CONFIDENT AND ADAPTIVE GENERATIVE SPEECH RECOGNITION VIA 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 risk control. This mechanism leverages ASR confidence scores and applies Learn then test (LTT) to control the expected relative word error rate degradation compared to the best achievable performance for a given model and hypothesis set. Experimental results demonstrate that our approach provides theoretical guarantees with high-probability bounds while matching or exceeding fixed-size correction baselines and requiring fewer hypotheses on average, achieving substantial computational savings under diverse acoustic conditions.

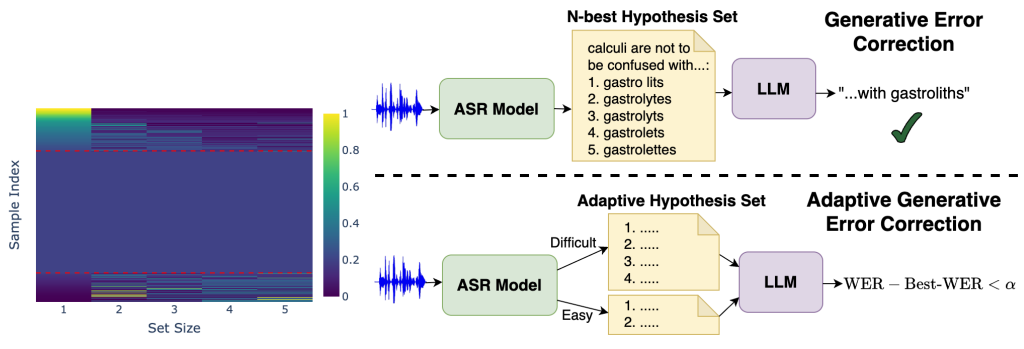
1 INTRODUCTION

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., 2024a). 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 Learn then test (LTT) (Angelopoulos et al., 2025), 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



(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 risk control to bound relative performance degradation from the oracle.

strategy yields smaller average set sizes, reducing computational costs, while empirically achieving comparable or lower WERs compared to fixed-N baselines on diverse benchmarks.

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 risk control to GER, providing theoretical guarantees alongside empirical validation of effective control over relative performance degradation while enabling principled uncertainty quantification in multi-hypothesis scenarios.
- We demonstrate substantial computational efficiency gains (up to 57.6% 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 risk control methods.

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 risk control methods attractive as a principled approach providing distribution-free uncertainty quantification.

Conformal Prediction and Risk Control Methods. Conformal prediction (CP) (Vovk et al., 2005; Angelopoulos et al., 2024a) provides a distribution-free framework for uncertainty quan-

tification 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, but requires bounded monotone loss functions to maintain distribution-free guarantees (Angelopoulos et al., 2024b). Learn then test (LTT) provides an alternative risk control approach that handles non-monotone loss functions through multiple hypothesis testing with family-wise error rate control, offering high-probability bounds without monotonicity assumptions.

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., 2024a; 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., 2024a; 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 - Learn then Test Framework

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. While standard CP controls miscoverage probability, many applications require control over more general risk measures. Conformal risk control (CRC)

extends CP to control bounded, monotone loss functions, but this monotonicity constraint limits applicability when loss functions exhibit non-monotone behavior with respect to the parameter.

The Learn then test (LTT) framework addresses these limitations by reformulating risk control as a multiple hypothesis testing problem, enabling finite-sample guarantees without monotonicity assumptions. 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. Given a parameterized prediction set function $\Gamma_\lambda : \mathcal{X} \to 2^{\mathcal{Y}}$ and bounded loss function $\ell : 2^{\mathcal{Y}} \times \mathcal{Y} \to [0, B]$, our goal is to find $\hat{\lambda}$ such that:

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

where $(X^{(m+1)}, Y^{(m+1)})$ is a new test point drawn exchangeably with the calibration data.

LTT operates on a discrete parameter grid $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ and associates each parameter λ_j with a null hypothesis $H_j : R(\lambda_j) > \alpha$, where $R(\lambda_j) = E[\ell(\Gamma_{\lambda_j}(X), Y)]$. Rejecting hypothesis H_j indicates that λ_j achieves the desired risk control. For each hypothesis, valid p-values are computed using the Hoeffding-Bentkus inequality (Appendix B.3) applied to the empirical risk:

$$\hat{R}_m(\lambda_j) = \frac{1}{m} \sum_{i=1}^{m} \ell(\Gamma_{\lambda_j}(X^{(i)}), Y^{(i)}). \tag{4}$$

A key advantage of LTT is its flexibility in handling non-monotone loss functions without requiring monotonicity enforcement. When combined with sequential testing procedures, LTT can avoid conservative multiplicity corrections by testing hypotheses in a predetermined sequence and stopping at the first rejection, achieving family-wise error rate control without the power loss of traditional methods like Bonferroni correction.

