probability (Freitag and Al-Onaizan, 2017). Although beam search produces high-probability outputs, its top hypotheses often differ only slightly (e.g., by punctuation or function words), leading to low diversity in H^{beam}.

Temperature sampling. To increase lexical and structural variety, we also draw hypotheses from a tempered categorical distribution

$$p_\tau(y_t \mid y_{<t}, \mathbf{x}) = \frac{p_\theta(y_t \mid y_{<t}, \mathbf{x})^{1/\tau}}{\sum_w p_\theta(w \mid y_{<t}, \mathbf{x})^{1/\tau}}, \tag{4}$$

with temperature $\tau > 0$ (Holtzman et al., 2019). Lower τ sharpens the distribution (less randomness), while higher τ flattens it, yielding more diverse but potentially less accurate sequences. By independently sampling until an EOS token, we obtain

$$H^{\text{sample}} = \{h_1^{\text{sample}}, \ldots, h_N^{\text{sample}}\}, \tag{5}$$

[4] https://github.com/openai/whisper

[5] https://github.com/facebookresearch/seamless_communication

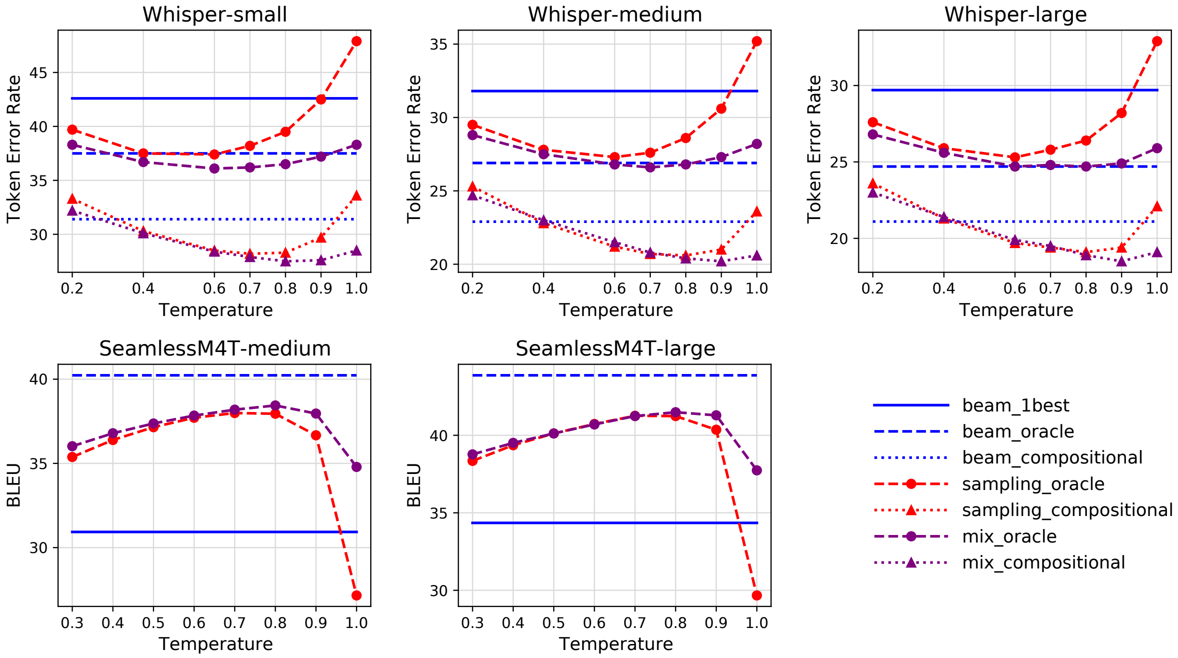

Figure 3: Average validation performance of beam, sampling, and beam–sampling mixture decoding at various temperatures.

where diversity arises naturally from stochastic choice at each step.

Beam–sampling mixture. Pure sampling may degrade 1-best accuracy, whereas pure beam search offers little variety. We therefore construct a *mixed* list

$$H^{\text{mix}} = \{h_1^{\text{beam}}\} \cup \{h_j^{\text{sample}} \,|\, j = 1, \dots, N-1\}, \quad (6)$$

retaining the highest-probability beam output for reliability and filling the remaining $N-1$ slots with temperature samples for diversity.

4.4 Optimising Sampling Temperature

Beam search (§4.3) returns exactly N hypotheses, whereas sampling requires choosing a *temperature* τ. We fix the list length to $N = 5$ for every utterance: (i) **beam** keeps the top–5 sequences, (ii) **sampling** draws 5 independent samples at temperature τ, and (iii) **mix** takes the 1-best beam hypothesis plus 4 temperature samples.

Evaluation metrics. Unlike the conventional practice of stripping punctuation, we retain all symbols so that the GER model can learn to correct fully-formatted ASR. Consequently, we measure **Token Error Rate (TER)** with a standard SacreBLEU (Post, 2018) tokenizer for ASR data. To assess the upper bound of an N-best list we report: (i) *oracle TER*, the TER of the single hypothesis $h_i \in H$ with lowest TER, and (ii) *compositional oracle TER* (Chen et al., 2023a), which greedily composes a new hypothesis by selecting from any h_i token by token, so as to minimize TER against the reference. For ST, we compute **oracle BLEU**: selecting the best hypothesis per utterance in terms of sentence-level BLEU, and calculating corpus-level BLEU.

Temperature optimization. Figure 3 shows validation results across different temperatures. ASR scores are the average of all 15 languages, and ST scores are the average of all 28 language pairs. We can observe that:

- Unsurprisingly, larger models generally yield lower TER and higher BLEU.
- The mixture strategy consistently beats pure sampling in oracle metrics for both tasks.
- Mixture outperforms beam search in oracle metrics for ASR but not for ST. However, later experiments still show that mixture is able to beat beam search on ST, which indicates that oracle BLEU may not be the best metric to estimate the N-best list quality for GER on ST.
- A general trend is that smaller models favor mixture decoding and larger models favor

