CONFIDENT AND ADAPTIVE GENERATIVE SPEECH RECOGNITION VIA CONFORMAL RISK CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic Speech Recognition (ASR) systems frequently produce transcription errors due to acoustic variability, which require post-processing correction methods. Recent approaches leverage Large Language Models (LLMs) for generative ASR error correction using N-best hypotheses but rely on fixed set sizes regardless of input complexity and do not provide performance guarantees. We propose an adaptive framework that dynamically determines the optimal number of hypotheses for each input using conformal risk control (CRC). This mechanism leverages ASR confidence scores and applies CRC to control the expected relative word error rate degradation compared to the best achievable performance for a given model and hypothesis set. Experimental results show that our approach matches or exceeds fixed-size correction baselines while requiring fewer hypotheses on average, maintaining robust performance under diverse acoustic conditions.

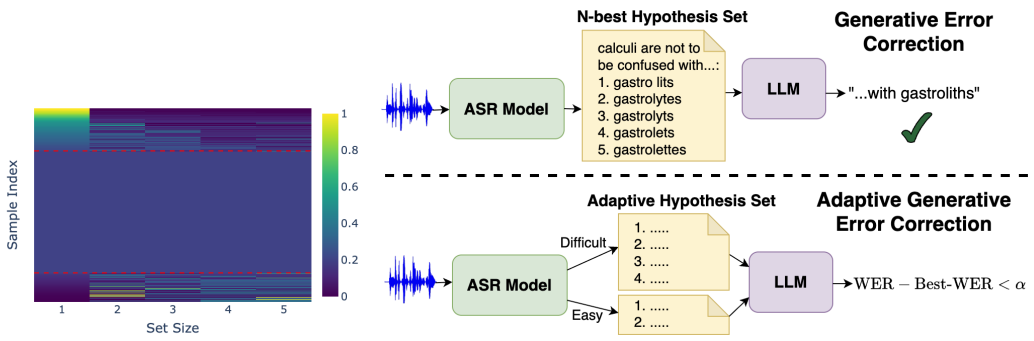
1 INTRODUCTION

ASR systems convert spoken language into text, enabling a wide array of applications from virtual assistants to transcription services (Kheddar et al., 2024). Over the past decade, deep learning advancements have propelled ASR performance, with models like Wav2Vec (Baevski et al., 2020) and Whisper (Radford et al., 2023) achieving remarkable accuracy on benchmark datasets through self-supervised learning and large-scale training. However, ASR remains challenged by real-world variability, including background noise, speaker accents, dialects, homophones, out-of-vocabulary words, and domain shifts, which often lead to transcription errors that degrade downstream tasks (Schneider et al., 2019).

To mitigate these issues, recent research (Yang et al., 2023; Ma et al., 2025; Mu et al., 2025) has explored integrating LLMs with ASR outputs for post-processing. A prominent approach involves generative error correction (GER), where the LLM receives a fixed-size set of hypotheses, produced by the ASR model, and is asked to provide improved transcriptions (Chen et al., 2023; Hu et al., 2024). Usually the LLM is fine-tuned on sequences of N-best hypotheses to learn mappings from noisy ASR outputs to ground-truth text, demonstrating noise-robust improvements.

Despite these advances, existing GER methods suffer from key limitations. They predominantly rely on a fixed hypothesis set size across all inputs, applying the same N value irrespective of whether the audio is simple (e.g., clear speech) or complex (e.g., accented or noisy), which can result in inefficient resource use—overloading the LLM with redundant hypotheses for straightforward cases or introducing low-quality hypotheses that may degrade correction performance, as can be seen in Fig. 1(a). Furthermore, these approaches lack statistical guarantees on the expected performance, such as bounding the gap to the oracle (best possible) transcription, leaving uncertainty in their reliability and their practical improvement.

To address these shortcomings, we propose an adaptive framework for hypothesis set construction in LLM-augmented ASR, as illustrated in Fig. 1(b). Instead of a static N, we dynamically form sets using a threshold rule over the likelihood scores of ASR hypotheses, ensuring only sufficiently plausible candidates are passed to the LLM. We tune these thresholds via conformal risk control (CRC) (Angelopoulos et al., 2024b), a distribution-free framework that we demonstrate provides risk control on the expected loss (e.g., word error rate (WER)) relative to the oracle performance. This adaptive strategy yields smaller average set sizes, reducing computational costs, while empirically achieving comparable or lower WERs compared to fixed-N baselines on diverse benchmarks.



(a) Motivating Example (b) Method Illustration

Figure 1: **(a)** WER performance patterns across hypothesis set sizes over **TedLium-3**. Samples are grouped by monotonicity: samples that improve with more hypotheses (top), show consistent performance (middle), or degrade with more hypotheses (bottom). **(b)** Comparison of standard GER using fixed 5-hypothesis sets versus our adaptive GER that dynamically selects variable-sized hypothesis sets with conformal risk control to bound relative performance degradation from the oracle.

Our main contributions can be summarized as follows:

- We propose an adaptive hypothesis selection framework that leverages ASR confidence scores to dynamically determine the optimal set sizes for each input, replacing the standard fixed-size approach with difficulty-aware resource allocation.
- We introduce the first application of CRC to GER, empirically demonstrating effective control over relative performance degradation while enabling principled uncertainty quantification in multi-hypothesis scenarios.
- We demonstrate substantial computational efficiency gains (up to 57.1% reduction in hypothesis usage) while maintaining correction performance across diverse acoustic conditions, validating both the empirical robustness and practical value of the proposed adaptive selection mechanism.

2 Related Work

Automatic Speech Recognition Error Correction. Language model rescoring has been extensively employed in ASR systems to enhance recognition accuracy, with external language models reranking N-best hypothesis lists to select optimal transcriptions (Song et al., 2021). Recent advances have moved beyond simple reranking toward generative error correction (GER), where LLMs synthesize improved transcriptions by leveraging complete N-best lists rather than merely selecting among existing candidates (Yang et al., 2023; Radhakrishnan et al., 2023; Yang et al., 2024; Ma et al., 2025; Liu et al., 2025; Ghosh et al., 2024; Mu et al., 2025). Contemporary benchmarks like HyPoradise (Chen et al., 2023) have formalized the hypotheses-to-transcription (H2T) mapping task, enabling systematic evaluation of LLM-based correction methods across diverse acoustic conditions. Our approach builds upon this foundation while introducing reliable and adaptive hypothesis selection via CRC.

