音声認識および音声翻訳における生成的誤り訂正のための多言語ベンチマーク

Zhengdong Yang[1], Zhen Wan[1], Sheng Li[2], Chao-Han Huck Yang[3], Chenhui Chu[1]
[1] 京都大学　[2] 東京科学大学　[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

概要

大規模言語モデル（LLMs）を用いた生成的誤り訂正（GER）は，音声モデルが生成した *N*-best リストを書き換えることで，従来の再スコアリングでは修正が困難であった認識・翻訳誤りを改善できる．しかし，既存の GER 研究は主に単言語自動音声認識（ASR）に焦点を当てており，多言語・マルチタスク設定での検討は限定的である．本研究では，15 言語および 28 言語対にまたがる ASR と音声翻訳（ST）を対象とした GER ベンチマーク **CoVoGER** を提案する．CoVoGER は，Whisper と SeamlessM4T の複数モデルサイズを用いて Common Voice 20.0 と CoVoST-2 をデコードし，ビームサーチと温度サンプリングを混合した 5-best リストを提供する．さらに，複数の LLM を用いてゼロショットおよび LoRA ファインチューニングで評価を行い，混合デコーディングが多くの設定で最良の GER 性能を示すことを確認した．

1　はじめに

自動音声認識（ASR）および音声翻訳（ST）のような音声テキスト変換システム [1, 2] は，音声アシスタントや字幕生成，多言語コミュニケーション支援など，実世界の多様な応用先に広く利用されている．しかし，最先端モデルであっても，雑音環境や訛りのある音声においては誤りが生じることがある．近年の大規模言語モデル（LLMs）[3, 4, 5, 6] の進展により，音声テキスト変換システムの出力を LLM によって修正・補正し，正確性と可読性の両方を向上させる新たな可能性が示されている．

生成的誤り訂正（GER）[7, 8] は，LLM を用いて音声テキスト変換システムの出力から得た *N*-best リストから最終出力を生成する枠組みである．GER は単なる再スコアリングにとどまらず、能動的な誤り訂正を可能にした．これにより，LLM は複数仮説から情報を統合し，言語知識や文脈推論に基づいて誤りを修正できる．

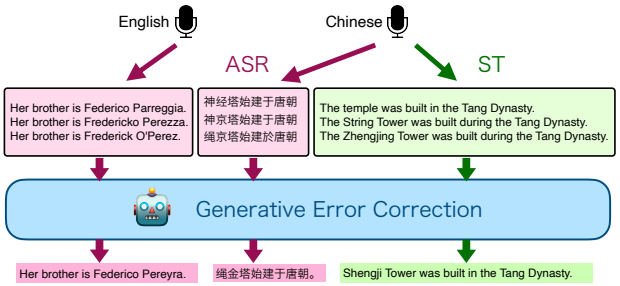

図 1　多言語・マルチタスク GER システムの例．

しかし，既存の GER 研究の多くは英語 ASR に限定されており [8, 9, 10]，非英語言語 [11, 12] や多言語設定 [13] に関する研究は断片的である．また，ASR と ST は個別に扱われることが多く，両者を横断した GER 学習の有効性は十分に検証されていない．さらに，GER の性能に大きく影響する音声モデルデコーディング設定についても，ビームサーチ以外の手法を含めた体系的な分析はほとんど行われていない．

これらの課題を踏まえ，本研究では以下の貢献をする．

- 多言語かつ複数の音声テキスト変換タスク（ASR および ST）を対象とする，GER における初の統合的ベンチマーク **CoVoGER** を提案する．
- デコーディング手法およびモデルサイズを含む音声モデルのデコーディング設定を体系的に検証し，GER 性能への影響を明らかにする．
- ゼロショットおよびファインチューニング設定において，複数の LLM を用いた大規模実験を行う．