N-best	GER	Ar	Ca	Cy	De	En	Et	Fa	Id	Ja	Lv	Sl	Sv	Tr	Zh	AVG
S+Beam	–	59.7	28.3	64.6	15.6	19.7	65.5	84.9	28.3	**40.0**	**55.1**	41.2	22.8	26.7	35.0	42.0
S+Beam	Qwen	63.8	**19.0**	64.5	**11.6**	16.5	65.3	77.0	19.7	49.8	55.2	**36.7**	**21.7**	**22.9**	17.0	38.6
S+Sample	Qwen	62.4	19.7	67.5	12.2	16.2	67.0	79.7	18.2	52.3	57.8	40.4	23.6	25.3	16.9	39.9
S+Mix	Qwen	**58.0**	19.2	**62.9**	**11.6**	**15.8**	**63.7**	**75.0**	**18.1**	46.1	**55.1**	40.3	**21.7**	23.7	**16.4**	**37.7**
M+Beam	–	**50.1**	21.1	**42.1**	10.2	16.2	**45.1**	56.8	18.8	**34.9**	**37.5**	30.2	15.4	20.8	28.5	30.6
M+Beam	Qwen	51.8	13.6	43.5	8.4	**13.0**	46.8	56.0	**12.4**	37.4	38.3	28.8	**14.6**	18.2	**13.7**	**28.3**
M+Sample	Qwen	55.7	14.9	47.8	8.9	13.6	49.4	58.1	14.1	49.3	40.2	29.3	16.4	21.3	14.0	30.9
M+Mix	Qwen	54.5	**13.5**	43.1	**8.3**	13.1	47.2	**55.1**	12.9	49.7	38.1	**27.9**	14.8	**18.0**	14.0	29.3
L+Beam	–	**48.4**	19.4	**38.6**	9.5	15.7	**41.4**	54.1	18.1	**33.0**	**35.1**	24.3	14.3	19.7	30.2	28.7
L+Beam	Qwen	50.1	**12.4**	40.1	**7.7**	**12.5**	44.6	55.2	12.8	39.7	35.4	22.7	**13.7**	**16.2**	**12.1**	**26.8**
L+Sample	Qwen	59.3	13.5	44.6	8.3	13.3	46.0	54.5	13.2	42.6	38.7	24.5	15.3	18.5	13.4	29.0
L+Mix	Qwen	57.1	12.5	41.3	7.8	**12.5**	44.6	**54.0**	**12.4**	44.8	36.4	**22.3**	14.3	17.1	12.5	27.8

Table 4: TER results of ASR GER with different first-pass decoding setups on test set. GER models are all LoRA fine-tuned on single task. "S," "M" and "L" stand for "small," "medium" and "large." The "AVG" column presents the average scores across all the languages.

N-best	GER	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	X–En
M+Beam	–	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+Beam	Qwen	**44.33**	**36.22**	45.75	**37.06**	**22.86**	**24.75**	**53.52**	19.88	26.49	**33.55**	35.78	**30.15**	**20.23**	**33.12**
M+Sample	Qwen	42.79	35.55	44.07	35.94	21.75	23.90	51.89	19.39	26.57	31.60	34.85	28.81	15.36	31.73
M+Mix	Qwen	43.96	36.15	45.50	36.72	22.54	24.34	52.40	**21.07**	**26.96**	31.83	**35.84**	29.13	18.75	32.71
L+Beam	–	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+Beam	Qwen	47.16	38.24	45.35	**39.20**	**26.73**	25.60	53.12	**22.73**	**31.48**	37.88	40.98	**32.13**	**20.94**	35.50
L+Sample	Qwen	46.60	37.46	53.12	38.19	25.97	25.54	53.21	21.61	31.30	38.40	40.24	31.01	19.40	35.54
L+Mix	Qwen	**47.92**	38.22	49.02	38.82	26.34	**25.64**	**54.00**	22.25	31.08	**38.74**	**41.25**	31.79	20.67	**35.85**

N-best	GER	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	En–X	AVG
M+Beam	–	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	29.90
M+Beam	Qwen	**22.98**	38.83	31.21	**33.89**	24.03	17.24	**37.27**	29.25	19.27	29.80	38.66	20.26	43.50	29.71	**31.42**
M+Sample	Qwen	22.29	38.02	29.81	33.20	22.85	16.51	36.59	30.70	18.21	28.58	38.35	19.18	43.68	29.07	30.40
M+Mix	Qwen	22.72	38.67	30.85	33.76	24.23	16.95	37.11	**30.75**	19.12	29.71	39.03	19.92	**44.32**	**29.78**	31.25
L+Beam	–	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	33.50
L+Beam	Qwen	25.18	39.70	32.87	36.01	27.30	18.29	38.70	32.55	21.54	33.72	41.03	22.26	46.67	31.99	33.75
L+Sample	Qwen	24.53	39.94	32.35	35.40	26.33	17.81	38.42	31.85	20.26	32.77	41.03	21.65	46.34	31.44	33.49
L+Mix	Qwen	**25.30**	40.86	33.38	36.09	27.71	18.30	39.12	**32.59**	21.48	34.06	41.90	22.34	**46.97**	**32.32**	**34.09**

Table 5: BLEU results of ST GER with different first-pass decoding setups on test set. GER models are all LoRA fine-tuned on single task. "M" and "L" stand for "medium" and "large." Columns "X-En," "En-X," and "AVG" present the average scores across any-to-English, English-to-any, and all the language pairs, respectively.

beam search, with sampling in the middle.

- For ASR, mixture and sampling have a larger advantage over beam on compositional TER than on oracle TER. In addition, the optimal τ for the compositional oracle is slightly higher than for the plain oracle. These observations likely suggest that compositional oracle favors diversity for compositional possibilities more than the accuracy of a single hypothesis.

Based on overall observation, we therefore set both τ^{ASR} and τ^{ST} to 0.8 for the following experiments.

5 Experimental Setups

5.1 GER Models

We evaluate 8 specific LLMs for our benchmark: 3 of them are from the Qwen2.5 family (Yang et al., 2024), including **Qwen2.5-7B-Instruct** (main model for investigations), **Qwen2.5-7B**, and **Qwen2.5-3B-Instruct**. Other LLMs include **Meta-Llama-3-8B-Instruct** (Grattafiori et al., 2024), **DeepSeek-R1-Distill-Llama-8B** (Guo et al., 2025), **Platypus2-7B** (Lee et al., 2023), **Falcon3-7B-Instruct** (**?**), and **GPT-4o** (Hurst et al., 2024) (commercial model for testing the performance upper-bound). GPT-4o cannot be finetuned with LoRA and is evaluated only in the zero-shot setting.

5.2 Parameter-Efficient Fine-Tuning

We use LoRA for PEFT and follow LitGPT[6]'s reference configuration with Rank $r = 8$, scaling $\alpha = 16$, LoRA dropout 0.05. Training runs for $25,000$ iterations for single-task training and $50,000$ iterations for multi-task training, with an effective batch size of 64. All experiments are conducted on one H-100 GPU with a single run.

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

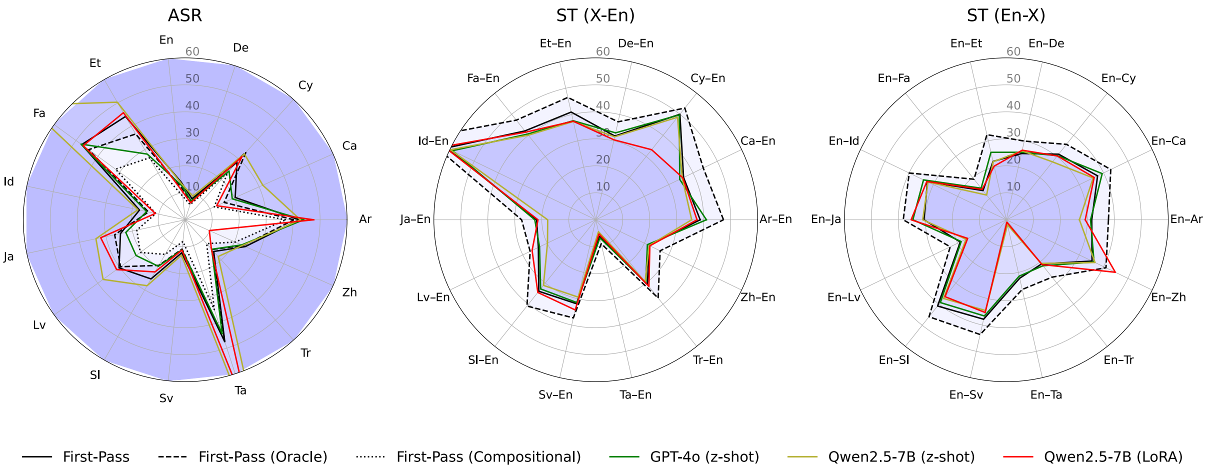

Figure 4: Comparison of GPT-4o and Qwen2.5-7B-Instruct models on different languages of Val-100 set. The left figure shows TER for ASR, the middle and right figures show BLEU for ST. N-best lists are large first-pass models decoded with mixture decoding. LoRA finetuning is conducted with single task. See the full results in Appendix A.

