In-Context Learning Boosts Speech Recognition via Human-like Adaptation to Speakers and Language Varieties

Anonymous ACL submission

Abstract

Human listeners readily adjust to unfamiliar speakers and language varieties through exposure, but do these adaptation benefits extend to state-of-the-art spoken language models (SLMs)? We introduce a scalable framework that allows for in-context learning (ICL) in Phi-4 Multimodal (Phi-4-MM) using interleaved task prompts and audio-text pairs, and find that as few as 12 example utterances (∼50 seconds) at inference time reduce word error rates by a relative 19.7% (1.2 pp.) on average across diverse English corpora. These improvements are most pronounced in low-resource varieties, when the context and target speaker match, and when more examples are provided—though scaling our procedure yields diminishing marginal returns to context length. Overall, we find that our novel ICL adaptation scheme (1) reveals a similar performance profile to human listeners, and (2) demonstrates consistent improvements to automatic speech recognition (ASR) robustness across diverse speakers and language backgrounds. While adaptation succeeds broadly, significant gaps remain for certain varieties, revealing where current models still fall short of human flexibility. We release our prompts and code on GitHub[1]

1 Introduction

Variation is inseparable from language—across and within accents, speakers, environments, and social settings; yet humans rapidly adapt at every level. This adaptability persists even when linguistic content is unpredictable; the mechanism is thought to involve fast (few-trial) re-weighting of acoustic–phonetic cues and recalibration of lexical priors (Sumner, 2011; Idemaru and Holt, 2014).

Automatic speech recognition (ASR) systems, in contrast, struggle whenever the test speaker, variety, or recording conditions diverge from the supervised training distribution. For example, word error rates (WERs) increase significantly in "non-standard" English varieties relative to high-resource, unmarked settings (Rogers et al., 2004; Ji et al., 2014; Graham and Roll, 2024). Traditional remedies, such as continued pre-training or supervised fine-tuning, are computationally expensive, cognitively implausible, and require often infeasible quantities of data (Azeemi et al., 2022; Nowakowski et al., 2023; Bartelds et al., 2023).

State-of-the-art ASR systems have taken many forms in recent years. Contrastive learning-based encoder models like Wav2Vec 2.0 (Baevski et al., 2020) or self-supervised models like HuBERT (Hsu et al., 2021) have been surpassed in performance by encoder-decoder models like Whisper (Radford et al., 2023). Most recently, a new class of spoken language models (SLMs) such as SALMONN (Tang et al., 2024), Qwen-Audio-Chat (Chu et al., 2023), and Phi-4-Multimodal (Phi-4-MM) (Abouelenin et al., 2025) has pushed encoder-decoder performance even higher—beyond human levels in many settings (Patman and Chodroff, 2024; Arora et al., 2025). Phi-4-MM, among the newest of these systems, has the capacity to enforce novel protocols, transcribe non-lexical features, and—for our purposes—interleave text and audio together in a way that facilitates text-guided audio prompting.

In this paper, we ask two questions: (1) *Can in-context learning (ICL) unlock human-like adaptation benefits in a state-of-the-art SLM?*, and (2) *If so, does this lead to state-of-the-art performance across diverse speakers and language varieties?* We craft a simple ICL prompting framework (fig. 1) in which the model is first exposed to a handful of labeled audio-transcript pairs from the target speaker, followed by an unlabeled continuation to transcribe. Applying this setup to Phi-4-MM, we find that just a few priming utterances reduce WERs by 5.4–36.4% (rel.) across four corpora spanning multiple English varieties (Kominek

[1] GitHub url will be shared following double-blind review.

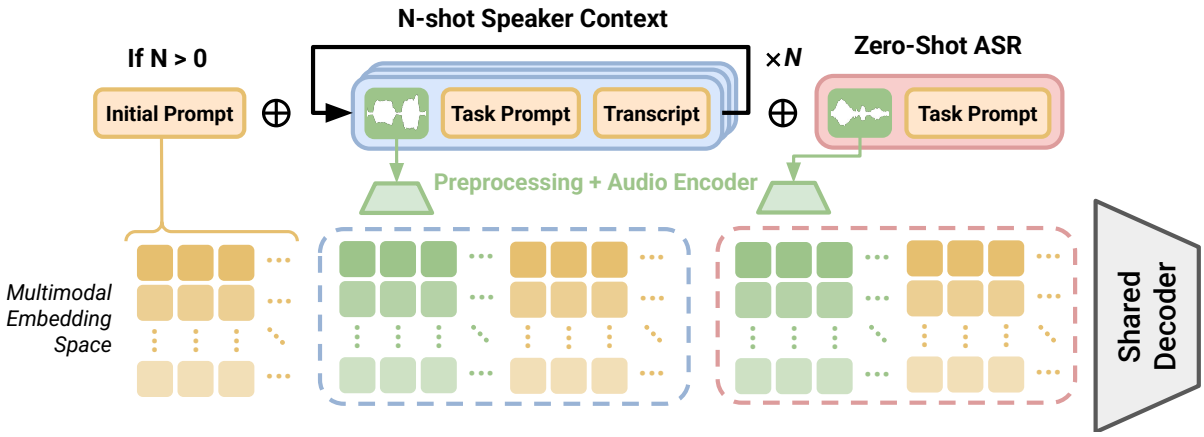

Figure 1: Our framework provides an initial description along with N transcribed examples (blue) before tasking the model to transcribe the final ASR objective audio (red). Phi-4-MM interleaves text (orange) with audio (green). These are projected into a multimodal embedding space, the context window of the shared decoder.

and Black, 2004; Zhao et al., 2018; Weinberger and Kunath, 2011; Byrne et al., 2014). Our findings demonstrate that (i) in-context learning significantly enhances ASR robustness, especially for low-resource varieties; (ii) this adaptation shows dynamic speaker- and variety-specific effects that evolve with context length; and (iii) prompt design plays a crucial role in maximizing these benefits for underrepresented varieties.

2 Background

Previous work has established that listeners recalibrate phonetic categories after minimal exposure to systematic variation, whether induced by foreign accents, coarticulation, or idiolectal quirks (Bradlow and Bent, 2008; Sidaras et al., 2009).