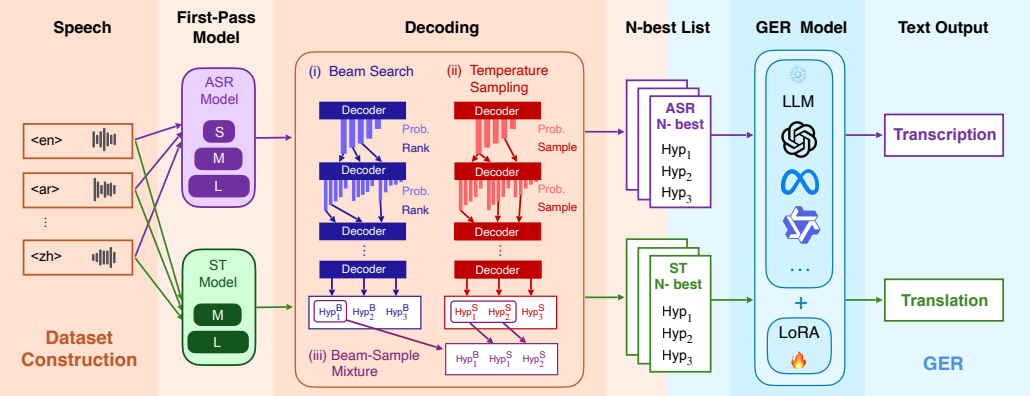

図 2　CoVoGER ベンチマークの全体像.

2　CoVoGER ベンチマーク

本節では，図 2 に示す概要とともに **CoVoGER** の構築手法を紹介する.

2.1　音声データセット・音声モデル

CoVoGER ベンチマークを構築するために，2 つの大規模公開データセットの音声をデコードする. 具体的には，ASR 用に **Common Voice 20.0**[1)]，ST 用に **CoVoST-2** [14] を用いる[2)].

CoVoGER における N-best リストは，公開されている最先端の基盤モデル 2 種類によって生成される. すなわち，ASR には **Whisper** [15]，ST には **SeamlessM4T** [16] を用いる. 各モデル系列について複数のモデルサイズを選択し，音声モデルの性能および仮説の多様性が下流の GER に与える影響を調べる.

2.2　音声モデルのデコーディング手法

音声モデル $p_\theta(\mathbf{y} \mid \mathbf{x})$ は，後段の GER モデルに入力される N-best リスト $H = \{h_1, \ldots, h_N\}$ を生成する. 本研究では，相補的な 2 つのデコーディング手法である **ビームサーチ** と **温度サンプリング** を検討する.

純粋なビームサーチは多様性に乏しい一方で，純粋なサンプリングは 1-best の精度を損ない得る. そこで，**混合**デコーディングによるリストの構築手法を提案する：信頼性のために確率の最も高いビーム出力を 1 つ保持し，残りの N–1 個を温度サンプリングで埋めることで，多様性を確保する. 本研究では，各発話についてリスト長を常に $N = 5$ に固定する. ビームサーチは厳密に N 個の仮説を返す一方で，サンプリングでは **温度** τ の選択が必要となる.

サンプリング温度の最適化.　図 3 は，異なる温度における検証セットの結果を示す. ASR のスコアは 15 言語の平均，ST のスコアは 28 言語対の平均である. ASR データに対しては SacreBLEU [17] の標準トークナイザを用いて **Token Error Rate（TER）** を測定する[3)]. さらに，N-best リストの上限性能として，**oracle TER** と **compositional oracle TER** [7] を報告する. ST については **oracle BLEU**[4)] を計算する.

図 3 は以下の結果を示した.

- 混合デコーディングは，両タスクにおいて oracle 指標において純粋なサンプリングを一貫して上回る.
- 混合デコーディングは ASR では oracle 指標でビームサーチを上回るが，ST ではそうではない[5)].
- 一般的な傾向として，サイズが小さいモデルは混合デコーディングを好み，サイズが大きいモデルはビームサーチを好み.
- ASR では，混合デコーディンおよびサンプリングは oracle TER よりも compositional TER において，ビームサーチに対する優位性が大きい. さらに，compositional oracle における最適な τ は，通常の oracle ににおける最適な τ よりもわずかに高い. これらの結果は，compositional oracle が

1）　https://commonvoice.mozilla.org/en/datasets
2）　データセットの詳細は付録 A を参照されたい.
3）　句読点を除去する慣行とは異なり，本研究ではすべての記号を保持し，GER モデルが完全に整形された ASR 出力を訂正できるようにする.
4）　各発話ごとに文レベル BLEU スコアが最高の仮説を選び，それに基づいてコーパスレベル BLEU スコアを算出する.
5）　ただし，後続の実験では混合デコーディングが ST でもビームサーチを上回ることが示されており，oracle BLEU は ST における GER のための N-best 品質推定指標として最適ではない可能性がある.