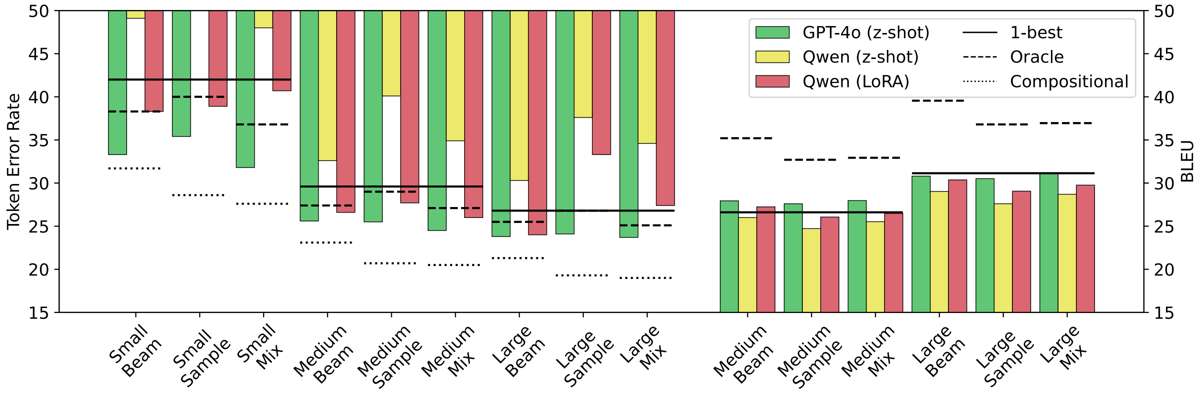

Figure 5: Comparison of GPT-4o and Qwen2.5-7B-Instruct models in different first-pass decoding setups on Val-100 set. The scores are the average across all the language or language pairs. See the full results in Appendix A.

5.3 Evaluation

Data splits. Besides the test sets in Tables 1 and 2, we create a *Val-100* subset by sampling 100 utterances per language from the validation set, which is specifically used for comparison with GPT-4o. This yields 1,500 utterances for ASR and 2,800 for ST, small enough for affordable lightweight evaluation, yet large enough to reflect full-set trends.

Metrics. For ASR, we compute **TER** using SacreBLEU tokenization, which keeps punctuation as tokens, crucial for our fully-formatted transcripts. We calculate TER based on WER implementation of *jiwer*.[7] ST output quality is measured with **SacreBLEU**.

6 Results and Analysis

6.1 First-Pass Decoding Setups

Tables 4 and 5 compare different first-pass setups using Qwen2.5-7B-Instruct with LoRA. For ASR, mixture decoding is unable to beat pure beam search on the larger Whisper models, mirroring the observation in Figure 3 that larger models favor beam search. For ST, the trend reverses: mixture decoding's advantage improves as the size of SeamlessM4T grows. One hypothesis is that the accuracy of ASR is high enough for N-best diversity to hurt the performance, but not for ST.

6.2 Comparison with GPT-4o

We conduct a comparison of **GPT-4o** with zero-shot and LoRA finetuning results of open-sourced models (represented by **Qwen2.5-7B-Instruct**) on Val-100. Figures 4 and 5 (ASR) present the perfor-

[7] https://github.com/jitsi/jiwer

GER	Ar	Ca	Cy	De	En	Et	Fa	Id	Ja	Lv	Sl	Sv	Tr	Zh	AVG
Q2.5-7B-i	58.8	13.6	40.4	8.1	12.8	45.2	56.5	12.5	42.6	37.8	22.4	14.2	17.2	12.6	28.2
Q2.5-7B	51.9	13.2	40.7	7.9	12.4	43.0	51.3	**12.4**	**37.2**	35.5	22.2	14.0	17.1	**12.5**	26.5
Q2.5-3B-i	61.6	14.5	41.3	8.5	13.5	45.5	62.9	13.2	45.0	36.8	23.5	14.6	18.1	17.1	29.7
L3-8B-i	49.4	12.5	38.8	**7.5**	**12.3**	**39.2**	**51.1**	12.6	44.0	34.4	**21.1**	13.5	**16.3**	15.2	**26.3**
DS-8B	58.6	12.9	**38.6**	8.0	13.3	41.1	52.8	13.5	46.0	35.7	22.2	13.9	17.7	14.7	27.8
P2-7B	**48.9**	**11.8**	40.1	7.6	12.6	41.4	**51.1**	13.0	40.8	**34.2**	22.0	**13.3**	18.2	14.7	26.4
F3-7B-i	53.4	14.0	40.2	9.0	13.0	43.0	55.5	14.9	48.8	36.8	23.9	15.0	20.2	21.8	29.3

GER	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	X–En
Q2.5-7B-i	47.56	37.95	51.12	38.67	26.96	26.06	54.02	23.58	**31.50**	38.53	41.25	32.14	19.03	36.03
Q2.5-7B	47.73	**38.41**	52.60	38.96	**27.10**	26.13	53.75	**23.96**	31.35	39.57	41.41	32.67	21.04	**36.51**
Q2.5-3B-i	47.42	37.91	51.49	38.45	26.29	25.90	53.21	21.89	31.19	38.62	40.80	31.76	20.29	35.79
L3-8B-i	47.86	38.28	**53.35**	**39.17**	26.95	26.09	54.29	23.52	31.49	**39.68**	41.10	31.83	18.54	36.32
DS-8B	**48.13**	37.95	52.47	38.17	26.53	25.81	**54.52**	23.33	31.02	38.59	40.54	31.41	19.55	36.00
P2-7B	48.00	38.20	52.31	38.83	26.90	**26.37**	53.63	22.57	31.28	38.58	**41.67**	**32.84**	**21.20**	36.34
F3-7B-i	47.67	38.06	50.97	38.58	26.31	26.10	52.22	22.49	30.37	37.88	40.39	32.00	20.50	35.66