Theorem 1 (LTT Finite-Sample Guarantee)**.** *For any FWER-controlling algorithm at level δ, the rejection set $\hat{\Lambda}$ satisfies:*

$$P\left(\sup_{\lambda \in \hat{\Lambda}}\{R(\lambda)\} \leq \alpha\right) \geq 1 - \delta \tag{5}$$

Unlike CRC, LTT handles arbitrary loss functions, including those exhibiting non-monotone relationships with the parameter λ. This flexibility makes LTT particularly suitable for applications where monotonicity assumptions are violated, such as scenarios where computational efficiency and prediction quality exhibit complex trade-offs.

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 LTT framework, established in § E.1. We define adaptive hypothesis sets, parametrized by λ as:

$$\Gamma_\lambda(\mathcal{H}_N) = \{(\hat{y}_1, c_1), \ldots, (\hat{y}_n, c_n)\}, \tag{6}$$

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

$$n = \min\left\{j : \sum_{i=1}^{j} s_i \geq \lambda\right\}, \tag{7}$$

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

$$\text{WER}(\hat{y}, y) = \frac{S(\hat{y}, y) + D(\hat{y}, y) + I(\hat{y}, y)}{W(\hat{y})} \tag{8}$$

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:

$$\ell(\Gamma_\lambda(\mathcal{H}_N), y) = \text{WER}(\mathcal{M}_{\text{H2T}}(\Gamma_\lambda(\mathcal{H}_N)), y) - \min_{j \in [N]} \text{WER}(\mathcal{M}_{\text{H2T}}(\mathcal{H}_j), y) \tag{9}$$

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 follows the LTT framework:

$$P\left(E[\ell(\Gamma_{\hat{\lambda}}(\mathcal{H}_N), Y)] \leq \alpha\right) \geq 1 - \delta, \tag{10}$$

where $\hat{\lambda}$ is selected from the rejection set obtained by the LTT procedure, providing finite-sample guarantees for expected performance degradation control. Our method is summarized in Algorithm 1.

Algorithm 1 Learn then Test Selection Procedure

Require: Calibration set $\{(H_N^{(i)}, y^{(i)})\}_{i=1}^m$, parameter grid $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$, error level δ

```
for j = 1 to k do
    Compute empirical risk R̂_m(λ_j) = (1/m) Σ_{i=1}^m ℓ(Γ_{λ_j}(H_N^(i)), y^(i))
    Calculate p-value p_j using Hoeffding-Bentkus inequality
    if p_j ≤ δ then
        return λ̂ = λ_j
    end if
end for
return failure (no valid λ found)
```

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

$$\mathbf{s} = \text{softmax}\left(\frac{\phi_\gamma(\mathbf{c})}{\tau}\right) \tag{11}$$

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.

It is important to note that our method is independent of the specific choice of the score used to define the adaptive set. Previous research in ASR has shown that likelihood values do not always provide a reliable measure of confidence (Li et al., 2021; Ravi et al., 2024) While Li et al. (2021); Ravi et al. (2024) focus on top-label calibration, approaches such as that of Popordanoska et al. (2022) offer canonical calibration, which enables the generation of confidence scores for several top hypotheses that can be seamlessly integrated with our framework. Nevertheless, for simplicity, we demonstrate our method using the more commonly available likelihood values.

5.2 Theoretical Considerations

While risk control frameworks provide principled methods for uncertainty quantification, our ASR application operates under conditions that require careful consideration of theoretical assumptions. We address these considerations and their practical implications.

Bounded loss. Risk control frameworks require bounded 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. While some risk control methods like CRC require monotone loss functions, the monotonicity violations in our application ($\sim 20\%$ of cases) represent precisely the efficiency opportunities our adaptive method exploits - scenarios where smaller hypothesis sets genuinely outperform larger ones. To maintain theoretical guarantees without monotonicity constraints, we employ the LTT framework, which handles non-monotone losses naturally through sequential hypothesis testing, ensuring our approach provides rigorous finite-sample bounds regardless of monotonicity violations.

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., 2024a), 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., 2024a)).

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 D.

Risk 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.

Table 1: LTT WER (%) results with LLaMA-2-7B fine-tuning. Baseline: Whisper's top-1 hypothesis. O_{llm}: post-LLM oracle. Our LTT method results with $\delta = 0.25$ across datasets. 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	LTT Method		α	Success Rate	O_{llm}
			Set Size	WER			
TedLium-3	8.0	6.06	$2.121_{57.58\%}$	$6.05_{-0.25\%}$	0.024	0.95	4.38
CHiME-4	11.49	6.24	$3.866_{22.68\%}$	$6.37_{+2.06\%}$	0.027	0.98	4.71
CommonVoice	14.1	8.42	$3.212_{38.09\%}$	$8.55_{+1.62\%}$	0.022	0.97	6.98

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 9. As secondary validation, we compute corpus-level 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 LTT framework by tracking the success rate of risk control across independent trials. For each dataset, we perform multiple calibration-test splits and measure the proportion of trials where the risk constraint $R(\hat{\lambda}) \leq \alpha$ is satisfied, validating the high-probability bound $P(R(\hat{\lambda}) \leq \alpha) \geq 1 - \delta$ empirically. Results demonstrate success rates consistently exceeding the theoretical minimum of $1 - \delta$, confirming 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 while maintaining the theoretical guarantees.