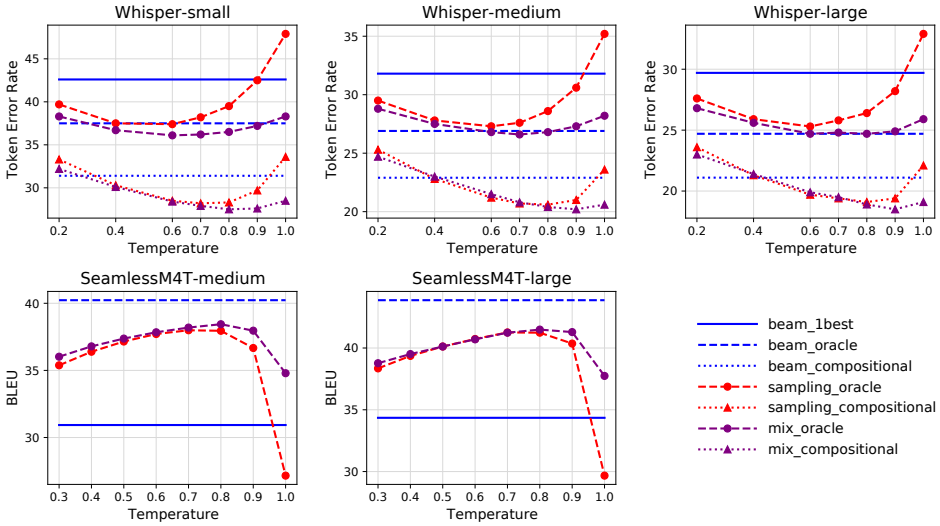

図 3　ビームサーチ，温度サンプリング，およびビーム・温度サンプリング混合デコーディングにおける，異なる温度で検証セットにおける平均性能.

単一仮説の正確性よりも，合成可能性における多様性をより強く好むことを示唆している.

以上の総合的な結果に基づき，以降の実験では τ^{ASR} および τ^{ST} をともに 0.8 に設定する.

3　実験

本ベンチマークでは，8 種類の LLM を GER モデルとして評価する．そのうち 3 つは Qwen2.5 系列 [18] であり，**Qwen2.5-7B-Instruct**，**Qwen2.5-7B**，**Qwen2.5-3B-Instruct** を含む．その他の LLM として，**Meta-Llama-3-8B-Instruct** [19]，**DeepSeek-R1-Distill-Llama-8B** [20]，**Platypus2-7B** [21]，**Falcon3-7B-Instruct** [22]，および商用モデルの **GPT-4o** [23] を評価する．GPT-4o は LoRA によるファインチューニング[6]ができないため，ゼロショット設定でのみ評価する.

テストセットに加え，GPT-4o との比較に特化した **Val-100** サブセットを作成する．具体的には，Common Voice 20.0 および CoVoST-2 の検証セットから各言語 100 発話をサンプリングする．ASR と ST の性能は **TER** と **SacreBLEU** により評価する.

3.1　GPT-4o との比較

Val-100 において，**GPT-4o** のゼロショット結果と，オープンソースモデル（代表として **Qwen2.5-7B-Instruct**）のゼロショットおよび LoRA ファインチューニング結果を比較する．図 4 は，音声モデルデコーディング設定別の性能比較を示す[7].

GPT-4o は両タスクで最良の性能を示し，LoRA を施した Qwen2.5-7B-Instruct を上回る．また，LoRA を用いた Qwen と異なり，GPT-4o では混合デコーディングが一貫して純粋なビームサーチを上回る．これは，混合デコーディングの多様性が LLM により良い訂正結果を見出させることを示している.

ゼロショット設定では，Qwen2.5-7B-Instruct は ASR において TER が悪い一方で，ST では妥当な BLEU を達成する．そこで、ASR は正確性に対する制約がより厳しく，過剰訂正の影響を受けやすい一方で，ST はより多様な表現を許容するためであると仮説を立てる．重要な点として，LoRA ファインチューニングは ASR と ST の両方，特に ASR において大幅な改善をもたらしており，学習が有効であることを裏付ける.