GER	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	En–X	AVG
Q2.5-7B-i	25.10	40.02	33.26	35.75	27.57	18.21	38.84	32.47	21.47	34.05	41.64	22.29	46.90	32.12	34.08
Q2.5-7B	25.26	40.46	33.60	35.97	27.77	18.54	39.01	32.53	21.71	34.33	41.96	22.58	**47.14**	32.15	34.33
Q2.5-3B-i	24.94	39.89	32.85	35.54	27.22	16.80	38.61	30.97	20.79	33.61	41.24	22.48	45.31	31.56	33.68
L3-8B-i	25.16	**41.39**	34.24	**36.14**	28.08	**21.52**	**39.40**	31.58	22.15	34.75	**42.59**	23.39	43.98	32.64	34.48
DS-8B	24.30	40.11	33.42	35.43	27.59	20.86	38.76	30.09	21.06	34.12	41.97	22.69	44.11	31.89	33.95
P2-7B	**25.68**	40.85	**36.54**	36.08	**29.07**	21.32	39.29	**32.93**	**24.46**	**35.28**	42.56	**23.66**	43.64	**33.18**	**34.76**
F3-7B-i	17.93	41.08	33.05	34.80	27.75	15.47	38.00	22.01	21.77	34.21	41.03	20.37	39.95	29.80	32.73

Table 6: GER model comparison on the test set with all the models being LoRA fine-tuned on multiple tasks (ASR and ST). The upper part presents TER for ASR, with the "AVG" column presenting the average scores across all the languages. The middle and lower parts present BLEU for ST, with columns "X-En," "En-X," and "AVG" presenting the average scores across any-to-English, English-to-any, and all the language pairs, respectively.

Whisper **(L+Beam)**	**Whisper** **(L+Sample)**	**Qwen2.5-7B-Instruct** **(GER for ASR)**
0.34	0.32	0.12
SeamlessM4T **(L+Beam)**	**SeamlessM4T** **(L+Sample)**	**Qwen2.5-7B-Instruct** **(GER for ST)**
0.09	0.09	0.10

Table 7: RTF of first-pass decoding and GER on the Val-100 set. "L" stand for "large."

mance comparison across different languages and different first-pass decoding setups, respectively.

GPT-4o delivers the best performance across both tasks, surpassing Qwen2.5-7B-Instruct with LoRA. Unlike for Qwen with LoRA, Beam–sampling mixture decoding consistently outperforms pure beam search for GPT-4o. This confirms that controlled diversity helps the LLM discover better corrections, in line with prior observations.

In the zero-shot setting, Qwen2.5-7B-Instruct exhibits poor TER on ASR but achieves reasonable BLEU on ST. We hypothesize that ASR's stricter correctness constraints make its outputs more vulnerable to over-correction, whereas ST tolerates more variation. Crucially, LoRA fine-tuning significantly improves both ASR and ST, especially ASR, validating the effectiveness of our training data. For both ASR and ST, Qwen2.5-7B-Instruct performs poorly in generating certain languages (Ta),[8] reflecting that open-source models still lack language coverage compared to commercial models like GPT-4o.

6.3 Multi-task Training and Benchmarking

For multi-task training, we select the mixture decoding with large first-pass models for both ASR and ST, and combine the ASR and ST data to create the new training set. The choice is based on the fact that GPT-4o performs best with these setups (Figure 5). Although Table 4 reveals that mixture decoding fails to outperform beam decoding with larger ASR models for Qwen2.5-7B-Instruct with LoRA, we argue that it is because Qwen cannot fully exploit this extra diversity like GPT-4o does. We therefore adopt the highest-potential first-pass setup to favor stronger LLMs.

Results are shown Table 6. Across both ASR and ST, the same trend holds within the Qwen-2.5 family: **Qwen2.5-7B** delivers the best performance, followed by **Qwen2.5-7B-Instruct**, with **Qwen2.5-3B-Instruct** trailing behind. These results indicate that (i) additional instruction tuning does not benefit GER on either task, and (ii) reducing the GER

[8] As most open-sourced models' capabilities for Tamil are extremely poor, we exclude "ta", "ta-en", and "en-ta" from the evaluation on the test set. (Tables 4, 5, and 6)

model size below 7 B parameters noticeably degrades performance.

When mixing the strongest Qwen2.5 model with other LLMs for comparison, results show that **Meta-Llama-3-8B-Instruct** attains the lowest average TER for ASR, while **Platypus2-7B** achieves the best average BLEU for ST. These two strong models also yield sufficient performance on the other task, both ranking 2nd. **Qwen2.5-7B** also shows a competitive and balanced capability, ranking 3rd on both tasks. **DeepSeek-R1-Distill-Llama-8B** and **Falcon3-7B-Instruct** are the weaker models among them, with the latter being the weakest, producing the poorest performance on ASR and ST (mainly for En-X).

The performance of Qwen2.5-7B-Instruct drops slightly compared with single-task LoRA ("L+Mix" rows in Tables 4 and 5). We attribute this to negative transfer: gradients from the ST objective encourage semantic paraphrasing, occasionally conflicting with the stricter accuracy required by ASR. Therefore, achieving universal speech-to-text GER models will require additional effort.

When comparing with the 1-best results (GER "-") in Tables 4 and 5, we can observe that most GER models outperform 1-best baselines, but there are still a few models that fail (2 for ASR, 1 for ST). Aside from multi-task negative transfer, another possible cause is that hallucinations occur for LoRA finetuned models, which hurts their performance (Details in Appendix B).

6.4 Inference Cost

Introducing an LLM for error correction raises concerns about additional inference cost. To quantify this impact, we measured the real-time factor (RTF) of first-pass decoding and GER on the Val-100 set in Table 7. Even after adding all the required time cost for the pipeline of GER with mixture decoding (Beam + Temperature + GER), the combined RTF stays well below 1.0, indicating that real-time processing is still achievable. Cost can be reduced further by re-using encoder states during the second-time first-pass decoding. Additionally, even if GER may not be ideal for live streaming due to possible latency, its accuracy gains still deliver clear value in offline scenarios.

7 Conclusion

We present CoVoGER, the first benchmark to unify multilingual, multitask GER for speech. By decoding Common Voice 20.0 and CoVoST 2 with multiple sizes of Whisper and SeamlessM4T, we generate and release N-best lists for 33 languages across ASR and ST—complete with oracle statistics and evaluation scripts. Our experiments demonstrate that (i) blending beam search with temperature sampling produces the most GER-friendly hypotheses, (ii) GPT-4o establishes a strong zero-shot upper bound across all languages, and (iii) joint ASR–ST GER fine-tuning reveals a trade-off between the two tasks, underscoring the need for future work to reconcile their objectives. CoVoGER thus provides an open test bed for investigating how LLMs can bridge the gap between first-pass speech models and human-level accuracy.

Limitations

- **Imbalanced language sizes.** CoVoGER inherits the distribution of Common Voice and CoVoST-2, with training utterances for different languages ranging from millions to thousands. We did not study how this imbalance affects GER training. Future work should explore per-language reweighting or curriculum sampling to mitigate this bias.

- **Coverage of first-pass decoding strategy.** We explore only beam search, temperature sampling, and their mixture. There are other decoding strategies, such as diverse beam (Vijayakumar et al., 2016) or nucleus sampling (Holtzman et al., 2019), that could be investigated as well.

- **Multi-task negative transfer with LoRA.** Our experiments on multi-task training show negative transfer between ASR and ST, which could be due to LoRA suffering measurable catastrophic forgetting. Strategies such as task-balanced sampling (Ruder, 2017), adapter routing (Pfeiffer et al., 2020), or multi-objective optimisation (Sener and Koltun, 2018) are necessary for addressing this issue.

