CONSTRUCTION CONFIDENCE NEED NOT BE VERIFICATION CONFIDENCE: ROUND-TRIP VERIFICATION AND SELECTIVE REPAIR FOR MASKED DIFFUSION SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked diffusion models can reduce computation by decoding in only a few iterative steps, but in speech generation this low-step regime can leave a minority of outputs with content failures, and which utterances fail can vary across sampling seeds. This suggests a selective strategy of spending extra computation only on outputs likely to fail. The challenge is identifying them. We study this problem in MaskGCT, a zero-shot text-to-speech system whose two stages are masked-diffusion decoders. The first maps text to semantic tokens, a discrete representation of what is said, and the second maps semantic tokens to acoustic tokens that are rendered into speech. The decoder therefore commits tokens, while the failure criterion is defined only after those tokens have been converted into a waveform. We define a content failure as rendered speech whose transcript has a word error rate (WER) above 10% against the target text, that is, speech whose content is unreliable rather than merely imperfect. This distinction gives rise to two kinds of confidence. Construction confidence comes from decoder-derived signals, such as token confidence and entropy, while the sequence is being committed. Verification confidence comes from checking the rendered speech for evidence that the output actually meets that criterion. Used as a failure detector on held-out data at an eight-step decoding budget, construction confidence falls well short of verification. A probing classifier trained on summaries of these signals (the learned probe) reaches AUC 0.629, whereas round-trip transcription with a separate small ASR verifier reaches 0.864. Round-trip transcription stays above every tested decoder-derived signal at the other decoding budgets we examined, and again in the acoustic stage. It also does so under failure labels from ASR families different from the verifier's, although the margin is smaller. Because a masked-diffusion sequence remains editable, verifier-localized spans can be re-masked and regenerated in place, whereas an autoregressive decoder would have to regenerate everything after them. This selective repair reduces transcript error relative to the unrepaired output across four ASR evaluators. These results show that, when a generator is judged only after its committed tokens have been transformed further, confidence used to construct those tokens need not be sufficient to verify the final output.

1 INTRODUCTION

Masked diffusion models generate through iterative parallel refinement, with the decoding schedule (which positions to commit, in what order, and over how many steps) chosen at inference time (Chang et al., 2022; Austin et al., 2021; Sahoo et al., 2024; Kim et al., 2025a; Ben-Hamu et al., 2025; Kim et al., 2025b; Wang et al., 2025a). Existing methods use confidence, entropy, margin, or stability to allocate this computation while the sequence is being constructed. Unlike an autoregressive prefix, a masked-diffusion sequence remains editable in principle: selected positions can be re-masked and predicted again from the surrounding context.

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.

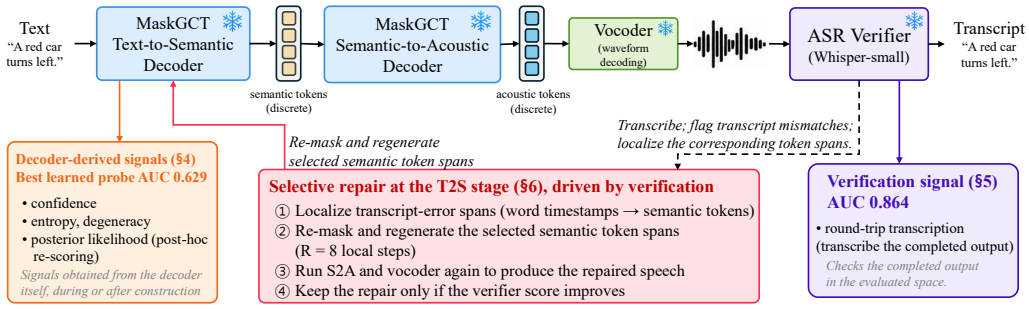


Figure 1: Overview. MaskGCT generates speech in two masked-diffusion stages (text to semantic tokens, semantic to acoustic tokens) followed by waveform decoding; snowflakes mark frozen weights. We compare two sources of evidence about whether a completed output failed: decoder-derived signals read from the text-to-semantic decoder (orange; Section 4) and round-trip transcription of the rendered speech by a separate ASR verifier (purple; Section 5). The verifier's transcript also localizes mismatched spans, which are re-masked and regenerated at the text-to-semantic stage and kept only if the verifier score improves (red; Section 6). AUC values are held-out English $s8$ results under Whisper-large-v3 labels.

At low decoding budgets, failures can vary across sampling seeds, creating an opportunity to allocate extra computation selectively rather than uniformly. The difficulty is deciding which completed outputs need it.

We distinguish two roles for confidence. *Construction confidence* concerns how computation should be allocated while a sample is being formed, whereas *verification confidence* concerns whether a completed output satisfies the criterion by which it will be judged. We study settings in which evaluation is performed on a downstream transformation of the decoder's committed representation, as in speech generation, where discrete tokens are rendered into a waveform. The empirical question is how much useful failure information they retain after the committed representation has been transformed into the final output that is actually evaluated.

Our testbed is MaskGCT (Wang et al., 2025b), a zero-shot text-to-speech system with two masked-diffusion stages, text-to-semantic (T2S) and semantic-to-acoustic (S2A), followed by waveform rendering. T2S commits semantic tokens before acoustic decoding and waveform rendering. Our concern, however, is whether the final speech preserves the intended linguistic content. We therefore evaluate content reliability on the rendered output, defining a content failure as a transcript word error rate (WER) above 10% against the target text.

Figure 1 summarizes the overall pipeline, including the decoder-derived signals used for analysis, output-space verification, and verification-conditioned selective repair.

On held-out English $s8$ outputs, a learned probe, a linear classifier trained on summary statistics of the decoder-derived signals, reaches an area under the ROC curve (AUC) of 0.629, whereas round-trip transcription with a separate small ASR verifier reaches 0.864. This gap persists across the other decoding budgets, the S2A stage, and alternative ASR failure labels.

Development analyses suggest a mismatch in what the two signal types capture. The tested decoder-derived signals vary largely between utterances, whereas low-step content failures also vary substantially across different realizations of the same utterance. This difference helps explain why construction signals are weaker completed-output detectors than round-trip verification.

We further use output-space verification to drive selective local repair. Repair improves transcript-defined content reliability relative to the same unrepaired bases across four ASR evaluators, while comparisons with longer denoising trajectories are less resolved. Under a hard domain shift, however, the verifier loses selectivity and repair can become harmful, showing that selective inference depends on the verifier remaining informative in the target domain.

Our contributions are threefold:

1. **Construction versus verification.** We distinguish confidence used to construct a masked-diffusion sample from evidence used to verify its completed output, and show that the tested decoder-derived signals are substantially weaker failure detectors than round-trip transcription across multiple T2S budgets, a secondary S2A stage, and alternative ASR failure labels.
2. **Diagnosis of the gap.** We show that the tested decoder-derived signals vary largely along an utterance-level axis, while content failure contains more realization-level variation, and use centering analyses to characterize this mismatch.
3. **Verification-conditioned selective inference.** We use output-space verification to drive local repair, obtain consistent held-out gains over the corresponding unrepaired bases across four ASR evaluators, and identify limits when the alternative is additional denoising or when verifier selectivity deteriorates under domain shift.

2 Related work

Inference-time computation and verification in masked diffusion. Masked diffusion models generate through iterative parallel refinement (Chang et al., 2022; Austin et al., 2021; Sahoo et al., 2024). Adaptive commitment, ordering, stability schedules, and remasking use internal scores to allocate inference-time computation while the sample remains editable (Ben-Hamu et al., 2025; Kim et al., 2025a;b; Wang et al., 2025a). These methods use such signals to guide sample construction. We instead ask whether the same signals can identify failures in completed outputs when evaluation is performed after the decoder's committed representation has been rendered.

Output-space checking and selective generation. ASR round trips have been used for best-of-N selection and distillation in autoregressive codec TTS (Asaria et al., 2026), while rendered-audio ASR and speaker-similarity metrics have served as preference or reinforcement signals (Du et al., 2024; Sun et al., 2025). We instead use round-trip transcription to verify completed outputs and compare its failure-detection signal directly with the decoder-derived signals available during construction.

Selective prediction and self-verification. Our detector analysis is related to selective prediction, calibration, and learned error prediction (Geifman & El-Yaniv, 2017; Hendrycks & Gimpel, 2017; Guo et al., 2017; Zadrozny & Elkan, 2002). ASR-based conclusions can also depend on the evaluator (Yu & Kang, 2026), motivating our use of multiple recognizers. Related questions arise in language models, where confidence, sampling consistency, and separately trained verifiers have been studied as ways to assess completed outputs (Kadavath et al., 2022; Kuhn et al., 2023; Farquhar et al., 2024; Cobbe et al., 2021; Lightman et al., 2023), while intrinsic self-correction without external feedback can fail (Huang et al., 2024). Our setting asks an analogous question in speech generation, where evaluation is performed on a rendered output downstream of the model's committed representation.

3 Experimental setup

System. MaskGCT (Wang et al., 2025b) maps text to 50-Hz semantic tokens in its text-to-semantic (T2S) stage, then to acoustic tokens from twelve residual vector-quantization codebooks in its semantic-to-acoustic (S2A) stage, followed by waveform decoding. The semantic tokens primarily encode phonetic content with reduced speaker-specific detail, so the T2S stage commits a content-bearing representation before acoustic rendering. Masked positions are predicted in parallel and committed by confidence with Gumbel perturbations (Chang et al., 2022). We vary the T2S step count while keeping the default S2A schedule fixed (Appendix L); sK denotes an initial T2S budget of K steps. S2A analyses use iteratively generated codebook 0, the first acoustic codebook in the residual stack, so their claims are limited to iteratively generated acoustic codebooks.

Failure definition and ASR roles. We report English WER, Mandarin CER, and content failure rate (CFR), defined as the fraction of utterances with transcript error above 10%; alternative thresholds are reported in Appendix A. CFR is transcript-defined and is not a human-intelligibility measure.

