SMALL LANGUAGE MODELS FOR LARGE SPEECH MODELS: LATENT ACOUSTIC ARBITRATION FOR ASR CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can improve automatic speech recognition (ASR) transcripts, but unrestricted rewriting may trade acoustic consistency for linguistic fluency. We ask whether post-ASR correction requires a text model comparable in scale to the speech model it assists. We propose Latent Acoustic Arbitration (LAA), a parameter-efficient framework in which a compact frozen text model provides linguistic preferences within a bounded speech-generated candidate space, while retained speech-side evidence determines which local changes are accepted. With Granite Speech 8B and a frozen Qwen3-0.6B text refiner, LAA improves the constrained language-selection baseline on all eight evaluated English ASR benchmarks. In a separate scaling analysis spanning refiners from 135M to 32B parameters—a 237× range—WER varies by only 0.14 percentage points. These results provide evidence that effective post-ASR arbitration does not require speech and text models of comparable scale, while controlled 2B and cross-backbone analyses delimit where the observed gains transfer.

1 INTRODUCTION

Post-ASR error correction is attractive because it can add a strong language prior without retraining the acoustic model. This idea extends a long line of language-model rescoring and Bayes-risk decoding for speech recognition (Jelinek, 2022; Mangu et al., 2000; Goel & Byrne, 2000). The same flexibility, however, is dangerous: a text-only language model can replace an acoustically plausible but unusual phrase with a more common, incorrect one. Recent work has therefore explored n-best hypotheses, constrained decoding, confidence filtering, and multi-stage correction rather than unrestricted text generation (Ma et al., 2025; Pu et al., 2023; Naderi et al., 2024).

This paper asks a more basic question: *does post-ASR correction require a language model comparable to, or larger than, the speech model it assists?* We hypothesize that it does not. If the text model only expresses preferences among hypotheses already proposed by the speech model, and final changes remain constrained by speech-side evidence, then a compact frozen refiner may be sufficient. Based on this idea, Latent Acoustic Arbitration (LAA) separates linguistic preference from acoustic verification: a frozen text model compares speech-generated candidates, while lightweight routers use retained, time-aligned speech evidence to decide which candidate-supported local changes to accept through selective arbitration rather than unrestricted text generation.

We evaluate LAA across Granite Speech, Whisper, and SeamlessM4T backbones and frozen text refiners ranging from 135M to 32B parameters, centering the main claim on a deliberately asymmetric configuration: Granite Speech 8B with Qwen3-0.6B. In this setting, LAA improves the constrained language-selection baseline on all eight benchmarks. A separate refiner-scaling analysis shows only a 0.14-point WER span across a 237× parameter range, indicating that a larger text model is not necessary for the observed arbitration benefit. Controlled smaller-model and cross-backbone analyses further delineate the mechanism's applicability and transfer boundaries.

Our contributions are:

- We introduce a constrained cross-modal correction framework that decouples linguistic preference from acoustic verification and applies local, selective post-ASR corrections

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

within a speech-generated candidate space, limiting the risk of acoustically unsupported free-form rewriting;

- We formulate an asymmetric cross-scale collaboration mechanism in which language proposes preferences while speech retains candidate generation and final acoustic authority, allowing a compact frozen text model to serve as a linguistic assistant without full-text generation or scale parity;
- We establish a systematic analysis of post-ASR arbitration across model scale, speech backbone, routing policy, candidate set, and acoustic alignment, characterizing the conditions and boundaries of this approach beyond performance in a single configuration.

2 Related Work

Unrestricted generative correction. Prompted language models and trained encoder-decoder systems can map ASR hypotheses to fluent corrected transcripts (Yang et al., 2023; Chen et al., 2023; La Quatra et al., 2024). The GenSEC challenge organizes this direction into a reproducible family of post-ASR language-modeling tasks and baselines (Yang et al., 2024). These methods demonstrate the value of linguistic context, but free-form generation can also introduce text unsupported by the speech model, especially when a plausible transcript is already available.

Constrained hypothesis selection. Classical ASR systems combine acoustic scores with statistical or recurrent language models to rescore decoder hypotheses (Chelba & Jelinek, 2000; Stolcke, 2002; Mikolov et al., 2010; Sundermeyer et al., 2012). Confusion-network consensus, minimum Bayes-risk decoding, minimum-phone-error objectives, and sampling-based expected-WER optimization make the decision boundary explicit rather than treating the top hypothesis as uniquely correct (Mangu et al., 2000; Goel & Byrne, 2000; Kumar & Byrne, 2004; Povey & Woodland, 2002; Shannon, 2017). Ma et al. (2025) revisit this bounded setting with large language models and constrained n-best or lattice decoding. LAA keeps this ownership: the text model ranks alternatives, but every accepted token must come from a speech-generated hypothesis.

Selective correction and acoustic verification. Unrestricted correction is particularly risky when the original transcript is already correct. Pu et al. (2023) use uncertainty estimation to identify less reliable utterances before multi-stage LLM correction, while Naderi et al. (2024) study sentence- and word-level confidence filtering and prompting constraints to reduce harmful edits. More broadly, calibration and selective-prediction research distinguishes confidence quality from raw task accuracy and exposes failures under distribution shift (Guo et al., 2017; El-Yaniv et al., 2010; Geifman & El-Yaniv, 2017; Hendrycks & Gimpel, 2016). Ensemble methods, trust scores, and verified-calibration analyses further show that uncertainty estimates require independent validation (Lakshminarayanan et al., 2017; Jiang et al., 2018; Kumar et al., 2019; Ovadia et al., 2019; Minderer et al., 2021). LAA complements confidence-based selection by using retained speech representations to verify candidate-supported local changes rather than relying on confidence alone.

Parameter-efficient cross-scale correction. Prior work establishes the value of generative correction, bounded hypothesis selection, and selective intervention, but does not determine whether the assisting text model must approach the speech model in scale. LAA targets this gap in three ways: the frozen text refiner remains inside the speech-generated candidate space, linguistic preference is separated from selective acoustic verification, and the evaluation explicitly tests whether a compact refiner can assist a substantially larger speech backbone. We avoid cross-paper WER comparisons because backbones, datasets, candidates, and training conditions differ; alternatives are compared only under a shared speech backbone within each controlled experiment.