Uncertainty Quantification in Language and Speech Processing. Uncertainty quantification has become critical for deploying natural language processing (NLP) and speech systems in high-stakes applications, with traditional approaches including ensemble methods, Monte Carlo dropout, and calibration techniques for well-calibrated probability estimates (Xiao et al., 2022). Speech processing faces unique challenges due to temporal audio signals and cascading recognition errors, leading to various approaches including acoustic confidence measures and neural uncertainty estimation (Wullach & Chazan, 2023; Rumberg et al., 2025). However, these methods often lack theoretical guarantees and may not generalize across acoustic conditions, making CRC attractive as a principled approach providing distribution-free uncertainty quantification.

Conformal Prediction and Conformal Risk Control. Conformal prediction (CP) (Vovk et al., 2005; Angelopoulos et al., 2024a) provides a distribution-free framework for uncertainty quantification that constructs prediction sets with guaranteed coverage under minimal exchangeability assumptions, without requiring distributional assumptions about models or data. The framework has found extensive applications across regression, classification, and structured prediction tasks,

including recent demonstrations in NLP for machine translation, text classification, and question answering (Campos et al., 2024). conformal risk control (CRC) extends CP's coverage guarantees to control expected loss functions beyond simple miscoverage, allowing practitioners to specify task-specific risk tolerances while maintaining distribution-free guarantees (Angelopoulos et al., 2024b). This generalization has established CRC as a versatile framework for applications requiring rigorous risk management where both performance and reliability are critical.

3 Problem Formulation

Consider an input audio signal $x \in \mathcal{X}$, and a corresponding transcription y. Possible transcription *hypotheses* are generated by a pre-trained ASR model using beam search decoding, and the top N are selected:

$$\mathcal{H}_N = \{(\hat{y}_1, c_1), (\hat{y}_2, c_2), \ldots, (\hat{y}_N, c_N)\} \tag{1}$$

where $\hat{y}_i$ represents the i-th hypothesis transcription and $c_i = \log p(y_i|x)$ denotes the log-likelihood score from the ASR model. The hypotheses are ranked by their scores in descending order such that $c_1 \geq c_2 \geq \ldots \geq c_N$, with higher scores indicating higher confidence.

The goal is to learn a mapping function $\mathcal{M}_{\text{H2T}}$ that predicts an improved transcription $\hat{y}^*$ from the N-best list:

$$\hat{y}^* = \mathcal{M}_{\text{H2T}}(\mathcal{H}_N; \theta) \tag{2}$$

where θ represents learnable parameters. While traditional language model rescoring approaches (Song et al., 2021) re-rank existing hypotheses to select the best candidate, generative error correction (GER)(Ma et al., 2025; Hu et al., 2024; Yang et al., 2023) represents the current state-of-the-art approach that can synthesize new transcriptions by leveraging information across all N-best hypotheses, potentially producing corrections that do not appear in the original hypothesis list.

LLMs have emerged as powerful tools for this task due to their ability to understand linguistic patterns and perform text generation. The common approaches involve either leveraging existing LLMs with various prompt engineering techniques (Chen et al., 2023; Yang et al., 2023) or fine-tuning a pre-trained LLM to learn the mapping from N-best hypotheses to ground-truth transcriptions (Hu et al., 2024; Radhakrishnan et al., 2023). The model receives the ranked hypotheses (optionally along with their confidence scores) as input and generates the corrected transcription autoregressively. The training process utilizes pairs $(\mathcal{H}_N, y)$, enabling the model to learn the relationship between ASR error patterns and optimal corrections across diverse acoustic conditions and speaking styles.

However, the conventional approach of using fixed-sized hypothesis sets overlooks a critical observation: not all audio segments require the same number of hypotheses for effective correction. In many cases, a smaller set is sufficient or even preferable for achieving optimal transcription quality. As illustrated in Fig. 1, there exist numerous instances where smaller hypothesis sets are sufficient and sometimes even yield better corrections than larger ones. This phenomenon suggests that additional hypotheses can introduce noise rather than useful signal, motivating the need for adaptive selection mechanisms that can dynamically determine the optimal number of hypotheses based on the specific characteristics of each audio segment.

4 Background - Conformal Risk Control

Conformal prediction (CP) is a distribution-free framework for uncertainty quantification that requires only the weak assumption of exchangeability between calibration and test data, without distributional assumptions about the underlying model or data generating process.A complete exposition is provided in Appendix A.

Consider a calibration dataset $\{(X^{(i)}, Y^{(i)})\}_{i=1}^{m}$, where $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ denote feature-response pairs. While standard CP controls miscoverage probability, many applications require control over more general risk measures. conformal risk control (CRC) extends CP to control the expectation of any bounded, monotone loss function $\ell : 2^{\mathcal{Y}} \times \mathcal{Y} \to [0, B]$:

$$E[\ell(\Gamma_\lambda(X^{(m+1)}), Y^{(m+1)})] \leq \alpha \tag{3}$$

where $\Gamma_\lambda : \mathcal{X} \to 2^{\mathcal{Y}}$ represents a parameterized prediction set function, $(X^{(m+1)}, Y^{(m+1)})$ is a new test point drawn exchangeably with the calibration data, and the expectation is taken over the randomness in both calibration data and the test point.

The key insight is that for monotone loss functions—where enlarging the prediction set cannot increase the loss—CRC maintains the distribution-free guarantees of standard CP while enabling control over task-specific risk measures. The CRC threshold selection procedure aims to find the optimal threshold:

$$\hat{\lambda} = \inf \left\{ \lambda : \frac{m}{m+1} \hat{R}_m(\lambda) + \frac{B}{m+1} \leq \alpha \right\}, \tag{4}$$

where the empirical risk is computed as:

$$\hat{R}_m(\lambda) = \frac{1}{m} \sum_{i=1}^{m} \ell(\Gamma_\lambda(X^{(i)}), Y^{(i)}), \tag{5}$$

representing the average empirical loss over the calibration set. For monotone loss functions, this threshold can be found efficiently by gradually adjusting λ until the risk constraint is satisfied.

CRC provides finite-sample guarantees that are tight up to $O(1/m)$ terms as stated in the following Theorem.