Ethical Considerations

This study exclusively uses publicly available datasets (Common Voice and CoVoST-2) for ASR and ST GER benchmarking, ensuring compliance with ethical and privacy standards. Our work does not involve any private or sensitive data collection. In addition, we confirm that the dataset and models used in our study were obtained and utilized in full compliance with their respective licenses and intended use guidelines.

Acknowledgment

This work was supported by the SPRING program of Kyoto University, JSPS KAKENHI Grant Number JP23K28144 and 24KJ1442, and NVIDIA Higher Education and Research Developer Program.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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 & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng. 2023a. Hyporadise: An open baseline for generative speech recognition with large language models. *Advances in Neural Information Processing Systems*, 36:31665–31688.

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. 2023b. 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.

Shanbo Cheng, Zhichao Huang, Tom Ko, Hang Li, Ningxin Peng, Lu Xu, and Qini Zhang. 2024. Towards achieving human parity on end-to-end simultaneous speech translation via llm agent. *arXiv preprint arXiv:2407.21646*.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. *arXiv preprint arXiv:1702.01806*.

Sreyan Ghosh, Mohammad Sadegh Rasooli, Michael Levit, Peidong Wang, Jian Xue, Dinesh Manocha, and Jinyu Li. 2024. Failing forward: Improving generative error correction for asr with synthetic data and retrieval augmentation. *arXiv preprint arXiv:2410.13198*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu 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. 2024a. Large language models are efficient learners of noise-robust speech recognition. *arXiv preprint arXiv:2401.10446*.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Dong Zhang, Zhehuai Chen, and Eng Siong Chng. 2024b. Gentranslate: Large language models are generative multilingual speech and machine translators. *arXiv preprint arXiv:2402.06894*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ariel N Lee, Cole J Hunter, and Nataniel Ruiz. 2023. Platypus: Quick, cheap, and powerful refinement of llms. *arXiv preprint arXiv:2308.07317*.

Sheng Li, Chen Chen, Chin Yuen Kwok, Chenhui Chu, Eng Siong Chng, and Hisashi Kawai. 2024. Investigating asr error correction with large language model and multilingual 1-best hypotheses. In *Proc. Interspeech*, pages 1315–1319.

Rao Ma, Mengjie Qian, Mark Gales, and Kate Knill. 2025. Asr error correction using large language models. *IEEE Transactions on Audio, Speech and Language Processing*.

Hermann Ney. 1999. Speech translation: Coupling of recognition and translation. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 1, pages 517–520. IEEE.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Amin Robatian, Mohammad Hajipour, Mohammad Reza Peyghan, Fatemeh Rajabi, Sajjad Amini, Shahrokh Ghaemmaghami, and Iman Gholampour. 2025. Gec-rag: Improving generative error correction via retrieval-augmented generation for automatic speech recognition systems. *arXiv preprint arXiv:2501.10734*.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.

Chenhui Shen, Liying Cheng, Lidong Bing, Yang You, and Luo Si. 2022. Sentbs: Sentence-level beam search for controllable summarization. *arXiv preprint arXiv:2210.14502*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ada Defne Tur, Adel Moumen, and Mirco Ravanelli. 2024. Progres: Prompted generative rescoring on asr n-best. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 600–607. IEEE.

Takuma Udagawa, Masayuki Suzuki, Gakuto Kurata, Nobuyasu Itoh, and George Saon. 2022. Effect and analysis of large-scale language model rescoring on competitive asr systems. *arXiv preprint arXiv:2204.00212*.

Takuma Udagawa, Masayuki Suzuki, Masayasu Muraoka, and Gakuto Kurata. 2024. Robust asr error correction with conservative data filtering. *arXiv preprint arXiv:2407.13300*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models (2016). *arXiv preprint arXiv:1610.02424*.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. Covost 2 and massively multilingual speech translation. In *Interspeech*, volume 2021, pages 2247–2251.

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)*, pages 6117–6121. IEEE.

An Yang, Baosong Yang, Beichen Zhang, and 1 others. 2024. Qwen2.5 technical report. arXiv:2412.15115. Alibaba Qwen Team.

Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. 2023. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Victor W Zue. 1985. The use of speech knowledge in automatic speech recognition. *Proceedings of the IEEE*, 73(11):1602–1615.

A Val-100 Results

Tables 8 and 10 present results evaluated on Val-100 set, with different first-pass decoding setups and GER models. We created Figures 4 and 5 based on these results.

We also conducted experiments using the task-activated in-context learning in Chen et al. (2023a) with GPT-4o on the Val-100 set for ASR GER. The N-best lists are decoded by Whisper-Large with Mixture decoding. Results are shown in Table 9. Compared with the Alpaca-style zero-shot prompt we used in the paper, task-activated prompting underperforms in the zero-shot case (Average TER of 25.6 compared to 23.7 in the 3rd row from the bottom of Table 8). However, it improves steadily with more demonstrations, echoing the trend reported by Chen et al. (2023a).

B Hallucination Analysis

We define hallucination in two cases:

- Empty-reference insertion: The first line catches any output when the reference is empty.
- Extreme mismatch: the number of word-level edit operations exceeds the number of reference characters (Character Error Rate (CER) > 1.0 for ASR task, Translation Error Rate (TER_{trn}) > 1.0 for ST task).