Table 1: ASR evaluation panel for the intervention experiments. Reference WER is measured on benchmark recordings using the unified evaluation path and is included to characterize ASR error levels on reference audio. Evaluators are reported separately rather than combined into a consensus label.

Evaluator	Architecture	Ref. WER	Verifier relation
HuBERT-large	SSL + CTC	5.69	cross-family
Whisper-large-v3	seq2seq	1.73	same-family
wav2vec2-Conformer	SSL + CTC	6.19	cross-family
Paraformer-en (FunASR)	supervised NAR	3.62	cross-family

Whisper-large-v3 (Radford et al., 2023) defines the main binary failure label, while Whisper-small (Radford et al., 2023) transcribes rendered speech to provide the round-trip verification score. The verifier score is the transcript error of Whisper-small against the target text, rather than a decoder-derived confidence quantity. Repair outcomes are evaluated separately with four ASR models (Table 1), with HuBERT-large (Hsu et al., 2021) prespecified as primary; Whisper-small is not in this panel. Because the main label and the verifier share the Whisper family, cross-family label and verifier controls are reported in Appendix C.

Data and statistics. We use Seed-TTS-Eval (Anastassiou et al., 2024): development/held-out splits contain $218/870$ English, $404/1616$ Mandarin, and $80/320$ hard-case utterances. All selection used only development data, and no held-out result was examined before final evaluation. Before held-out evaluation, we fixed the primary evaluator, CFR thresholds, decoder-derived summaries, learned probe, and the AUC criterion of 0.75 (acceptable discrimination; Hosmer et al., 2013). Each condition uses five sampling seeds from one checkpoint. Comparisons are paired by utterance using 4000 utterance-cluster bootstrap resamples (Field & Welsh, 2007), with Holm correction (Holm, 1979) within prespecified families. We call a difference *resolved* when its 95% interval excludes zero.

T2S decoding budgets. Detection uses $s8$ and $s16$; repair uses $s16$ and $s32$ bases, with $s64$ and $s128$ as additional-denoising comparators. Here sK denotes an initial T2S budget of K steps, whether or not repair follows. This design separates two questions: whether the verifier can rank failures, and whether using that ranking to trigger repair improves outputs generated at a fixed initial T2S budget. The $s64$ and $s128$ conditions provide additional-denoising comparators for asking whether repair offers a different use of inference-time computation. We use *operating point* to denote the initial T2S step count (e.g., $s16$), whether or not subsequent repair is applied.

Cost. We report average wall-clock cost per utterance, including verification, localization, all repair attempts (accepted or rejected), re-rendering, and re-verification when invoked; one-time model loading is excluded (Appendix L).

Decoder-derived signals. At each position's commitment step, we record confidence (top-1 probability), entropy, and degeneracy (one minus the probability mass of the top-50 tokens) from the raw pre-filter distribution. We also evaluate posterior likelihood, a post-hoc decoder-derived score that re-scores masked positions in the completed sequence. Position-level values are pooled using rules selected on the development split and frozen, and a linear learned probe combines five summary statistics (mean, 10th and 90th percentiles, maximum, variance) of each available signal, giving 20 features at $s8$ and 15 in the prespecified $s16$ probe (Appendix D).

4 Decoder-derived signals fall short for completed-output verification

Both decoder-derived and round-trip detectors are evaluated by utterance-level AUC for the content-failure labels defined in Section 3. Figure 2 reports the primary held-out English $s8$ T2S comparison and a prespecified secondary replication in S2A codebook 0. In both stages, the upper confidence bounds of the tested decoder-derived detectors remain below the prespecified criterion of 0.75,

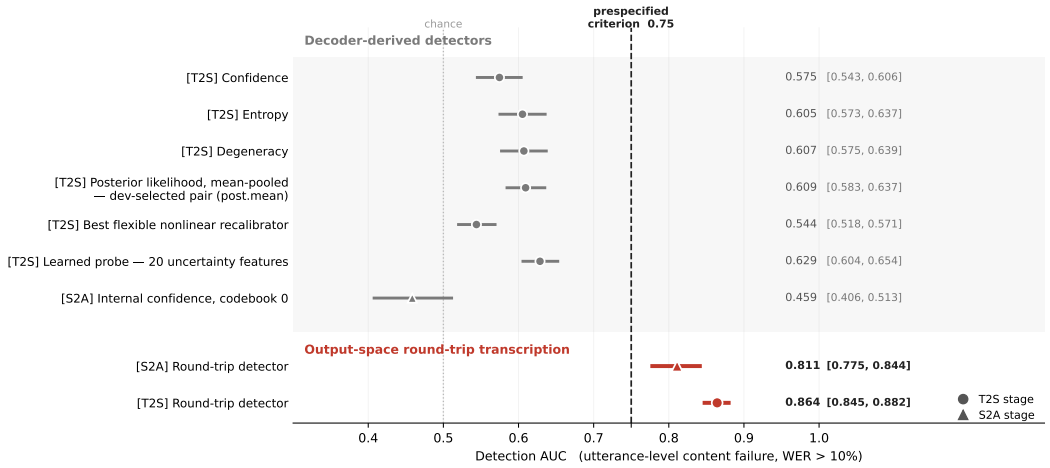


Figure 2: Held-out detection AUC for transcript-defined content failure (WER $>$ 10%, Whisper-large-v3 labels; 95% utterance-cluster bootstrap intervals). T2S uses $s8$ and S2A uses codebook 0. T2S pooling, recalibration, and probe choices were fixed on the development split and applied without re-selection.

whereas the round-trip lower bounds exceed it. The S2A evaluation was executed after unblinding and is therefore treated as secondary.

4.1 Decoder-derived readouts across budgets, stages, and ASR pairings

At $s8$, the tested scalar decoder-derived readouts, including posterior likelihood re-scored after generation, reach AUC 0.575–0.609 on held-out data, and development-selected pooling and nonlinear recalibration do not improve on them (best recalibrator 0.544; Appendix D). The learned probe is stronger at 0.629 [$0.604, 0.654$], but round-trip transcription reaches 0.864 [$0.845, 0.882$]; their paired AUC difference is $+0.236$ [+0.207, +0.263].

The pattern persists at larger T2S budgets. At $s16$, the prespecified 15-feature probe reaches 0.515 [$0.481, 0.549$]; adding posterior-likelihood features in a post-hoc control changes this only to 0.516 [$0.482, 0.551$]. Round-trip transcription remains strong at 0.847 [$0.814, 0.877$], similar in magnitude to its $s8$ estimate. A post-hoc $s32$ control shows the same ordering, with AUC 0.591 [$0.551, 0.632$] for the learned probe and 0.832 [$0.795, 0.868$] for round-trip transcription.

The ordering is also robust to the ASR models used for verification and failure labels. Across the tested Whisper-small, Whisper-medium, and data2vec-audio-base-960h verifiers under Whisper-large-v3 and HuBERT-large labels, round-trip detection exceeds the learned probe in every pairing at all three T2S budgets, with paired differences excluding zero, although its absolute AUC depends on the verifier–label combination (Appendix C).

The same gap appears in S2A codebook 0. Commitment-time confidence reaches AUC 0.459 [$0.406, 0.513$], whereas round-trip transcription reaches 0.811 [$0.775, 0.844$], a paired difference of $+0.352$ [+0.287, +0.417].

4.2 The signals and failures vary along different axes

Failure includes substantial realization-level variation. On the development split, its intraclass correlation (ICC), the share of variance attributable to between-utterance differences rather than within-utterance variation across seeds, is 0.27 [$0.18, 0.35$] at $s8$ and rises to 0.50 [$0.34, 0.61$] at $s16$, indicating that realization-level variation decreases as the denoising budget increases.

The decoder-derived signals vary much more strongly between utterances. Confidence, entropy, and degeneracy have ICCs of 0.74–0.75 at $s8$ and 0.81–0.87 at $s16$, compared with failure ICCs of 0.27 and 0.50. Round-trip transcription instead has ICCs of 0.25 and 0.46, closer to the corresponding

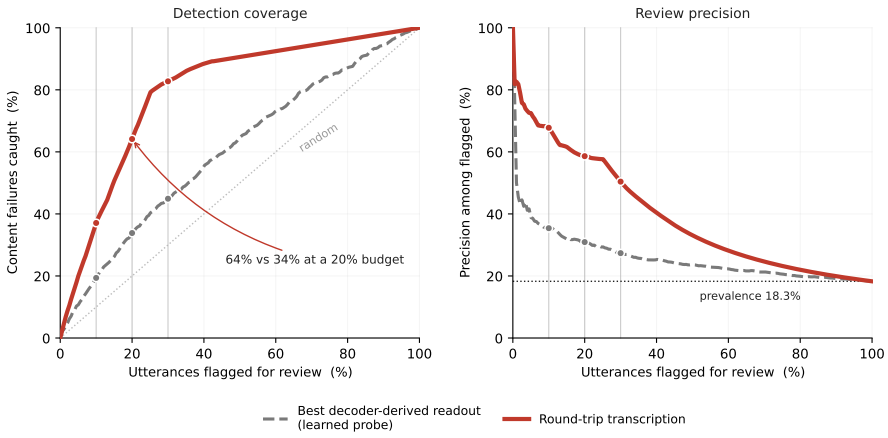


Figure 3: Review-budget trade-off on the English held-out $s8$ condition ($n = 4350$, content-failure prevalence 18.28%; WER $> 10\%$, Whisper-large-v3 labels). Curves trace review thresholds, with markers at budgets of 10, 20, and 30% of utterances. This is a post-hoc descriptive analysis and does not define the frozen repair-policy threshold.

failure ICCs. Consistently, subtracting each utterance's across-seed mean modestly improves several decoder-derived detectors while substantially reducing round-trip detection AUC (Appendix K). This diagnostic supports a mismatch between the variation captured by the tested decoder-derived signals and realization-level failure.