Theorem 1 (CRC Finite-Sample Guarantee)**.** *Under the exchangeability assumption and for bounded monotone loss functions, the set predictor $\Gamma_{\hat{\lambda}}$ selected by the CRC procedure satisfies:*

$$\alpha - \frac{2B}{m+1} \leq E[\ell(\Gamma_{\hat{\lambda}}(X^{(m+1)}), Y^{(m+1)})] \leq \alpha. \tag{6}$$

Note that CRC reduces to standard CP when the loss function is the miscoverage indicator. CRC has been applied to areas such as such as medical diagnosis, autonomous driving, ordinal classification, and ranked retrieval systems (Andéol et al., 2023; Xu et al., 2023; 2024; Overman et al., 2024).

5 METHOD

5.1 ADAPTIVE HYPOTHESIS SELECTION VIA CONFORMAL RISK CONTROL

Building on the GER framework, presented in § 3, we propose an adaptive selection mechanism that dynamically estimates the optimal number of hypotheses for each input sample. Rather than using a fixed set of size N, our approach selects the minimal subset size n^* that maintains correction performance while reducing computational cost.

Adaptive hypothesis set. We formulate an adaptive hypotheis selection problem within the CRC framework, established in § 4. We define adaptive hypothesis sets, parametrized by λ as:

$$\Gamma_\lambda(\mathcal{H}_N) = \{(\hat{y}_1, c_1), \ldots, (\hat{y}_n, c_n)\}, \tag{7}$$

where n is the adaptive set size determined according to λ:

$$n = \min \left\{ j : \sum_{i=1}^{j} s_i \geq \lambda \right\}, \tag{8}$$

and $\mathbf{s} = (s_1, \ldots, s_N)$ represents normalized nonconformity scores derived from ASR confidence scores $\mathbf{c} = (c_1, \ldots, c_N)$.

The enhanced pipeline becomes $\hat{y}^* = \mathcal{M}_{\text{H2T}}(\Gamma_{\hat{\lambda}}(\mathcal{H}_N); \theta)$, where $\hat{\lambda}$ is the calibrated threshold for controlling expected performance degradation. This approach maintains compatibility with any pre-trained H2T model while reducing computational overhead through principled uncertainty quantification.

Risk function and CRC. Our loss is defined with respect to word error rate (WER), which is a standard metric used for evaluating ASR performance. The WER quantifies transcription

Algorithm 1 Adaptive Selection Procedure with CRC Calibration Framework

Require: Calibration set $\{(\mathcal{H}_N^{(i)}, y^{(i)})\}_{i=1}^m$ and a test sample $\mathcal{H}_N^{(m+1)}$
1: **Calibration Phase:**
2: **for** $\lambda \in \Lambda$ (candidate threshold values) **do**
3: Compute $\ell(\Gamma_\lambda(\mathcal{H}_N^{(i)}), y^{(i)})$ for all $i \in [m]$
4: Estimate $\hat{R}_m(\lambda) = \frac{1}{m}\sum_{i=1}^m \ell(\Gamma_\lambda(\mathcal{H}_N^{(i)}), y_i)$
5: **end for**
6: Select $\hat{\lambda} = \inf\left\{\lambda : \frac{m}{m+1}\hat{R}_m(\lambda) + \frac{B}{m+1} \leq \alpha\right\}$
7: **Test Phase:**
8: Compute normalized scores $\mathbf{s} = \text{softmax}(\phi_\gamma(\mathbf{c}^{(m+1)})/\tau)$
9: Select $n^* = \min\{n : \sum_{i=1}^n s_i \geq \hat{\lambda}\}$
10: **Return** hypothesis set $\mathcal{H}_{n^*}^{(m+1)} = \left\{\left(\hat{y}_1^{(m+1)}, c_1^{(m+1)}\right), \ldots, \left(\hat{y}_{n^*}^{(m+1)}, c_{n^*}^{(m+1)}\right)\right\}$

accuracy by measuring the minimum number of word-level edits required to transform the predicted transcription into the ground truth:

$$\text{WER}(\hat{y}, y) = \frac{S(\hat{y}, y) + D(\hat{y}, y) + I(\hat{y}, y)}{W(\hat{y})} \tag{9}$$

where $S(\hat{y}, y)$, $D(\hat{y}, y)$, and $I(\hat{y}, y)$ represent the number of substitutions, deletions, and insertions, respectively, and $W(y)$ is the total number of words in the reference transcription.

Rather than controlling absolute WER, which requires domain-specific thresholds, we control the *per-sample* relative degradation from the best achievable performance:

$$\ell(\Gamma_\lambda(\mathcal{H}_N), y) = \text{WER}(\mathcal{M}_{\text{H2T}}(\Gamma_\lambda(\mathcal{H}_N)), y) - \min_{j \in [N]} \text{WER}(\mathcal{M}_{\text{H2T}}(\mathcal{H}_j), y) \tag{10}$$

where $\mathcal{H}_j = \{(\hat{y}_1, c_1), \ldots, (\hat{y}_j, c_j)\}$ denotes the top-j hypothesis set.

This loss function exhibits predominantly monotonic behavior, where enlarging the hypothesis set typically does not worsen performance. Our adaptive selection can identify cases where smaller sets are sufficient or even beneficial. In the worst-case scenario, selecting all N hypotheses converges to the standard fixed-N baseline performance, ensuring no performance degradation from existing methods. Finally, our risk control objective adapts the CRC framework:

$$E[\ell(\Gamma_{\hat{\lambda}}(\mathcal{H}_N), Y)] \leq \alpha, \tag{11}$$

where $\hat{\lambda}$ is calibrated according to the CRC procedure, for controlling expected performance degradation. Our method is summarized in Algorithm 1.

Score definition. The selection mechanism relies on a composite score derived from ASR log-likelihoods, designed to adapt flexibly to varying dataset characteristics:

$$\mathbf{s} = \text{softmax}\left(\frac{\phi_\gamma(\mathbf{c})}{\tau}\right) \tag{12}$$

Here, ϕ_γ denotes an adaptive normalization function and τ is a temperature parameter. The function ϕ_γ interpolates between two transformation regimes through a single parameter γ, enabling the score to adjust to dataset-specific speech quality. To prevent redundancy, penalties are applied when the ASR system generates repeated hypotheses. Further details on the adaptive normalization, design rationale, and repetition handling are provided in the Appendix B.1.

