Beyond WER: Entity and Disfluency Recall in Accented Conversational ASR

Fiza Husain [1], *Ankit Pandey* [1], *Yash Singh* [1]

[1] Stimuler, Bengaluru, India

fiza@stimuler.tech, ankit@stimuler.tech, yash@stimuler.tech

Abstract

ASR systems optimised for Word Error Rate (WER) often miss named entities and filled pauses in accented conversational English, both critical for language-learning feedback. We present a three-stage pipeline for speakers from India, Indonesia, and Latin America: (1) heuristic SQL filters curating entity-rich training data at $\sim$2.8$\times$ the entity density of random sampling, (2) regional LoRA adapters fine-tuned on Qwen2.5-Omni-3B producing both verbatim and corrected transcripts in a single forward pass, and (3) a six-category error taxonomy validated by an LLM-based judge (83.8% agreement, 210 human-labelled samples). The pipeline achieves 80–85% entity recall (up from 53–55%), 76–86% filler recall (up from $<$5%), and 6–10% WER across 6 k test utterances, outperforming Whisper and a commercial ASR on entity recall while matching a zero-shot 30B model with 10$\times$ fewer parameters. Paired bootstrap tests confirm that curation alone accounts for 2.8–4.2 pp of entity recall gain ($p < 0.0001$).

Index Terms: speech recognition, named entity recognition, accented speech, disfluency detection, parameter-efficient fine-tuning

1. Introduction

Speech is fundamental to human communication and for English language learners practicing conversational fluency, accurate speech recognition is not merely about minimising Word Error Rate (WER) or Character Error Rate (CER); it is about sustaining trust in the learning loop. In conversational language-learning platforms, ASR must faithfully capture what a learner actually said—including disfluencies, cultural references, and regionally specific entities—so that meaningful feedback can follow. In practice, learners need two complementary views of their speech: (i) a **verbatim transcript** that retains fillers such as *uh* and *um* for fluency self-assessment, and (ii) a **corrected transcript** in which entities are properly spelled and minor structural errors are smoothed, enabling comprehension-focused feedback. Commercial ASR systems, however, are optimised for aggregate WER and provide no mechanism to control this trade-off.

Evaluating several widely used ASR systems on conversational English from India, Indonesia, and Latin America exposes a stark mismatch between headline WER and downstream utility. Baseline systems achieve 13–21% WER—figures that might appear acceptable—yet entity recall ranges from only 53–55%, and filler recall is near zero ($<$5%). The disconnect arises because WER penalises a harmless tense shift (”was” $\rightarrow$ ”is”) and a garbled *“Agatha Christie”* $\rightarrow$ *“agatha cristee”* equally, despite vastly different pedagogical consequences.

Recent work has introduced entity-focused ASR benchmarks [1] and demonstrated that entity errors propagate irrecoverably to downstream NER systems [2]. Post-hoc LLM correction pipelines attempt to mitigate these errors: Chen et al. [3] use N-best hypothesis re-ranking, while Pusateri et al. [4] retrieve entity candidates from a vector database. These approaches leave the upstream acoustic model unchanged and add inference-time complexity. A complementary line applies LoRA [5] to adapt ASR models for accented speech [6], yet existing studies sample training data randomly without regard for entity density.

We do not propose a new architecture or training objective. Instead, we validate that a simple, reproducible, three-stage pipeline yields large, statistically significant gains on entity recall and filler preservation across three diverse speaker populations. Our contributions are:

1. **Heuristic entity-rich data curation.** Lexical filters applied to existing transcripts surface utterances with $\sim$2.8$\times$ higher entity density than random sampling. Paired bootstrap tests confirm that models trained on this curated data improve entity recall by 2.8–4.2 percentage points over random selection ($p < 0.0001$ across all regions), establishing data curation as a significant and cost-effective lever for entity-focused ASR.
2. **Regional LoRA adaptation with dual-output generation.** Fine-tuning Qwen2.5-Omni-3B with standard rank-32 LoRA on 10 k curated samples per region yields 80–85% entity recall (up from 53–55%) and 76–86% filler recall (up from $<$5%), while reducing WER to 6–10% - a 53–60% relative reduction. The model produces both verbatim and corrected transcripts in a single forward pass.
3. **A diagnostic error taxonomy.** We introduce a six-category error framework - spanning phonetic confusion, hallucination, acoustic out-of-vocabulary errors, named-entity misspellings, non-actionable audio, and incomplete transcription - validated via an LLM-based judge which enables systematic diagnosis of residual errors.