Finally, development controls over alternative commitment orders, heuristic schedules, and an exact-token oracle show little headroom within the tested confidence- and exact-match-based rule class (Appendix J). Within this tested class, changing construction order therefore does not close the verification gap.

5 Output-space verification

Having established that round-trip transcription is a stronger completed-output detector, we now ask how selectively it can direct additional computation: how many content failures can be caught when only a fraction of outputs is sent for review, and where that selectivity breaks down.

Given a generated utterance, we transcribe the rendered waveform with Whisper-small and compare the transcript with the conditioning text. The resulting round-trip error is an utterance-level detection score. At this stage, detection operates at the utterance level: it determines whether a generated utterance should receive additional computation, rather than identifying a specific error span. We rank utterances by round-trip error and flag those above a chosen threshold. Varying the threshold changes the review budget, defined as the fraction of utterances routed for further checking or repair rather than accepted as-is. For each review budget, we measure the fraction of content failures caught (Figure 3).

We next characterize round-trip detection in precision–recall terms at $s8$ and $s16$. Because AUPRC depends on failure prevalence (i.e., the fraction of outputs labeled as content failures), each AUPRC value is interpreted together with the corresponding prevalence. On held-out English $s8$ outputs, round-trip verification reaches AUPRC 0.592 $[0.547, 0.638]$ at failure prevalence 18.28%. At $s16$, where failure prevalence is 9.52%, AUPRC is 0.469 $[0.400, 0.545]$. At a fixed recall of 80%, precision is 0.563 at $s8$ and 0.288 at $s16$. Thus, when the threshold is set to catch 80% of failures, the flagged set contains a smaller fraction of true failures at $s16$, although failures are also less prevalent overall.

For illustration, we consider a 20% review budget, selecting the 20% of outputs with the highest detection scores for each detector. The 20% point is the middle entry of the descriptive grid $\{5, 10, 20, 30, 50\}\%$ used in this analysis (Appendix C).

At this review budget, round-trip verification catches 64.1% of all content failures. Of the outputs sent for review, 58.6% are actual failures, while the remaining unreviewed outputs have a failure rate of 8.2%. Under the same review budget, the learned probe catches only 33.8% of failures; 30.9% of its reviewed outputs are failures, and the unreviewed set retains a 15.1% failure rate. Thus, for the same amount of review, round-trip verification directs review toward more failures and leaves a lower failure rate among the outputs that are not reviewed. Across the tested review-budget grid $\{5, 10, 20, 30, 50\}\%$, round-trip verification catches a larger fraction of content failures than the learned probe at every budget.

The English review-budget analysis shows that round-trip verification can concentrate many failures into a relatively small reviewed set. It does not, however, assign a nonzero error score to every failure. In a post-hoc analysis across $s8$, $s16$, and $s32$, the verifier assigns zero round-trip error to 11–22$\%$ of failures defined by Whisper-large-v3 and 35–57$\%$ of failures defined by HuBERT-large, with larger miss fractions at higher T2S budgets (Appendix C). For these outputs, the verifier transcript matches the target even though the failure-label ASR marks the output as a content failure. A threshold on the verifier score therefore cannot selectively identify them.

A separate limitation appears when the frozen threshold is transferred across domains. On ordinary Mandarin, the realized first-round flag rate—the fraction actually flagged under the frozen threshold—is 75.4%. On the held-out hard subset, it rises to 94.8%, so nearly every output is sent for additional computation. The policy therefore retains substantially less selectivity than in the English setting, especially on the hard subset. Repair outcomes under these transfer conditions are examined in Section 6.

6 VERIFICATION-CONDITIONED SELECTIVE INFERENCE

Section 5 showed that round-trip verification can identify completed outputs that are more likely to contain content failures. We now ask whether that signal can support selective intervention rather than detection alone. We first test whether verifier-triggered local repair reduces transcript-defined failures relative to the unrepaired base and measure its computational cost (Section 6.1). We then examine whether the repair benefit depends on where the verifier localizes the error and check for changes in automatic quality measures. Finally, we test how the same frozen policy behaves when transferred to Mandarin conditions in which verifier selectivity changes.

For each base output flagged by the frozen round-trip threshold, we use word-level timestamps from the verifier transcript to map transcript mismatches to the corresponding semantic-token spans. We enlarge each localized semantic-token span by a fixed margin on both sides, then re-mask and regenerate the tokens within the enlarged span using eight local denoising steps while keeping all other tokens fixed. After each repair attempt, we re-evaluate the regenerated output with the round-trip verifier. The repaired candidate is accepted only if its round-trip error is lower than that of the current output; otherwise, the current output is retained. For $s16$ and $s32$ bases, the frozen policy allows at most two such repair rounds (s16 + R8$\times$2 and s32 + R8$\times$2).

6.1 FAILURE REDUCTION AND COST

All sixteen held-out comparisons, spanning four ASR evaluators, two metrics (WER and CFR), and two base operating points ($s16$ and $s32$), favor repair and survive Holm correction (Figure 4). At the $s32$ base, the primary HuBERT-large evaluator, which is outside the verifier's Whisper family, shows WER and CFR reductions of $0.39\,pp$ [+0.26, +0.54] and $1.20\,pp$ [+0.74, +1.66], respectively. Thus, the within-base repair gain is also observed under an evaluator outside the verifier's ASR family.

The gain comes with additional computation. At s32+R8$\times$2, detection, localization, repair attempts, re-rendering, and re-verification add 1.34 s per utterance, corresponding to about 24% of the total cost. The realized flag rates at $s32$, defined as the fractions of all utterances flagged in each round, are 25.9% in round 1 and 18.9% in round 2. Among the flagged candidates, 37.7% are accepted in round 1 and 13.8% in round 2. Direct comparisons with longer-denoising baselines at $s64$ and $s128$ are less decisive (Appendix I).

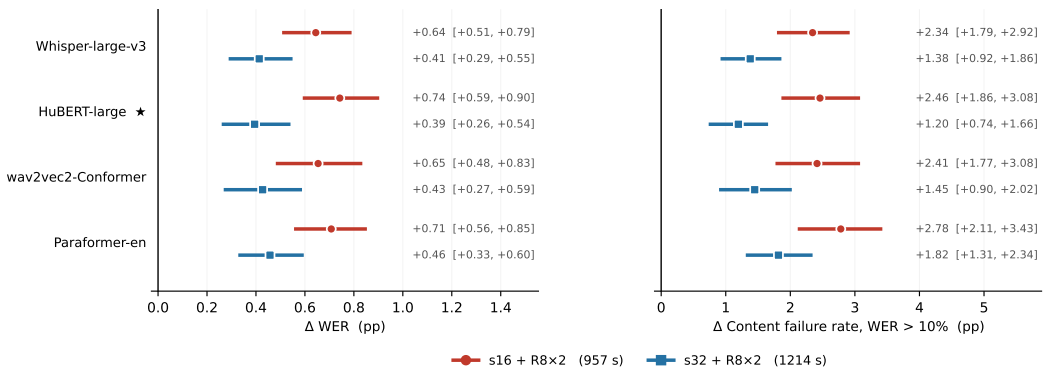


Figure 4: Held-out effect of round-trip-guided local repair relative to the corresponding unrepaired base. Points show paired differences (base − repair; positive favors repair) in WER and content failure rate (CFR; WER $> 10\%$) at the $s16$ and $s32$ base operating points across four ASR evaluators; bars show 95% utterance-cluster bootstrap intervals. The star marks HuBERT-large, the prespecified primary evaluator. All sixteen comparisons favor repair and survive Holm correction (Section 6.1).

6.2 Localization and automatic quality checks

A prespecified held-out ablation tests whether the location chosen for repair matters. We compare ASR-localized repair with a random span matched to the same total masked extent, a span selected from the decoder's lowest-confidence region, and full-utterance repair with eight or thirty-two denoising steps. Across the eight combinations of four evaluators and two metrics, every comparison with the random-span baseline favors ASR localization and excludes zero before correction; six survive Holm correction. All eight comparisons with the decoder low-confidence baseline also favor ASR localization, with five surviving Holm correction. The prespecified primary comparison, HuBERT-large CFR against the random-span baseline, reaches $p_{\text{Holm}} = 0.055$ and therefore does not meet the corrected significance threshold. Taken together, these results support ASR-based localization, although the prespecified primary criterion is not met. Full-utterance, acceptance-rate, and cost results are reported in Appendix C.

Because the acceptance rule keeps a repaired candidate only when its transcript-based verifier score improves, we also check whether repair changes other automatically measured properties of the speech. We examine F0, energy, pauses, speaker similarity, boundary discontinuity, and UTMOS (Chen et al., 2022; Desplanques et al., 2020; Saeki et al., 2022). At $s32$, none of these measures shows a resolved change; at $s16$, energy variation decreases slightly. These measures serve as automatic quality checks rather than evidence of perceptual non-inferiority. Complete analyses are reported in Appendices G and C.

6.3 Transfer under verifier shift

The same frozen repair policy has different effects when transferred to Mandarin conditions. Using the base-minus-repair convention of Figure 4, the CER difference is $+0.127\,pp$ [+0.075, +0.179] on held-out ordinary Mandarin but $-0.577\,pp$ [-0.929, -0.252] on the held-out hard subset.

The verifier also becomes much less selective under this transfer. With the same frozen threshold, the first-round flag rate rises to 75.4% on ordinary Mandarin and 94.8% on the hard subset. In a post-hoc analysis of the hard subset, only the quartile with the highest verifier error shows a resolved CER degradation; the intervals for the other three quartiles include zero. This pattern is consistent with verifier mismatch contributing to the observed failure, but does not establish it as the sole cause. Detailed stratified results are reported in Appendix C.11; circularity and overlap controls are reported in Appendices E and B.