N-best	GER	Ar	Ca	Cy	De	En	Et	Fa	Id	Ja	Lv	Sl	Sv	Ta	Tr	Zh	AVG
S+Beam	–	48.2	28.4	57.4	11.2	16.4	67.8	107.6	28.9	37.1	46.6	40.5	17.8	62.0	24.9	34.4	42.0
S+Beam	(Oracle)	53.4	24.2	58.9	12.9	15.6	61.6	76.7	21.2	35.1	50.5	34.1	18.6	61.8	20.3	29.8	38.3
S+Beam	(Compositional)	49.7	19.7	46.3	10.2	12.3	51.4	56.8	18.3	26.7	41.6	29.2	15.1	53.3	17.2	27.6	31.7
S+Beam	GPT-4o	43.2	25.9	48.3	9.2	14.5	56.9	61.8	21.9	28.9	35.7	30.4	15.6	57.6	20.1	29.6	33.3
S+Beam	Qwen (z-shot)	45.1	28.9	52.7	11.1	15.5	69.4	129.4	36.5	32.2	46.3	37.5	18.7	158.2	25.0	29.5	49.1
S+Beam	Qwen (LoRA)	50.3	16.7	54.4	8.0	13.0	68.3	65.8	18.0	37.0	66.8	36.1	18.1	85.5	21.3	14.6	38.3
S+Sample	(Oracle)	54.0	26.2	63.4	13.3	16.2	64.4	72.2	24.3	35.2	54.9	38.7	20.5	64.5	23.5	29.4	40.0
S+Sample	(Compositional)	47.4	17.5	41.7	9.5	11.0	45.9	52.5	17.2	22.7	36.3	26.7	13.7	46.7	17.1	22.6	28.6
S+Sample	GPT-4o	43.4	25.7	46.5	10.1	17.5	50.9	94.4	25.9	28.7	34.2	34.0	15.2	54.9	20.3	29.4	35.4
S+Sample	Qwen (z-shot)	50.9	33.0	107.7	12.9	19.9	98.5	230.8	41.9	56.5	69.6	78.0	19.5	101.8	29.5	38.5	65.9
S+Sample	Qwen (LoRA)	54.8	17.8	58.2	9.0	14.5	69.9	71.7	18.3	35.0	48.3	40.4	19.8	88.1	23.0	15.4	38.9
S+Mix	(Oracle)	51.4	23.7	55.9	12.0	15.0	59.8	67.5	22.1	33.3	49.5	34.8	18.4	59.7	21.0	28.2	36.8
S+Mix	(Compositional)	46.5	16.9	38.6	9.2	10.9	44.2	50.1	16.9	22.8	34.8	25.8	13.3	45.1	16.4	23.0	27.6
S+Mix	GPT-4o	42.6	23.8	41.3	8.7	15.3	48.1	66.8	22.7	28.4	32.1	28.2	13.9	55.0	19.1	31.5	31.8
S+Mix	Qwen (z-shot)	49.1	32.4	71.4	11.8	19.4	74.2	125.0	36.6	43.9	49.5	38.6	19.3	91.6	29.8	27.8	48.0
S+Mix	Qwen (LoRA)	48.7	16.2	53.4	8.7	13.4	68.5	121.3	16.9	42.6	48.5	38.0	18.7	81.2	21.5	13.3	40.7
M+Beam	–	48.2	20.3	32.6	9.9	16.7	48.2	53.6	16.2	25.7	32.2	28.8	13.3	54.4	16.6	27.5	29.6
M+Beam	(Oracle)	42.9	17.5	37.4	8.0	12.6	41.8	49.7	15.2	29.0	32.8	23.2	11.4	52.7	15.4	23.2	27.4
M+Beam	(Compositional)	40.9	14.3	30.5	6.9	10.1	34.4	39.7	11.2	21.9	27.2	19.3	9.2	46.0	13.2	21.2	23.1
M+Beam	GPT-4o	45.4	18.8	29.3	7.9	15.6	38.4	46.2	14.2	22.8	25.7	20.2	11.5	50.7	13.7	23.9	25.6
M+Beam	Qwen (z-shot)	46.2	25.7	34.8	10.1	15.8	51.4	54.8	15.2	28.7	35.0	26.2	13.3	90.4	16.9	24.3	32.6
M+Beam	Qwen (LoRA)	43.9	12.8	33.8	6.8	12.4	51.7	49.6	12.3	24.4	33.7	23.9	12.2	54.9	14.8	11.4	26.6
M+Sample	(Oracle)	43.8	18.7	42.4	8.5	13.2	44.3	50.6	15.2	28.9	35.6	25.9	12.7	55.0	17.4	22.9	29.0
M+Sample	(Compositional)	39.4	13.0	27.1	6.4	9.2	29.8	35.7	10.9	19.1	23.0	18.1	8.7	38.7	13.1	18.0	20.7
M+Sample	GPT-4o	45.2	21.0	28.9	7.3	13.9	33.9	51.8	16.9	22.5	22.7	20.9	12.7	47.0	14.1	23.5	25.5
M+Sample	Qwen (z-shot)	47.9	26.4	53.7	8.9	68.4	14.9	87.8	23.7	31.0	37.2	28.6	24.2	97.2	25.9	26.1	40.1
M+Sample	Qwen (LoRA)	48.9	13.1	36.8	6.9	11.9	54.6	54.0	12.9	26.0	32.5	26.8	14.3	47.7	17.0	12.6	27.7
M+Mix	(Oracle)	42.3	17.4	37.2	8.0	12.4	40.8	47.3	14.2	27.9	32.6	23.3	11.9	52.3	16.1	22.2	27.1
M+Mix	(Compositional)	39.1	12.9	25.6	6.4	9.3	29.3	34.9	10.9	19.2	22.8	17.5	8.8	38.9	12.8	18.5	20.5
M+Mix	GPT-4o	43.6	18.3	28.2	7.7	13.6	32.8	46.9	13.6	22.2	23.2	18.5	12.6	49.3	13.8	23.5	24.5
M+Mix	Qwen (z-shot)	48.6	22.4	40.7	10.2	52.5	60.9	15.8	20.0	29.8	32.8	27.6	22.2	97.4	18.7	24.1	34.9
M+Mix	Qwen (LoRA)	42.6	12.2	34.1	7.4	12.4	51.7	50.3	11.4	48.7	31.0	23.6	12.3	47.0	14.8	11.2	26.0
L+Beam	–	42.2	20.3	28.7	8.2	14.0	44.2	47.4	17.3	24.5	30.3	25.3	12.3	47.5	15.8	24.4	26.8
L+Beam	(Oracle)	41.4	16.1	33.7	7.3	12.1	37.9	46.3	13.3	28.1	29.7	19.0	10.3	47.7	14.4	24.7	25.5
L+Beam	(Compositional)	39.6	12.9	27.6	6.3	9.7	31.2	35.9	11.3	20.8	24.0	15.9	8.2	41.1	12.4	23.0	21.3
L+Beam	GPT-4o	42.2	19.0	26.0	6.7	12.8	34.9	43.7	13.6	23.1	24.5	19.6	10.8	44.5	15.0	20.0	23.8
L+Beam	Qwen (z-shot)	42.6	50.6	31.2	9.1	13.9	48.5	54.0	15.0	24.6	33.2	25.3	13.0	84.8	17.2	21.5	30.3
L+Beam	Qwen (LoRA)	44.4	12.7	30.9	6.7	11.0	46.7	47.5	10.9	32.4	30.2	24.4	13.2	98.7	16.0	9.4	24.0
L+Beam	(Oracle)	41.8	16.9	38.1	7.5	12.4	39.9	47.3	14.5	27.4	32.3	21.8	11.5	50.9	15.6	24.1	26.8
L+Sample	(Compositional)	37.7	11.9	24.8	5.9	8.7	26.9	32.8	11.1	18.1	20.8	15.5	7.9	35.2	11.9	19.9	19.3
L+Sample	GPT-4o	42.7	21.4	26.0	6.9	12.8	28.3	56.3	14.3	23.6	25.4	28.6	10.9	45.6	14.5	24.9	24.1
L+Sample	Qwen (z-shot)	41.9	35.3	37.3	9.3	14.4	52.9	87.6	19.5	34.7	43.9	35.7	13.6	94.1	19.4	24.4	37.6
L+Sample	Qwen (LoRA)	46.6	14.8	35.8	7.5	11.1	53.2	45.7	13.2	26.0	35.1	23.3	11.5	150.4	14.5	10.3	33.3
L+Mix	(Oracle)	40.3	15.8	33.7	7.2	11.8	36.7	44.2	13.9	26.7	29.6	19.5	10.9	47.4	14.8	23.5	25.1
L+Mix	(Compositional)	37.3	11.8	23.5	5.9	8.8	26.5	31.9	11.1	18.3	20.6	14.8	8.1	35.0	11.9	20.3	19.0
L+Mix	GPT-4o	43.9	19.1	24.3	7.3	12.8	28.1	47.7	14.6	22.5	22.5	19.9	11.2	44.4	14.7	22.9	23.7
L+Mix	Qwen (z-shot)	42.7	31.5	32.8	9.2	13.5	50.3	72.3	17.9	33.8	37.6	28.2	13.2	95.2	18.3	23.3	34.6
L+Mix	Qwen (LoRA)	47.6	12.8	31.5	6.6	10.7	45.8	46.7	11.3	32.0	31.4	22.3	11.3	76.9	14.5	9.8	27.4