3.2　マルチタスク学習とベンチマーク

マルチタスク学習では，ASR と ST の両方に対して，「Large」サイズの音声モデルと混合デコーディングを選択し，ASR と ST のデータを結合して新たな訓練データを構築する．GPT-4o がこれらの設定で最良の性能を示した（図 4）という結果に基づき，より強力な LLM に有利となる「潜在力の高い」音声モデルのディコーディング設定を採用する.

結果を表 1 に示す．ASR と ST の両方で，Qwen2.5 系列内の傾向は一貫しており，**Qwen2.5-7B** が最良，

6)　ファインチューニングの詳細は付録 B を参照されたい.

7)　言語別の性能比較は付録 C を参照されたい.

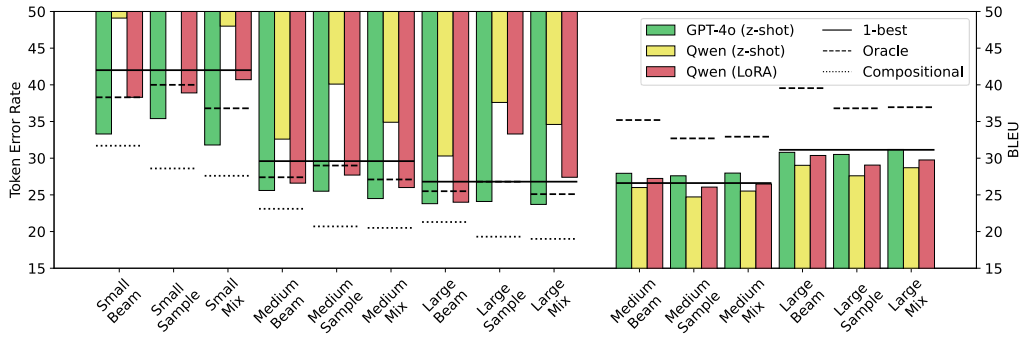

図 4　Val-100 セットにおける，異なる音声モデルのデコーディング設定においての GPT-4o と Qwen2.5-7B-Instruct の比較. スコアは，各言語（または言語対）にわたる平均である.

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

表 1　マルチタスク（ASR ＋ ST）において LoRA によりファインチューニングされたすべてのモデルについて，テストセット上での GER モデル比較.

次いで **Qwen2.5-7B-Instruct**，**Qwen2.5-3B-Instruct** は最下位となる. これらの結果は，(i) 追加の instruction tuning はいずれのタスクにおいても GER を改善しないこと，(ii) GER モデルは 7B パラメータ未満のサイズにおいては性能が顕著に低下することを示している.

最強の Qwen2.5 モデルと他の LLM を比較すると，ASR では **Meta-Llama-3-8B-Instruct** が平均 TER を最小とし，ST では **Platypus2-7B** が平均 BLEU で最良となる. これら 2 つのモデルはもう一方のタスクでも十分な性能を示し，いずれも 2 位に位置する. **Qwen2.5-7B** は両タスクで 3 位となる. **DeepSeek-R1-Distill-Llama-8B** と **Falcon3-7B-Instruct** は相対的に弱く，後者は最も性能が低く，ASR と ST の両方（特に En-X）で最悪の結果を示す.

4　おわりに

本研究では，音声に対する多言語・マルチタスク GER を統合的に扱う初のベンチマークである CoVoGER を提案した. Common Voice 20.0 および CoVoST 2 を，複数のモデルサイズの Whisper と SeamlessM4T でデコードすることで，ASR と ST にまたがる 33 言語に対する N-best リストを生成する. 実験結果から，(i) ビームサーチと温度サンプリングの混合でコーディングは GER に最も適した仮説集合が得られること，(ii) GPT-4o が全言語において強力なゼロショットの上限性能を示すこと，が明らかとなった.

謝辞

本研究は, 京都大学 SPRING プログラム, JSPS 科研費（課題番号 JP23K28144 および JP24KJ1442），ならびに NVIDIA Higher Education and Research Developer Program の支援を受けて実施された.