5.2 Theoretical Considerations

While CRC provides a principled framework for risk control, our ASR application operates under conditions that slightly deviate from the strict theoretical assumptions. We address these deviations and their practical implications.

Bounded loss. The CRC framework requires bounded, monotone loss functions. Our loss function satisfies boundedness through clipping: we enforce $\ell(\Gamma_\lambda(\mathcal{H}_N), y) \leq B$ where B is set based on validation set statistics such that violations are rare and have negligible impact.

Monotonicity. Strict monotonicity is violated in approximately 20% of samples, where smaller hypothesis sets occasionally outperform larger ones. In addition, 95% of consecutive pairwise comparisons maintain monotonicity, indicating predominantly monotonic behavior with minor violations. To address this, we evaluated the monotonizing procedure proposed by Angelopoulos et al. (2024b) for handling non-monotone loss functions. This approach constructs $\tilde{\ell}_i(\lambda) = \sup_{\lambda' \geq \lambda} \ell_i(\lambda')$ to enforce monotonicity with asymptotic guarantees. However, empirical results showed degraded performance across datasets, likely because monotonizing eliminates the beneficial non-monotonic cases that our adaptive method is designed to exploit. By forcing conservative behavior, this procedure prevents identification of scenarios where smaller sets genuinely outperform larger ones—precisely the phenomenon enabling our computational savings. We therefore use the loss as is, as it demonstrates empirical robustness and maintains effective risk control in practice despite the slight assumption violations. This suggests that the theoretical worst-case scenarios may not reflect typical ASR behavior, where monotonicity violations often signal exploitable efficiency opportunities rather than problematic cases.

6 EXPERIMENTAL SETUP

Datasets and Benchmark We evaluate our approach on three datasets from the HyPoradise benchmark (Chen et al., 2023), spanning different acoustic difficulty levels based on average WER performance:

- **TedLium-3** (Hernandez et al., 2018) (avg. WER $\sim 8\%$) contains TED Talk recordings with diverse noise, accents, and topics. Following HyPoradise protocol, we sample $50,000$ utterances: $47,500$ for training/validation, and $2,500$ for calibration/test.
- **CHiME-4** (Vincent et al., 2017) (avg. WER $\sim 11\%$) contains far-field noisy recordings across different environments. We use the complete train split ($9,600$ utterances) for train/validation and test-real split ($1,320$ utterances) for calibration/test. Data was obtained from RobustGER (Hu et al., 2024), which provides the required ASR likelihood scores.
- **CommonVoice** (Ardila et al., 2020) (avg. WER $\sim 14\%$) contains multilingual recordings from diverse speakers with different accents. We select $50,000$ samples from train-en split using $47,500$ samples for train/validation, and $2,500$ samples for calibration/test.

ASR Hypothesis Generation. We employ Whisper models (Radford et al., 2023) for N-best hypothesis generation via beam search, removing repetitive utterances and selecting top-5 ($N = 5$) hypotheses by posterior probability. TedLium-3 and CommonVoice use Whisper-base (beam-width is 60, following HyPoradise (Chen et al., 2023)), while CHiME-4 uses Whisper-Large-v2 (beam-width is 50, following RobustGER (Hu et al., 2024)).

LLM and Training. We fine-tune LLaMA-2-7B (Touvron et al., 2023) using LoRA (Hu et al., 2022) for efficient H2T mapping. The model generates corrected transcriptions from N-best inputs via standard next-token prediction. Training details, hyperparameters, and computational requirements are in Appendix C.

CRC Calibration. We use the validation data to determine both the target risk levels (α) and the dataset-specific score function parameters (γ and τ), based on the empirical performance and score discriminability patterns. The selection methodology and theoretical considerations are discussed in Appendix B.

6.1 EVALUATION METRICS

Performance Measurements. We evaluate our approach using WER as the primary metric, as described in § 5. We employ two complementary WER calculation methodologies. The primary approach performs instance-level computation followed by averaging across samples, directly corresponding to the defined loss function 10. As secondary validation, we compute corpus-level

Table 1: WER (%) results with LLaMA-2-7B fine-tuning. Baseline: Whisper's top-1 hypothesis. O_{llm}: post-LLM oracle. Our method results represent one operating point from Figure 2. Subscript percentages denote relative WER change vs. vanilla GER (WER column) and relative size reduction vs. constant $N = 5$ (size column).

Test Set	Baseline	GER	Our Method		α(Target WER)	O_{llm}
			Set Size	WER		
TedLium-3	8.0	6.04	$2.145_{57.1\%}$	$5.96_{-1.3\%}$	$1.785_{(6.115)}$	4.33
CHiME-4	11.49	6.38	$3.8_{24.0\%}$	$6.55_{+2.7\%}$	$1.9_{(6.63)}$	4.73
CommonVoice	14.1	8.46	$3.1_{38.0\%}$	$8.5_{+0.5\%}$	$1.7_{(8.65)}$	6.95

WER through concatenation of all predictions and references using the `evaluate`[1] package, which reduces sensitivity to sample length variability and enables comparison with prior works using corpus-level conventions.

Risk Control Validation. Beyond standard WER evaluation, we validate the empirical effectiveness of our CRC framework by tracking whether average relative WER degradation remains below the specified target α, demonstrating effective risk control in practice. We compare against all constant set sizes $(1-5)$ to show that our adaptive method achieves superior performance-efficiency trade-offs across all possible fixed-size baselines.

Experimental Protocol. To ensure statistical reliability, we perform $T = 50$ independent trials with resampled calibration/test splits, allocating 30-40% of test samples for calibration Among the selected α levels, we report only configurations achieving 100% calibration success rate. Final results represent mean values across all trials.