Experimental Protocol. To ensure statistical reliability, we perform $T = 50$ independent trials with resampled calibration/test splits, allocating 30-50% of test samples for calibration. We set $\delta = 0.25$ for the LTT framework to account for relatively small calibration set sizes, preventing overly conservative p-value thresholds in the Hoeffding-Bentkus inequality while maintaining robust theoretical guarantees, resulting in empirical success rates above 75%. Among the selected α levels, we report only configurations achieving 100% calibration success rate. Final results represent mean values across all trials.

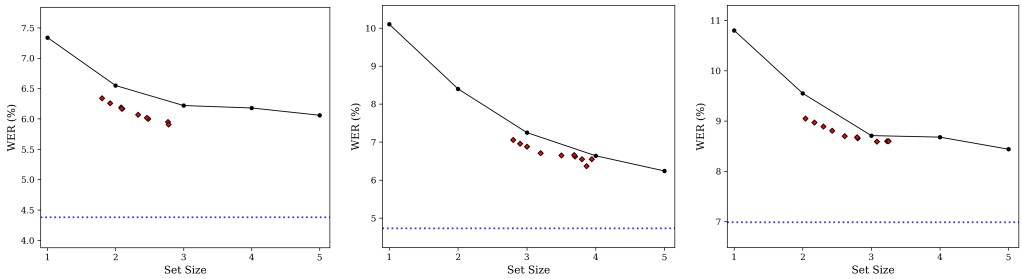
7 Results and Analysis

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

[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).

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 57.6% reduction in average set size while improving performance. CHiME-4 demonstrates computational savings of 23% with a modest performance trade-off of 2% relative increase in WER. CommonVoice exhibits 38% computational reduction with minimal performance impact. 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 theoretical guarantees hold empirically across all trials, with the high-probability bound ($P(R(\hat{\lambda}) \leq \alpha) \geq 1 - \delta = 0.75$, in our case) consistently satisfied in practice. This confirms the effectiveness of our approach in providing rigorous risk control - a property absent in prior methods and made possible through the LTT framework.

Our complementary corpus-level WERs are presented in Table C.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.

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.

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...			

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 risk control methods configurations including absolute WER targets, coverage-based objectives for samples below specified WER thresholds, 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. 9. 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. C.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.

Scalability to Larger Models and Zero-Shot Settings. To evaluate generalizability beyond our LLaMA-2-7B baseline, we conducted experiments with LLaMA-2-13B (fine-tuned) and GPT-3.5-turbo (zero-shot prompting). Results demonstrate that our framework maintains consistent performance-efficiency trade-offs across both larger model scales and deployment scenarios where fine-tuning is not feasible. The computational savings persist relative to the increased inference

costs, confirming that our adaptive selection mechanism provides value across different model architectures and prompting paradigms. This addresses practical deployment considerations where model size and training constraints vary significantly. Detailed results are provided in Appendix C.2.

Cross-Domain Extension. We extended our framework to speech translation tasks using the GenTranslate paradigm on multilingual datasets. with some methodological adaptation. Results show successful transfer with substantial computational savings while maintaining competitive translation quality. This cross-domain validation demonstrates broader applicability beyond ASR to generative error correction scenarios involving N-best hypothesis integration, directly addressing the broader impact potential in the GER community. Full methodology and results in Appendix C.3.

CRC Implementation. We also implemented the CRC framework as an alternative risk control method. While CRC lacks theoretical guarantees due to monotonicity violations in our application ($\sim 20\%$ of cases), it achieves similar empirical performance to our LTT approach. This demonstrates that both frameworks effectively exploit the same underlying adaptive selection patterns, with LTT providing the additional benefit of rigorous theoretical validation. The CRC implementation and comparative analysis are detailed in Appendix E.

Theoretical Bound Validation. To examine the gap between theoretical high-probability guarantees (75%) and observed empirical performance (95-98%), we conducted experiments with larger calibration sets. Results confirm that empirical performance approaches theoretical bounds as calibration data increases, validating that our method maintains the required statistical guarantees while the observed gap reflects conservative finite-sample bounds. Detailed analysis is provided in Appendix C.4.

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 LTT to dynamically determine optimal hypothesis set sizes, providing rigorous theoretical guarantees with high-probability bounds while demonstrating substantial computational savings and maintaining competitive performance across datasets with diverse acoustic conditions. Validation across larger language models, zero-shot settings, and cross-domain speech translation tasks confirms the framework's broad applicability and robustness across different deployment scenarios.