7 DISCUSSION

Construction and verification answer different questions. Decoder-derived confidence serves construction while a sample remains editable, but it is not necessarily a strong indicator of whether the rendered output satisfies the final evaluation criterion. The learned probe does not close this gap, suggesting that the tested summaries do not recover the information provided by output-space verification. A post-hoc development-split probe over T2S hidden states did not exceed the summary probe; richer cross-stage and trajectory features remain untested.

Selective repair supports a narrower claim than strategy ranking. Verifier-guided repair improves the corresponding unrepaired base across the held-out evaluator panel, and the localization controls support acting on ASR-identified spans, although the preregistered primary localization criterion is not met. Comparisons with longer denoising trajectories are less decisive, so our results support verifier-guided improvement of a fixed base rather than a general ranking of repair against additional denoising. Because the verifier is part of the inference procedure, evaluation with recognizers outside its family is informative but does not provide perceptual ground truth.

Verification must remain informative in the target domain. The English review-budget results show that round-trip verification can concentrate failures into a smaller reviewed set, but this selectivity degrades under domain shift and repair can become harmful. Practical use therefore requires target-domain monitoring and a fallback when flag rates become extreme. More broadly, our evidence concerns generators whose committed representation is evaluated only after a downstream transformation; whether the same distinction holds when the generated sequence itself is directly scored remains open.

8 LIMITATIONS

Our evidence comes from one public MaskGCT checkpoint. The decoder-derived readouts reported here cover scalar uncertainty summaries, fixed pooling and recalibration, and a learned probe over those summaries. A post-hoc development-split probe over T2S hidden states did not exceed the summary probe; richer attention, cross-stage, and trajectory-based representations remain outside the reported scope. The S2A result is also a within-system replication rather than evidence from an independently trained model.

Verification and evaluation are transcript-based and depend on the ASR models used. Round-trip performance varies with verifier–label pairing, although its ordering above the decoder-derived signals held in every pairing tested, and the Mandarin transfer results show that a frozen verifier can lose selectivity under domain shift. The four evaluation recognizers therefore provide complementary automatic measurements rather than ground truth; detailed cross-family and transfer controls are reported in Appendix C.

Finally, the repair acceptance rule optimizes transcript error rather than human perception. Automatic measures of prosody, speaker similarity, boundary continuity, and predicted naturalness show no resolved change at $s32$ and only a slight decrease in energy variation at $s16$, but these measures cannot establish perceptual non-inferiority. We therefore claim improvements only in transcript-defined content reliability, not in human intelligibility, naturalness, speaker identity, or overall perceptual quality.

9 CONCLUSION

We distinguish confidence used to construct a masked-diffusion sample from evidence used to verify its completed output. In MaskGCT, the tested decoder-derived signals are substantially weaker content-failure detectors than round-trip transcription, a pattern that persists across T2S budgets, a secondary S2A replication, and alternative ASR labels and verifiers. Using the verifier to trigger and localize repair improves the corresponding unrepaired bases across four evaluators, while transfer results show that the benefit can deteriorate when verifier selectivity changes across domains. These results support output-space verification as a useful basis for selective inference when the evaluation criterion is defined downstream of the model's committed representation.

10 AI USE STATEMENT

Hypotheses, experimental design, pre-registration decisions, and the interpretation of results are the authors'. The generation, verification, and repair pipeline was implemented by the authors. Large language models (LLMs) were used in four ways. (i) Research execution in the analysis stage: LLM-based sessions implemented and ran the analysis scripts listed in Appendix M, including the post-hoc controls in Appendix C and the automated numeric audit, under author-specified protocols; confirmatory analyses followed the frozen pre-registered protocol, while post-hoc analyses are identified as such in the paper. Every reported number is traced to a generated result file, and all outputs were verified by the authors against source files. (ii) Drafting: LLMs drafted portions of the manuscript from author-specified content and result files. (iii) Writing assistance: LLMs aided language editing, clarity, and consistency checking, with limited rewriting. (iv) Retrieval: LLMs assisted the search for potentially relevant related work. The authors reviewed and take full responsibility for all claims, numbers, and text. No LLM was used to fabricate or select results.

REFERENCES

Philip Anastassiou, Jiawei Chen, Jitong Chen, et al. Seed-TTS: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Ali Asaria, Tony Salomone, and Deep Gandhi. Reliable neural-codec text-to-speech by ASR self-verification and distillation: Near-zero catastrophic failures across models and codecs. *arXiv preprint arXiv:2606.18323*, 2026.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas Nolte, and Brian Karrer. Accelerated sampling from masked diffusion models via entropy bounded unmasking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proc. Interspeech*, pp. 3830–3834, 2020. doi: 10.21437/Interspeech.2020-2650.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. CosyVoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Christopher A. Field and Alan H. Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B*, 69(3):369–390, 2007.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, Hoboken, NJ, 3 edition, 2013.

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

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham M. Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30749–30768. PMLR, 2025a.

Seo Hyun Kim, Sunwoo Hong, Hojung Jung, Youngrok Park, and Se-Young Yun. KLASS: KL-guided fast inference in masked diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 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 (ICML)*, 2023.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Interspeech*, pp. 4521–4525, 2022. doi: 10.21437/Interspeech.2022-439.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4135.

Xiaohui Sun, Ruitong Xiao, Jianye Mo, Bowen Wu, Qun Yu, and Baoxun Wang. F5R-TTS: Improving flow matching based text-to-speech with group relative policy optimization. *arXiv preprint arXiv:2504.02407*, 2025.

Guanghan Wang, Yair Schiff, Subham Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. MaskGCT: Zero-shot text-to-speech with masked generative codec transformer. In *International Conference on Learning Representations (ICLR)*, 2025b.

Taehyung Yu and Seongjae Kang. Best-of-n TTS evaluation is confounded by ASR family alignment. *arXiv preprint arXiv:2607.08256*, 2026.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

A THRESHOLD SENSITIVITY

Content failure rate at four thresholds, all operating points and all four evaluators. Utterance-level cluster bootstrap, 4000 resamples, seed 20260729; one resample is shared across conditions so that differences are paired.

Thr.	Condition	Whisper-large-v3	HuBERT-large	wav2vec2-Conformer	Paraformer-en
5%	16 steps	19.54	44.31	52.29	32.66
	32 steps	18.90	43.49	52.94	32.84
	64 steps	18.26	40.55	50.28	29.54
	128 steps	17.25	39.63	49.27	27.52
	16 + one repair round	16.88	42.48	50.55	29.54
	16 + two repair rounds	16.24	42.48	50.37	29.08
	32 + one repair round	16.42	42.11	50.92	30.64
	32 + two repair rounds	15.78	41.74	50.73	30.37
10%	16 steps	8.72	25.05	32.66	15.23
	32 steps	8.26	24.86	32.11	16.61
	64 steps	8.44	21.38	30.09	13.85
	128 steps	7.25	21.56	28.17	12.66
	16 + one repair round	7.43	23.30	30.73	13.03
	16 + two repair rounds	6.88	23.30	30.28	13.03
	32 + one repair round	6.61	23.94	30.46	14.50
	32 + two repair rounds	6.33	23.58	29.82	14.04
20%	16 steps	2.02	9.36	14.31	3.76
	32 steps	1.38	8.72	13.58	4.86
	64 steps	0.92	7.16	11.93	3.58
	128 steps	1.38	8.26	11.38	2.66
	16 + one repair round	1.10	8.53	13.67	2.94
	16 + two repair rounds	0.83	8.35	13.39	2.94
	32 + one repair round	0.64	8.26	12.02	3.39
	32 + two repair rounds	0.46	8.17	11.65	3.39
50%	16 steps	0.18	0.73	0.64	0.18
	32 steps	0.09	0.92	0.64	0.28
	64 steps	0.28	0.73	0.92	0.00
	128 steps	0.18	0.18	0.37	0.18
	16 + one repair round	0.18	0.73	0.92	0.18
	16 + two repair rounds	0.09	0.55	0.92	0.18
	32 + one repair round	0.09	0.83	0.64	0.28
	32 + two repair rounds	0.09	0.73	0.64	0.37

Repair against its own base.

Thr.	Contrast	Whisper-large-v3	HuBERT-large	wav2vec2-Conformer	Paraformer-en
5%	16, one round	**+2.66**	**+1.83**	**+1.74**	**+3.12**
	16, two rounds	**+3.30**	**+1.83**	**+1.93**	**+3.58**
	32, one round	**+2.48**	**+1.38**	**+2.02**	**+2.20**
	32, two rounds	**+3.12**	**+1.74**	**+2.20**	**+2.48**
10%	16, one round	**+1.28**	**+1.74**	**+1.93**	**+2.20**
	16, two rounds	**+1.83**	**+1.74**	**+2.39**	**+2.20**
	32, one round	**+1.65**	**+0.92**	**+1.65**	**+2.11**
	32, two rounds	**+1.93**	**+1.28**	**+2.29**	**+2.57**
20%	16, one round	**+0.92**	**+0.83**	+0.64	**+0.83**
	16, two rounds	**+1.19**	**+1.01**	**+0.92**	+0.83
	32, one round	**+0.73**	+0.46	**+1.56**	**+1.47**
	32, two rounds	**+0.92**	+0.55	**+1.93**	**+1.47**
50%	16, one round	+0.00	+0.00	-0.28	+0.00
	16, two rounds	+0.09	+0.18	-0.28	+0.00
	32, one round	+0.00	+0.09	+0.00	+0.00
	32, two rounds	+0.00	+0.18	+0.00	-0.09

Positive values favour repair; bold entries exclude zero. The point estimate favours repair in 52 of 64 cells, and resolution degrades as the threshold rises: 16/16 and 16/16 cells exclude zero at 5% and 10%, 12/16 at 20%, and 0/16 at 50%.