3 Latent Acoustic Arbitration

3.1 Problem Formulation

Let x denote an audio utterance, $d(x)$ the speech model's draft transcript, and

$$H_N(x) = \{h_1, \ldots, h_N\}$$

the ordered set of N hypotheses produced from the same waveform. The evaluated setting uses $N = 4$. A frozen text refiner assigns each candidate a score $S_{\text{text}}(h \mid d(x))$, yielding the constrained baseline

$$b(x) = \arg \max_{h \in H_N(x)} S_{\text{text}}(h \mid d(x)).$$

This baseline can change the draft only by selecting a speech-generated hypothesis. By contrast, free correction maps $d(x)$ to an arbitrary text sequence, allowing linguistic fluency to override the evidence represented by the finite candidate set generated from the utterance.

LAA begins from $b(x)$ and decomposes its differences from each alternative candidate into span proposals. It retains time-aligned encoder frames $z_{1:T}(x)$, where T varies with the utterance duration, and uses a phoneme CTC head to score the acoustic compatibility of the candidates. A lightweight two-stage router combines language, speech, span, and phoneme-compatibility features to decide which proposals should be verified and applied. All large speech and text backbones remain frozen; only the phoneme CTC head and router are trained. The acoustic representation is therefore a selective verification channel rather than a second end-to-end speech-language model.

3.2 The LAA Framework

Figure 1 summarizes the proposal, routing, acoustic-check, and merge stages. For an alternative candidate $h \in H_N(x) \setminus \{b(x)\}$, let $\mathcal{D}(b, h)$ denote the set of token spans on which h differs from the baseline $b(x)$. A *proposal* $p = (h, s)$ consists of a candidate hypothesis h and a differing span $s \in \mathcal{D}(b, h)$, and represents copying span s from h into $b(x)$. The phoneme CTC head supplies a matched acoustic gap

$$\Delta_{\text{ac}}(p, x) = S_{\text{ph}}(h, z_{1:T}(x)) - S_{\text{ph}}(b(x), z_{1:T}(x)),$$

where $S_{\text{ph}}(\cdot, z_{1:T}(x))$ measures phoneme compatibility between a transcript and the time-aligned encoder frames. This gap quantifies whether the alternative candidate has stronger acoustic support than the baseline over the complete candidate transcript and retained encoder sequence.

LAA applies a two-stage routing policy to each proposal. Stage 1 estimates whether proposal p overlaps a likely transcription error. For proposals that pass Stage 1, Stage 2 estimates whether the replacement is likely to be beneficial using Stage 1 features together with the phoneme gap and span-level verification signals. With fixed thresholds τ_1 and τ_2, a proposal is eligible for application only if

$$R_1(p) \geq \tau_1, \qquad B_2(p, \Delta_{\text{ac}}) \geq \tau_2, \qquad \Delta_{\text{ac}}(p, x) > 0.$$

Overlapping eligible proposals are resolved deterministically by a fixed priority rule, and the remaining non-overlapping spans are inserted into $b(x)$. The resulting transcript may combine non-overlapping spans from multiple candidates, but every inserted token originates from a speech-generated hypothesis. Arbitration therefore remains bounded by hypotheses proposed from the same waveform and verified against its retained time-aligned encoder evidence.

The *offset control* uses exactly the same candidate text, Stage 1 calls, and Stage 2 routing decisions as the matched condition, but replaces the final positive acoustic-gap check with phoneme scores from a different utterance. Because Stage 2 is held fixed and already uses matched phoneme-gap features, this comparison diagnoses sensitivity of the final acceptance gate to sample alignment; it does not isolate every causal contribution of acoustic evidence in the full router. We use "Baseline" in result tables for language-only constrained selection within $H_N(x)$, "free correction" for unrestricted generation, and "LAA" for the matched selective-arbitration route with sample-aligned acoustic evidence throughout the reported controlled component analyses.

3.3 Training and Inference

We train the phoneme CTC head on frozen frame-level Granite encoder outputs from 10,000 LibriSpeech train-other utterances. Standard matched-utterance CTC loss is combined with contrastive margins that favor correct transcripts over offset utterances or controlled phoneme-level perturbations. This training enables the acoustic-gap signal Δ_{ac} to distinguish sample-aligned evidence from generic score biases that do not reflect utterance alignment in the training data.

The span router is trained separately on a 400-utterance source-calibration set with disjoint training and development splits. Stage 1 supervision labels proposals that overlap a reference transcription

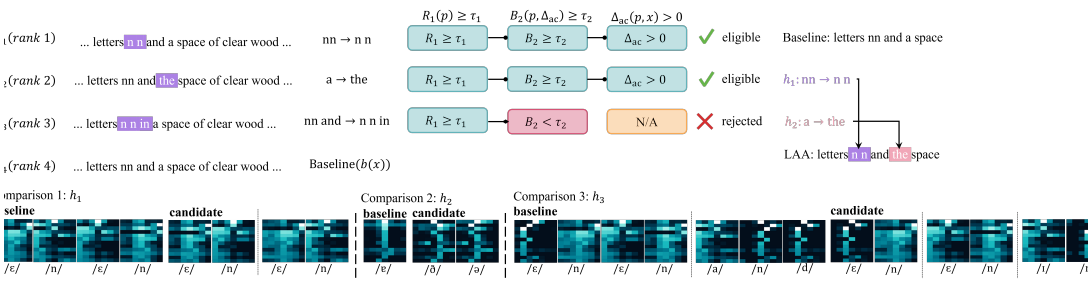

Figure 1: LAA overview. The frozen text refiner selects a baseline from four speech-generated hypotheses. Differing spans become proposals; Stage 1 screens likely errors, Stage 2 estimates replacement benefit, and the final positive acoustic-gap check retains proposals with matched speech support. Accepted non-overlapping spans are merged into the baseline. The local CTC tiles are explanatory excerpts only; the gate uses the complete-transcript phoneme score.

error; Stage 2 supervision labels risk-positive proposals that also have a positive matched acoustic gap and reduce the local reference error. The router sees only source-domain references during training; target-benchmark references are never exposed to the routing policy at inference.

Thresholds are calibrated on the router development split and then fixed for all target benchmarks. The Stage 1 cutoff is selected according to a predefined recall operating point, while the Stage 2 threshold is chosen to maximize net local gain on the source-development split, with harmful edits and routing load used as tie-breakers. After calibration, the phoneme heads, router weights, and routing thresholds remain fixed during evaluation, with no per-benchmark tuning.

At inference, the speech model generates hypotheses via deterministic beam search with four beams, producing the draft $d(x)$ and candidate set $H_4(x)$. The frozen text refiner scores each candidate and selects the baseline $b(x)$. For the LAA route, the system constructs span proposals from $\mathcal{D}(b, h)$ for each alternative h, applies the two frozen routing stages, checks the matched acoustic gap $\Delta_{\text{ac}}(p, x) > 0$, resolves overlaps, and inserts accepted spans. The constrained N-best baseline stops after selecting $b(x)$; the offset condition reuses the matched routing decisions but performs the final acoustic check with phoneme scores from a paired different utterance, probing sample alignment only at that final gate while leaving candidate selection and upstream routing unchanged.