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 N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

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*, 2024a. URL https://arxiv.org/abs/2401.10446.

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

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

Qiujia Li, David Qiu, Yu Zhang, Bo Li, Yanzhang He, Philip C Woodland, Liangliang Cao, and Trevor Strohman. Confidence estimation for attention-based sequence-to-sequence models for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6388–6392. IEEE, 2021.

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.

Rao Ma, Mengjie Qian, Mark Gales, and Kate Knill. Asr error correction using large language models. *arXiv preprint arXiv:2409.09554*, 2024.

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.

Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. A consistent and differentiable lp canonical calibration error estimator. *Advances in Neural Information Processing Systems*, 35: 7933–7946, 2022.

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.

Nagarathna Ravi, Vipul Arora, et al. Teles: Temporal lexeme similarity score to estimate confidence in end-to-end asr. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 4399–4408, 2024.

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:

$$P(Y^{(m+1)} \notin C(X^{(m+1)})) \leq \alpha \tag{12}$$

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 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 proved insufficient, creating overly homogeneous score distributions that degraded selection quality. Consequently, we developed the two-level normalization strategy with parameter γ, enabling adaptive score transformation based on dataset difficulty while maintaining robustness across parameter choices.

B.1.2 Function Design

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

$$\phi_\gamma(\mathbf{c}) = (1 - \gamma) \cdot (-\mathbf{c}^{-1}) + \gamma \cdot \mathbf{c} \tag{13}$$

where $\mathbf{c}^{-1}$ denotes element-wise reciprocal operation, i.e., $(\mathbf{c}^{-1})_i = 1/c_i$ for each component i.

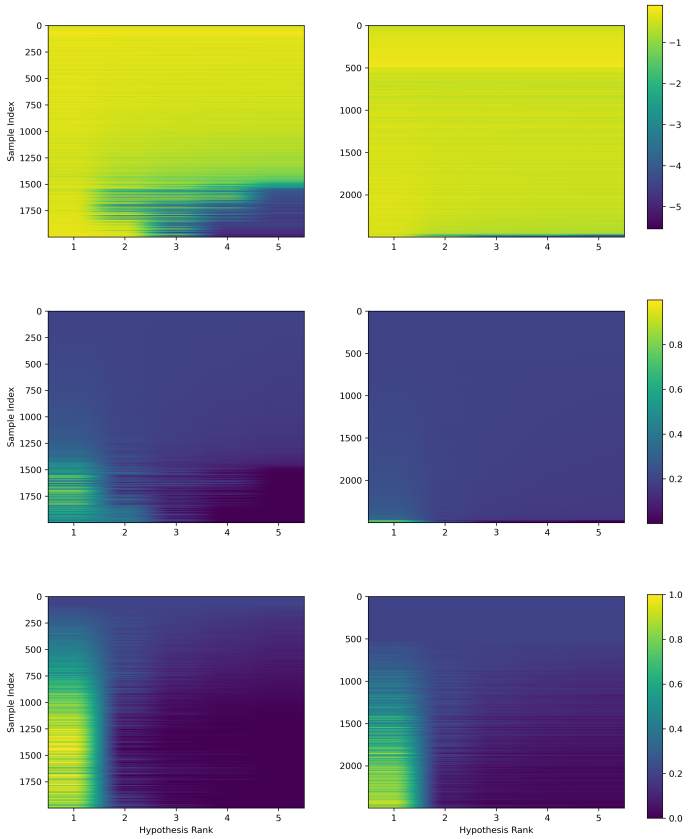


Figure B.1: ASR confidence score distributions for TedLium-3 (left) and CommonVoice (right) across processing stages: raw scores (top), softmax normalization (middle), and full transformation (bottom).

The identity transformation ($\gamma = 1$) preserves natural ASR score differences for high-SNR conditions, while the reciprocal transformation ($\gamma = 0$) amplifies small differences between compressed scores for challenging acoustic conditions. The parameter γ controls smooth transitions between these regimes.

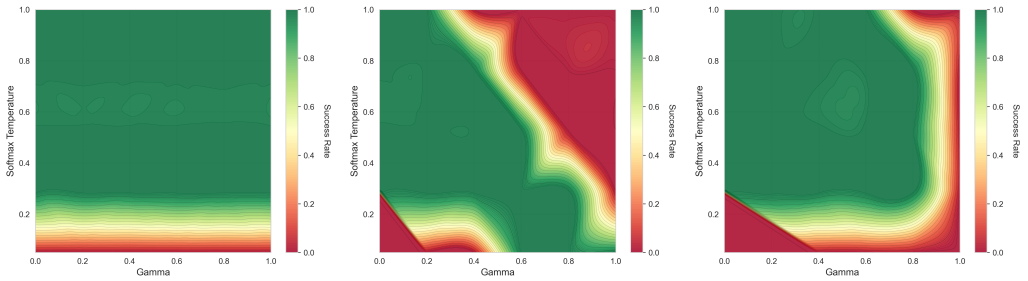
B.1.3 Parameter Selection Strategy

We base parameter selection on signal-to-noise ratio characteristics and empirical validation.