Phonetic variation is not merely a barrier to overcome but serves as a necessary resource for adaptation. Sumner (2011) showed that listeners exposed to variable voice onset times (VOTs) from French-accented English speakers successfully shifted their phonetic boundaries, while those exposed to invariant VOTs did not adapt. This demonstrates that variation is beneficial—indeed necessary—for robust speech perception. Moon and Sumner (2013) extended this work by showing that learned sub-lexical contrasts generalize across speakers of different non-native accents, with learned cues proving dominant enough to improve word recognition when paired with native contrasts. Work by de Marneffe et al. (2011) revealed that lexical frequency alone provides limited benefits—successful adaptation requires the interaction of phonetic variation with lexical context, not mere repetition.

Pre–deep learning pipelines relied on maximum a posteriori (MAP) adaptation and feature–space transforms such as fMLLR or i–vectors. Neural end–to–end models revived interest through layer–wise re–training, LHUC, and meta-learning (Klejch et al., 2018). Yet these methods require either dozens of utterances per speaker or backpropagation at test time. More lightweight ideas use context biasing or rescoring with personalized language models, but benefits remain inconsistent across domains (Prabhavalkar et al., 2023).

Inspired by text LLM control, researchers have explored prefix tuning, adapters, and LoRA injections to steer multilingual ASR without updating the core model (Le et al., 2021; Roll, 2025). Works such as Le et al. (2021) show that a few frozen vectors per language can close the gap to full fine-tuning on talker–independent tasks, while scaling negligibly in parameters. However, most studies optimize on supervised validation sets and do not test zero–shot adaptation at inference.

We are not the first to provide labeled exemplars directly at inference time. Early sequence–to–sequence ASR treated preceding audio–transcript pairs as an additional context window (Kim et al., 2023). Whisper's dense logits make such prompting tricky, but Wang et al. (2024) and Chen et al. (2024) independently showed sizable WER drops by concatenating audio–text pairs, especially for dialectal Chinese. Later, COSMIC introduced instruction tuning to reinforce the format, while Phi-4-MM extends the paradigm to low–footprint models. Our work focuses on the specific schema for implementing ICL in SLMs like Phi-4-MM, detailing the interleaving of task prompts, ground truth transcriptions, and audio exemplars within a shared

context window while studying the effectiveness, scaling, and cognitive plausibility of ICL.

Evidence from bilingual production suggests that talker-specific traits such as speaking rate (Bradlow et al., 2017; Graham and Nolan, 2019) or tonal structure (Graham and Post, 2018) carry over from L1 to L2 . For ASR, cross–lingual prompts or multilingual adapter stacks can leverage high-resource L1 data to bootstrap L2 decoding (Hsu et al., 2024). Our work intersects these lines by probing whether an English–centric SLM can nevertheless exploit talker-specific cues shared across dialects and second languages. To our knowledge, this is the first study to apply in-context learning for speaker adaptation in ASR across multiple speech corpora, and the first to apply these paradigms in multimodal language models.

3 Data

Our experiments leverage four English speech corpora that collectively span diverse speaker demographics, accent varieties, and speech contexts. This selection enables comprehensive evaluation of in-context adaptation across different types of linguistic variation while maintaining experimental rigor through controlled comparisons.

3.1 L2-ARCTIC

L2-ARCTIC (Zhao et al., 2018) contains high-quality recordings from 24 non-native English speakers representing six major world languages: Hindi, Korean, Mandarin, Spanish, Arabic, and Vietnamese. Each first language group includes two male and two female speakers, providing balanced gender representation. Each speaker recorded approximately one hour of read speech consisting of 1,132 phonetically balanced sentences adapted from the CMU ARCTIC prompt set (Kominek and Black, 2004).

3.2 CMU-Arctic

CMU-Arctic (Kominek and Black, 2004) is comprised of approximately 18 hours of phonetically balanced American English read speech across 18 speakers. Each speaker read the same set of approximately 1,200 utterances designed for comprehensive coverage of American English phonetic contexts. While featuring primarily American English speakers, the corpus also includes speakers with German, Indian, and other regional backgrounds, providing some accent diversity within the "native" category.

3.3 Hispanic-English Corpus (HEC)

The Hispanic-English Corpus (Byrne et al., 2014) contains approximately 30 hours of bilingual speech data from 22 Spanish heritage speakers residing in the United States. Speakers were adult native Spanish speakers from Central and South America who had lived in the United States for at least one year. For this study, we use only the English read speech portions to maintain consistency with other corpora.

3.4 Speech Accent Archive (SAA)

The Speech Accent Archive (Weinberger and Kunath, 2011) contains approximately 23 hours of English speech from over 2,500 speakers representing more than 200 first language backgrounds worldwide. All speakers read the identical 69-word paragraph beginning with "Please call Stella...". This uniform elicitation enables systematic comparison across accent types while controlling for lexical and syntactic factors. Given the identical elicitation paragraph, we utilized SAA to benchmark 0-shot ASR performance disparities across a wide range of accents within the Phi-4-MM specifically, and not to evaluate the proposed ICL framework.

3.5 Data Selection Rationale

These four corpora were selected to provide complementary perspectives on accent adaptation while enabling rigorous experimental control. The shared elicitation materials between L2-ARCTIC and CMU-Arctic enable direct comparison of adaptation effects for native versus non-native speakers under identical linguistic conditions. Together, the corpora span native American English, major world language varieties, Spanish heritage varieties, and global accent diversity, providing comprehensive coverage of English pronunciation variation.

For this study, we filtered speakers to ensure adequate context examples for few-shot evaluation, including only speakers with at least 13 valid utterances (minimum 2.5 seconds duration) and varieties represented by at least two speakers. This ensured that both test and context utterances could be drawn from the same variety with sufficient speech material to construct few-shot prompts of varying lengths. For each test utterance, we randomly sampled a fixed number of non-overlapping utterances from either the same speaker or a different speaker of the same variety, depending on the experimental condition. This procedure was repeated for all shot

count conditions (0-12). After filtering, our analysis included 15 speakers from CMU-Arctic, 14 speakers from L2-ARCTIC, 7 speakers from HEC, and the full SAA corpus for zero-shot evaluation of the Phi-4 model.

4 Model: Phi-4-MM

Phi-4-MM builds on a frozen Phi-4-Mini-Instruct core by integrating dedicated encoders for vision and audio via lightweight LoRA, enabling unified text generation from multimodal inputs (Abouelenin et al., 2025). The model supports up to 128 thousand tokens of context and generates outputs in dozens of languages.