The loss at the highest threshold is a counting limit rather than a reversal: at 50% the base failure rate falls to 0.09–0.92% depending on the evaluator, i.e. roughly one to ten utterances out of 1090, and every point estimate is within ± 0.2 percentage points of zero. We therefore state that the within-base repair conclusion holds at the 5% and 10% thresholds, holds for most evaluators at 20%, and is not resolved at 50%.

B HELD-OUT SPLIT OVERLAP

The hard case list shares 200 of its 400 prompt identifiers with the main Mandarin set; 160 carry a modified, harder target sentence and 40 are identical items. Because the two files were split independently, a small number of held-out items appear in the opposite development split.

Held-out split	Exact duplicates	Prompt-shared
Hard ($n = 320$)	7 (2.19%)	22
Mandarin ($n = 1616$)	8 (0.5%)	23

The 15-item exclusion list (SHA-256 prefix `f04cc2d0c9f2b526`) was frozen before held-out evaluation. Prompt-shared items with different target sentences are retained because they constitute a different task.

The English held-out split contains *no* excluded items: all 15 duplicates fall in the Mandarin and hard splits, so the overlap concerns those two only. We therefore report the Mandarin and hard results with and without the excluded items; the English results are unaffected by construction.

Split	Items	n	ΔCER (pp)	ΔCFR $> 10\%$ (pp)
Mandarin	all	1616	$+0.126$ [+0.075, +0.180]	$+0.681$ [+0.347, +1.015]
Mandarin	duplicates removed	1608	$+0.127$ [+0.075, +0.179]	$+0.684$ [+0.336, +1.020]
Hard	all	320	-0.562 [-0.895, -0.240]	-1.750 [-3.188, -0.375]
Hard	duplicates removed	313	-0.577 [-0.929, -0.252]	-1.789 [-3.259, -0.447]

Positive values favour repair. Removing the duplicates moves no point estimate by more than 0.02 pp and changes no sign or significance verdict. The Mandarin gain and the hard-set loss both replicate on held-out data.

C HELD-OUT EVALUATION

C.1 PROTOCOL

The held-out split was declared the primary result split before the evaluation was run (`frozen_policy.yaml`, section 0). All development-split choices—the signal–aggregation pair, the recalibrator family, and the probe specification—were frozen to `dev_selections_frozen.json` and *applied* to the held-out data; nothing was re-selected. The probe protocol necessarily differs between splits: on the development split it is evaluated by repeated grouped cross-validation, whereas on held-out data it is trained once on the full development split and evaluated once. The held-out protocol is the stricter of the two.

English held-out: $n = 4350$ (870 utterances $\times$ 5 seeds), content-failure prevalence 18.28%.

C.2 OPERATING POINTS

Area under the ROC curve is scale-free and therefore hard to read operationally. The table below fixes recall at 80% and reports what each detector costs at that recall.

Detector	AUPRC	Precision	Flagged
Confidence	0.240 [0.212, 0.275]	0.197	74.3%
Entropy	0.276 [0.244, 0.315]	0.199	73.7%
Degeneracy	0.278 [0.245, 0.316]	0.199	73.6%
Posterior likelihood	0.272 [0.241, 0.308]	0.205	71.5%
Learned probe	0.291 [0.259, 0.327]	0.213	68.6%
Round-trip detector	0.592 [0.547, 0.638]	0.563	26.0%

To reach 80% recall the decoder-derived signals must flag roughly three quarters of all utterances; the round-trip detector reaches the same recall while flagging about one quarter.

C.3 PAIRED DETECTOR DIFFERENCES

Differences are computed on the same bootstrap resamples, so they are paired and do not rely on comparing separate confidence intervals.

Contrast	ΔAUC
Round-trip − Confidence	$+\mathbf{0.290}$ [+0.255, +0.325]
Round-trip − Entropy	$+\mathbf{0.259}$ [+0.224, +0.294]
Round-trip − Degeneracy	$+\mathbf{0.257}$ [+0.222, +0.292]
Round-trip − Posterior likelihood	$+\mathbf{0.255}$ [+0.225, +0.283]
Round-trip − Learned probe	$+\mathbf{0.236}$ [+0.207, +0.263]

C.4 SECOND AND THIRD BASE CONDITIONS

The detection result is not specific to the 8-step base. Removing posterior-likelihood features from the held-out 8-step probe leaves AUC 0.621 [0.597, 0.646], versus 0.629 [0.604, 0.654] with all 20 features; their paired difference is $+0.007$ [-0.004, +0.019]. At a 16-step base the same 15 features reach 0.515 [0.481, 0.549]. The bases differ in failure prevalence, so this isolates the feature set rather than the cause of the change.

Detector (16-step base)	AUC
Degeneracy, s16	0.491 [0.440, 0.544]
Entropy, s16	0.487 [0.436, 0.541]
Confidence, s16	0.476 [0.428, 0.529]
Learned probe, s16 (15 feat.)	0.515 [0.481, 0.549]
Round-trip detector, s16	0.847 [0.814, 0.877]

All four decoder-derived intervals contain 0.5. Round-trip detection at $s16$ has AUPRC 0.469 $[0.400, 0.545]$ at failure prevalence 9.52% and exceeds the fifteen-feature probe by $+0.332$ [+0.287, +0.375] in paired AUC, compared with the corresponding $s8$ round-trip–probe difference of $+0.236$ [+0.207, +0.263]. Its AUC remains similar in magnitude to the $s8$ round-trip estimate. The larger paired gap is caused by the lower internal estimate, not by a higher external estimate. At 80% recall, precision is 0.288 at $s16$ versus 0.563 at $s8$; the comparison should be read with the corresponding prevalences (9.52% and 18.28%).

Held-out $s32$ detection. This post-hoc control was executed after unblinding with frozen detector definitions and no re-selection; the probe was trained once on the development split at 32 steps and applied once to held-out outputs.

Detector ($s32$, Whisper-large-v3 label)	AUC
Confidence	0.466 $[0.412, 0.520]$
Entropy	0.490 $[0.432, 0.548]$
Degeneracy	0.495 $[0.437, 0.555]$
Posterior likelihood, mean-pooled	0.540 $[0.490, 0.590]$
Flexible nonlinear recalibrator	0.541 $[0.498, 0.585]$
Learned probe (20 feat.)	0.591 $[0.551, 0.632]$
Round-trip detector	0.832 $[0.795, 0.868]$

Detection-label family control. The table reports the learned probe and round-trip detector under four failure labels. Paired differences are shown where the supplied canonical control reports a same-resample contrast; no interval is reconstructed from marginal AUCs.

Base	Label	Probe AUC	Round-trip AUC	Paired Δ
$s8$	Whisper	0.629 $[0.604, 0.654]$	0.864 $[0.845, 0.882]$	$+0.236$ [+0.207, +0.263]
$s8$	HuBERT	0.600 $[0.579, 0.621]$	0.724 $[0.705, 0.742]$	$+0.124$ [+0.099, +0.149]
$s8$	Conformer	0.597 $[0.577, 0.617]$	0.682 $[0.664, 0.700]$	$+0.085$ [+0.060, +0.109]
$s8$	Paraformer-en	0.629 $[0.608, 0.649]$	0.790 $[0.774, 0.806]$	$+0.162$ [+0.137, +0.187]
$s16$	Whisper	0.515 $[0.481, 0.549]$	0.847 $[0.814, 0.877]$	$+0.332$ [+0.287, +0.375]
$s16$	HuBERT	0.526 $[0.503, 0.548]$	0.659 $[0.636, 0.683]$	$+0.133$ [+0.101, +0.166]
$s16$	Conformer	0.534 $[0.514, 0.554]$	0.636 $[0.615, 0.658]$	$+0.102$ [+0.073, +0.131]
$s16$	Paraformer-en	0.534 $[0.508, 0.559]$	0.751 $[0.724, 0.778]$	$+0.217$ [+0.179, +0.256]
$s32$	Whisper	0.591 $[0.551, 0.632]$	0.832 $[0.795, 0.868]$	$+0.240$ [+0.191, +0.289]
$s32$	HuBERT	0.521 $[0.496, 0.547]$	0.632 $[0.607, 0.658]$	$+0.111$ [+0.076, +0.146]
$s32$	Conformer	0.563 $[0.539, 0.586]$	0.615 $[0.593, 0.638]$	$+0.052$ [+0.020, +0.084]
$s32$	Paraformer-en	0.559 $[0.528, 0.592]$	0.732 $[0.703, 0.760]$	$+0.173$ [+0.132, +0.213]

Prevalence differs by label ($s32$: 8.5% under Whisper-large-v3 and 26.6% under HuBERT-large), reflecting the evaluators' reference-recording error floors (1.7% versus 5.7%). The Whisper-label advantage therefore combines a same-family effect with a lower label-noise floor; the two are not separable in this design. The $s16$ paired intervals were added from the 2026-09-22 control regeneration.

Verifier-family and verifier-size controls. Post-hoc; executed after unblinding; frozen definitions; no re-selection. The round-trip detector is recomputed with two alternative verifiers on the same rendered held-out audio: data2vec-audio-base-960h, from outside the Whisper family, and Whisper-medium, a larger member of it. Paired differences are on identical bootstrap resamples.[1]

[1]We define ASR family by fine-tuning data, objective, and output vocabulary. data2vec-audio-base-960h and HuBERT-large-ls960-ft share LibriSpeech-960h CTC character-level fine-tuning and are treated as one family for the HuBERT label; Whisper-small, Whisper-medium, and Whisper-large-v3 share Whisper's weakly supervised objective and are one family for the Whisper label. Our pre-registered family assignment used pre-training objective only and would have classified data2vec as cross-family for both labels; the result corrected that.