Figure B.1 illustrates score evolution across processing stages for different acoustic conditions. TedLium-3's high-SNR conditions produce naturally discriminative scores, requiring only sharpening via low temperature ($\tau = 0.05$) while preserving relationships ($\gamma = 1.0$). CommonVoice's challenging conditions with compressed distributions require reciprocal amplification ($\gamma = 0.0$) followed by moderate temperature ($\tau = 1.0$). CHiME-4 represents intermediate conditions requiring balanced transformation ($\gamma = 0.5$).

To validate this SNR-based rationale, we conducted systematic grid testing across the parameter space. Figure B.2 shows success regions where the method achieves valid risk control with better performance-compute trade-offs than fixed baselines.

The heat-maps confirm that successful regions align with our SNR-based parameter selection, with edge cases (high/low SNR) showing concentrated success areas and intermediate conditions requiring parameters within mid-value ranges.



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

Figure B.2: Parameter selection grid test results showing success regions across (γ, τ) parameter space. Success is defined as achieving valid risk control with superior performance-compute trade-offs compared to fixed baselines. Heat-maps confirm that successful regions align with SNR-based parameter selection strategy.

B.1.4 Robustness Analysis

Grid testing across uniformly sampled parameter combinations demonstrates method robustness. Success criteria require: (1) valid λ selection with controlled risk, and (2) superior performance-compute trade-offs compared to fixed-set baselines.

Results show overall success rate of 70% (61/88 combinations), with per-dataset rates: TedLium-3: 80% (24/30), CHiME-4: 62% (18/29), CommonVoice: 66% (19/29). This indicates parameter selection is an optimization step rather than a critical requirement, providing practitioners flexibility while maintaining performance guarantees.

B.1.5 Automated Parameter Selection

To automate parameter selection for deployment across unseen acoustic conditions, we investigated the relationship between score distribution characteristics and working point success. Our intuition is that meaningful working points should exhibit score vectors that respond differently across varying input difficulties, with entropy serving as a natural measure of this distributional behavior.

Analysis confirmed that score vector entropy serves as a reliable predictor of parameter effectiveness - successful working points exhibit distinct entropy patterns compared to failed configurations. Based on this observation, we developed a simple entropy-based rule: thresholding score vector entropy at 4.85 effectively discriminates between suitable and unsuitable parameter combinations.

We validated this rule on previously unseen parameter pairs, achieving prediction precision exceeding 80% and F1 scores above 75%. This enables practitioners to assess parameter suitability on new datasets without extensive manual calibration, providing a practical deployment pathway that transforms manual parameter tuning into principled selection with automated validation.

It should be noted that parameter sensitivity increases when targeting tighter performance bounds (lower α and δ values), as the method operates in narrower feasible regions where precise parameter selection becomes more critical.

B.1.6 Handling Hypothesis Repetitions

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

$$s_{i,r} = s_i \cdot \beta^r \tag{14}$$

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 TARGET RISK

B.2.1 RISK TARGET CALIBRATION

We establish target risk levels based on the performance range achievable through fixed hypothesis set selection. The feasible degradation range spans from fixed-1-hypothesis (worst case) to fixed-5-hypotheses (best case) performance. Since our adaptive method dynamically selects smaller sets with minimal degradation, targeting risk bounds within this empirically-derived range represents the natural operating regime. These empirically-derived ranges are $[1.7, 3.0]$ for TedLium-3, $[1.53, 5.76]$ for CHiME-4, and $[1.46, 3.81]$ for CommonVoice. We uniformly sample target risk values within the validated degradation range. The specific choice within this range determines the desired performance-compute trade-off: values closer to 1-hypothesis' risk prioritize computational efficiency, while values approaching 5-hypotheses' risk emphasize performance preservation. The two risk–control frameworks require different α–calibration approaches. For CRC, we directly use validation–derived values as the target expected risk levels, since CRC bounds $\mathbb{E}[R(\hat{\lambda})] \leq \alpha$. On the other hand, LTT provides high–probability bounds $\Pr\Big(R(\hat{\lambda}) \leq \alpha\Big) \geq 1 - \delta$, meaning that α represents approximately the $(1-\delta)_{th}$ percentile of the empirical risk distribution rather than its expectation. To achieve equivalent average performance, LTT therefore requires α values that are higher than the target expected empirical risk. Due to small calibration set sizes, the Hoeffding-Bentkus inequality produces conservative p-values, leading to more conservative set selections with higher empirical coverage. In practice, LTT typically yields valid selections for α values from the upper 75% of the feasible range, while CRC can achieve valid selections across the full range due to its expectation-based formulation. We report results only for α values that demonstrate successful risk control within the validated range, ensuring both theoretical validity and practical utility across different performance-efficiency preferences.