For speech/audio, Phi-4-MM accepts 80-dimensional log-Mel filter-bank frames and processes them through a convolutional front end followed by Conformer blocks (Gulati et al., 2020). A two-layer projector then maps encoded audio into the text embedding space, where modality-specific LoRA adapters interface with the frozen layers.

Pre-training aligns the audio encoder and frozen text decoder using approximately 2 million hours of anonymized speech–text pairs spanning eight languages (Chinese, English, French, German, Italian, Japanese, Portuguese, Spanish). This stage uses only paired ASR data to teach the model cross-modal semantic alignment.

Instruction fine-tuning After the pre-training phase, Phi-4-MM is fine-tuned on roughly 100 million curated speech and audio samples—covering ASR, speech translation, question answering, summarization, and broader audio understanding—across the same eight languages. Maximum audio lengths vary by task (from 30 seconds for ASR to 30 minutes for summarization), ensuring the model learns both short-form and long-form speech processing in diverse linguistic contexts.

5 Methods

5.1 In-Context Learning Framework

We introduce a novel prompting framework that enables Phi-4-MM to perform fast, low-data adaptation through ICL. Our approach leverages the multimodal capabilities of the model by interleaving transcribed audio-text pairs as exemplars before presenting target audio for transcription. Unlike traditional ASR adaptation methods that require parameter updates or extensive speaker-specific data, our approach achieves adaptation purely through prompt engineering at inference time.

The framework operates by providing N audio-transcript example pairs ("shots") followed by a target audio segment to be transcribed. We systematically evaluated 0 through 12 in-context examples to capture both initial adaptation effects and scaling effects. Each prompt includes a series of <|user|> audio inputs paired with <|assistant|> transcriptions, followed by an unlabeled test audio segment. Full prompt templates and token formatting are provided in Appendix 7.

5.2 Prompt Design and Speaker Context Conditions

We developed two prompting strategies to investigate format specificity effects. For zero-shot evaluation, we employed both a standard prompt ("Transcribe the audio clip into text") and a variation explicitly mentioning non-native speech. Our few-shot framework follows a structured conversation format that begins with explicit instructions, includes model acknowledgment, presents each audio-transcript pair individually, and concludes with the transcription request. The variation prompt includes explicit "Transcription:" markers designed to provide clearer structural cues that may benefit lower-resource varieties.

To investigate adaptation specificity, we examined two context conditions that map onto human perceptual learning paradigms: same-speaker (within-talker evidence from the identical individual as the target) and different-speaker (within-variety evidence from other speakers of the same language variety).

5.3 Model Configuration and Preprocessing

All experiments used Phi-4-MM with greedy decoding to ensure deterministic outputs suitable for controlled evaluation. Technical implementation details, including a minor code adjustment to the model's `num_logits_to_keep` parameter to ensure correct behavior with our generation settings, are provided in Appendix 7.

Audio preprocessing involved resampling to 16 kHz, normalization to float32 to preserve dynamic range and numerical precision during downstream processing, and filtering clips shorter than 2.5 seconds. This duration cutoff ensures coverage of the "few-trials" regime documented in human adaptation literature while maintaining sufficient acoustic information for analysis. Standard text preprocessing included lowercasing, punctuation removal, and whitespace normalization to ensure fair

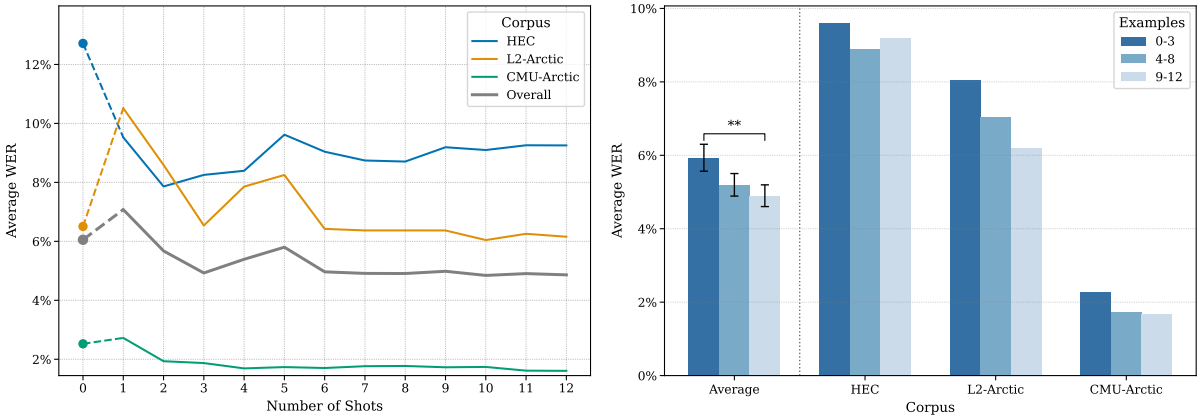

Figure 2: In-context learning consistently reduces WERs across all corpora with diminishing returns. (**Left**) WER trajectories by shot count show rapid initial improvement plateauing around 6-10 examples. (**Right**) Aggregated results across shot buckets (0-3, 4-8, 9-12) demonstrate strictly decreasing WERs with more examples. CMU-Arctic (native speakers) achieves lowest WERs across all conditions.

and consistent word error rate (WER) calculation. Comprehensive preprocessing specifications are detailed in Appendix 7.

5.4 Experimental Design and Evaluation

We applied strict filtering criteria to ensure robust evaluation: varieties required at least two speakers, speakers needed at least 13 valid utterances (enabling 12-shot evaluation), and we limited analysis to 50 utterances per speaker. This was to maintain a consistent evaluation budget across speakers and prevent over-representation of any individual voice. Context examples were selected using controlled randomization with fixed seeds for reproducibility, excluding examples with identical transcripts to the test audio.

Our primary evaluation metric was WER, computed using the `jiwer` library. Results were aggregated across trial, speaker, language variety, and corpus levels to provide comprehensive analysis of adaptation effects. We conducted up to 50 trials per speaker per condition, with experiments running on NVIDIA A100 GPUs requiring approximately 8-12 hours per corpus for complete evaluation.