Our results suggest that for accented conversational speech, *what* you train on matters at least as much as *how* you train – and that lightweight curation is a practical first step toward entity-aware ASR without the inference-time cost of post-hoc LLM correction.

2. Related Work

2.1. Parameter-Efficient Fine-Tuning for ASR

Large pre-trained models such as Whisper [7] and Wav2Vec2 [8] have shifted focus to efficient adaptation. LoRA [5] injects trainable low-rank matrices into frozen Transformer layers; applied to Whisper for multilingual ASR [9] and

code-switching [10], rank-16 to rank-64 adapters match full fine-tuning at a fraction of the parameters. Memory-efficient implementations such as Unsloth [11] further lower the hardware barrier. However, none of these studies target entity recall as a training objective. We pair standard LoRA (rank-32) with a data curation step that enriches the training set for entity density.

2.2. Entity-Aware Speech Recognition

Named entity recognition from speech has evolved from broadcast-news systems [12] to analyses of ASR-NER error propagation: Szymański et al. [2] showed that entity errors propagate irrecoverably to downstream NER taggers. WhisperNER [13] jointly optimises transcription and entity tagging but requires entity-type prompts at inference. Contextual biasing injects entity lists during decoding [14]; retrieval-augmented post-processing corrects names via an LLM [4]; and Ling and Ye [15] use LLM feedback as a reward signal for RL-based ASR fine-tuning. All add inference-time complexity or require an auxiliary LLM, and none recover from acoustic-level failures. Liang et al. [16] augment entity-rich data via speech editing but lack the prosodic variation of real conversational audio. Our data-centric approach reduces entity errors at training time without entity-type labels or an auxiliary LLM at serving time.

2.3. ASR for Accented and Non-Native Speech

Multilingual models such as Whisper and SeamlessM4T [17] have advanced accented-speech recognition, yet evaluations remain dominated by aggregate WER. Graham and Roll [18] document persistent WER gaps for non-native accents. Bagat et al. [6] propose accent-specific LoRA experts on L2-ARCTIC [19], showing that per-accent adapters outperform standard LoRA and full fine-tuning—motivating our per-region design. Few studies report entity recall, despite entity words carrying the highest communicative load in learner utterances.

2.4. Disfluency Modelling in ASR

Filled pauses constitute 5–10% of spontaneous speech [20] and serve as communicative signals [21], yet Amann et al. [22] show Whisper correctly transcribes only 56% of disfluent words. L2 learners produce fillers at more than twice the native rate [23], making filler frequency a key fluency indicator [24]. Recent work recovers fillers via modified CTC forced alignment [22] or domain-specific fine-tuning [25], but neither targets L2 speech nor pairs filler preservation with entity-aware transcription.

3. Methodology

We describe a three-stage pipeline for adapting a multimodal speech model to accented, entity-rich conversational English. Each stage uses standard techniques; the contribution lies in their systematic combination and empirical validation. Figure 1 provides an overview.