B.2.2 PERFORMANCE TRADE-OFFS

The target risk α directly controls the performance-compute trade-off by determining acceptable expected relative WER degradation from oracle performance. Lower α values (tighter bounds) require larger hypothesis sets to achieve the target risk level, due to average WER monotonicity on average, while higher α values (looser bounds) enable smaller sets with acceptable performance degradation.

Table B.1: Effect of risk tolerance α on performance-compute trade-offs (CommonVoice dataset).

Risk Tolerance (α)	Avg. Set Size	WER (%)
0.21 (tighter)	3.73	8.53
0.22	3.49	8.59
0.23	3.14	8.66
0.24 (looser)	2.99	8.73

Table B.1 demonstrates this relationship empirically. As α increases, average set sizes decrease from while WER increases as well. The different operating points in Figure 2 correspond to varying α selections within these validated ranges.

B.3 OTHER PARAMETERS SELECTION

We set additional framework parameters based on validation analysis to ensure robust performance across datasets. The repetition penalty $\beta = 1.25 - 1.5$ handles duplicate hypotheses by applying exponential decay to repeated entries, preventing overweighting of redundant information. The loss bound $B = 1.25$ accounts for rare cases exceeding 100% relative WER degradation, satisfying the bounded loss requirement for theoretical guarantees. These parameters were determined through empirical validation to maintain stability across varying hypothesis quality distributions while preserving the risk control framework's theoretical foundations.

B.4 HOEFFDING-BENTKUS P-VALUE COMPUTATION

For LTT implementation, we compute valid p-values using the Hoeffding-Bentkus inequality. Given empirical risk $\hat{R}_m(\lambda_j)$ on the calibration set, the p-value for hypothesis $H_j : R(\lambda_j) > \alpha$ is:

$$p_j^{HB} = \min\left\{\exp\{-nh_1(\hat{R}_m(\lambda_j) \wedge \alpha, \alpha)\}, eP\left(\text{Bin}(n,\alpha) \geq \lfloor n\hat{R}_m(\lambda_j)\rfloor\right)\right\} \tag{15}$$

where $h_1(a,b) = a\log(a/b) + (1-a)\log((1-a)/(1-b))$ and n is the calibration set size. This provides finite-sample valid p-values without distributional assumptions, enabling the sequential testing procedure in Algorithm 1.

C ABLATION STUDY

Table C.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 E.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	$2.21_{55.8\%}$	$5.05_{0.0\%}$	3.03
CHiME-4	6.37	$3.8_{24.0\%}$	$6.6_{+3.6\%}$	4.78
CommonVoice	7.8	$3.07_{38.6\%}$	$7.95_{+1.9\%}$	6.31

C.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.

C.2 SCALABILITY TO LARGER LANGUAGE MODELS

To evaluate the generalizability of our adaptive framework beyond the LLaMA-2-7B baseline, we conducted experiments with LLaMA-2-13B (fine-tuned) and GPT-3.5-turbo (zero-shot prompting).

Table C.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

These experiments assess whether the computational efficiency gains and adaptive selection benefits persist with more capable and resource-intensive models. All experiments use the LTT calibration procedure with dataset-specific α values, as detailed in Section B.1.6.

C.2.1 LLaMA-2-13B Results

We fine-tuned LLaMA-2-13B using identical training procedures, hyperparameters, and evaluation protocols as our main experiments. Table C.3 presents representative working points for each dataset.

Table C.3: LLaMA-2-13B adaptive selection results. Results shown for one representative working point per dataset; multiple valid working points exist across the parameter space.

Test Set	Baseline	GER	Our Method		α	δ	O_{llm}
			Set Size	WER			
TedLium-3	8.0	6.47	$2.40_{51.9\%}$	$6.46_{-0.01\%}$	2.1	0.25	4.69
CHiME-4	11.49	7.92	$3.88_{22.4\%}$	$8.16_{+3.06\%}$	1.25	0.25	7.26
CommonVoice	14.1	8.10	$3.50_{30.1\%}$	$8.28_{+2.27\%}$	2.0	0.25	6.50

The results demonstrate that our framework maintains its value proposition with larger models: TedLium-3 achieves 51.9% computational savings with identical performance, while CHiME-4 and CommonVoice show modest performance trade-offs (3.06% and 2.27% relative WER increase) for substantial efficiency gains (22.4% and 30.1% hypothesis reduction). These results confirm that the performance of the adaptive selection mechanism generalizes across model scales.

C.2.2 GPT-3.5-Turbo Zero-Shot Results

To assess applicability without fine-tuning, we evaluated GPT-3.5-turbo using zero-shot prompting based on templates from HyPoradise (Chen et al., 2023) and (Ma et al., 2024).