The experimental design systematically tested all combinations of shot counts (0-12), speaker conditions (same/different), and prompt types (standard/variation) across the four speech corpora. This comprehensive approach enables detailed analysis of how adaptation benefits vary across different linguistic populations and experimental conditions. Speaker-level results comparing 0-shot and 12-shot performance are presented in Appendix 7 and grouped results are shown in fig. 2.

6 Results

Our experiments demonstrate that providing in-context audio-transcript examples consistently improves ASR performance across all tested corpora, with the magnitude and pattern of improvements varying systematically across speaker populations and experimental conditions.

6.1 In-Context Learning Effectiveness

Figure 2 shows that in-context learning produces substantial and generally consistent improvements across all corpora with 9-12 examples significantly better than 0-3 at a 95% confidence level (two-sample t-test). The left panel reveals characteristic diminishing returns: the largest gains occur between 0 and 1 shots, with performance improvements plateauing around 6-10 examples. CMU-Arctic consistently achieves the lowest WERs across all shot conditions, reflecting the high-resource nature of standard American English in ASR training data. Baseline (0-shot) WERs vary widely across corpora, ranging from 2.5% (CMU-Arctic) to 12.7% (HEC). The HEC average is heavily influenced by a single outlier speaker (Speaker 7, Table 3 in Appendix 7) with an exceptionally high 0-shot WER of 63.9%; excluding this speaker, the 0-shot average for the remaining HEC speakers is approximately 4.2%. This outlier also significantly impacts the 12-shot HEC average (41.0% WER for Speaker 7–see section 7). Non-native speakers in SAA reached 11.4% compared to just 1.2% for native speakers.

Aggregated across shot ranges, performance follows a clear hierarchy: 9-12 shots outperform 4-8

Table 1: Phi-4-MM performance by shot count and corpus. Our ICL method improves performance across all corpora and nearly all language varieties, with an average WER decrease of 19.7% rel. (1.2 pp.). Zero-shot performance on SAA (section 3.4) highlights high-low resource discrepancies found between the other corpora.

Corpus / *Variety*	***N*-shot WER**													0 → 12-shot
	0	1	2	3	4	5	6	7	8	9	10	11	12	
CMU-Arctic	**2.5**	**2.7**	**1.9**	**1.9**	**1.7**	**1.7**	**1.7**	**1.8**	**1.8**	**1.7**	**1.7**	**1.6**	**1.6**	**-0.9 (-36.1%)**
German	*3.5*	***3.6***	*2.9*	*2.6*	*2.7*	*2.5*	*2.5*	*2.8*	*2.6*	*2.5*	*2.6*	*2.3*	***2.3***	*-1.2 (-35.2%)*
Indian	*3.2*	***4.6***	*2.1*	*2.2*	*2.0*	*2.0*	*2.0*	*1.9*	*2.1*	*2.0*	*2.1*	***1.8***	*1.9*	*-1.2 (-39.2%)*
U.S.	***2.0***	*1.8*	*1.7*	*1.8*	*1.5*	*1.6*	*1.6*	*1.7*	*1.6*	*1.6*	*1.7*	***1.4***	*1.5*	*-0.4 (-21.9%)*
Other	***2.1***	*1.6*	*1.6*	*1.4*	*1.2*	*1.3*	*1.2*	*1.3*	*1.3*	*1.3*	*1.2*	*1.3*	***1.2***	*-0.9 (-44.5%)*
L2-Arctic	**6.5**	**10.5**	**8.6**	**6.5**	**7.9**	**8.3**	**6.4**	**6.4**	**6.4**	**6.4**	**6.0**	**6.3**	**6.2**	**-0.3 (-5.4%)**
Hindi	*4.0*	***7.1***	*3.9*	*4.0*	*3.7*	*3.8*	*3.8*	*3.7*	*3.6*	*3.6*	*3.7*	***3.4***	*3.5*	*-0.5 (-13.7%)*
Korean	*4.2*	*8.5*	*14.9*	*4.5*	*14.9*	***15.2***	*4.3*	*3.9*	*3.9*	*3.7*	***3.4***	*3.8*	*3.9*	*-0.3 (-7.9%)*
Mandarin	*7.4*	***11.4***	*8.3*	*7.7*	*7.6*	*8.1*	*7.7*	*7.7*	*7.7*	*7.7*	***7.6***	*7.8*	*7.7*	*+0.3 (+3.9%)*
Spanish	*6.5*	***10.1***	*7.0*	*6.3*	*6.1*	*6.0*	*6.2*	*6.2*	*6.3*	*6.2*	***5.7***	*5.9*	*5.9*	*-0.6 (-9.0%)*
Vietnamese	*11.3*	***17.2***	*13.1*	*11.3*	*10.8*	*12.6*	*11.1*	*11.1*	*11.2*	*11.4*	***10.6***	*11.4*	*10.7*	*-0.5 (-4.9%)*
HEC	**12.7**	**9.5**	**7.9**	**8.3**	**8.4**	**9.6**	**9.0**	**8.7**	**8.7**	**9.2**	**9.1**	**9.3**	**9.3**	**-3.5 (-27.2%)**
SAA[2]	**4.7**												**Avg**[3]**:**	**- 1.2 (-19.7%)**
Native	*1.2*													
Non-Native	*11.4*													

shots, which in turn outperform 0-3 shots across all corpora. The asymptotic shape indicates that approximately 25-30 seconds of transcribed audio captures most adaptation benefits available through ICL (see fig. 2).

6.2 Corpus-Level Performance Patterns

Table 1 reveals several consistent patterns across corpora and varieties. Nearly all speaker groups achieve their highest WERs in the 0-1 shot conditions, with the notable exception of L2-Arctic Korean speakers who show elevated error rates before improving substantially. Most varieties reach their highest accuracies with 10-12 examples.

Low-resource varieties generally experience large absolute and relative improvements. For HEC, the substantial average absolute WER reduction reported in Table 1 (-3.5 points from 0 to 12 shots) is largely driven by the aforementioned outlier speaker's improvement (-22.9 points). Within CMU-Arctic, speakers with German and Indian backgrounds show relative gains of 35.2% and 39.2% respectively (0 to 12 shots), compared to 21.9% for US English speakers. The "Other" category (see Section 3.2 for details) achieves the largest relative improvement of 44.5%. Similarly, in L2-Arctic, most non-native varieties achieve gains, with Hindi (-13.7%) and Spanish (-9.0%) showing notable improvements, while Korean shows a more modest -7.9% gain.