3.1. Heuristic Entity-Rich Data Curation

Randomly sampled conversational audio is entity-sparse: only 24–28% of utterances in our production logs contain at least one proper noun. To concentrate supervision on the tokens where ASR systems fail most consequentially, we apply lightweight lexical heuristics to existing transcripts stored in a BigQuery data warehouse[1]. Specifically, we retain utterances that satisfy *any* of the following SQL-level filters applied to the corrected transcript field: (a) **Consecutive capitalised words** — matches sequences such as *Taylor Swift* or *Machu Picchu*, (b) **Acronyms** — matches tokens of two or more uppercase letters such as *MBA* or *UNESCO*, (c) **Honorifics and titles** — matches prefixes such as *Mr, Dr, Prof, President,* etc, followed by a space, or (d) **Mid-sentence capitalisation** — matches capitalised words that do not follow sentence-ending punctuation, capturing entity mentions embedded in running speech. All filters require a minimum transcript length of four words. This curation yields datasets with 70–78% entity density ($\sim$2.8$\times$ the random baseline), reducing annotation effort by $\sim$65%. We apply filters independently per region and sample 10 k utterances per region from the filtered pool.

3.2. Reference Transcription

Manual inspection of disagreements between professional annotators and model predictions revealed that Gemini 2.5 Pro [26] produced transcripts closer to the intended speech—particularly for regionally specific entities (local food names, band names, cultural references) where annotators unfamiliar with the speaker's cultural context introduced errors. We therefore adopted Gemini 2.5 Pro as the reference transcriber, prompting it with raw audio. Within each region, 250 utterances (750 total) were independently verified by a human annotator as a gold-standard validation subset.

3.3. Dual-Output Training Format

The model produces two outputs in a single forward pass via a structured JSON response: an **original transcript** that preserves filled pauses from a closed set (*um*, *uh*, *ah*, *er*, *hm*) at their spoken positions, accent-influenced pronunciations, and false starts; and a **corrected transcript** that replaces mispronounced words with their intended forms, resolves entity spelling, and removes stuttering-induced repetitions while retaining the speaker's original content words and sentence structure. A boolean comprehensibility flag gates both outputs: when the audio is too degraded, the model returns empty strings rather than hallucinating.

This multi-task formulation serves two purposes. First, the original transcript supports fluency self-assessment while the corrected version feeds the downstream dialogue system. Second, jointly learning corrected entity spellings provides an auxiliary training signal that improves entity recognition in the original transcript through shared acoustic representations. **All metrics reported in this paper are computed on the original (verbatim) transcript**, which is the harder evaluation target.

3.4. Regional LoRA Adaptation

We fine-tune Qwen2.5-Omni-3B [27] separately for each region using LoRA [5]. Training a dedicated adapter per region allows the model to specialise in the phonetic, prosodic, and lexical characteristics of each speaker population (e.g., retroflex consonants in Indian English, vowel shifts in Latin American English) without cross-region interference.

LoRA configuration. We use rank $r = 32$ and scaling factor $\alpha = 32$, applied to the attention projection matrices of the LLM component. All original model weights are frozen; only the low-rank matrices are updated.

[1] Training data is drawn from proprietary production logs and cannot be released. Training code will be made publicly available.