Table C.4: GPT-3.5-turbo zero-shot adaptive selection results. Results shown for one representative working point per dataset.

Test Set	Baseline	GER	Our Method		α	δ	O_{llm}
			Set Size	WER			
CommonVoice	14.1	11.73	$2.31_{42.3\%}$	$11.81_{+0.67\%}$	1.5	0.25	10.51
CHiME-4	11.49	9.77	$2.19_{56.1\%}$	$9.89_{+1.17\%}$	1.65	0.25	8.61

Our adaptive method achieves substantial computational savings (42-56% hypothesis reduction) with minimal performance impact (+0.67-1.17% relative WER increase) in zero-shot settings

Experimental notes: For CommonVoice, we evaluated hypothesis sets in the range $[1, 4]$ as preliminary tests showed performance degradation beyond 4 hypotheses; the GER baseline (11.73%) represents the best fixed size in this range. TedLium-3 was excluded from this evaluation because, consistent with findings in (Ma et al., 2024), all zero-shot working points performed worse than the baseline across all hypothesis sizes - likely due to the dataset's initially strong baseline performance, making the comparison uninformative.

These experiments validate three key findings: (1) the adaptive framework generalizes across model scales with consistent behavior, (2) zero-shot prompting achieves substantial efficiency gains (42-56% reduction) despite smaller absolute WER differences, and (3) computational savings persist relative to the increased inference cost of larger models.

C.3 Extension to Speech Translation Tasks

To demonstrate the broader applicability of our adaptive hypothesis selection framework, we evaluated its performance on speech translation tasks using the GenTranslate paradigm Hu et al. (2024b). This cross-domain validation examines whether our method maintains its computational efficiency benefits when applied to translation scenarios involving N-best hypothesis integration.

Dataset Selection and Monotonicity Validation. We selected three language pairs from the FLEURS X→En speech translation dataset (fr→en, cy→en, ar→en) based on a critical prerequisite: monotonic performance improvement with increasing hypothesis set sizes. Using the published GenTranslate checkpoint [2], we validated that BLEU scores followed the expected ordering $\text{BLEU}(N = 5) > \text{BLEU}(N = 4) > ... > \text{BLEU}(N = 1)$ on average across these language pairs. This monotonicity condition ensures a meaningful performance-compute trade-off exists, validating the potential utility of adaptive selection.

Methodological Adaptation for Speech Translation. A key challenge emerged in adapting our framework to speech translation tasks: BLEU, the standard evaluation metric, operates at the corpus level and provides limited meaningful information at the instance level required for our risk-based selection mechanism. To address this, we employed TER (Translation Edit Rate) for instance-level risk computation and adaptive set selection, while reporting final results using corpus-level BLEU for comparability with existing work. This approach leverages TER's established validity at the instance level while maintaining evaluation consistency. We validated that TER and BLEU preserve relative ordering (with inverse correlation) across our test sets.

Table C.5: Speech translation results with adaptive hypothesis selection on FLEURS X→En test sets compared to fixed N=5 baseline. Our method achieves substantial computational savings while maintaining competitive translation quality.

Task	GER		Adaptive Selection			α	Success Rate	$\text{TER}_{O_{llm}}$
	TER (%)	BLEU	Avg. Size	BLEU	TER (%)			
fr→en	4.62	37.50	$3.21_{35.8\%}$	$37.23_{-0.70\%}$	4.67	5	0.97	4.24
cy→en	5.1	33.89	$2.62_{47.7\%}$	$33.39_{-1.47\%}$	5.3	6.9	0.96	4.67
ar→en	5.21	34.47	$1.72_{65.5\%}$	$33.44_{-2.98\%}$	5.29	5.65	0.98	4.81

Implementation Details. Due to smaller test set sizes (400-1000 samples) compared to ASR experiments, we increased the error tolerance δ to 0.3 to prevent overly conservative bounds in the LTT framework. All other methodological components, including parameter selection and risk calibration procedures, remained consistent with our ASR implementation. The same LTT sequential testing approach was applied with TER-based loss functions for hypothesis set selection.

Summary. This cross-domain validation demonstrates that our adaptive hypothesis selection framework generalizes effectively beyond ASR to speech translation tasks, achieving substantial computational savings (36-66% hypothesis reduction) while maintaining competitive translation quality. These results confirm the broader applicability of our approach across generative error correction scenarios involving N-best hypothesis integration, addressing the computational efficiency challenges inherent in LLM-based post-processing systems.

[2] https://huggingface.co/PeacefulData/GenTranslate

C.4 THEORETICAL HIGH-PROBABILITY GUARANTEE ANALYSIS

Our LTT framework provides finite-sample guarantees $P(R(\hat{\lambda}) \leq \alpha) \geq 1 - \delta$ through the Hoeffding-Bentkus inequality. In practice, we observe empirical success rates (95-98%) that consistently exceed the theoretical bounds of $1 - \delta = 0.75$. To validate that this behavior reflects conservative finite-sample bounds, we conducted systematic experiments varying calibration set sizes and δ values.