A critical finding is that baseline disparities persist despite adaptation. SAA illustrates this most starkly: in zero-shot conditions, native speakers achieve 1.2% WERs while non-native speakers reach 11.4% WERs—a nearly 10-fold difference despite all speakers reading identical text. L2-Arctic Mandarin is the only variety that performs slightly worse (+3.9%) at 12 shots compared to zero-shot. This small negative change falls within expected noise levels and does not contradict the overall adaptation trend. We hypothesize that bilingual interference or mismatches between orthography and pronunciation may underlie the weaker or inconsistent adaptation patterns observed in these speaker groups.

6.3 Speaker Context Specificity

Table 2 examines whether adaptation benefits vary when using examples from the same speaker versus different speakers of the same variety. The results reveal a nuanced pattern: no difference appears in the 1-3 shot range, but same-speaker examples provide a substantial 1.1 percentage point advantage (19.6% relative improvement) specifi-

[2] *Native* and *Non-native* are distinctions which overlook the complexity of many language-learning trajectories, however they manifest the wide gaps in ASR performance.

[3] Average over speakers. (See Appendix 7)

cally when 4-6 examples are provided. This advantage disappears entirely at higher shot counts, with different-speaker examples slightly outperforming same-speaker context at 10-12 shots.

This pattern suggests two distinct adaptation mechanisms operating at different scales: initial speaker-specific acoustic calibration that benefits from idiosyncratic features, followed by variety-level adaptation that emphasizes shared phonetic patterns across speakers within accent groups.

Table 2: Comparison of same-speaker versus different-speaker context performance across shot groups. Same-speaker context shows a notable benefit in the 4-6 shot range.

	Speaker Condition		
Shot Group	*Same*	*Different*	*Same* Advantage[4]
1–3	5.6	5.7	0.1 (+1.8%)
4–6	4.5	5.6	**1.1 (+19.6%)**
7–9	4.6	4.5	-0.1 (-2.2%)
10–12	4.5	4.3	**-0.2 (-4.7%)**

6.4 Prompt Sensitivity and Format Effects

Figure 3 plots the impact of prompt wording on WERs. We tested exactly **four** lightweight templates (fully specified in Appendix 7): for *zero-shot* inference a **standard** instruction taken from the model card ("Transcribe the audio clip into text.") versus a **variation** that pre-labels the clip as coming from a "non-native English speaker"; for *few-shot* (N-shot) inference a standard template that simply concatenates each audio–text pair and a variation that additionally pre-generates the token "`Transcription:`" before each exemplar. These manipulations isolate two hypothesized helpers—*task framing* through explicit accent information and *I/O scaffolding* through consistent answer delimiters.

In the zero-shot setting, the non-native framing yielded a *small but consistent* improvement across all corpora (blue bars in Figure 3), lowering WERs by 0.1–0.3 pp even for native speech. The effect suggests that the phrase "non-native" activates a broader acoustic–phonetic prior learned during instruction tuning, making the decoder slightly more tolerant of unexpected phone–letter mappings.

For few-shot adaptation, injecting the "`Transcription:`" tag proved the larger lever. It reduced early-shot volatility (<4 examples) and delivered up to 0.9pp average WER gains in L2-Arctic and 0.6pp in HEC, while leaving high-resource CMU-Arctic essentially unchanged. Together, the two **variation** templates confirmed our a priori expectation: explicit accent cues help immediately, and explicit answer markers help the model exploit sparse context more reliably, especially for under-represented varieties.

[4]The "Same Advantage" is calculated as (Different WER - Same WER). The relative percentage in parentheses is calculated as (Absolute Advantage / Different WER) * 100%.

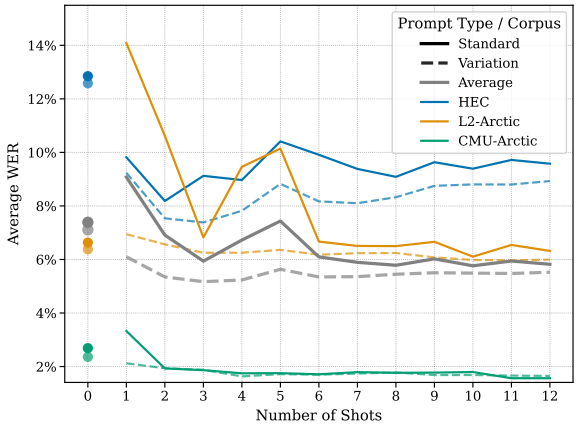

Figure 3: Prompt format sensitivity across corpora. Including explicit "Transcription:" markers reduces WERs in low-resource corpora (HEC, L2-Arctic). In zero shot settings, marginal gains are induced simply by informing the model that it's transcribing non-native speech.

7 Discussion and Conclusion

This study set out to answer two primary research questions:

(1) Can ICL unlock human-like adaptation benefits in a state-of-the-art SLM?

Yes. Across four corpora that span native, heritage, and second-language English, supplying even a few transcribed examples produces the steep, rapidly-saturating learning curve that psycholinguistic work reports for humans exposed to unfamiliar talkers. Performance gains arise fastest in the first few trials (∼25–30 s of speech), taper off thereafter, and follow two recognizable phases: (i) a *speaker-specific* phase in which 4–6 same-speaker examples yield a ∼20% relative WER reduction, and (ii) a *variety-general* phase in which larger shot counts confer similar benefits even when examples come from different speakers of the same group.

(2) If so, does this human-like adaptation translate into state-of-the-art recognition across diverse speakers and language varieties?

Largely, but not uniformly. After 9–12 shots,

Phi-4-MM achieves or exceeds state-of-the-art WERs on high-resource American English (<2%), while delivering sizeable absolute drops (1–3.5 pp.; 10–45% relative) for low-resource varieties such as Spanish-heritage, L1 Hindi, and L1 Korean.