Table 8: TER scores on Val-100 for ASR. “S,” “M” and “L” stand for “small,” “medium” and “large.” “Qwen” stands for Qwen2.5-7B-Instruct. The “AVG” column presents the average scores across all the languages.

n-shot	Ar	Ca	Cy	De	En	Et	Fa	Id	Ja	Lv	Sl	Sv	Ta	Tr	Zh	AVG
n=0	41.9	19.5	26.9	7.4	13.6	38.2	49.6	15.7	26.0	26.4	20.8	11.3	47.2	15.0	24.9	25.6
n=1	49.6	19.1	20.6	7.7	12.5	22.9	44.2	18.2	23.5	23.7	14.3	14.4	46.2	14.1	14.0	23.0
n=5	45.7	17.6	19.6	6.9	11.9	22.4	47.0	11.9	33.2	19.1	13.6	16.1	40.9	14.7	9.9	22.0
n=10	52.1	17.2	18.3	6.4	11.6	21.9	48.4	11.3	25.4	18.9	13.3	14.2	41.0	14.2	10.1	21.6

Table 9: TER results using task-activated in-context learning with GPT-4o on the Val-100 set for ASR GER. The N-best lists are decoded by Whisper-Large with Mixture decoding.

So, the hallucination rate is the percentage of sentence pairs where either the reference is empty but the system still outputs tokens, or the sentence-level CER or TER$_{\text{trn}}$ exceeds 1.0. With this definition, we conducted an analysis on the Qwen model outputs in Table 6.

To prevent hallucination on ASR tasks, the model is strong in well-represented Latin languages and acceptable in Chinese, but it needs a targeted adaptation for scripts that diverge in segmentation or writing direction, as shown in Table 11.

For translation tasks (as shown in Table 12), in terms of language resource levels, the pattern is much like in ASR; by the writing system, the riskiest directions are from English into non-Latin scripts or highly agglutinative languages, while translating into English from languages with simpler morphology and scripts is relatively safe.

In summary, at the sentence level, hallucinations are quite alarming: they either never occur or wreck the entire sentence. Throughout the test set, the reduction of hallucinations can achieve a minimum

N-best	GER	Ar–En	Ca–En	Cy–En	De–En	Et–En	Fa–En	Id–En	Ja–En	Lv–En	Sl–En	Sv–En	Ta-En	Tr–En	Zh–En	X–En
M+Beam	-	33.69	35.24	41.48	27.64	34.24	36.03	56.66	17.69	5.84	25.40	21.41	4.18	27.46	22.49	27.82
M+Beam	(Oracle)	46.03	44.17	52.81	33.38	40.32	52.05	76.85	25.19	9.48	36.72	33.93	8.61	36.73	27.23	37.39
M+Beam	GPT-4o	34.87	34.01	43.93	29.22	33.74	37.08	59.19	18.70	6.61	25.33	25.46	5.32	28.48	22.00	28.85
M+Beam	Qwen (z-shot)	34.57	34.64	39.08	29.33	33.45	34.07	57.15	18.83	5.84	25.27	24.76	5.33	28.07	22.80	28.08
M+Beam	Qwen (LoRA)	35.87	34.68	41.50	30.03	33.39	39.05	59.95	20.23	6.29	28.33	28.21	5.72	30.29	21.37	29.64
M+Sample	(Oracle)	40.91	43.96	47.94	32.99	39.51	44.04	67.89	22.90	9.06	32.73	28.45	8.51	34.93	26.35	34.30
M+Sample	GPT-4o	33.03	33.97	44.65	29.10	33.33	34.04	56.58	18.33	6.73	27.72	21.55	5.12	30.05	19.37	28.11
M+Sample	Qwen (z-shot)	29.90	34.18	39.11	28.20	32.12	33.35	55.12	15.68	4.67	25.07	18.75	3.08	27.25	18.98	26.10
M+Sample	Qwen (LoRA)	34.22	33.15	31.70	31.12	32.01	36.28	58.73	16.21	6.97	28.02	25.02	5.34	31.45	20.45	27.90
M+Mix	(Oracle)	41.79	43.84	47.19	32.44	40.10	45.37	68.03	23.28	8.94	33.18	28.99	8.63	35.02	27.06	34.56
M+Mix	GPT-4o	35.15	33.78	45.30	28.41	31.29	36.97	58.82	18.63	7.47	28.85	22.13	4.60	30.77	21.20	28.81
M+Mix	Qwen (z-shot)	34.13	34.85	38.63	27.86	31.08	35.39	55.57	15.56	5.23	26.07	21.11	3.38	29.76	21.79	27.17
M+Mix	Qwen (LoRA)	34.27	35.34	35.86	31.19	32.64	37.02	58.19	18.21	7.45	27.98	26.91	4.99	32.79	18.71	28.68
L+Beam	-	38.70	36.12	50.01	31.83	40.94	42.11	61.35	21.90	23.61	33.87	31.92	6.34	30.72	21.95	33.67
L+Beam	(Oracle)	51.14	46.33	61.81	37.64	47.58	50.33	76.55	31.36	31.03	43.58	40.67	8.99	41.59	25.59	42.44
L+Beam	GPT-4o	37.10	35.55	49.79	32.01	38.28	37.56	57.53	21.82	22.88	30.79	31.07	7.14	30.89	20.27	32.33
L+Beam	Qwen (z-shot)	37.50	35.29	50.15	31.59	39.51	39.89	56.33	22.22	21.71	30.42	31.21	5.11	31.25	20.08	32.30
L+Beam	Qwen (LoRA)	38.64	36.19	47.66	33.05	38.06	42.80	57.15	23.91	26.25	33.60	31.78	6.81	32.39	20.56	33.49
L+Sample	(Oracle)	48.31	43.39	53.63	37.52	46.12	46.33	70.77	27.50	27.65	40.64	37.42	8.89	35.93	26.73	39.34
L+Sample	GPT-4o	40.63	33.82	49.56	30.40	35.90	39.16	59.56	20.47	24.43	30.84	33.15	6.13	30.86	20.91	32.56
L+Sample	Qwen (z-shot)	35.48	34.58	48.50	28.61	36.51	35.52	55.98	17.62	19.37	29.27	27.10	2.84	28.54	18.56	29.89
L+Sample	Qwen (LoRA)	24.07	34.94	33.30	29.12	35.70	43.02	61.41	18.81	28.44	33.48	34.70	5.78	32.25	23.05	31.29
L+Mix	(Oracle)	47.34	44.22	52.95	37.23	46.50	47.22	70.27	27.47	26.99	40.88	37.29	9.09	37.09	26.47	39.36
L+Mix	GPT-4o	41.01	34.53	49.71	33.04	37.64	40.60	58.96	22.59	23.12	32.77	31.46	6.90	29.73	21.24	33.09
L+Mix	Qwen (z-shot)	35.55	35.57	48.61	31.53	37.35	40.09	59.80	17.81	19.85	31.09	29.39	4.53	29.35	21.97	31.61
L+Mix	Qwen (LoRA)	37.52	35.83	33.21	30.38	37.46	41.54	62.46	21.40	26.18	34.44	34.26	5.40	31.60	22.25	32.42