4 Experimental Protocol

4.1 Models and Baselines

Our main evaluation uses Granite Speech 8B (Saon et al., 2025) with a frozen Qwen3-0.6B text refiner (Team et al., 2025), creating a deliberately asymmetric setting in which a compact language model assists a substantially larger speech model. Granite Speech 2B is used as a controlled diagnostic setting for detailed component, routing, and candidate-width analyses. We additionally evaluate Whisper small and large-v3 (Radford et al., 2023) and SeamlessM4T medium and v2 (Duquenne et al., 2023) to examine sensitivity across speech architectures. All speech backbones and text refiners remain frozen. Per deployment, the trainable phoneme head contains 208,112 parameters and the two logistic routers contain 23 parameters, for 208,135 trainable parameters in total. Additional refiner analyses compare six frozen text models, and a separate scaling analysis spans 135M to 32B parameters. These counts exclude all frozen backbone and refiner parameters.

We compare four core conditions: unrestricted correction, constrained N-best selection (Baseline), an offset-audio diagnostic, and LAA. The Baseline isolates language-only candidate selection, whereas the offset condition tests sensitivity to sample-aligned acoustic evidence at the final acceptance stage. Under Granite Speech 8B, we further compare LAA with MBR, STAR, Prompt Projector, MossFormerGAN, and DeepFilterNet3 within the same evaluation.

Table 1: Primary same-backbone comparison with Granite Speech 8B across eight benchmarks. Values are WER (%), reported as mean ± sample SD over three runs. Overall WER is the equally weighted mean across benchmarks. Lower is better; bold marks the lowest mean in each column.

Method	AMI	CHiME-6	FLEURS	LS-clean
LAA	**14.85**±0.03	**68.80**±0.02	8.03±0.00	1.25±0.01
MBR	16.23±0.22	73.08±0.36	8.31±0.11	1.43±0.04
STAR	15.41±0.04	83.47±1.44	**7.90**±0.02	1.15±0.01
Prompt Projector	20.68±2.49	290.40±77.84	8.65±0.14	1.73±0.05
MossFormerGAN	15.49±0.00	112.67±2.87	7.91±0.00	**1.13**±0.01
DeepFilterNet3	34.91±0.06	170.80±4.50	74.01±0.00	1.43±0.01

Method	LS-other	SPGI	TED-3	VoxPopuli	Overall WER ↓
LAA	3.27±0.00	5.31±0.03	11.49±0.03	6.07±0.01	**14.885**±0.006
MBR	3.84±0.08	6.06±0.14	11.33±0.05	6.86±0.12	15.893±0.070
STAR	**3.16**±0.00	5.35±0.03	10.99±0.02	**5.89**±0.02	16.664±0.180
Prompt Projector	4.09±0.07	**4.43**±0.14	**10.46**±0.16	6.88±0.20	43.413±9.900
MossFormerGAN	3.19±0.01	5.53±0.00	10.88±0.00	6.04±0.00	20.355±0.360
DeepFilterNet3	4.04±0.00	15.41±0.38	10.93±0.02	6.20±0.00	39.716±0.620

4.2 Datasets and Evaluation

We evaluate eight English ASR benchmarks spanning meetings, noisy conversational speech, read speech, financial calls, talks, and parliamentary speech: AMI (Carletta et al., 2005), CHiME-6 (Watanabe et al., 2020), FLEURS English (Conneau et al., 2023), LibriSpeech test-clean and test-other (Panayotov et al., 2015), SPGISpeech (O'Neill et al., 2021), TED-LIUM 3 (Hernandez et al., 2018), and the English evaluation split of VoxPopuli (Wang et al., 2021).

We report WER independently for each benchmark and do not pool utterances across corpora. For within-experiment summaries, Overall WER is the equally weighted mean of the eight benchmark WERs, and Overall Gain is Baseline Overall WER minus LAA Overall WER. Each Overall value is first computed within a phoneme-head seed and then summarized across the three seeds with sample SD; deterministic Baselines are broadcast across seeds before paired Gain is computed. In the Granite component analysis, FLEURS uses 647 scoreable examples and each other benchmark uses 1,000. Because the backbone, refiner, scaling, and candidate-width analyses use separately fixed input sets, absolute WER values are interpreted only within the corresponding experiment.

The LibriSpeech test-other slice used in the scaling analysis was repeatedly inspected during method development and is therefore treated as development evidence rather than an untouched final test. Training and evaluation examples remain disjoint within each experiment.

5 Results

5.1 Scale Mismatch and Model Sensitivity

Table 1 gives the primary same-backbone comparison under Granite Speech 8B. LAA obtains the lowest Overall WER, while the strongest alternative varies across individual benchmarks; its advantage is aggregate rather than uniform across domains. The 0.138-point gain over Baseline measures the complete LAA route relative to constrained selection using the same Qwen3-0.6B refiner and should not be interpreted as a standalone text-model effect in isolation.

Appendix Table 2 tests the central scale-mismatch claim. With Granite Speech 8B and Qwen3-0.6B, LAA has lower WER than Baseline on all eight benchmarks, showing that the text refiner need not approach the speech model's scale to provide useful linguistic preferences. The same table also defines the transfer boundary: regressions for several Whisper and SeamlessM4T settings show that backbone compatibility remains part of the interface and that scale asymmetry alone does not guarantee improvement across the speech architectures evaluated in this study.