The asymptotic nature of performance gains, with the largest improvements occurring in the first few examples before plateauing around 6–10 shots, closely parallels psychometric curves observed in human speech perception studies (Bradlow and Bent, 2008). This convergence suggests that ICL accesses fundamental mechanisms of acoustic-phonetic recalibration, potentially involving rapid re-weighting of phonetic features based on observed speaker-specific patterns. The fact that approximately 25–30 seconds of transcribed audio captures most available adaptation benefits mirrors the “fast” adaptation documented in human perceptual learning paradigms.

A notable effect in the results (Table 1) is the transient increase in WERs for several varieties when transitioning from zero-shot to a single in-context example. We hypothesize that this reflects an initial phase of task adaptation: While Phi-4-MM is instruction-tuned, it was not explicitly trained on ICL for the ASR task, let alone our novel protocol. The first audio-transcript pair may therefore present a dual challenge: the model must first recognize and assimilate the novel task structure itself (essentially learning the "rules" of this ICL interaction) before it can effectively leverage the example’s content for acoustic-lexical recalibration towards the target speaker or variety. Once this foundational task understanding is established, subsequent exemplars can more directly contribute to the targeted adaptation, leading to the WER reductions observed with additional examples. Given the strength of this effect in HEC and L2-ARCTIC, there may be interaction effects between task-learning and the out-of-distribution nature of some English varieties, however, analyses across additional corpora would be required to fully disentangle these effects.

The observation that ASR performance is notably sensitive to the prompt offers intriguing parallels to human speech perception. Research by D’Onofrio (2015) demonstrates that human listeners’ categorization of ambiguous speech sounds is shaped by explicit social information provided about the speaker. The "non-native" prompt appears to prime the ASR system, potentially by activating or re-weighting internal models suited for greater acoustic-phonetic variability or by adjusting decision thresholds, even when such characteristics are not objectively present in the target audio.

The magnitude of improvements for several low-resource varieties is particularly striking. For example, in CMU-Arctic, the ’Other’ subgroup (speakers with Scottish, Canadian, and Israeli backgrounds, as detailed in Section 3.2) achieved a 44.5% relative WER reduction (0 to 12 shots), and the ’Indian’ subgroup saw a 39.2% reduction. Spanish heritage speakers (HEC) experienced an average relative reduction of 27.2%, though this figure is significantly influenced by an outlier speaker improving from a very high baseline (see Section 6). Other L2-Arctic varieties such as Hindi English (-13.7%) and Korean English (-7.9%) also benefited, albeit with more modest relative reductions (Table 1). These gains help narrow the performance gap compared to high-resource US English. This finding connects to evidence that frequency alone provides limited benefits—our results show that meaningful adaptation requires quality variation, not mere repetition (de Marneffe et al., 2011). The disproportionate gains for underrepresented speakers suggest that ICL may help mitigate biases inherent in training distributions dominated by high-resource varieties.

We found that same-speaker examples provide a significant advantage at 4–6 shots (about 1.1 percentage points), but this benefit disappears entirely with longer contexts. This pattern aligns with multi-level adaptation processes in human perception (Sidaras et al., 2009; Moon and Sumner, 2013). Adaptation appears to begin with speaker-specific cues and then shift toward variety-level generalization, as the model learns to extract features that transcend individual speaker characteristics.

More importantly, these disproportionate gains for low-resource varieties offer a scalable pathway toward more equitable speech technology, requiring no additional training data or computational resources beyond slightly longer inference contexts. Just as humans cope effortlessly with variation at every level, frontier SLMs show emergent robustness that can be unlocked purely through prompt engineering—offering a practical tool to improve ASR equity for speakers and varieties historically underserved by speech technology. Future work should extend this framework to spontaneous speech, cross-lingual settings, and streaming applications while probing the precise mechanisms of variety-specific adaptation.

Limitations

While this study demonstrates the significant potential of ICL for ASR adaptation within a frontier model, its limitations define key avenues for future research. First, generalizability is constrained: our experiments used a single model (Phi-4-MM) and focused on read English speech. Second, the ICL methodology has scope for expansion. Our reliance on accurately transcribed context, while establishing the potential of ICL with quality signals, may not always be practical. Future efforts should explore unsupervised or self-transcription for context generation and active context selection. Additionally, we explored adaptation mainly within the same speaker or variety and tested limited prompt variations, which inhibited the engineering goal to make speech recognition more robust and restricted the scope of investigations into human-model perceptual convergence.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*.

Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. 2022. Representative subset selection for efficient fine-tuning in self-supervised speech recognition. *arXiv preprint arXiv:2203.09829*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. *arXiv preprint arXiv:2305.10951*.

Ann R. Bradlow and Tessa Bent. 2008. Perceptual adaptation to non-native speech. *Cognition*, 106(2):707–729.

Ann R. Bradlow, Midam Kim, and Michael Blasingame. 2017. Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate. *The Journal of the Acoustical Society of America*, 141(2):886–899.

W Byrne, E Knodt, J Bernstein, and F Emami. 2014. Hispanic–english database ldc2014s05.

Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C. Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg. 2024. SALM: Speech-Augmented Language Model with in-Context Learning for Speech Recognition and Translation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13521–13525. ISSN: 2379-190X.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Marie-Catherine de Marneffe, John Jr. Tomlinson, Marisa Tice, and Meghan Sumner. 2011. The interaction of lexical frequency and phonetic variation in the perception of accented speech. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society (CogSci 2011)*, pages 3575–3580. Cognitive Science Society.

Annette D'Onofrio. 2015. Persona-based information shapes linguistic perception: Valley girls and california vowels. *Journal of Sociolinguistics*, 19(2):241–256.

Calbert Graham and Francis Nolan. 2019. Articulation rate as a metric in spoken language assessment. In *Proceedings of Interspeech 2019*, pages 3564–3568. ISCA.

Calbert Graham and Brechtje Post. 2018. Second language acquisition of intonation: Peak alignment in american english. *Journal of phonetics*, 66:1–14.

Calbert Graham and Nathan Roll. 2024. Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2).

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Ming-Hao Hsu, Kuan Po Huang, and Hung-yi Lee. 2024. Meta-whisper: Speech-based meta-icl for asr on low-resource languages. *arXiv preprint arXiv:2409.10429*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Kaori Idemaru and Lori L Holt. 2014. Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 40(3):1009.

Caili Ji, John J. Galvin, Yi-ping Chang, Anting Xu, and Qian-Jie Fu. 2014. Perception of Speech Produced by Native and Nonnative Talkers by Listeners with Normal Hearing and Listeners with Cochlear Implants. *Journal of Speech, Language, and Hearing Research*, 57(2):532–554. Publisher: American Speech-Language-Hearing Association.