参考文献

[1] Victor W Zue. The use of speech knowledge in automatic speech recognition. **Proceedings of the IEEE**, Vol. 73, No. 11, pp. 1602–1615, 1985.

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

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

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

[6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. **arXiv preprint arXiv:2309.16609**, 2023.

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

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

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

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

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

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

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

[14] Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. Covost 2 and massively multilingual speech translation. In **Interspeech**, Vol. 2021, pp. 2247–2251, 2021.

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

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

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

[18] An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. arXiv:2412.15115, December 2024. Alibaba Qwen Team.

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

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

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

[22] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. **arXiv preprint arXiv:2311.16867**, 2023.

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

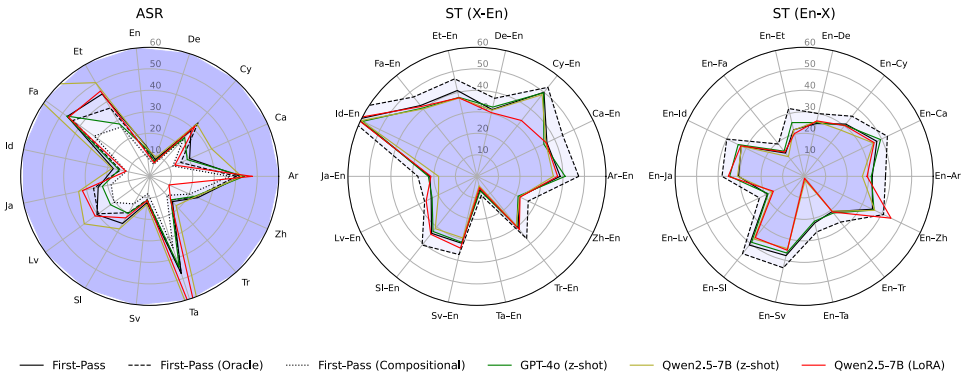

図 5　Val-100 セットにおける GPT-4o と Qwen2.5-7B-Instruct の言語別比較. 左図は ASR の TER，中図および右図は ST の BLEU を示す. *N*-best リストは，「Large」サイズの音声モデルを混合デコーディングで生成したものを用いる. LoRA によるファインチューニングは単一タスクで実施する.

	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

表 2　デコーディング設定において，ASR データセット Common Voice 20.0 からデコードされた *N*-best リスト数.

	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

表 3　デコーディング設定において，ST データセット CoVoST 2 からデコードされた *N*-best リスト数.「En-X」行は 14 の英語 → X 方向を集約したものである.

A　データセットの詳細

表 2 表 3 は，これら 2 つのデータセットからデコードされた *N*-best リストの量をまとめたものである.

一部の音声セグメントは両データセットに重複して含まれるため，データリークを防ぐ目的で以下のフィルタリングを行う：片方のデータセットの検証または テストセットに含まれる発話は，もう一方のデータセットの訓練データから除去する. また，一方のテストセットに含まれる発話は，他方の検証セットから削除する.

B　ファインチューニングの詳細

ファインチューニングにはLoRA を用い, LitGPT[8] の推奨設定に従って，Rank $r = 8$，スケーリング $\alpha = 16$，LoRA dropout 0.05 を採用する. 学習は，単一タスク学習では 25,000 iteration，マルチタスク学習では 50,000 iteration 実行し，バッチサイズは 64 とする. すべての実験は H-100 GPU 1 枚で行い，各条件につき 1 回の実行とする.

C　GPT-4o との比較：言語別

図 5 は，Val-100 において，**GPT-4o** と **Qwen2.5-7B-Instruct** 言語別の性能比較を示す.「Large」サイズの音声モデルと混合デコーディングを用いた ASR/ST のいずれにおいても，Qwen2.5-7B-Instruct は特定の言語（Ta）で生成性能が低い.[9] これは，GPT-4o のような商用モデルと比べて，オープンソースモデルでは依然として言語カバレッジが不十分であることを示唆している.

8）https://github.com/Lightning-AI/litgpt

9）多くのオープンソースモデルにおいてタミル語の性能が極めて低いため，テストセット評価では「ta」,「ta-en」,「en-ta」を除外する（表 1 を参照）.