7 Results and Analysis

7.1 WER and Set Size Tradeoffs

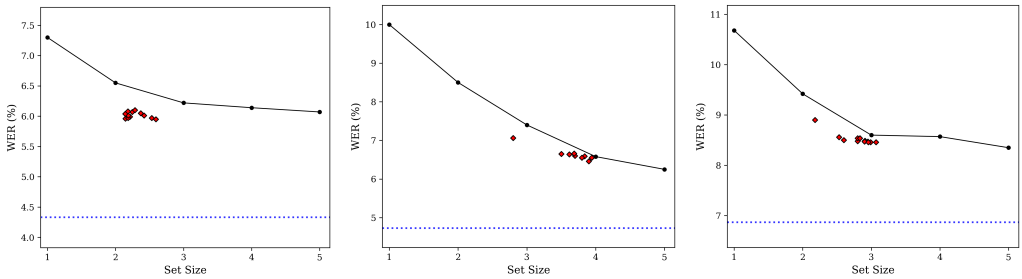
Table 1 presents experimental results across datasets. The baseline performance corresponds to Whisper's top-1 hypothesis, establishing the initial recognition accuracy before post-processing. The GER results demonstrate the effectiveness of the fine-tuned LLaMA-2-7B model when provided with a fixed set of top-5 hypotheses. For reference, we include the oracle bound O_{llm}, which represents the best possible performance when the LLM receives the optimal number of hypotheses for each sample (between $1-5$).

The results show that GER achieves substantial improvements over the baseline across all conditions, with gains varying according to dataset difficulty. Our adaptive selection framework demonstrates superior computational efficiency while preserving or enhancing correction quality. Our method dynamically determines the optimal number of hypotheses for each input, as reflected in the average set sizes reported. These results validate the effectivness of our approach across diverse acoustic conditions. On TedLium-3, the method achieves a 57.1% reduction in average set size while improving performance (5.96% vs 6.04% WER). CHiME-4 demonstrates computational savings of 24% with a modest performance trade-off of 2.7% relative increase in WER. CommonVoice exhibits 38% computational reduction with minimal performance impact (0.5% relative increase). Notably, the adaptive selection mechanism occasionally outperforms the fixed-size baseline, indicating that excessive hypotheses can introduce noise rather than useful information for error correction.

Regarding the reliability of our selection mechanism, we find that the obtained WER remains below the target, empirically confirming the effectiveness of our approach in providing guaranteed performance—a property absent in prior methods and made possible through the CRC framework.

Our complementary corpus-level WERs are presented in Table D.1. Though this metric produces different absolute values but maintain the same relative trends and ordering compared to constant set sizes, confirming the robustness of our findings across evaluation methodologies.

[1] https://pypi.org/project/evaluate/



(a) TedLium-3 (b) CHiME-4 (c) CommonVoice

Figure 2: Performance-compute trade-off analysis across datasets. Each plot shows WER vs. set size for constant set sizes (connected line), oracle performance level (vertical line), and our adaptive method performance (points).

Figure 2 illustrates the performance-compute trade-off characteristics of our adaptive approach compared to fixed set sizes across all datasets. Each subplot displays the WER performance curve for constant set sizes $N = 1$ through $N = 5$, with vertical reference line indicating the O_{llm} oracle performance bound. We observe that our method's operating points consistently demonstrate better tradeoffs relative to the fixed-set performance curve, achieving computational efficiency gains while maintaining competitive or even improved error rates.

7.2 Analysis

To better understand the adaptive selection mechanism's behavior, we examine representative cases that illustrate when different set sizes are optimal. Table 2 presents three scenarios with complete hypothesis lists, ASR scores, and LLM predictions that demonstrate the correlation between score distributions and optimal set sizes.

Case 1: Full Set Required (Common Voice): When ASR scores exhibit narrow gaps (-0.42 to -0.51), our method correctly identifies the need for comprehensive information. The LLM progressively refines its prediction across set sizes, ultimately achieving perfect accuracy with the complete hypothesis set by correctly generating "gastroliths" rather than the various incorrect alternatives ("gallstones," "gastrolytes"). The compressed score distribution leads our normalization to select larger sets, aligning with the empirical benefit of additional hypotheses.

Case 2: Single Hypothesis Optimal (TedLium-3): When the top hypothesis achieves perfect accuracy and exhibits substantial score separation (-0.21 vs -0.31), additional hypotheses degrade performance from 0% to 21% WER. In this case, the discriminative score gap correctly signals high confidence in the first hypothesis, leading our method to favor minimal sets. This demonstrates that additional hypotheses can introduce harmful noise.

Case 3: Performance Plateau (CHiME-4): When WER remains constant (6.25%) across all set sizes, our method demonstrates computational efficiency potential. While the tight score clustering (-0.46 to -0.49) would typically lead our normalization to select larger sets, this case illustrates where our approach provides a safety net—in the worst case, we select all 5 hypotheses and achieve identical performance to the baseline, but when score normalization successfully identifies the plateau, we achieve the same WER with reduced computational cost.

These examples demonstrate how our adaptive selection responds to different ASR confidence patterns: discriminative scores enable efficient small sets, while compressed scores lead to more comprehensive hypothesis selection. This validates our approach of dynamically adjusting set sizes based on the underlying score distributions rather than using fixed configurations.

7.3 Ablation Studies

We briefly report several ablation studies that we performed to validate different aspects of our proposed framework.

Alternative Problem Formulations. We evaluated multiple CP and CRC configurations including absolute WER targets, coverage-based objectives for samples below specified WER thresholds,

Table 2: Representative examples showing the relationship between ASR score distributions and optimal set sizes. Case 1 demonstrates progressive improvement with larger sets, Case 2 shows degradation beyond the optimal single hypothesis, and Case 3 illustrates performance plateau enabling computational savings.

Case	Hypotheses	Score	LLM Predictions by Set Size	WER per Size (%)
Case 1: Full Set	H1: calculi are not to be confused with gastro lits	-0.42	Size 1: ...with gallstones	12.5
	H2: calculi are not to be confused with gastrolytes	-0.44	Size 2: ...with gastrolytes	12.5
	H3: calculi are not to be confused with gastrolyts	-0.47	Size 3: ...with gastrolytes	12.5
	H4: calculi are not to be confused with gastrolets	-0.50	Size 4: ...with gastrolytes	12.5
	H5: calculi are not to be confused with gastrolettes	-0.51	Size 5: ...with gastroliths	0.0
	GT: calculi are not to be confused with gastroliths			
Case 2: Single Opt.	H1: ...medical team assign of the ship...	-0.21	Size 1: ...team a sign of...	0.0
	H2: ...medical team a sign of the ship...	-0.31	Size 2-5: ...team assigned to...	21
	H3: ...medical team assigned to the ship...	-0.37		21
	H4: ...medical team assigned of the ship...	-0.41		21
	H5: ...medical team assigned the ship...	-0.43		21
	GT: ...medical team a sign of the ship...			
Case 3: Plateau	H1: ...new york state sold about seventy seven million of...	-0.46	All sizes: separately new york	6.25
	H2: ...new york state sold about seventy seven million in...	-0.47	state sold about seventy seven	6.25
	H3: ...here it states all about seventy seven million in...	-0.47	point one million dollars in	6.25
	H4: ...new york state sold about seventy seven million of...	-0.49	certificates of participation	6.25
	H5: ...new york state sold about seventy seven million dollars in...	-0.49		6.25
	GT: ...seventy seven point one million dollars of...			