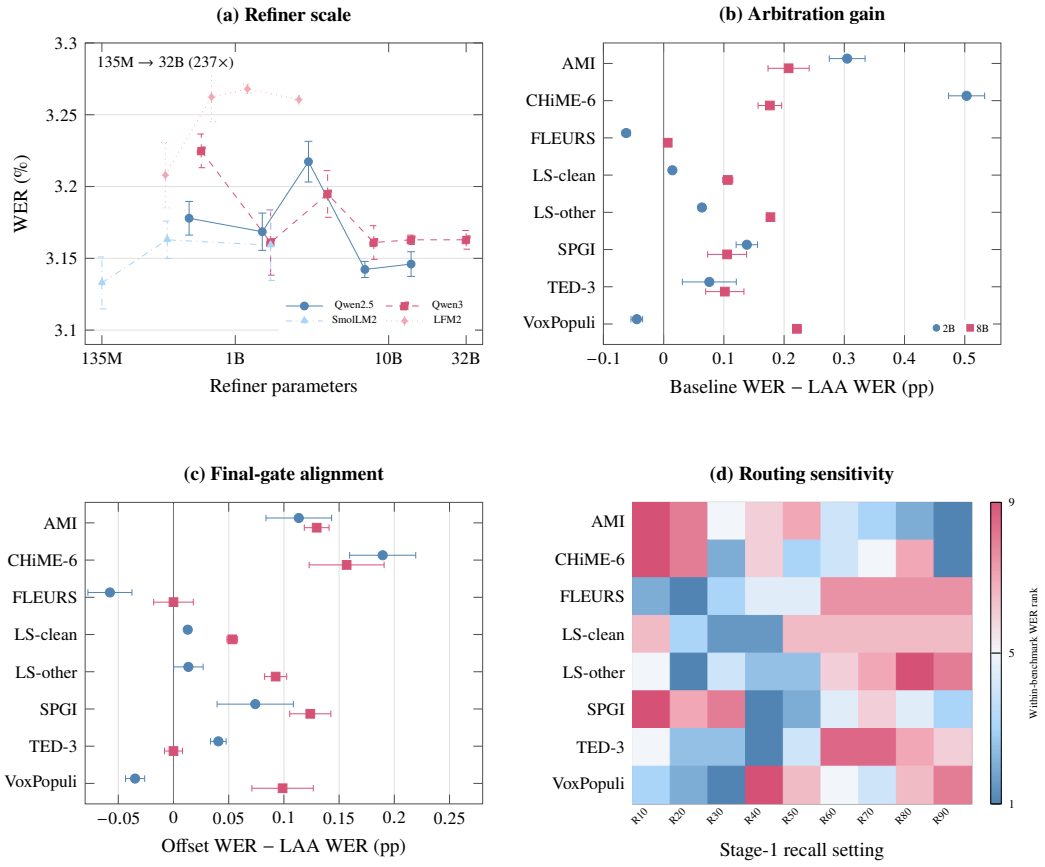

Figure 2: LAA diagnostics. (a) Refiner scaling on a separate Granite Speech 2B LS-other subset; absolute WER is not comparable across panels. (b) Paired Baseline−LAA gain. (c) Paired Offset−LAA difference at the fixed final acoustic gate, not the full acoustic module's causal effect. Circles and squares denote 2B and 8B; error bars are sample SD. (d) Within-benchmark WER rank of R10–R90 for Granite Speech 2B; rank 1 is the lowest WER in that benchmark.

Appendix Table 3 evaluates six frozen refiners with Granite Speech 2B and 8B at R40. Excluding the Qwen3-0.6B row in the 2B block, the remaining 88 Baseline–LAA pairs contain 85 improvements, one tie, and two regressions; both regressions use Granite Speech 8B with SmolLM2-360M. This broad but incomplete stability shows that *refiner scale does not determine correction quality*: compact Qwen3-0.6B remains competitive with larger refiners, supporting parameter efficiency without constituting a general model ranking across the evaluated refiner families.

5.2 Component Decomposition

Appendix Table 5 presents the four-condition decomposition for both Granite backbones; Figure 2(b,c) shows the corresponding paired differences. The headline Granite Speech 8B configuration improves Baseline on all eight benchmarks, reducing Overall WER from $15.023_{\pm 0.000}$ to $14.885_{\pm 0.006}$ (paired gain $0.138_{\pm 0.006}$ points), with the largest benchmark reduction being 0.22 points on VoxPopuli. Granite Speech 2B serves as the finer-grained analysis setting: LAA improves Baseline on six benchmarks and reduces Overall WER from $16.641_{\pm 0.000}$ to $16.517_{\pm 0.006}$ (gain $0.124_{\pm 0.006}$), while regressing on FLEURS and VoxPopuli. For both backbones, Baseline outperforms unrestricted correction on all eight benchmarks. The matched 2B route beats its offset diagnostic on six benchmarks and loses on two; the 8B route is lower on six and tied in mean on FLEURS and TED-LIUM 3. Baseline comparisons measure the incremental change from the complete arbitration route, whereas matched–offset differences characterize only the final acoustic gate under fixed upstream routing in the controlled analyses reported here.

5.3 Sensitivity and Parameter Diagnostics

Figure 2 combines four complementary analyses. In (a), 18 frozen refiners span 135M–32B parameters on an independent LibriSpeech test-other input, yet three-head R40 LAA WER spans only 0.14 points and the 135M SmolLM2 refiner has the lowest mean. Appendix Table 4 gives the complete model-level values. This near scale invariance supports parameter efficiency on that input, not a general model ranking. Panel (b) connects that scale result to the main comparison: LAA improves Baseline on all eight Granite Speech 8B benchmarks and six of eight 2B benchmarks. Panel (c) narrows the acoustic claim: matched scoring beats the offset final-gate control on six benchmarks for each backbone (with two 8B mean ties), but upstream routing is fixed and already contains matched acoustic features. Panel (d) shows that R40 is not uniformly WER-optimal: the best recall setting varies by benchmark, and the R10–R90 range is 0.003–0.315 WER points under the shared fixed-policy evaluation.

Appendix Table 6 retains the absolute R10–R90 values and the two controls omitted from Figure 2(d). R40 is the unified routing operating point fixed for the main experiments; the sweep is a sensitivity analysis rather than a post-hoc selection procedure. R30 has the lowest observed Overall WER ($16.506_{\pm 0.010}$), while R40 obtains $16.517_{\pm 0.006}$; the 0.011-point mean difference is small, and the best setting varies across benchmarks. All-span verification is worse than R40 on all eight benchmarks. Removing Stage 2 from the R40 call set worsens six benchmarks (by up to 0.30 points on VoxPopuli) and improves two (by up to 0.08 points on CHiME-6). These results characterize sensitivity around the fixed R40 point without claiming that R40 is generally optimal.

Appendix Table 7 and Figure 3 analyze candidate-set width using $N \in \{2, 4, 8, 16\}$. The lowest final Overall WER is obtained at $N = 4$ ($14.654_{\pm 0.014}$), while $N = 8$ yields a nearly identical $14.661_{\pm 0.010}$ and only a marginally larger arbitration gain ($0.160_{\pm 0.010}$ versus $0.158_{\pm 0.014}$). The differences are small relative to the observed seed variation and do not support a robust ranking between the two settings. Since $N = 4$ uses half the candidate budget of $N = 8$, we retain $N = 4$ as the default candidate width. We make no claim about latency or compute savings, which were not measured separately for each N, and leave end-to-end cost comparison to future work.