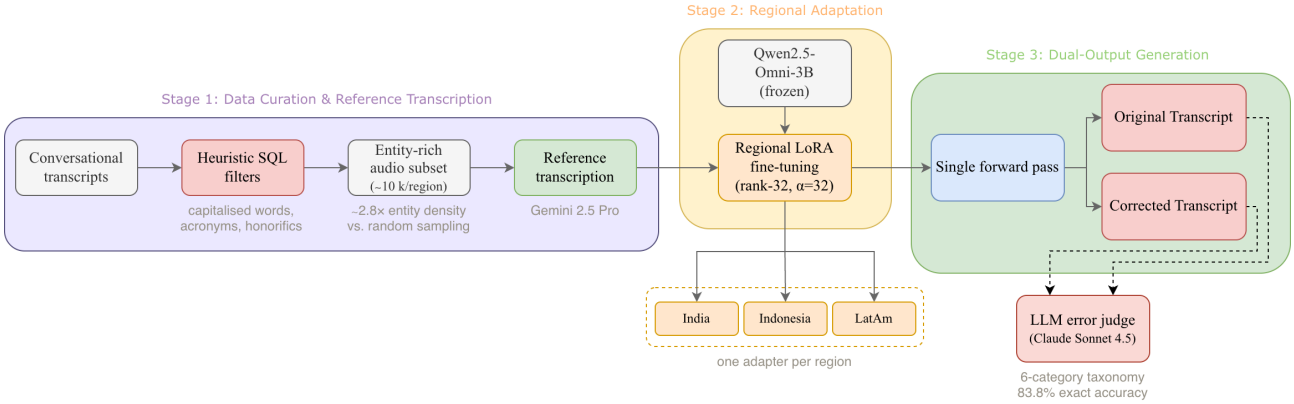

Figure 1: *Overview of the proposed three-stage pipeline: Stage 1: Heuristic SQL filters select entity-rich utterances (∼2.8 times enrichment); reference transcripts are generated by Gemini 2.5 Pro (750 human-verified samples). Stage 2: Per-region LoRA adapters are trained on Qwen2.5-Omni-3B. Stage 3: A single forward pass produces both a verbatim transcript and a corrected transcript. Dashed arrows show the diagnostic evaluation path via the LLM-based error judge.*

Training details. Each regional adapter is trained for 2 epochs on 9 k utterances (90% of the 10 k curated set; the remaining 10% is held out for validation). We use the AdamW-8bit optimiser [28] with learning rate 5×10^{-5}, cosine scheduling, 10% warmup ratio, weight decay of 0.01, per-device batch size of 4, and maximum sequence length of 2 048 tokens. Training is performed using the SFTTrainer framework with Unsloth's [11] memory-efficiency patches (gradient checkpointing, optimised embedding handling), enabling fine-tuning on a single GPU. Best checkpoints are selected by validation loss, evaluated every 300 steps. Each regional adapter trains in approximately 6 hours (2 epochs) on a single NVIDIA A100-40GB GPU. At inference time, adapters are served via vLLM [29], achieving a P95 latency of ∼800 ms. **All-region adapter.** In addition to the three regional adapters, we train a fourth *all-geo* adapter on a geographically balanced mixture of data from remaining regions, targeting speakers outside the three primary regions.

3.5. Error Taxonomy and LLM-Based Judge

To diagnose residual errors and guide future data collection, we develop a six-category taxonomy: **C1** phonetic substitution (e.g., *tea break* → *the break*), **C2** named entity error (any proper noun misspelled or garbled; takes priority), **C3** acoustic gibberish/OOV, **C4** hallucination (inserted content not in audio), **C5** omission (missing content words), and **C6** non-actionable audio (too degraded for any transcriber). C1–C4 represent actionable errors; C5–C6 are potentially unrecoverable.

We operationalise this as an LLM-based judge that receives a reference and ASR transcript and returns a structured JSON with the primary error category. To select the judge, we benchmarked 18 LLMs against a 210-sample human-labelled set stratified across regions. Claude Sonnet 4.5 [30] achieved the highest agreement: 83.8% exact accuracy, with F1 of 93.1% on entity errors (n=63) and 85.2% on phonetic substitutions (n=65). The judge is used as a *diagnostic tool*; core quantitative claims do not depend on its accuracy.

3.6. Evaluation Protocol

Test Set. Each regional adapter is evaluated on a held-out test set of ∼2 k utterances that is fully non-overlapping with the 10 k training pool. Test utterances are drawn from the same production distribution and time period but are never seen during training or validation.

Metrics. We report four complementary metrics:

- **WER / CER** — standard word- and character-level error rates, computed against Gemini 2.5 Pro references.
- **Entity recall** — fraction of ground-truth named entities correctly present in the transcript, measured via exact string match after case normalisation.
- **Filler recall** — fraction of ground-truth filled pauses (*uh*, *um*, *ah*) correctly preserved in the verbatim transcript.