We modified the splits to enable larger calibration sets for statistical analysis. For CommonVoice, we reran the complete pipeline with modified train-test split, while for TedLium-3 and CHiME-4 we re-split the existing test sets (CHiME-4 also expanded with its 'simu' set samples). All operating points maintained superior performance-efficiency trade-offs compared to fixed-size baselines.

Table C.6: Theoretical vs. empirical performance on modified-split CommonVoice.

Test Set Size	$\delta = 0.1$ (90%)	$\delta = 0.15$ (85%)	$\delta = 0.2$ (80%)
15k (7.5k for Calib)	90%	86%	83%
15k (4.5k for Calib)	95%	90%	89%

Table C.7: Cross-dataset validation of performance behavior.

Dataset	Size	$\delta = 0.3$ (70%)	$\delta = 0.2$ (80%)	$\delta = 0.15$ (85%)
TedLium-3	2.5k (2k for calib)	78%	87%	89%
CHiME-4 (expanded)	2.96k (1.48k for calib)	83%	91%	93%

Tables C.6 and C.7 demonstrate consistent patterns across all datasets: empirical performance approaches theoretical bounds as calibration data increases, while maintaining appropriate response to δ variations. This validates that our method maintains the required statistical guarantees and confirms that the observed behavior in our main experiments stems from conservative Hoeffding-Bentkus bounds with small calibration sets.

D LLM TRAINING CONFIGURATION DETAILS

D.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.

D.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:"*

D.3 Computational Requirements

Model training is conducted on a single NVIDIA RTX 6000 Ada GPU with 48GB 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.

E Conformal Risk Control Approach

E.1 Background

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. Conformal risk control (CRC) extends CP to control the expectation of any bounded, monotone loss function $\ell : 2^{\mathcal{Y}} \times \mathcal{Y} \to [0, B]$:

$$E[\ell(\Gamma_\lambda(X^{(m+1)}), Y^{(m+1)})] \leq \alpha \tag{16}$$

under the same notations represented in 5.

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:

$$\hat{\lambda} = \inf\left\{\lambda : \frac{m}{m+1}\hat{R}_m(\lambda) + \frac{B}{m+1} \leq \alpha\right\}, \tag{17}$$

where the empirical risk is computed as:

$$\hat{R}_m(\lambda) = \frac{1}{m}\sum_{i=1}^{m} \ell(\Gamma_\lambda(X^{(i)}), Y^{(i)}), \tag{18}$$

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 2 (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:*

$$\alpha - \frac{2B}{m+1} \leq E[\ell(\Gamma_{\hat{\lambda}}(X^{(m+1)}), Y^{(m+1)})] \leq \alpha. \tag{19}$$

Note that CRC reduces to standard CP when the loss function is the miscoverage indicator. CRC has been applied to areas 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).

E.2 CRC Implementation and Empirical Analysis

We initially explored conformal risk control (CRC) as our primary theoretical framework for adaptive hypothesis selection. Our CRC implementation follows Algorithm E.1, using the relative WER degradation loss function defined in Equation 9 with calibrated thresholds to control expected performance degradation.

Algorithm E.1 Adaptive Selection Procedure with CRC Calibration Framework

Require: Calibration set $\{(\mathcal{H}_N^{(i)}, y^{(i)})\}_{i=1}^m$ and a test sample $\mathcal{H}_N^{(m+1)}$

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

Table E.1 presents our CRC experimental results across the three datasets. The method achieves substantial computational savings while maintaining competitive performance. These results validate the practical effectiveness of adaptive selection based on this method.

However, our analysis revealed that strict monotonicity is violated in approximately 20% of samples, where smaller hypothesis sets occasionally outperform larger ones. While 95% of consecutive pairwise comparisons maintain monotonicity, indicating predominantly monotonic behavior, these violations present a theoretical challenge for CRC's formal guarantees. We evaluated the monotonizing procedure proposed by Angelopoulos et al. (2024b), which constructs $\tilde{\ell}_i(\lambda) = \sup_{\lambda' \geq \lambda} \ell_i(\lambda')$ to enforce monotonicity. However, empirical results showed degraded performance across datasets, as monotonizing eliminates precisely the beneficial cases where smaller sets genuinely outperform larger ones—the phenomenon enabling our computational savings.

These findings indicate that monotonicity violations in ASR often signal exploitable efficiency opportunities rather than problematic cases. While our CRC implementation demonstrates strong empirical performance and effective risk control in practice, the theoretical violations prevent us from claiming formal statistical guarantees. Therefore, we present this CRC approach as an additional empirical demonstration that complements our theoretically rigorous LTT framework, which naturally handles non-monotone losses while providing formal $P(R(\hat{\lambda}) \leq \alpha) \geq 1 - \delta$ guarantees.

Table E.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