and bounded-WER hypothesis guarantees, following approaches from prior ASR uncertainty quantification works (Ernez et al., 2023). These alternatives consistently yielded inferior empirical performance compared to our relative loss, defined in Eq. 10. Absolute WER targets operate at a global level without instance-specific optimization, while bounded-WER guarantees (Ernez et al., 2023) showed poor correlation between hypothesis quality and final LLM output quality, validating our relative degradation formulation that adapts to each sample's achievable performance range.

Training Set Size Analysis: We examined our choice of training the LLM with constant-5 hypothesis sets, while evaluating with variable set sizes. To this end, we conducted comprehensive ablation experiments training separate LLaMA-2-7B models on fixed set sizes (1-5 hypotheses), as well as dynamic sizes, then evaluating each model across all possible test set sizes. The results are reported in Tab. D.2. This 6×5 result matrix reveals that while specific combinations (e.g., train-3/test-3) occasionally outperformed the baseline, the constant-5 trained model achieved optimal average WER across all test configurations. These results confirm that our adaptive approach provides genuine improvements over the best achievable fixed-size baseline, establishing the validity of our comparative framework.

8 Conclusion and Future Work

This work presents an adaptive framework for hypothesis selection in generative ASR error correction, addressing computational inefficiency through principled uncertainty quantification. Our method employs CRC to dynamically determine optimal hypothesis set sizes, demonstrating substantial computational savings while maintaining competitive performance across datasets with diverse acoustic conditions. The framework requires only calibration without model retraining, enabling straightforward adoption in existing systems.

Future work could investigate confidence-driven adaptive compute allocation in multi-model systems, including reasoning and agent-based applications, where similar mechanisms for identifying and reducing computational costs may achieve comparable performance with greater efficiency.

References

Léo Andéol, Thomas Fel, Florence De Grancey, and Luca Mossina. Confident object detection via conformal prediction and conformal risk control: an application to railway signaling. In *Conformal and Probabilistic Prediction with Applications*, pp. 36–55. PMLR, 2023.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024a.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024b.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4211–4215, 2020.

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

Margarida M. Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516, 2024. doi: 10.1162/tacl_a_00715.

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*, 36:31665–31688, 2023.

Fares Ernez, Alexandre Arnold, Audrey Galametz, Catherine Kobus, and Nawal Ould-Amer. Applying the conformal prediction paradigm for the uncertainty quantification of an end-to-end automatic speech recognition model (wav2vec 2.0). In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pp. 16–35. PMLR, 2023.

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.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation. *arXiv preprint arXiv:1805.04699*, 2018.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and Eng Siong Chng. Large language models are efficient learners of noise-robust speech recognition. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.10446.

Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. Automatic speech recognition using advanced deep learning approaches: A survey. *Information fusion*, 109:102422, 2024.

Yanyan Liu, Minqiang Xu, Yihao Chen, Liang He, Lei Fang, Sian Fang, and Lin Liu. Denoising ger: A noise-robust generative error correction with llm for speech recognition. *arXiv preprint arXiv:2509.04392*, 2025.

R. Ma, M. Qian, M. Gales, and K. Knill. Asr error correction using large language models. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1389–1401, 2025. doi: 10.1109/TASLPRO.2025.3551083.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

Bingshen Mu, Kun Wei, Pengcheng Guo, and Lei Xie. Mixture of lora experts with multi-modal and multi-granularity llm generative error correction for accented speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

William Overman, Jacqueline Vallon, and Mohsen Bayati. Aligning model properties via conformal risk control. *Advances in Neural Information Processing Systems*, 37:110702–110722, 2024.

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.

Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A. Kiani, David Gomez-Cabrero, and Jesper N. Tegner. Whispering llama: A cross-modal generative error correction framework for speech recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10007–10016, Singapore, 2023. Association for Computational Linguistics.

Lars Rumberg, Christopher Gebauer, and Jörn Ostermann. Aggregation-free uncertainty estimation for ctc-based automatic speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *Proceedings of Interspeech*, pp. 3465–3469, 2019.

Yuanfeng Song, Di Jiang, Xuefang Zhao, Qian Xu, Raymond Chi-Wing Wong, Lixin Fan, and Qiang Yang. L2rs: A learning-to-rescore mechanism for hybrid speech recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3474–3482. ACM, 2021.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 46:535–557, 2017.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Tomer Wullach and Shlomo E Chazan. Don't be so sure! boosting asr decoding via confidence relaxation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13780–13788, 2023.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 7273–7284, 2022.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Conformal risk control for ordinal classification. In *Uncertainty in Artificial Intelligence*, pp. 2346–2355. PMLR, 2023.

Yunpeng Xu, Mufang Ying, Wenge Guo, and Zhi Wei. Two-stage risk control with application to ranked retrieval. *arXiv preprint arXiv:2404.17769*, 2024.

C.-H. Huck Yang, Y. Gu, Y.-C. Liu, S. Ghosh, I. Bulyko, and A. 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, 2023. doi: 10.1109/ASRU57964.2023.10389673.

C.-H. Huck Yang et al. Large language model based generative error correction: A challenge and baselines for speech recognition, speaker tagging, and emotion recognition. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 371–378, 2024. doi: 10.1109/SLT61566.2024.10832176.

A CONFORMAL PREDICTION FRAMEWORK

Let $\mathcal{X}$ denote the input space and $\mathcal{H}$ the output space. Consider a calibration set $\{(X^{(i)}, Y^{(i)})\}_{i=1}^{m}$ where $(X^{(i)}, Y^{(i)}) \in \mathcal{X} \times \mathcal{H}$, and a new test point $(X^{(m+1)}, Y^{(m+1)})$. Conformal prediction requires that the calibration data and test point are exchangeable, meaning the joint distribution remains invariant under permutations.