Baselines. We compare against five systems spanning the efficiency–quality spectrum: Parakeet TDT-CTC 110M [31](the previously deployed production model), Whisper [7], Qwen2.5-Omni-3B (un-fine-tuned), AssemblyAI Universal-3-Pro [32](a commercial API-based system), and Qwen3-Omni-30B [33] (a mixture-of-experts model with 3B active parameters, evaluated zero-shot to probe the effect of scale without task-specific adaptation). All baselines are evaluated on identical test audio with the same reference transcripts.

Statistical testing. To assess whether gains from heuristic curation are significant, we conduct paired bootstrap tests [34] (10 000 iterations) comparing the entity-heuristic (EH) and random-sample (RS) adapters on entity recall. We report the observed difference, 95% bootstrap confidence intervals, and p-values.

4. Results

4.1. Main Results

Table 1 presents results for all models across the three target regions. We highlight three key findings.

Entity recall. The fine-tuned models (qwen-ft-eh) achieve 80–85% entity recall, a 26–29 pp improvement over Parakeet (53–55%). Gains over stronger baselines are modest but consis-

Table 1: *ASR performance across three regions. Best result per metric is **bold**; second best is underlined. All metrics computed on the verbatim transcript against Gemini 2.5 Pro references. Qwen3-Omni-30B uses MoE (3B active parameters) and is evaluated zero-shot.*

Region	Model	WER↓	CER↓	Entity Recall↑	Filler Recall↑
India	Parakeet	13.01	7.11	53.59	0.37
	Qwen (base)	11.02	6.28	65.34	5.33
	Universal-3-Pro	7.22	3.58	78.64	25.37
	Whisper	9.63	5.97	78.59	1.63
	Qwen3-Omni-30B	7.13	4.02	76.60	**91.27**
	Qwen-ft-eh	**5.95**	**3.18**	**79.52**	76.79
	Qwen-ft-rs	6.51	3.46	74.42	87.94
Indonesia	Parakeet	18.18	10.02	55.23	4.15
	Qwen (base)	13.20	7.57	70.81	16.76
	Universal-3-Pro	9.87	5.27	80.95	32.18
	Whisper	12.70	7.69	79.16	7.12
	Qwen3-Omni-30B	8.43	4.74	82.18	**92.95**
	Qwen-ft-eh	**7.36**	**4.28**	**84.60**	85.68
	Qwen-ft-rs	7.67	4.36	81.73	87.14
LatAm	Parakeet	21.12	12.21	54.31	1.18
	Qwen (base)	16.76	9.85	68.79	10.36
	Universal-3-Pro	12.96	7.75	78.63	17.61
	Whisper	16.44	10.57	76.92	2.66
	Qwen3-Omni-30B	11.79	6.98	78.82	**82.50**
	Qwen-ft-eh	**10.00**	**5.83**	**81.87**	76.47
	Qwen-ft-rs	10.75	6.27	79.05	76.51

tent: qwen-ft-eh exceeds Universal-3-Pro by up to 4pp, Qwen3-Omni-30B zero-shot by 2–3 pp, and Whisper by 1–5 pp. The fine-tuned 3B model matches or exceeds the 30B model despite 10× fewer total parameters.

WER. Qwen-ft-eh reduces WER to 6.0–10.0%, a 53–60% relative reduction over Parakeet. Relative to the unfine-tuned Qwen2.5-Omni-3B, WER drops by 40–46%, confirming the gain stems from adaptation rather than base model selection.

Filler recall. Most baselines discard filled pauses: Parakeet achieves 0.4–4.2% and Whisper 1.6–7.1%. Qwen3-Omni-30B achieves the highest filler recall (82.5–93.0%), likely due to its larger capacity. Our fine-tuned models achieve 76–86%—slightly below Qwen3-Omni-30B but from a model with 10× fewer parameters.

4.2. Effect of Data Curation: EH vs. RS Ablation

The central empirical claim of this paper is that heuristic data curation (EH) produces better entity recall than random sampling (RS), even when model architecture, training recipe, and data volume are held constant. Paired bootstrap tests (10 000 iterations) confirm statistically significant gains across all three regions: India (+4.19 pp, $p < 0.0001$), Indonesia (+2.84 pp, $p < 0.0001$), and Latin America (+2.76 pp, $p < 0.0001$), with 95% confidence intervals excluding zero. The effect is largest for India, possibly because Indian English entity names are more distinct from the base model's training distribution.