Changhun Kim, Joonhyung Park, Hajin Shim, and Eunho Yang. 2023. Sgem: Test-time adaptation for automatic speech recognition via sequential-level generalized entropy minimization. *arXiv preprint arXiv:2306.01981*.

Ondřej Klejch, Joachim Fainberg, and Peter Bell. 2018. Learning to adapt: a meta-learning approach for speaker adaptation. *arXiv preprint arXiv:1808.10239*.

John Kominek and Alan W Black. 2004. The cmu arctic speech databases. In *SSW*, pages 223–224.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. *arXiv preprint arXiv:2106.01463*.

Kyuwon Moon and Meghan Sumner. 2013. The learning and generalization of contrasts consistent or inconsistent with native biases. In *INTERSPEECH*, pages 2103–2107.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. 2023. Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, 60(2):103148.

Chloe Patman and Eleanor Chodroff. 2024. Speech recognition in adverse conditions by humans and machines. *JASA Express Letters*, 4(11).

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:325–351.

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.

Catherine L. Rogers, Jonathan Dalby, and Kanae Nishi. 2004. Effects of Noise and Proficiency on Intelligibility of Chinese-Accented English. *Language and Speech*, 47(2):139–154. Publisher: SAGE Publications Ltd.

Nathan Roll. 2025. Polyprompt: Automating knowledge extraction from multilingual language models with dynamic prompt generation. *arXiv preprint arXiv:2502.19756*.

Sabrina K. Sidaras, Jessica E. D. Alexander, and Lynne C. Nygaard. 2009. Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America*, 125(5):3306–3316.

Meghan Sumner. 2011. The role of variation in the perception of accented speech. *Cognition*, 119(1):131–136.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Siyin Wang, Chao-Han Yang, Ji Wu, and Chao Zhang. 2024. Can whisper perform speech-based in-context learning? In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13421–13425. IEEE.

Steven H Weinberger and Stephen A Kunath. 2011. The speech accent archive: towards a typology of english accents. *Language & Computers*, 73(1).

Guanlong Zhao, Evgeny Chukharev-Hudilainen, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Ricardo Gutierrez-Osuna, and John Levis. 2018. L2-arctic: A non-native english speech corpus.

Appendix: Supplementary Experimental Details

A.1 Model and Generation Parameters

Model Configuration:

- **Model:** `microsoft/Phi-4-multimodal-instruct`
- **Processor:** `AutoProcessor.from_pretrained('microsoft/Phi-4-multimodal-instruct', trust_remote_code=True)`
- **Loading:** `AutoModelForCausalLM.from_pretrained('microsoft/Phi-4-multimodal-instruct', trust_remote_code=True, torch_dtype='auto', attn_implementation='flash_attention_2')`

Generation Configuration:

- `max_new_tokens`: 1200
- `do_sample`: False (greedy decoding)
- `num_beams`: 1
- `num_logits_to_keep`: 1 (explicitly set at multiple levels)

A.2 Datasets and Comprehensive Preprocessing

A.2.1 Dataset Sources

- **L2-Arctic:** `NathanRoll/l2-arctic-dataset-250`
- **HEC (HISP-ENG):** `NathanRoll/hisp-eng`
- **CMU-Arctic:** `NathanRoll/cmu-arctic`

A.2.2 Audio Preprocessing Pipeline

Target Specifications:

- Sample Rate: 16,000 Hz (resampled using `librosa.resample`)
- Format: 32-bit float (`np.float32`)
- Duration: Minimum 2.5 seconds (shorter clips filtered out)

Normalization Algorithm Steps:

- **Integer handling:** Convert integer types by dividing by max value for that dtype
- **Float handling:** Convert directly to `np.float32`
- **FLAC bug detection:** Check for max > 0.99 and min > -0.5, indicating missing negative values
- **Bug correction:** Flip values above 0.9 threshold if bug detected
- **Range clipping:** Clip extreme values exceeding ± 1.1 to [-1.0, 1.0]

Resampling Configuration:

- Primary method: `librosa.resample` with default parameters
- Fallback method: `librosa.resample` with `res_type='kaiser_fast'` if primary fails
- All resampling errors are logged and re-raised for debugging

A.2.3 Text Normalization

Normalization Steps:

1. Convert to lowercase
2. Remove punctuation: `. , ? ! ; : " ' ( ) [ ]` (each replaced with space)
3. Normalize whitespace: Multiple spaces collapsed to single spaces, leading/trailing whitespace removed

Implementation Logic: Convert text to lowercase, iterate through punctuation list replacing each with space, then split and rejoin to normalize whitespace.

A.2.4 Dataset Filtering Criteria

Variety-Level Filtering:

- Varieties must have $\geq$2 speakers
- For HEC dataset: exclude samples with `variety == 'unknown'`

Speaker-Level Filtering:

- Speakers must have $\geq$(max_shots + 1) valid utterances
- Maximum 50 samples per speaker used (selected via shuffling with speaker-specific seed)

Sample-Level Filtering:

- Duration $\geq$2.5 seconds
- Valid audio array present
- Valid transcript field present and non-empty

Variety Mapping Details:

- **CMU-Arctic:** Based on speaker ID mapping to variety (see `CMU_ARCTIC_VARIETIES`)
- **HEC:** Based on speaker origin mapping (see `HISP_ENG_ORIGINS`)
- **L2-Arctic:** Uses `l1` field directly

A.3 Experimental Design and Context Selection

A.3.1 Random Seed Management

Global Seed: 42 (default, configurable via command line)

Seed Hierarchies:

- **Speaker-level shuffling:** `global_seed + hash(f"{variety}_{speaker}") % 10000`
- **Trial-level context selection:** `global_seed + hash(f"{speaker}_{trial_idx}") % 10000`
- **Different-speaker selection:** Same as trial-level but includes variety information

This hierarchical seeding ensures:

1. Reproducible speaker orderings
2. Consistent context selection across runs
3. Deterministic different-speaker selection

A.3.2 Context Example Selection Algorithm

Same-Speaker Condition Logic:

- Build candidate list excluding current test sample
- Filter out samples with identical normalized transcripts
- Use trial-specific random seed for selection
- Sample n_shots examples without replacement

Different-Speaker Condition Logic:

- Select random other speaker from same variety
- Collect samples from selected speaker
- Filter out samples with identical transcripts to test audio
- Sample n_shots examples using same random seed strategy

A.3.3 Trial Generation Process

Trial Count Calculation:

- Maximum trials per speaker: min(pool_size - n_shots, max_trials)
- Pool size limit: 50 samples per speaker
- Test samples drawn sequentially from shuffled pool

Quality Control:

- Skip trials where insufficient context examples available
- Skip samples without valid transcript fields
- Handle all exceptions gracefully with detailed logging

A.4 In-Context Learning Prompts

The following prompt structures are used for the Phi-4 model, where `<|user|>`, `<|assistant|>`, `<|audio_N|>`, and `<|end|>` are special model tokens.