Base	Label	Verifier	Align.	Round-trip AUC	Δ vs. probe	Δ vs. Whisper-small
$s8$	Whisper	Whisper-small	yes	0.864 [0.845, 0.882]	+0.236 [+0.207, +0.263]	—
$s8$	Whisper	data2vec	no	0.782 [0.758, 0.805]	+0.154 [+0.123, +0.185]	−0.082 [-0.106, -0.058]
$s8$	Whisper	Whisper-medium	yes	0.888 [0.870, 0.905]	+0.259 [+0.229, +0.289]	+0.023 [+0.006, +0.041]
$s8$	HuBERT	Whisper-small	no	0.724 [0.705, 0.742]	+0.124 [+0.099, +0.149]	—
$s8$	HuBERT	data2vec	yes	0.861 [0.847, 0.874]	+0.261 [+0.238, +0.284]	+0.137 [+0.116, +0.158]
$s8$	HuBERT	Whisper-medium	no	0.701 [0.682, 0.720]	+0.101 [+0.076, +0.126]	−0.023 [-0.034, -0.011]
$s16$	Whisper	Whisper-small	yes	0.847 [0.814, 0.877]	+0.332 [+0.287, +0.375]	—
$s16$	Whisper	data2vec	no	0.728 [0.689, 0.767]	+0.216 [+0.166, +0.265]	−0.119 [-0.162, -0.076]
$s16$	Whisper	Whisper-medium	yes	0.871 [0.844, 0.895]	+0.355 [+0.310, +0.399]	+0.025 [-0.005, +0.054]
$s16$	HuBERT	Whisper-small	no	0.659 [0.636, 0.683]	+0.133 [+0.101, +0.166]	—
$s16$	HuBERT	data2vec	yes	0.835 [0.818, 0.852]	+0.305 [+0.278, +0.334]	+0.176 [+0.151, +0.202]
$s16$	HuBERT	Whisper-medium	no	0.634 [0.612, 0.658]	+0.108 [+0.076, +0.140]	−0.025 [-0.039, -0.010]
$s32$	Whisper	Whisper-small	yes	0.832 [0.795, 0.868]	+0.240 [+0.191, +0.289]	—
$s32$	Whisper	data2vec	no	0.708 [0.660, 0.753]	+0.116 [+0.057, +0.176]	−0.124 [-0.177, -0.073]
$s32$	Whisper	Whisper-medium	yes	0.875 [0.840, 0.906]	+0.283 [+0.234, +0.333]	+0.043 [+0.006, +0.079]
$s32$	HuBERT	Whisper-small	no	0.632 [0.607, 0.658]	+0.111 [+0.076, +0.146]	—
$s32$	HuBERT	data2vec	yes	0.840 [0.822, 0.858]	+0.319 [+0.288, +0.350]	+0.208 [+0.180, +0.236]
$s32$	HuBERT	Whisper-medium	no	0.615 [0.591, 0.641]	+0.094 [+0.059, +0.129]	−0.017 [-0.031, -0.002]

Switching the verifier out of the Whisper family lowers round-trip AUC under Whisper-large-v3 labels and raises it under HuBERT-large labels; verifier size changes it by at most about 0.04. Under every pairing the round-trip detector remains above the learned probe.

Held-out $s16$ posterior likelihood and 20-feature probe. Post-hoc; executed after unblinding; frozen definitions; no re-selection. Posterior likelihood was re-scored at $s16$ after the sealed evaluation and the 20-feature probe was trained once on the development split; the fifteen-feature values are retained for comparison. Paired Δ is round-trip minus the listed detector.

Detector ($s16$)	Whisper label	Paired Δ	HuBERT label	Paired Δ
Posterior likelihood, mean-pooled	0.553 [0.509, 0.600]	+0.293 [+0.243, +0.341]	0.525 [0.498, 0.553]	+0.134 [+0.098, +0.170]
Learned probe (20 feat.)	0.516 [0.482, 0.551]	+0.330 [+0.286, +0.374]	0.526 [0.504, 0.548]	+0.133 [+0.101, +0.165]
Learned probe (15 feat.)	0.515 [0.481, 0.549]	+0.332 [+0.287, +0.375]	0.526 [0.503, 0.548]	+0.133 [+0.101, +0.166]
Round-trip detector	0.847 [0.814, 0.877]	—	0.659 [0.636, 0.683]	—

Held-out verifier miss share. Post-hoc; executed after unblinding; frozen definitions; no re-selection. A failure is missed when the verifier assigns it zero error. Stochastic and intrinsic failures are typed by 5-seed agreement on the held-out split (16-seed typing on the development split is reported in Appendix D); intervals are Wilson.

Base	Label	Failures	Missed, all	Missed, stochastic (n)	Missed, intrinsic (n)
$s8$	Whisper	795	10.8% [8.8, 13.2]	12.3% [10.0, 15.1] (625)	5.3% [2.8, 9.8] (170)
$s8$	HuBERT	1798	35.4% [33.3, 37.7]	33.0% [30.4, 35.7] (1193)	40.2% [36.3, 44.1] (605)
$s16$	Whisper	414	17.6% [14.3, 21.6]	22.3% [17.7, 27.6] (269)	9.0% [5.3, 14.7] (145)
$s16$	HuBERT	1273	50.3% [47.5, 53.0]	48.7% [45.3, 52.2] (798)	52.8% [48.3, 57.3] (475)
$s32$	Whisper	371	22.1% [18.2, 26.6]	23.8% [18.5, 30.0] (206)	20.0% [14.6, 26.8] (165)
$s32$	HuBERT	1155	56.5% [53.6, 59.3]	55.7% [51.9, 59.5] (655)	57.4% [53.0, 61.7] (500)

Review-budget grid. Budget grid $\{5, 10, 20, 30, 50\}\%$ fixed on 2026-08-01 after unblinding; 20% is the reporting midpoint and was not chosen to match prevalence. Detection scores and labels are frozen. Values below use linear interpolation within tied score blocks at review-budget boundaries, matching the tie-handling rule used for the canonical held-out analysis.

Detector ($s8$, Whisper-large-v3 label), failures caught	5%	10%	20%	30%	50%
Round-trip detector	19.9%	37.1%	64.1%	82.7%	90.6%
Learned probe	10.7%	19.4%	33.8%	44.9%	64.5%

C.5 Single-run gains largely flatten beyond 32 steps

Evaluator	16	32	64	128	16→32	32→64	64→128
Whisper-large-v3	3.04	2.75	2.58	2.64	+0.289	+0.168	−0.055
HuBERT-large	7.91	7.29	7.13	7.01	+0.620	+0.166	+0.117
wav2vec2-Conformer	9.94	9.25	9.31	9.03	+0.696	−0.059	+0.274
Paraformer-en	4.95	4.43	4.15	4.17	+0.512	+0.287	−0.023

Word error rate (%) at each step count, then paired differences between adjacent settings (positive favours the larger step count). Bold marks intervals excluding zero. Beyond 32 steps only one of four evaluators resolves a further gain, and no evaluator resolves one between 64 and 128 steps.

C.6 HELD-OUT 16-CELL COMPARISON FAMILIES

The following tables report every cell in the three prespecified held-out families. Superscript $^{\checkmark}$ marks an interval excluding zero before correction; $^{\checkmark\checkmark}$ marks survival after Holm correction across all 16 cells in that table.

Table 2: Held-out repair versus its identical base. The multiplicity family is all 16 cells (four evaluators $\times$ two metrics $\times$ two operating points); 16/16 exclude zero before correction and 16/16 survive Holm.

Metric	Condition	Whisper-large-v3	HuBERT-large	wav2vec2-Conformer	Paraformer-en
WER	s16+R8×2	$+0.64^{\checkmark\checkmark}$ [+0.51, +0.79]	$+0.74^{\checkmark\checkmark}$ [+0.59, +0.90]	$+0.65^{\checkmark\checkmark}$ [+0.48, +0.83]	$+0.71^{\checkmark\checkmark}$ [+0.56, +0.85]
WER	s32+R8×2	$+0.41^{\checkmark\checkmark}$ [+0.29, +0.55]	$+0.39^{\checkmark\checkmark}$ [+0.26, +0.54]	$+0.43^{\checkmark\checkmark}$ [+0.27, +0.59]	$+0.46^{\checkmark\checkmark}$ [+0.33, +0.60]
CFR	s16+R8×2	$+2.34^{\checkmark\checkmark}$ [+1.79, +2.92]	$+2.46^{\checkmark\checkmark}$ [+1.86, +3.08]	$+2.41^{\checkmark\checkmark}$ [+1.77, +3.08]	$+2.78^{\checkmark\checkmark}$ [+2.11, +3.43]
CFR	s32+R8×2	$+1.38^{\checkmark\checkmark}$ [+0.92, +1.86]	$+1.20^{\checkmark\checkmark}$ [+0.74, +1.66]	$+1.45^{\checkmark\checkmark}$ [+0.90, +2.02]	$+1.82^{\checkmark\checkmark}$ [+1.31, +2.34]

Table 3: Held-out repair versus $s64$. The multiplicity family is all 16 cells; 3/16 exclude zero before correction and 1/16 survives Holm. The sole survivor is wav2vec2-Conformer WER at s32+R8×2 ($p_{\mathrm{Holm}} = 0.024$).

Metric	Condition	Whisper-large-v3	HuBERT-large	wav2vec2-Conformer	Paraformer-en
WER	s16+R8×2	$+0.19^{\checkmark}$ [+0.03, +0.34]	-0.04 [-0.32, +0.24]	$+0.02$ [-0.32, +0.34]	-0.09 [-0.33, +0.14]
WER	s32+R8×2	$+0.25^{\checkmark}$ [+0.06, +0.45]	$+0.23$ [-0.02, +0.49]	$+0.49^{\checkmark\checkmark}$ [+0.18, +0.80]	$+0.17$ [-0.05, +0.39]
CFR	s16+R8×2	$+0.41$ [-0.32, +1.10]	-0.92 [-2.11, +0.28]	-0.46 [-1.91, +0.99]	-0.67 [-1.84, +0.46]
CFR	s32+R8×2	$+0.44$ [-0.25, +1.15]	$+0.53$ [-0.60, +1.70]	$+0.69$ [-0.83, +2.18]	$+0.53$ [-0.53, +1.56]

Table 4: Held-out repair versus $s128$. The multiplicity family is all 16 cells; 4/16 exclude zero before correction and 3/16 survive Holm. All three survivors are Whisper-large-v3 cells: WER at s32+R8×2 and CFR at both operating points.