N-best	GER	En–Ar	En–Ca	En–Cy	En–De	En–Et	En–Fa	En–Id	En–Ja	En–Lv	En–Sl	En–Sv	En-Ta	En–Tr	En–Zh	En–X	AVG
M+Beam	-	26.47	35.44	28.65	23.51	18.31	12.20	29.81	30.41	14.26	36.19	32.26	16.80	20.48	30.77	25.40	26.61
M+Beam	(Oracle)	35.40	44.67	34.41	30.71	25.26	17.40	40.19	37.02	18.10	45.98	41.37	26.99	27.02	37.62	33.01	35.20
M+Beam	GPT-4o	28.07	36.78	30.49	24.29	21.35	13.81	30.82	33.58	14.77	37.97	33.86	16.62	21.85	34.36	27.04	27.94
M+Beam	Qwen (z-shot)	24.48	33.36	26.66	24.94	17.39	11.28	29.92	32.51	14.69	33.34	33.45	1.75	17.90	33.32	23.93	26.00
M+Beam	Qwen (LoRA)	26.44	35.48	26.94	24.35	16.72	12.01	31.25	33.23	12.11	35.03	32.97	2.26	20.55	38.44	24.84	27.24
M+Sample	(Oracle)	32.00	41.45	33.05	28.06	24.99	18.57	34.47	37.70	19.14	43.06	38.28	23.91	24.55	36.19	31.10	32.70
M+Sample	GPT-4o	26.81	38.69	29.30	25.37	22.11	16.24	29.58	32.89	15.93	36.49	33.51	17.58	20.87	33.93	27.09	27.60
M+Sample	Qwen (z-shot)	24.21	35.48	26.04	22.77	17.79	13.18	26.73	29.54	13.56	32.26	31.95	2.09	17.65	33.32	23.33	24.72
M+Sample	Qwen (LoRA)	23.20	34.91	27.81	23.93	17.13	10.62	30.32	33.63	13.58	32.61	33.21	2.52	18.00	37.61	24.22	26.06
M+Mix	(Oracle)	33.33	41.53	33.27	28.48	24.96	18.41	35.38	37.49	18.03	44.14	38.28	23.60	24.31	36.83	31.29	32.93
M+Mix	GPT-4o	26.63	37.25	30.29	24.76	20.45	15.68	31.28	33.25	15.90	38.14	32.55	18.52	20.23	34.94	27.13	27.97
M+Mix	Qwen (z-shot)	25.43	35.56	26.93	24.59	17.16	12.73	28.31	30.42	13.23	34.40	30.41	1.76	20.05	33.06	23.86	25.52
M+Mix	Qwen (LoRA)	24.84	34.71	27.16	24.15	17.14	11.37	29.29	33.84	13.72	31.92	30.73	1.82	19.37	39.88	24.28	26.48
L+Beam	-	31.66	37.53	31.09	25.18	22.19	13.84	32.45	30.53	18.95	40.87	37.80	22.22	20.94	35.25	28.61	31.14
L+Beam	(Oracle)	40.77	46.13	37.23	33.26	32.02	20.01	44.25	35.99	24.88	50.06	47.15	30.46	29.29	41.71	36.66	39.55
L+Beam	GPT-4o	30.55	39.24	32.04	27.00	23.59	14.38	34.07	33.02	18.61	38.80	37.80	21.19	23.04	36.27	29.26	30.80
L+Beam	Qwen (z-shot)	28.12	36.64	26.62	26.38	22.26	12.36	31.32	30.81	16.05	36.48	36.41	1.94	19.38	35.99	25.77	29.03
L+Beam	Qwen (LoRA)	30.13	37.10	28.14	26.73	21.06	14.64	32.48	35.41	17.45	36.76	34.57	1.19	22.02	43.86	27.25	30.37
L+Sample	(Oracle)	36.96	42.60	34.50	31.36	31.94	18.58	40.56	37.87	23.11	45.83	43.09	25.51	27.05	40.56	34.25	36.80
L+Sample	GPT-4o	27.53	38.11	32.07	26.03	23.44	14.00	33.23	33.99	15.96	39.28	36.98	19.55	20.72	37.35	28.45	30.51
L+Sample	Qwen (z-shot)	27.82	35.93	28.82	25.49	20.19	11.38	32.67	31.62	12.70	36.25	34.53	1.88	20.13	35.12	25.32	27.60
L+Sample	Qwen (LoRA)	29.94	37.43	28.55	26.61	20.09	13.09	31.79	35.88	14.45	37.32	34.25	2.04	20.44	43.61	26.82	29.06
L+Mix	(Oracle)	37.84	42.95	35.75	29.91	32.35	19.22	39.94	38.20	23.08	45.96	43.61	26.49	27.19	40.98	34.53	36.95
L+Mix	GPT-4o	31.22	39.43	30.39	25.66	25.64	14.97	34.45	34.64	18.58	39.23	36.56	21.42	21.79	36.08	29.29	31.19
L+Mix	Qwen (z-shot)	27.08	35.71	27.45	26.19	22.06	11.72	32.21	30.99	16.58	37.29	34.49	1.91	20.68	36.49	25.78	28.70
L+Mix	Qwen (LoRA)	29.24	36.14	30.27	26.44	20.37	14.41	32.70	35.29	15.90	36.44	35.24	0.92	21.24	44.82	27.10	29.76

Table 10: BLEU scores on Val-100 for ST. “M” and “L” stand for “medium” and “large.” “Qwen” stands for Qwen2.5-7B-Instruct. Columns “X-En,” “En-X,” and “AVG” present the average scores across any-to-English, English-to-any, and all the language pairs, respectively.

De	**Id**	**Ca**	**Sv**	**En**	**Zh**	**Sl**	**Lv**	**Ja**	**Cy**	**Et**	**Ar**	**Fa**	**Ta**
0.4	0.2	0.9	0.2	1.1	0.5	0.2	0.9	3.6	1.2	0.4	4.6	6.7	11.3

Table 11: ASR Sentence-level hallucination rates (%) (percentage of hallucinated sentences) for each language.

Sl→En	**En→Sl**	**Tr→En**	**En→Tr**	**Sv→En**	**En→Sv**	**Lv→En**	**En→Lv**	**Fa→En**	**En→Fa**
0.2	3.9	0.9	8.2	1.2	4.7	1.3	5.6	1.9	8.3
De→En	**En→De**	**Id→En**	**En→Id**	**En→Et**	**Et→En**	**Cy→En**	**En→Cy**	**Ja→En**	**En→Ja**
3.8	3.4	3.3	2.4	5.3	5.6	5.2	4.1	11.5	5.0
En→Ar	**Ar→En**	**Zh→En**	**En→Zh**	**Ta→En**	**En→Ta**				
5.0	5.7	6.2	8.1	38.9	17.2				

Table 12: ST Sentence-level hallucination rates (%) for each translation direction.

overall increase in absolute precision of 1%. For low-resource languages, the improvement is even greater, typically greater than 10%.