6 Conclusion

This paper asks whether post-ASR correction requires a text model comparable in scale to the speech model it assists. Our results suggest that it does not. With Granite Speech 8B, LAA using a frozen Qwen3-0.6B refiner improves the constrained language-selection Baseline on all eight evaluated benchmarks, while a separate 135M–32B refiner sweep shows only limited sensitivity to text-model scale. These results indicate that useful linguistic guidance can come from a substantially smaller text model when candidate generation remains speech-owned and local changes are selectively verified by acoustic evidence under the evaluated benchmark conditions.

The observed gains are not universal. Granite Speech 2B improves on six of eight benchmarks, and several Whisper and SeamlessM4T configurations do not improve, indicating that backbone compatibility remains important. Our current evidence also does not establish compute-optimal routing or candidate width. Future work should test the interface on additional speech backbones, untouched evaluation sets, and native end-to-end cost measurements.

AI USE STATEMENT

This research was conducted under the direction of the authors. Generative AI tools were used to assist with literature discovery; discussion and refinement of research hypotheses, methodology, and experimental design; development of the theoretical framing; mathematical derivations and proof drafts; method implementation and software-code editing and inspection; organization, analysis, and interpretation of results; preparation of tables and figures; and manuscript drafting, translation, and language editing. The authors defined the research question and scope, selected the experimental protocols and analyses, made all final methodological and interpretive decisions, and determined which AI-generated suggestions to accept, revise, or reject. All AI-assisted material included in the submission was subject to substantive human review and revision. This review covered theoretical assumptions and proof logic, experimental comparisons and statistical interpretation, and consistency among the text, equations, tables, and figures. AI outputs were treated as provisional assistance rather than as evidence. The authors take responsibility for the final manuscript, mathematical statements, experimental claims, citations, code, and research artifacts.

REPRODUCIBILITY STATEMENT

We fix model revisions, the four-candidate ordering used by the main comparisons, LAA thresholds, and evaluation inputs. The candidate-width analysis evaluates $N = 1, 2, 4, 8, 16$, with direct ASR as the $N = 1$ control. The component, policy, and candidate-width analyses use the same three phoneme-head seeds, although the candidate-width study has its own fixed input set. The accompanying bundle provides machine-readable results, seed-level records, source-file hashes, and scripts that regenerate the reported Overall metrics and figures.

REFERENCES

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pp. 28–39. Springer, 2005.

Ciprian Chelba and Frederick Jelinek. Structured language modeling. *Computer Speech & Language*, 14(4):283–332, 2000.

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.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805. IEEE, 2023.

Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, et al. Seamlessm4t—massively multilingual and multimodal machine translation, 2023.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Vaibhava Goel and William J Byrne. Minimum bayes-risk automatic speech recognition. *Computer Speech & Language*, 14(2):115–135, 2000.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *International conference on speech and computer*, pp. 198–208. Springer, 2018.

Frederick Jelinek. *Statistical methods for speech recognition*. MIT press, 2022.

Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. *Advances in neural information processing systems*, 31, 2018.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in neural information processing systems*, 32, 2019.

Shankar Kumar and Bill Byrne. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the human language technology conference of the North American chapter of the association for computational linguistics: HLT-NAACL 2004*, pp. 169–176, 2004.

Moreno La Quatra, Valerio Mario Salerno, Yu Tsao, and Sabato Marco Siniscalchi. Flanec: Exploring flan-t5 for post-asr error correction. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 608–615. IEEE, 2024.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Rao Ma, Mengjie Qian, Mark Gales, and Kate Knill. Asr error correction using large language models. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1389–1401, 2025.

Lidia Mangu, Eric Brill, and Andreas Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14 (4):373–400, 2000.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pp. 1045–1048. Makuhari, 2010.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, 34:15682–15694, 2021.

Maryam Naderi, Enno Hermann, Alexandre Nanchen, Sevada Hovsepyan, and Mathew Magimai Doss. Towards interfacing large language models with asr systems using confidence measures and prompting. *arXiv preprint arXiv:2407.21414*, 2024.

Patrick K O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D Shulman, et al. Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. *arXiv preprint arXiv:2104.02014*, 2021.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Daniel Povey and Philip C Woodland. Minimum phone error and i-smoothing for improved discriminative training. In *2002 IEEE international conference on acoustics, speech, and signal processing*, volume 1, pp. I–105. IEEE, 2002.

Jie Pu, Thai-Son Nguyen, and Sebastian Stüker. Multi-stage large language model correction for speech recognition. *arXiv preprint arXiv:2310.11532*, 2023.

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.

George Saon, Avihu Dekel, Alexander Brooks, Tohru Nagano, Abraham Daniels, Aharon Satt, Ashish Mittal, Brian Kingsbury, David Haws, Edmilson Morais, Gakuto Kurata, Hagai Aronowitz, Ibrahim Ibrahim, Jeff Kuo, Kate Soule, Luis Lastras, Masayuki Suzuki, Ron Hoory, Samuel Thomas, Sashi Novitasari, Takashi Fukuda, Vishal Sunder, Xiaodong Cui, and Zvi Kons. Granite-speech: open-source speech-aware llms with strong english asr capabilities, 2025. URL https://arxiv.org/abs/2505.08699.

Matt Shannon. Optimizing expected word error rate via sampling for speech recognition. *arXiv preprint arXiv:1706.02776*, 2017.

Andreas Stolcke. Srilm-an extensible language modeling toolkit. In *Proc. ICSLP 2002*, pp. 901–904, 2002.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Proc. Interspeech 2012*, pp. 194–197, 2012.

Qwen Team et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 6(7):13, 2025.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 993–1003, 2021.

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, et al. Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *arXiv preprint arXiv:2004.09249*, 2020.

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

Chao-Han Huck Yang, Taejin Park, Yuan Gong, Yuanchao Li, Zhehuai Chen, Yen-Ting Lin, Chen Chen, Yuchen Hu, Kunal Dhawan, Piotr Żelasko, 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. IEEE, 2024.

A SUPPLEMENTARY RESULTS

The following tables provide the full values underlying the refiner-scaling, component, routing-policy, and candidate-width analyses summarized in the main text.

Table 2: Speech-backbone sensitivity with Qwen3-0.6B at R40 and $N = 4$. Values are WER (%), reported independently for each benchmark. Lower is better; bold compares Baseline with LAA.