Metric	Condition	Whisper-large-v3	HuBERT-large	wav2vec2-Conformer	Paraformer-en
WER	s16+R8×2	$+0.24^{\checkmark}$ [+0.07, +0.41]	-0.16 [-0.43, +0.11]	-0.26 [-0.59, +0.08]	-0.07 [-0.30, +0.17]
WER	s32+R8×2	$+0.30^{\checkmark\checkmark}$ [+0.13, +0.48]	$+0.11$ [-0.13, +0.36]	$+0.21$ [-0.12, +0.55]	$+0.19$ [-0.03, +0.42]
CFR	s16+R8×2	$+1.10^{\checkmark\checkmark}$ [+0.37, +1.84]	-0.69 [-1.98, +0.53]	-0.97 [-2.41, +0.48]	-0.25 [-1.38, +0.90]
CFR	s32+R8×2	$+1.13^{\checkmark\checkmark}$ [+0.39, +1.89]	$+0.76$ [-0.41, +1.93]	$+0.18$ [-1.31, +1.68]	$+0.94$ [-0.09, +1.98]

C.7 EQUIVALENT STEPS AT MATCHED CFR (WER $> 10\%$)

Evaluator	Operating point	CFR-equivalent steps	P(cheaper)
Whisper-large-v3	s16+R8×2	≥ 128 [53, ≥ 128]	1.00
Whisper-large-v3	s32+R8×2	≥ 128 [57, ≥ 128]	1.00
HuBERT-large	s16+R8×2	30.0 [24, ≥ 128]	0.30
HuBERT-large	s32+R8×2	≥ 128 [50, ≥ 128]	0.98
wav2vec2-Conformer	s16+R8×2	42.0 [25, ≥ 128]	0.61
wav2vec2-Conformer	s32+R8×2	≥ 128 [49, ≥ 128]	0.97
Paraformer-en	s16+R8×2	44.7 [28, ≥ 128]	0.82
Paraformer-en	s32+R8×2	≥ 128 [54, ≥ 128]	1.00

Paired inversion and interpolation. For each bootstrap replicate we resample utterances once and recompute both the single-run step–failure curve and the method's failure rate on that same resample, so the target is paired rather than fixed at the point estimate. The failure measure is the content failure rate at a word error rate above 10%. Between adjacent step counts we interpolate linearly in failure rate and geometrically in step count, that is, linearly in $\log$ steps; no spline or monotone smoother is applied. Curves are used as observed and are not monotonicised, which matters here because the measured curve is non-monotone in 40–97% of replicates depending on the evaluator. Where the target line crosses the curve more than once (6–27% of replicates across the two operating points) we take the first crossing in increasing step order, that is, the smallest equivalent step count. Replicates whose target is not reached within the measured range are right-censored at 128 steps and retained in the denominator; they are never dropped and never treated as infinite. P(cheaper) is the fraction of replicates in which the modelled cost of the CFR-equivalent single run exceeds the measured cost of the method, $N(c_0 + c_1 e) > C_{\text{method}}$, which for the 32-step operating point reduces to $e > 49.5$ steps. Censored replicates satisfy this inequality because the capped 128-step single run already costs more than the method.

Implementation sketch.

```
for b = 1,...,B:
    S <- resample utterances once with replacement
    curve <- [(16,CFR(s16,S)), (32,CFR(s32,S)),
              (64,CFR(s64,S)), (128,CFR(s128,S))]
    target <- CFR(method,S)                  # paired target
    if target >= curve[1].failure:
        e <- 16; censored <- false
    else if target < min(curve.failure):
        e <- 128; censored <- true
    else:
        i <- first adjacent pair that crosses target
        if no such i:
            e <- 128; censored <- true
        else:
            f <- (curve[i].failure - target) /
                 (curve[i].failure - curve[i+1].failure)
            e <- exp(log(step[i]) +
                     f*(log(step[i+1]) - log(step[i])))
    cheaper[b] <- (N*(c0 + c1*e) > C_method)
P(cheaper) <- mean(cheaper)
```

The left-bound branch returning 16 steps is part of the implementation but is not taken by any replicate in these data. The affine single-run cost model was calibrated on the 218-utterance development set; method costs use measured component accounting, including verifier calls. Both sides scale linearly with utterance count, so the comparison event is invariant to that normalization. The asymmetry between the modelled single-run cost and the measured method cost is examined in Appendix L, where refitting the cost constants changes P(cheaper) by at most 0.01.

Observed inversion branches. The following audit summarizes the 32-step operating point. "Non-monotone" counts replicates whose curve rises somewhere as the step count increases; "multiple crossings" counts replicates where the target meets the curve more than once.

Evaluator	Interpolated	Censored	Non-monotone	Multiple crossings	P(cheaper)
Whisper-large-v3	11.3%	88.7%	97.0%	11.1%	0.9988
HuBERT-large	24.1%	75.9%	73.1%	14.0%	0.9775
wav2vec2-Conformer	46.0%	54.0%	39.7%	6.2%	0.9695
Paraformer-en	18.1%	81.8%	76.5%	14.4%	0.9978

The first-crossing rule is conservative for the cost claim: choosing the smallest equivalent step gives the cheapest compatible single run and therefore makes the method harder to prefer. Between 54 and 89% of replicates are right-censored, meaning that a 128-step single run does not reach the method's failure rate at all. Those replicates remain in both numerator and denominator. Because the true equivalent step is then at least 128 and a 128-step run already costs more than the method, their contribution does not depend on interpolation beyond the measured range. The minimum $s32$ value across evaluators is 0.9695. A separate audit script, `src/eval/equivstep_audit.py`, recomputes the inversion from the stored inputs and reproduces all eight stored P(cheaper) values to four decimal places.

C.8 Does the acoustic quality change account for the error-rate change?

Between 32 and 128 steps UTMOS improves by $+0.0106$ [+0.0029, +0.0179]. The interval excludes zero, but the change is 0.023 standard deviations of the between-utterance UTMOS distribution ($\sigma = 0.457$).

For that improvement to account for the concurrent error-rate reduction, UTMOS would have to predict word error rate far more strongly than it does. We use the cross-sectional association between word error rate and UTMOS within a fixed condition as a deliberately generous sensitivity estimate. It is not a formal causal bound: difficult utterances can score poorly on both measures for shared reasons, while other confounding or measurement error could act differently.

Evaluator	Required |slope|	Cross-sectional slope	Explained in calculation
Whisper-large-v3	10.7	-1.08 [-1.86, -0.33]	17.4%
HuBERT-large	26.8	-3.35 [-4.83, -1.84]	18.1%
wav2vec2-Conformer	20.3	-3.26 [-4.92, -1.63]	24.2%
Paraformer-en	25.0	-2.27 [-3.27, -1.28]	13.1%

Required values are slope magnitudes in percentage points of word error rate per UTMOS point; the cross-sectional slopes are reported with their negative sign. Granting this deliberately generous calculation in full, the UTMOS shift explains 13–24% of the concurrent error-rate reduction under the calculation.

A second check does not rely on that regression at all. If the acoustic account held, the utterances whose content failures are resolved over this range should gain more UTMOS than the rest. They do not.

Evaluator	Resolved	ΔUTMOS difference
Whisper-large-v3	14	$+0.0009$ [-0.0829, +0.0746]
HuBERT-large	37	$+0.0119$ [-0.0205, +0.0495]
wav2vec2-Conformer	63	-0.0004 [-0.0276, +0.0261]
Paraformer-en	28	$+0.0500$ [-0.0041, +0.1055]

Difference between the resolved utterances and all others. No interval excludes zero.

C.9 Flag and acceptance rates

Condition	Flagged (% all)	Accepted (% flagged)	Accepted (n)	Rejected (n)
16 steps, round 1	29.31%	42.12%	537	738
16 steps, round 2	20.25%	16.57%	146	735
32 steps, round 1	25.86%	37.69%	424	701
32 steps, round 2	18.94%	13.83%	114	710

Acceptance rate is the share of flagged utterances whose regeneration was kept. It falls by more than half from the first round to the second, which is why we stop at two rounds.

C.10 PRESPECIFIED HELD-OUT LOCALIZATION ABLATION

P2 compares five first-round interventions on the same 870 held-out English utterances at five seeds. A repairs the ASR-identified span; B repairs a random span with exactly the same mask-union size; C repairs a span selected from the decoder's lowest-confidence region; D1 masks the full utterance for $R = 8$; and D2 masks the full utterance for the base step count, $R = 32$. A, B, C, and D1 use the same number of forward passes because mask size does not change the forward count. The preregistration was fixed before this run (SHA-256 prefix `8e08ee3595cec350`), after the primary held-out results had been unblinded.

Descriptively, all 24 equal-forward-pass directions across B, C, and D1 favor A. The prespecified 24-cell Holm family instead comprises A-versus-B, A-versus-C, and A-versus-D2 (eight evaluator–metric cells each); D1 is an uncorrected equal-forward-pass reference outside this family. For A versus B, all eight evaluator–metric contrasts exclude zero before correction and six survive Holm correction. For A versus C, all eight favor A and five survive. The preregistered primary cell, HuBERT-large CFR for B minus A, reaches $p_{\mathrm{Holm}} = 0.055$ and therefore does not meet the threshold fixed in advance. We report localization as supported but not established by that criterion.

All eight D1 minus A contrasts favor A and six exclude zero before correction. D2 is a deliberately stronger reference: full-mask regeneration at $R = 32$ beats A in three of eight contrasts after correction, while costing 1002 s versus 585–729 s for the span-repair conditions, or 1.37–1.71× as much. Across the eight evaluator–metric cells, A recovers 57.7%–84.8% of D2's point-estimate improvement over the same base.