This ablation isolates the contribution of data curation: both EH and RS use the same base model, LoRA configuration, hyperparameters, and dual-output format. The RS adapter itself achieves 74–82% entity recall—far above all baselines—making the additional curation gain all the more meaningful. EH also achieves equal or better WER than RS (Table 1), confirming that entity-focused curation does not degrade general transcription quality.

4.3. Discussion

The gap between WER and entity recall across all baselines underscores that WER is necessary but insufficient for language-learning applications: Parakeet achieves 13–21% WER yet misses nearly half of all entities, while filler recall below 5% severely limits fluency assessment.

Qwen3-Omni-30B achieves comparable entity recall and higher filler recall *without task-specific training*, but requires 10× more parameters. Our 3B pipeline demonstrates that data curation paired with LoRA can match or exceed not only a commercial API but also a much larger model. The pipeline also yields consistent gains across all three regions despite distinct phonetic profiles, validating the per-region adapter approach.

5. Conclusion

We presented a three-stage pipeline for entity-aware conversational ASR targeting accented English from India, Indonesia, and Latin America. Heuristic SQL filters enrich training data to ∼2.8× the entity density of random sampling; regional LoRA adapters raise entity recall to 80–85% (from 53–55%) and filler recall to 76–86% (from <5%), while reducing WER to 6–10% at a P95 latency of ∼800 ms when served via vLLM. Bootstrap tests confirm that curation alone accounts for 2.8–4.2 pp of entity recall gain ($p < 0.0001$). The fine-tuned 3B model outperforms Whisper and a commercial ASR on entity recall while matching a zero-shot 30B model with 10× fewer parameters. Several limitations qualify these findings. First, test-set references are silver-standard (Gemini 2.5 Pro) rather than fully human-verified, though spotchecks on 250 samples per region show high agreement. Second, there remains a filler recall gap vs. the 30B model suggesting room for dual-output training optimisation. Third, evaluation is limited to English; extending to multilingual conversational settings—particularly code-switching, which is common across all three regions—is an important direction for future work.

6. Generative AI Use Disclosure

Gemini 2.5 Pro was used to generate reference transcripts (Section 3.2). Claude Sonnet 4.5 was used as the LLM-based error judge (Section 3.5). Generative AI tools were used for editing and polishing the manuscript; all authors reviewed and take responsibility for the final content.

7. References

[1] M. Del Rio, N. Delworth, R. Westerman, M. Huang, N. Bhandari, J. Palakapilly, Q. McNamara, J. Dong, P. Zelasko, and M. Jetté, "Earnings-21: A practical benchmark for ASR in the wild," *arXiv preprint arXiv:2104.11348*, 2021.

[2] P. Szymański, L. Augustyniak, M. Morzy, A. Szymczak, K. Surdyk, and P. Żelasko, "Why aren't we NER yet? artifacts of ASR errors in named entity recognition in spontaneous speech transcripts," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 1746–1761.

[3] C. Chen, Y. Hu, C.-H. H. Yang, S. M. Siniscalchi, P.-Y. Chen, and E.-S. Chng, "Hyporadise: An open baseline for generative speech recognition with large language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 31 665–31 688, 2023.

[4] E. Pusateri, A. Walia, A. Kashi, B. Bandyopadhyay, N. Hyder, S. Mahinder, R. Anantha, D. Liu, and S. Gondala, "Retrieval augmented correction of named entity speech recognition errors," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "LoRA: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[6] R. Bagat, I. Illina, and E. Vincent, "Mixture of LoRA experts for low-resourced multi-accent automatic speech recognition," in *Proc. Interspeech*, 2025.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[9] Z. Song, J. Zhuo, Y. Yang, Z. Ma, S. Zhang, and X. Chen, "LoRA-Whisper: Parameter-efficient and extensible multilingual ASR," in *Proc. Interspeech*, 2024, pp. 3934–3938.