Speech backbone	AMI		CHiME-6	
	Baseline	LAA	Baseline	LAA
Granite Speech 8B	15.06±0.00	**14.85**±0.03	68.97±0.00	**68.80**±0.02
Granite Speech 2B	16.88±0.00	**16.57**±0.03	81.14±0.00	**80.64**±0.03
SeamlessM4T-med	57.11±0.29	**56.55**±0.22	103.53±0.30	**102.02**±0.17
SeamlessM4T-v2	93.96±0.09	**93.64**±0.13	153.25±0.88	**152.99**±0.78
Whisper large-v3	**27.27**±0.00	**27.27**±0.00	**74.60**±0.00	**74.60**±0.00
Whisper small	**85.90**±0.00	**85.90**±0.00	**157.11**±0.00	**157.11**±0.00

Speech backbone	FLEURS		LS-clean	
	Baseline	LAA	Baseline	LAA
Granite Speech 8B	8.04±0.00	**8.03**±0.00	1.36±0.00	**1.25**±0.01
Granite Speech 2B	**6.65**±0.00	6.71±0.01	1.30±0.00	**1.29**±0.00
SeamlessM4T-med	8.34±0.00	**8.31**±0.00	3.88±0.01	**3.86**±0.00
SeamlessM4T-v2	7.82±0.00	**7.76**±0.00	**2.67**±0.01	2.69±0.01
Whisper large-v3	**4.58**±0.00	**4.58**±0.00	**3.88**±0.00	**3.88**±0.00
Whisper small	**6.44**±0.00	**6.44**±0.00	**3.73**±0.00	**3.73**±0.00

Speech backbone	LS-other		SPGI	
	Baseline	LAA	Baseline	LAA
Granite Speech 8B	3.45±0.00	**3.27**±0.00	5.42±0.00	**5.31**±0.03
Granite Speech 2B	3.94±0.00	**3.87**±0.01	5.37±0.00	**5.23**±0.02
SeamlessM4T-med	**7.47**±0.44	7.92±0.01	11.92±0.01	**11.55**±0.03
SeamlessM4T-v2	**5.20**±0.38	5.62±0.00	14.31±0.09	**14.07**±0.05
Whisper large-v3	**5.64**±0.10	5.77±0.00	**4.37**±0.00	**4.37**±0.00
Whisper small	**8.08**±0.27	8.39±0.00	**21.34**±0.00	**21.34**±0.00

Speech backbone	TED-3		VoxPopuli	
	Baseline	LAA	Baseline	LAA
Granite Speech 8B	11.59±0.00	**11.49**±0.03	6.30±0.00	**6.07**±0.01
Granite Speech 2B	11.73±0.00	**11.65**±0.04	**6.12**±0.00	6.17±0.01
SeamlessM4T-med	23.16±0.00	**22.71**±0.05	9.81±0.02	**9.76**±0.01
SeamlessM4T-v2	24.79±0.00	**24.14**±0.02	**8.66**±0.04	8.68±0.02
Whisper large-v3	**12.59**±0.00	**12.59**±0.00	**11.60**±0.00	**11.60**±0.00
Whisper small	**11.80**±0.00	**11.80**±0.00	**87.62**±0.00	**87.62**±0.00

Table 3: Frozen text-refiner sensitivity with Granite Speech 8B (first block) and 2B (second block) at R40. Values are WER (%), reported independently for each benchmark. The 2B Qwen3-0.6B row follows the component-analysis setting. Lower is better; bold compares Baseline with LAA.

LLM/refiner	AMI		CHiME-6		FLEURS		LS-clean	
	Baseline	LAA	Baseline	LAA	Baseline	LAA	Baseline	LAA
Granite Speech 8B backbone								
Qwen3-4B	14.99±0.00	**14.90**±0.03	68.89±0.00	**68.81**±0.06	8.13±0.00	**8.02**±0.02	1.43±0.00	**1.24**±0.02
Mistral-7B	19.07±0.00	**16.31**±0.06	71.27±0.00	**69.35**±0.16	9.97±0.00	**8.42**±0.01	5.66±0.00	**2.92**±0.01
Phi-3.5-mini	20.10±0.00	**16.41**±0.07	71.58±0.00	**68.54**±0.11	9.44±0.00	**8.42**±0.02	4.98±0.00	**2.73**±0.08
LFM2-350M	16.25±0.00	**15.33**±0.01	69.97±0.00	**68.93**±0.05	8.31±0.00	**8.07**±0.02	2.19±0.00	**1.53**±0.01
SmolLM2-360M	15.11±0.00	**14.90**±0.05	**68.48**±0.00	68.57±0.03	**8.00**±0.00	8.05±0.01	1.27±0.00	**1.22**±0.00
Qwen3-0.6B	15.06±0.00	**14.85**±0.03	68.97±0.00	**68.80**±0.02	8.04±0.00	**8.03**±0.00	1.36±0.00	**1.25**±0.01
Granite Speech 2B backbone								
Qwen3-4B	15.08±0.00	**15.06**±0.04	68.68±0.00	**68.29**±0.06	8.44±0.00	**8.35**±0.01	1.54±0.00	**1.37**±0.01
Mistral-7B	19.48±0.00	**16.45**±0.08	71.29±0.00	**68.76**±0.18	10.53±0.00	**8.98**±0.02	5.55±0.00	**2.95**±0.03
Phi-3.5-mini	22.78±0.00	**18.26**±0.14	71.11±0.00	**68.86**±0.17	10.27±0.00	**8.86**±0.00	5.04±0.00	**3.02**±0.04
LFM2-350M	16.40±0.00	**15.30**±0.03	69.66±0.00	**68.60**±0.05	8.73±0.00	**8.36**±0.02	2.23±0.00	**1.59**±0.03
SmolLM2-360M	15.12±0.00	**15.02**±0.04	68.25±0.00	**68.08**±0.04	**8.28**±0.00	**8.28**±0.01	1.39±0.00	**1.34**±0.00
Qwen3-0.6B	16.88±0.00	**16.57**±0.03	81.14±0.00	**80.64**±0.03	**6.65**±0.00	6.71±0.01	1.30±0.00	**1.29**±0.00