A.4.1 Zero-Shot Prompts

Standard Prompt: `<|user|><|audio_1|>Transcribe the audio clip into text.<|end|><|assistant|>`

Variation Prompt (Non-Native Focus): `<|user|><|audio_1|>Transcribe the audio clip from a non-native English speaker into text.<|end|><|assistant|>`

A.4.2 Few-Shot Prompt Structure

Initial Instruction Block:

- User message: Explains providing N examples from non-native speaker, followed by new audio from same/different speaker
- Assistant acknowledgment: Confirms understanding and intent to use examples for transcription

Dynamic Elements:

- `{num_shots_text}`: "an example" (1-shot) or "N examples" (N-shot)
- `{speaker_reference}`: "the same speaker" or "a different speaker"
- `{pronoun_text}`: "it" (1-shot) or "them" (N-shot)

Example Block Structure:

- **Standard:** User provides audio, assistant responds with transcript
- **Variation:** Same as standard but assistant response prefixed with "Transcription: "

Final Query Block: User provides final audio with speaker reference, assistant begins response (with "Transcription: " prefix for variation prompt).

A.5 Computational Requirements and Implementation

Hardware Specifications:

- GPU: NVIDIA A100 (required for flash attention)
- Memory: Minimum 40GB GPU memory recommended
- Runtime: 8-12 hours per corpus for complete evaluation (0-12 shots)

Software Dependencies:

- torch, peft, torchvision, backoff, flash-attn
- tqdm, jiwer, librosa, transformers, datasets

Model Memory Management:

- Model loaded with `torch_dtype='auto'` for optimal precision/memory trade-off
- Flash attention implementation used to reduce memory footprint
- Single inference batch processing (no batching across utterances)

A.6 Statistical Analysis and Data Collection

A.6.1 Trial Collection

Number of Trials:

- Default maximum: 50 trials per speaker per condition
- Actual trials: min(available_samples - n_shots, max_trials)
- Zero-shot: All valid samples used (up to 50)

Data Validation:

- WER calculated using `jiwer` library with normalized texts
- All results stored with full precision (no rounding during intermediate calculations)
- Individual trial results preserved in addition to averages

A.6.2 Result Aggregation

Speaker-Level Results Format:

- Variety identification
- Run counts per shot condition
- Average WER per shot condition
- Complete list of individual WER values

Corpus-Level Results:

- Weighted averages across all speakers
- Total sample counts per condition
- Preservation of speaker-level breakdowns

A.7 Reproducibility Checklist

1. Environment Setup:

- Use identical package versions (see dependencies list)
- Set all random seeds (Python, NumPy, PyTorch)

2. Data Processing:

- Apply exact audio normalization pipeline (including FLAC bug correction)
- Use identical text normalization (case, punctuation, whitespace)
- Apply same filtering criteria (duration, variety, speaker counts)

3. Experimental Configuration:

- Use hierarchical random seeding as specified
- Maintain exact prompt structure (including special tokens)
- Follow context selection algorithm precisely

4. Model Configuration:

- Use greedy decoding (do_sample=False)
- Set num_logits_to_keep=1 at all levels
- Use flash attention implementation

5. Evaluation:

- Calculate WER using jiwer with normalized texts
- Aggregate results maintaining full precision
- Store individual trial values, not just averages

A.8 Extended Speaker-Level Results

The following table provides complete speaker-level results for 0-shot and 12-shot conditions across all corpora, enabling verification of reported aggregate statistics.

Table 3: Complete speaker-level Word Error Rates (WER) for 0-shot and 12-shot conditions. All WERs are percentages.

Dataset (Corpus)	Speaker	0-shot WER (%)	12-shot WER (%)
CMU-Arctic	aew	1.5	0.7
CMU-Arctic	ahw	2.8	2.0
CMU-Arctic	aup	3.1	2.1
CMU-Arctic	axb	3.8	2.4
CMU-Arctic	bdl	1.4	0.9
CMU-Arctic	clb	1.9	0.7
CMU-Arctic	eey	1.9	2.6
CMU-Arctic	fem	4.1	2.5
CMU-Arctic	gka	1.9	1.3
CMU-Arctic	ksp	3.4	2.4
CMU-Arctic	ljm	2.2	1.8
CMU-Arctic	lnh	2.2	1.0
CMU-Arctic	rms	2.0	0.7
CMU-Arctic	slp	3.9	1.8
CMU-Arctic	slt	1.7	1.1
HEC	0	4.6	7.6
HEC	1	3.9	3.3
HEC	18	6.7	4.9
HEC	3	2.3	1.9
HEC	4	4.7	3.8
HEC	6	3.0	2.3
HEC	7	63.9	41.0
L2-Arctic	ASI	3.5	3.3
L2-Arctic	BWC	9.4	10.3
L2-Arctic	EBVS	8.8	7.5
L2-Arctic	ERMS	6.5	6.1
L2-Arctic	HJK	4.1	4.0
L2-Arctic	HQTV	17.9	15.8
L2-Arctic	LXC	6.4	6.1
L2-Arctic	MBMPS	5.5	5.1
L2-Arctic	NCC	6.4	6.7
L2-Arctic	NJS	5.1	4.9
L2-Arctic	PNV	4.6	5.6
L2-Arctic	RRBI	4.9	3.5
L2-Arctic	TNI	3.7	3.5
L2-Arctic	YKWK	4.2	3.7
Grand Average (across speakers)		**6.1**	**4.9**