[10] H. Yang, Y. Peng, H. Huang, and S. Li, "Adapting whisper for parameter-efficient code-switching speech recognition via soft prompt tuning," *arXiv preprint arXiv:2506.21576*, 2025.

[11] D. Han and M. Han, "Unsloth," 2023. [Online]. Available: https://github.com/unslothai/unsloth

[12] F. Kubala, R. Schwartz, R. Stone, and R. Weischedel, "Named entity extraction from speech," in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*. Citeseer, 1998, pp. 287–292.

[13] G. Ayache, M. Pirchi, A. Navon, A. Shamsian, G. Hetz, and J. Keshet, "WhisperNER: Unified open named entity and speech recognition," *arXiv preprint arXiv:2409.08107*, 2024.

[14] I. Williams, A. Kannan, P. S. Aleksic, D. Rybach, and T. N. Sainath, "Contextual speech recognition in end-to-end neural network systems using beam search." in *Interspeech*, 2018, pp. 2227–2231.

[15] S. Ling and G. Ye, "Customizing speech recognition model with large language model feedback," *arXiv preprint arXiv:2506.11091*, 2025.

[16] Z. Liang, Z. Song, Z. Ma, C. Du, K. Yu, and X. Chen, "Improving code-switching and named entity recognition in ASR with speech editing based data augmentation," *arXiv preprint arXiv:2306.08588*, 2023.

[17] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman *et al.*, "SeamlessM4T: massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.

[18] C. Graham and N. Roll, "Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits," *JASA Express Letters*, vol. 4, no. 2, 2024.

[19] G. Zhao, E. Chukharev-Hudilainen, S. Sonsaat, A. Silpachai, I. Lucic, R. Gutierrez-Osuna, and J. Levis, "L2-ARCTIC: A non-native English speech corpus," Corpus and accompanying paper, 2018.

[20] E. Shriberg, "To 'errrr' is human: ecology and acoustics of speech disfluencies," *Journal of the international phonetic association*, vol. 31, no. 1, pp. 153–169, 2001.

[21] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[22] R. Amann, Z. Li, B. Bruno, and J. Niehues, "Augmenting automatic speech recognition models with disfluency detection," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 224–231.

[23] S. Götz, *Fluency in Native and Nonnative English Speech*. Amsterdam: John Benjamins Publishing Company, 2013.

[24] P. Tavakoli, "Assessment of second language fluency," *Language Teaching*, pp. 1–17, 2025.

[25] E. Akinrintoyo, N. Abdelhalim, and N. Salomons, "WhisperD: Dementia speech recognition and filler word detection with whisper," *arXiv preprint arXiv:2505.21551*, 2025.

[26] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[27] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang *et al.*, "Qwen2. 5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[28] T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer, "8-bit optimizers via block-wise quantization," *arXiv preprint arXiv:2110.02861*, 2021.

[29] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the 29th symposium on operating systems principles*, 2023, pp. 611–626.

[30] Anthropic, "Introducing Claude Sonnet 4.5," 2025. [Online]. Available: https://www.anthropic.com/news/claude-sonnet-4-5

[31] E. Harper, S. Majumdar, N. R. Koluguri, D. Rekesh, S. Kriman, H. Huang, O. Hrinchuk, K. Puvvada, J. Balam, and B. Ginsburg, "Parakeet-TDT CTC-110M," NVIDIA NGC Model Catalog, 2024, accessed: 2026-02-23. [Online]. Available: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/parakeet-tdt_ctc-110m

[32] AssemblyAI, "Introducing universal-3 pro: A new class of speech language model optimized for voice AI," 2025. [Online]. Available: https://www.assemblyai.com/blog/introducing-universal-3-pro

[33] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu *et al.*, "Qwen3-omni technical report," *arXiv preprint arXiv:2509.17765*, 2025.

[34] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 388–395.