Conformal prediction is a distribution-free statistical framework that provides uncertainty quantification for machine learning predictions with finite-sample guarantees. Given a calibration dataset separate from training data, CP constructs prediction sets that satisfy coverage properties regardless of the underlying model architecture or data distribution.

For the test input $X^{(m+1)}$ with unknown label $Y^{(m+1)}$, the goal is to construct a prediction set $C(X^{(m+1)})$ such that:

$$P(Y^{(m+1)} \notin C(X^{(m+1)})) \leq \alpha \tag{13}$$

where α is a user-specified significance level (e.g., 0.1 for 90% coverage).

The framework relies on nonconformity scores $s_i : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that measure how atypical a prediction is for a given input. For any pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the nonconformity score $s(x, y)$ should reflect their agreement, with lower scores indicating better agreement. For the calibration set, where true labels are known, $s_i = s(X^{(i)}, Y^{(i)})$ quantifies the disagreement between the model's prediction and the true label. The prediction set is then constructed by including all labels whose nonconformity scores fall below a data-dependent threshold.

B CONFORMAL RISK CONTROL IMPLEMENTATION DETAILS

B.1 SCORE FUNCTION DESIGN

B.1.1 MOTIVATION

Our analysis revealed that score distributions vary significantly across datasets with different noise characteristics. Higher signal-to-noise ratio conditions (e.g., TedLium-3) produce more discriminative ASR confidence scores, while challenging acoustic environments (e.g., CommonVoice) yield compressed score distributions. Temperature-only adaptation (means extreme values for the compressed distributions) proved insufficient, creating overly homogeneous score distributions that degraded selection quality. Consequently, we developed the two-level normalization strategy with the additional parameter γ, enabling adaptive score transformation based on dataset difficulty. This parameterization directly influences the resulting score distributions and, subsequently, the selected set sizes, proving essential for robust performance across diverse acoustic conditions.

B.1.2 FUNCTION DESIGN

The normalization function $\phi_\gamma(\mathbf{c})$ smoothly interpolates between two transformation regimes based on a single parameter $\gamma \in [0, 1]$:

$$\phi_\gamma(\mathbf{c}) = w_{\text{inv}}(\gamma) \cdot f_{\text{inv}}(\mathbf{c}) + w_{\text{id}}(\gamma) \cdot f_{\text{id}}(\mathbf{c}) \tag{14}$$

where:

- $f_{\text{id}}(\mathbf{c}) = \mathbf{c}$ (identity transformation)
- $f_{\text{inv}}(\mathbf{c}) = -1/\mathbf{c}$ (reciprocal transformation)

The identity transformation $f_{\text{id}}(\mathbf{c}) = \mathbf{c}$ preserves the natural ASR score differences, suitable for high-SNR conditions where scores are discriminative. The reciprocal transformation $f_{\text{inv}}(\mathbf{c}) = -1/\mathbf{c}$ amplifies small differences between compressed scores, beneficial for challenging acoustic conditions where ASR confidence variations are minimal.

The parameter $\gamma \in [0, 1]$ controls the interpolation: $\gamma = 1$ corresponds to pure identity (high-SNR), while $\gamma = 0$ corresponds to pure reciprocal transformation (low-SNR). The linear weights $w_{\text{id}}(\gamma) = \gamma$ and $w_{\text{inv}}(\gamma) = (1 - \gamma)$ ensure smooth transitions between regimes.

B.1.3 Chosen Values

We select γ and temperature parameters based on dataset discriminability characteristics. Easier conditions with discriminative ASR scores use $\gamma \approx 1$ (preserving natural score differences), while challenging acoustic conditions require $\gamma \approx 0$ (amplifying small differences through reciprocal transformation). The temperature parameter balances the softmax distribution accordingly.

Table B.1 presents the optimal parameters determined for each dataset:

Dataset	γ	Temperature (τ)
TedLium-3	1.0	0.05
CHiME-4	0.5	1.0
CommonVoice	0.0	1.0

Table B.1: Score function parameters for each dataset.

These parameter selections reflect the varying signal-to-noise characteristics across datasets. TedLium-3, with the highest SNR from clean TED talk recordings, uses $\gamma = 1.0$ to preserve the naturally discriminative ASR confidence scores, combined with low temperature ($\tau = 0.05$) to create sharp selection boundaries. CHiME-4, representing intermediate SNR with moderate noise environments, employs $\gamma = 0.5$ to balance between preserving and amplifying score differences. CommonVoice, with the most challenging acoustic conditions and lowest effective SNR due to diverse speaker accents and recording qualities, requires $\gamma = 0.0$ to maximally amplify small confidence variations through reciprocal transformation. The higher temperature values ($\tau = 1.0$) for CHiME-4 and CommonVoice create smoother selection boundaries appropriate for noisier confidence estimates. In practice, we explore dynamic ranges around these base values rather than single fixed parameters, enabling the multiple operating points with different performance-efficiency trade-offs presented in Figure 2.

B.1.4 Handling Hypothesis Repetitions

When the ASR system produces fewer than N unique hypotheses, repeated hypotheses receive exponentially decaying scores to avoid overweighting redundant information:

$$s_{i,r} = s_i \cdot \beta^r \tag{15}$$

where $R_{i,r}$ is the adjusted score for hypothesis i with repetition count r, and $\beta \in (0, 1)$ is the decay factor. This mechanism is based on the assumption that repeated hypotheses provide no additional information for well-calibrated models, which we validate empirically in our experiments.

B.2 Risk Target Calibration

To establish achievable risk targets α for our CRC framework, we perform preliminary analysis on the validation set to characterize the performance gap between optimal and fixed 5-hypothesis selection. We compute the relative WER degradation statistics across multiple random validation subsets, measuring the difference between the best achievable performance (oracle selection) and constant 5-hypothesis performance for each subset.