LLM/refiner	LS-other		SPGI		TED-3		VoxPopuli	
	Base.	LAA	Base.	LAA	Base.	LAA	Base.	LAA
Granite Speech 8B backbone								
Qwen3-4B	3.53±0.00	**3.30**±0.01	5.38±0.00	**5.33**±0.05	11.46±0.00	**11.45**±0.01	6.07±0.00	**6.04**±0.01
Mistral-7B	8.04±0.00	**5.14**±0.02	8.05±0.00	**6.02**±0.06	13.38±0.00	**12.16**±0.06	8.74±0.00	**6.92**±0.07
Phi-3.5-mini	7.60±0.00	**4.73**±0.08	13.12±0.00	**5.94**±0.05	17.18±0.00	**12.42**±0.09	8.73±0.00	**6.80**±0.04
LFM2-350M	4.13±0.00	**3.54**±0.01	5.56±0.00	**5.21**±0.03	11.98±0.00	**11.73**±0.04	6.66±0.00	**6.24**±0.04
SmolLM2-360M	3.40±0.00	**3.25**±0.01	5.40±0.00	**5.33**±0.07	11.56±0.00	**11.47**±0.03	6.09±0.00	**6.01**±0.03
Qwen3-0.6B	3.45±0.00	**3.27**±0.00	5.42±0.00	**5.31**±0.03	11.59±0.00	**11.49**±0.03	6.30±0.00	**6.07**±0.01
Granite Speech 2B backbone								
Qwen3-4B	4.12±0.00	**3.90**±0.01	5.35±0.00	**5.30**±0.04	11.59±0.00	**11.57**±0.01	6.15±0.00	**6.10**±0.00
Mistral-7B	8.08±0.00	**5.39**±0.02	7.81±0.00	**5.89**±0.02	13.52±0.00	**12.23**±0.06	9.13±0.00	**7.09**±0.03
Phi-3.5-mini	7.49±0.00	**5.42**±0.05	15.63±0.00	**5.96**±0.06	16.29±0.00	**12.33**±0.02	8.54±0.00	**6.85**±0.01
LFM2-350M	4.70±0.00	**4.23**±0.01	5.60±0.00	**5.25**±0.02	12.18±0.00	**11.83**±0.05	6.91±0.00	**6.37**±0.01
SmolLM2-360M	4.02±0.00	**3.91**±0.00	5.34±0.00	**5.24**±0.02	11.73±0.00	**11.63**±0.05	6.20±0.00	**6.14**±0.01
Qwen3-0.6B	3.94±0.00	**3.87**±0.01	5.37±0.00	**5.23**±0.02	11.73±0.00	**11.65**±0.04	**6.12**±0.00	6.17±0.01

Table 4: Frozen-refiner scaling with Granite Speech 2B on a separate LibriSpeech test-other subset. LAA WER (%) is mean ± sample SD over three phoneme-head seeds at R40. These absolute values are not comparable with the other experiments. Lower is better; bold marks the lowest mean.

Frozen text refiner	LAA WER (%)	Frozen text refiner	LAA WER (%)
Qwen2.5-0.5B	3.18±0.01	Qwen3-0.6B	3.22±0.01
Qwen2.5-1.5B	3.17±0.01	Qwen3-1.7B	3.16±0.02
Qwen2.5-3B	3.22±0.01	Qwen3-4B	3.19±0.02
Qwen2.5-7B	3.14±0.01	Qwen3-8B	3.16±0.01
Qwen2.5-14B	3.15±0.01	Qwen3-14B	3.16±0.00
LFM2-350M	3.21±0.02	Qwen3-32B	3.16±0.01
LFM2-700M	3.26±0.02	SmolLM2-135M	**3.13**±0.02
LFM2-1.2B	3.27±0.00	SmolLM2-360M	3.16±0.01
LFM2-2.6B	3.26±0.00	SmolLM2-1.7B	3.16±0.02

Table 5: Component decomposition with Granite Speech 8B and Granite Speech 2B.

Benchmark	Speech backbone	Unconstrained	Offset	Baseline	LAA
AMI	Granite Speech 8B	15.96±0.00	14.98±0.03	15.06±0.00	**14.85**±0.03
	Granite Speech 2B	17.74±0.00	16.68±0.06	16.88±0.00	**16.57**±0.03
CHiME-6	Granite Speech 8B	73.83±0.00	68.95±0.04	68.97±0.00	**68.80**±0.02
	Granite Speech 2B	81.71±0.00	80.83±0.02	81.14±0.00	**80.64**±0.03
FLEURS	Granite Speech 8B	8.18±0.00	8.03±0.02	8.04±0.00	**8.03**±0.00
	Granite Speech 2B	6.80±0.00	6.65±0.02	**6.65**±0.00	6.71±0.01
LS-clean	Granite Speech 8B	1.43±0.00	1.31±0.00	1.36±0.00	**1.25**±0.01
	Granite Speech 2B	1.83±0.00	1.30±0.01	1.30±0.00	**1.29**±0.00
LS-other	Granite Speech 8B	3.64±0.00	3.36±0.01	3.45±0.00	**3.27**±0.00
	Granite Speech 2B	4.23±0.00	3.89±0.01	3.94±0.00	**3.87**±0.01
SPGI	Granite Speech 8B	5.46±0.00	5.44±0.04	5.42±0.00	**5.31**±0.03
	Granite Speech 2B	5.53±0.00	5.31±0.05	5.37±0.00	**5.23**±0.02
TED-3	Granite Speech 8B	11.79±0.00	11.49±0.04	11.59±0.00	**11.49**±0.03
	Granite Speech 2B	12.03±0.00	11.69±0.04	11.73±0.00	**11.65**±0.04
VoxPopuli	Granite Speech 8B	6.32±0.00	6.17±0.02	6.30±0.00	**6.07**±0.01
	Granite Speech 2B	6.48±0.00	6.13±0.02	**6.12**±0.00	6.17±0.01

WER (%), mean ± SD. Qwen3-0.6B, R40, $N = 4$. Bold indicates the lower mean between Baseline and LAA within each row. Unconstrained: free-form correction; Offset: offset-audio control. Each backbone uses its own fixed candidate and reference set.

Table 6: LAA policy sensitivity for Granite Speech 2B + Qwen3-0.6B across all eight benchmarks. The frozen router, candidates, and three phoneme-head seeds match the controlled Granite Speech 2B analysis; its R40 row reproduces their LAA values. Overall is computed within each seed before the sample SD. R10–R90 vary only the source-development Stage-1 recall cutoff. The two controls respectively remove Stage 2 from the R40 call set and send every span to verification. Lower is better.

Variant	AMI	CHiME-6	FLEURS	LS-clean
w/o Stage-2	16.613±0.080	80.563±0.023	6.857±0.028	1.266±0.034
All-span verification	16.646±0.114	82.164±0.119	6.815±0.011	1.304±0.006
R10	16.779±0.019	80.733±0.085	6.661±0.004	1.291±0.003
R20	16.652±0.019	80.707±0.041	6.634±0.004	1.290±0.003
R30	16.561±0.054	80.557±0.020	6.675±0.004	1.288±0.005
R40	16.571±0.030	80.642±0.030	6.707±0.007	1.288±0.005
R50	16.607±0.024	80.602±0.030	6.707±0.007	1.291±0.007
R60	16.522±0.034	80.629±0.023	6.792±0.011	1.291±0.007
R70	16.513±0.034	80.635±0.020	6.792±0.011	1.291±0.007
R80	16.493±0.044	80.668±0.030	6.792±0.011	1.291±0.007
R90	16.464±0.039	80.531±0.030	6.792±0.011	1.291±0.007

Variant	LS-other	SPGI	TED-3	VoxPopuli	Overall
w/o Stage-2	4.044±0.019	5.266±0.035	11.913±0.123	6.465±0.017	16.623±0.037
All-span verification	3.975±0.017	5.254±0.009	11.678±0.107	6.183±0.021	16.752±0.019
R10	3.890±0.009	5.326±0.014	11.667±0.039	6.151±0.005	16.562±0.017
R20	3.867±0.009	5.296±0.005	11.663±0.039	6.134±0.020	16.530±0.010
R30	3.877±0.007	5.298±0.004	11.663±0.035	6.131±0.028	**16.506**±0.010
R40	3.875±0.006	5.232±0.018	11.653±0.045	6.169±0.010	16.517±0.006
R50	3.875±0.006	5.238±0.023	11.665±0.048	6.160±0.011	16.518±0.004
R60	3.905±0.015	5.251±0.017	11.711±0.051	6.158±0.015	16.532±0.009
R70	3.915±0.017	5.255±0.015	11.711±0.051	6.156±0.017	16.534±0.010
R80	3.925±0.015	5.251±0.015	11.699±0.058	6.160±0.017	16.535±0.013
R90	3.919±0.015	5.249±0.010	11.689±0.062	6.165±0.017	16.513±0.015

Table 7: Candidate-width sensitivity for Granite Speech 2B with Qwen3-0.6B under fixed R40 routing. $N = 1$ is the direct-ASR control. Overall values are computed within each seed before the sample SD, and Gain is the paired Baseline−LAA difference. Lower WER and higher Gain are better; bold marks the best displayed mean.

Benchmark	Route	$N = 1$	$N = 2$	$N = 4$	$N = 8$	$N = 16$
AMI	Baseline	16.41±0.00	15.03±0.00	15.10±0.00	15.13±0.00	15.20±0.00
	LAA	16.41±0.00	**14.84**±0.03	14.88±0.03	14.93±0.02	15.30±0.03
CHiME-6	Baseline	83.34±0.00	68.19±0.00	68.60±0.00	68.60±0.00	68.62±0.00
	LAA	83.34±0.00	67.91±0.09	67.71±0.12	**67.52**±0.08	68.21±0.25
FLEURS	Baseline	6.64±0.00	6.62±0.00	6.62±0.00	6.63±0.00	6.65±0.00
	LAA	6.64±0.00	**6.61**±0.01	6.70±0.00	6.74±0.03	6.80±0.02
LS-clean	Baseline	1.31±0.00	1.30±0.00	1.32±0.00	1.32±0.00	1.32±0.00
	LAA	1.31±0.00	**1.28**±0.01	1.29±0.01	1.30±0.01	1.33±0.01
LS-other	Baseline	3.94±0.00	3.90±0.00	3.93±0.00	3.95±0.00	3.98±0.00
	LAA	3.94±0.00	**3.82**±0.01	3.85±0.00	3.90±0.01	3.98±0.01
SPGI	Baseline	5.46±0.00	5.46±0.00	5.44±0.00	5.40±0.00	5.37±0.00
	LAA	5.46±0.00	5.36±0.02	5.26±0.03	5.24±0.05	**5.21**±0.09
TED-3	Baseline	11.56±0.00	**11.39**±0.00	11.42±0.00	11.44±0.00	11.47±0.00
	LAA	11.56±0.00	11.40±0.01	11.45±0.02	11.51±0.03	11.64±0.05
VoxPopuli	Baseline	**6.04**±0.00	6.07±0.00	6.07±0.00	6.09±0.00	6.11±0.00
	LAA	**6.04**±0.00	6.10±0.02	6.10±0.03	6.17±0.03	6.23±0.05
Overall WER	Baseline	16.838±0.000	14.744±0.000	14.812±0.000	14.821±0.000	14.839±0.000
	LAA	16.838±0.000	14.664±0.009	**14.654**±0.014	14.661±0.010	14.837±0.019
Gain ↑	B.−LAA	0.000±0.000	0.081±0.009	0.158±0.014	**0.160**±0.010	0.002±0.019

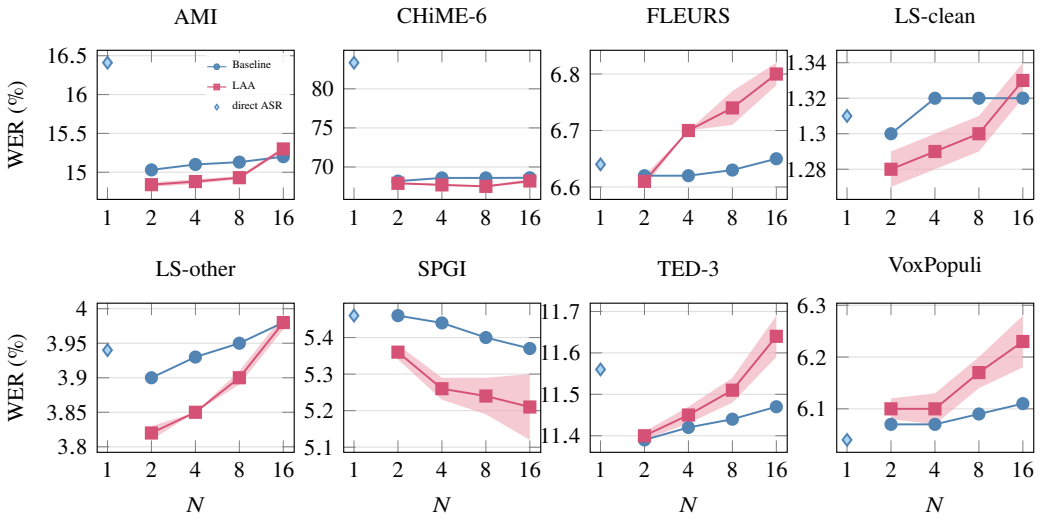

Figure 3: Candidate-width sensitivity. Baseline and LAA lines show $N = 2, 4, 8, 16$; the isolated diamond at $N = 1$ is the direct-ASR control. Shaded bands show sample SD over three phoneme-head seeds for LAA; Baseline and direct ASR are deterministic.