From this empirical distribution of performance gaps, we select the 90th percentile as our initial risk target α. This serves as a conservative starting point that we can then adjust based on dataset-specific characteristics: moving slightly downward for more aggressive operating points when validation statistics support tighter bounds, or upward for more conservative settings in challenging acoustic conditions. These adjustments yield the different operating points presented in our results. The validation-derived statistics inform both the selection of achievable α values and the calibration of dataset-specific parameters, providing robustness across varying acoustic conditions while

preventing calibration failures that could compromise the empirical risk control properties of our method.

Table B.2 presents the empirical degradation statistics that inform our risk target selection. The chosen α ranges closely align with the 90th-95th percentile statistics: TedLium-3 uses $\alpha \in [1.75, 1.85]$ (90th-95th percentiles: 1.77-1.79), CommonVoice operates at $\alpha \in [1.695, 1.725]$ (90th-95th percentiles: 1.71-1.87), and CHiME-4 employs $\alpha \in [1.9, 2.0]$ (90th-95th percentiles: 1.87-1.96). These operating ranges demonstrate how dataset-specific statistics guide the selection of achievable risk targets while maintaining conservative bounds. Note that for CHiME-4, due to significant distribution shift between training ($\sim$25% WER), test-real ($\sim$11.5% WER), and dev-real ($\sim$9% WER) subsets, we computed weighted statistics using an external validation subset (dev-real) to better estimate the degradation distribution for the test-real evaluation set.

Dataset	99th	95th	90th
TedLium-3	1.83	1.79	1.77
CHiME-4	2.05	1.96	1.87
CommonVoice	2.0	1.87	1.71

Table B.2: Relative WER degradation statistics across datasets (percentage points).

Additionally, we set implementation parameters based on validation analysis: repetition penalty $\beta = 1.25 - 1.5$ for handling duplicate hypotheses and loss bound $B = 1.25$ across all datasets to account for rare cases exceeding 100% relative WER degradation. These parameters were determined through empirical validation to ensure robust performance across varying hypothesis quality distributions and maintain the bounded loss requirements for our empirical risk control framework.

C LLM Training Configuration Details

C.1 Hyperparameters

We train using AdamW optimizer, effective batch size 32 (achieved through batch size 8 with 4-step gradient accumulation), and cosine learning rate scheduler (with 0.05 warmup ratio). The LoRA configuration uses rank $r = 16$ and scaling parameter $\alpha = 32$, implemented via the PEFT library (Mangrulkar et al., 2022).

Dataset-specific hyperparameters accommodate varying dataset sizes: learning rate range from 5e-5 to 1e-4, dropout rates range from 0.05-0.1, training epochs from 5-10, with larger datasets requiring higher values for both parameters to achieve optimal convergence.

C.2 Prompt Template

The training utilizes the following prompt template:

> *"Correct this speech recognition transcript using the hypotheses below. Provide ONLY the corrected transcript, nothing more.*
> *###Hypotheses:*
> *- {1st ∼ 5th utterances}*
> *###Corrected-transcript:"*

C.3 Computational Requirements

Model training is conducted on a single NVIDIA H100 GPU with 80GB memory. Training duration varies by dataset size: CHiME-4 requires approximately 1 hour due to its smaller scale (9,600 samples), while TedLium-3 and CommonVoice each require 3-4 hours given their larger training sets (47,500 samples each). The LoRA parameterization significantly reduces computational overhead compared to full fine-tuning, enabling efficient adaptation while maintaining the frozen backbone parameters.

D Ablation Study

Table D.1: Corpus-level WER (%) results with LLaMA-2-7B fine-tuning. Our method results represent one operating point from Figure 2. Results show consistent trends with instance-level averaging (Table 1) despite different absolute values, demonstrating robustness across evaluation methodologies. Subscript percentages denote relative WER change vs. vanilla GER and relative size reduction vs. constant $N = 5$.

Test Set	**GER**	**Our Method**		$\mathbf{O}_{llm}$
		Set Size	**WER**	
TedLium-3	5.05	$2.21_{55.8\%}$	$5.05_{0.0\%}$	4.33
CHiME-4	6.37	$3.8_{24.0\%}$	$6.6_{+3.6\%}$	4.73
CommonVoice	7.8	$3.07_{38.6\%}$	$7.95_{+1.9\%}$	6.95

D.1 Analysis of Training Set Size Effects

Note: This ablation study uses a simplified experimental setup with different hyperparameters and dataset splits compared to the main experiments, but demonstrates consistent patterns that validate our core findings.

The ablation results reveal several key patterns that validate our experimental design. The constant-5 training approach achieves the lowest average WER (7.79%) across all test configurations, confirming its superiority as a baseline model. While diagonal elements (matching train/test sizes) occasionally show local optima—such as train-3/test-3 achieving 6.58% versus the train-5/test-3 result of 6.74%—these improvements are marginal and inconsistent across the full evaluation matrix.

Models trained on smaller hypothesis sets exhibit clear performance degradation when tested on larger sets, as expected. The train-1 model struggles significantly with multi-hypothesis inputs, achieving 11.65% WER on 5-hypothesis tests compared to 6.38% for the train-5 model. This demonstrates the importance of exposure to diverse hypothesis patterns during training.

The dynamic training model, despite having access to variable set sizes during training, underperforms the constant-5 baseline (8.43% vs 7.79% average WER). This degraded performance likely stems from the increased complexity of learning hypothesis-to-transcription mappings across varying input lengths simultaneously, creating a more challenging optimization landscape that prevents the model from fully mastering any single configuration. The model must learn to handle the variability in input structure while maintaining transcription quality, leading to suboptimal specialization compared to the focused constant-5 training regime.

These results establish that our adaptive approach provides genuine improvements over the best achievable fixed-size baseline, validating the comparative framework used throughout our main experiments.

Table D.2: Training Set Size Ablation Study: WER (%) across different training and test configurations on CHiME-4 dataset

Train \ Test	**1-hyp**	**2-hyp**	**3-hyp**	**4-hyp**	**5-hyp**	**Average**
Train-1	**10.32**	10.85	11.12	11.38	11.65	11.06
Train-2	10.72	**8.55**	8.92	9.15	9.41	9.35
Train-3	10.89	8.95	**6.58**	6.89	7.12	8.09
Train-4	10.95	9.12	6.89	**6.52**	6.71	8.04
Train-5	10.48	8.69	6.74	6.64	**6.38**	**7.79**
Dynamic	11.23	9.45	7.32	7.18	6.95	8.43