Condition	Accepted (% flagged)	Accepted (n)	Rejected (n)
A: ASR span	37.69%	424	701
B: random, union-matched	29.16%	328	797
C: decoder low-confidence	29.24%	329	796
D1: full mask, $R = 8$	28.98%	326	799
D2: full mask, $R = 32$	41.78%	470	655

Localization also changes the verifier's acceptance decision: A is accepted 8.5 *pp* more often than B, C, or D1 despite the matched forward-pass budget. This verifier-aligned rate is a mechanism diagnostic rather than independent evidence of cross-evaluator quality.

An earlier random-span control matched summed span length rather than the size of the resulting mask union, which agreed exactly for 98.2% of development utterances. P2 instead matches mask-union size. The equality holds for all 1125 flagged held-out outputs, with zero maximum difference, and an assertion on this equality remained active throughout the run. This correction explains why the earlier development repetitions were not used to adjudicate localization.

C.11 HELD-OUT HARD-SET TRANSFER RESULT

The frozen policy worsens held-out hard-set CER by 0.577 *pp* [-0.929, -0.252] after duplicate removal. The verifier flags 94.81% of utterances in round 1 (mean verifier error 0.795); round-2 flag and acceptance rates are 42.81% and 33.43%. The loss is concentrated in the top quartile of verifier error:

Quartile	Mean verifier error	ΔCER (pp)	95% CI
Q1	0.071	+0.084	[-0.014, +0.225]
Q2	0.181	+0.043	[-0.158, +0.240]
Q3	0.354	−0.032	[-0.489, +0.455]
Q4	2.575	−2.342	[-3.484, -1.210]

Only Q4 excludes zero; the other quartile intervals include zero. This post-unblinding descriptive analysis shows that the observed loss is concentrated in the extreme-verifier-error stratum and sup-

ports verifier mismatch as a plausible mechanism. The extreme flag rate could trigger a fallback, but calibrating such a rule requires independent target domains.

C.12 Full standard set

Development and held-out splits together comprise the complete English Seed-TTS-Eval set, allowing direct comparison with published numbers.

Evaluator	16	32	64	128	16+R	32+R
Whisper-large-v3	2.98	2.68	2.56	2.57	2.34	2.25
HuBERT-large	7.79	7.27	6.93	6.85	7.08	6.87
wav2vec2-Conformer	9.77	9.22	9.15	8.83	9.16	8.74
Paraformer-en	4.85	4.51	4.11	4.06	4.16	4.03

Word error rate (%) on $n = 1088$ utterances, five seeds. +R denotes two repair rounds.

D Development-split decoder-derived diagnostic details

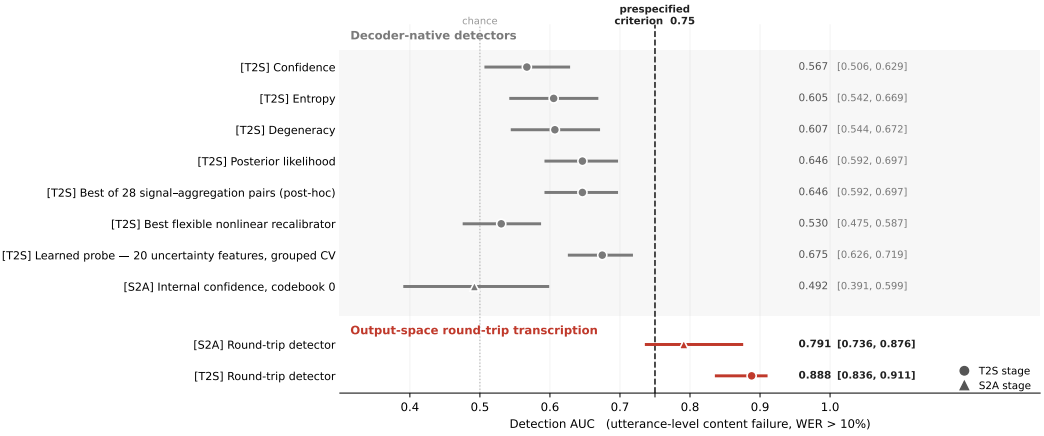


Figure 5: Development-split detection results. Unlike the held-out panel in Figure 2, this diagnostic figure includes the optimistic post-hoc full-sample sweep over 28 signal–aggregation pairs. Development values are retained for auditability and are not used as the main result.

Learned readout. The canonical grouped-CV probe uses decoder-uncertainty features only. It reaches 0.675 $[0.626, 0.719]$ with 20 features at 8 steps and 0.577 $[0.497, 0.655]$ with 15 effective features at 16 steps, where posterior likelihood was not logged. Adding utterance length changes the 8-step result by $+0.010$ [-0.004, +0.025]. At matched recall, probe flag rates are 62.2% and 67.8% at 8 and 16 steps, compared with 29.2% and 24.8% for round-trip transcription.

Calibration and redundancy. Confidence, entropy, and degeneracy reach pairwise correlations up to $|r| = 0.995$, with VIF 240–412. Confidence quintiles have non-monotone failure rates of 28.4%, 18.3%, 10.1%, 20.2%, and 14.7%. Binned, spline, decision-tree, gradient-boosted, and MLP recalibrators do not improve the raw ranking.

Aggregation and local effects. Seven pooling rules are applied to eligible signals using disjoint selection and reporting partitions; the full-sample optimistic sweep is shown only as a diagnostic upper estimate. ASR-localized error spans have lower confidence and higher degeneracy, but standardized effects are below 0.1 and do not compose into reliable utterance decisions.

S2A and commitment order. S2A codebook 0 replicates the internal–external gap. Confidence-first commitment does not significantly outperform random ordering, and increasing the iterative budget does not yield a resolved gain.

E CIRCULARITY AND CROSS-FAMILY CONTROLS

The verifier and the primary evaluator are different models, but both transcribe audio, so we report four controls that bound how much of the measured gain could come from that shared modality. Detection runs timestamp-free: timestamp-free and timestamped modes yield identical transcripts, verifier scores, and flag decisions in the audit (Pearson and Spearman 1.000; Jaccard 1.000), while timestamp-free detection reduces cost from 174 s to 66.9 s per 218 utterances.

Verifier-as-evaluator inflation. Running the pipeline with the verifier set to the evaluator itself (Whisper-large-v3, i.e.
verifier $=$ evaluator) raises the measured gain from $+1.983$ to $+2.854$ [+2.41, +3.32] percentage points — an inflation of $+0.87$ points, a factor of 1.44. All figures in the main text use the honest configuration, in which the verifier is a smaller model from a different checkpoint.

Third-evaluator control. We repeat the measurement with wav2vec2-base (CTC head), a model that is *not* part of the four-evaluator panel and shares no pretraining lineage with Whisper. It gives $+2.12$ [+1.64, +2.62] percentage points, agreeing in direction and magnitude with the $+1.98$ measured by Whisper-large-v3, with consistent sign across all five seeds. This is a different model from the panel's wav2vec2-Conformer.

Reference-controlled recovery. Absolute gains are not comparable across evaluators with different floors, so we express each gain as the fraction of the headroom above that evaluator's error on the original recordings.

Evaluator	Reference	Base	Headroom	Recovered
Whisper-large-v3	1.787	5.321	3.534	56.1%
wav2vec2-base	9.481	16.151	6.670	31.8%

Note. Reference and base rates in this table come from the original seed-tts-eval scoring path, whose text normalisation differs slightly from the panel above (Whisper 1.787 versus 1.727). The two paths were cross-checked and agree to within 0.2 points; we keep each table internally consistent rather than mixing them.

The naive relative improvement differs by a factor of three between these two evaluators (37.3% versus 13.1%); after controlling for the reference floor the ratio falls to 1.76. Much of the apparent dilution is therefore a property of the recognizer rather than of the method — but not all of it, and we report the residual difference rather than treating the control as removing it.

Mandarin cross-family detection. This measurement uses an $s8$ base with a flag rate of 83.9%, where base failure mass is far higher than at the $s32$ operating point reported in the main text. Pairing the multilingual Whisper-small verifier, with thresholds calibrated separately by language ($\mathrm{sel_wer} > 0$ for English and $\mathrm{sel_wer} > 0.20$ for Mandarin), with a Mandarin evaluator of a different lineage and architecture variant (Paraformer-zh, a SeACo-Paraformer; the panel's Paraformer-en is a distinct English model) removes any shared-family explanation. On 404 utterances $\times$ 5 seeds, CER improves from 5.329% to 3.245%, a gain of $+2.084$ [+1.807, +2.382] percentage points. Its larger magnitude than the held-out $s32$ gain is consistent with the dose–response pattern in Appendix H; this control establishes lineage robustness, not Mandarin cost efficiency.

F DEVELOPMENT HARD-SET DIAGNOSIS

The Seed-TTS-Eval hard set is a separate Mandarin benchmark of tongue twisters and repetition-inflated variants, not a stratum of the English data. We evaluate 80 held-in utterances $\times$ 5 seeds at a base of 32 steps.

F.1 BASE-TO-REPAIR, FIVE FAILURE THRESHOLDS

Metric	Δ (pp)	95% CI
CER	-0.051	[-0.424, +0.274]
CFR $> 5\%$	+0.500	[-1.250, +2.250]
CFR $> 10\%$	**+1.500**	[+0.500, +2.750]
CFR $> 20\%$	-0.500	[-2.000, +1.000]
CFR $> 50\%$	-0.250	[-0.750, +0.000]

Positive values favour repair. Only the 10% threshold excludes zero; because five thresholds are examined, we do not report this single crossing as an isolated finding.

On the held-out hard split ($n = 320$) the method is significantly *worse* than its base at both character error rate and content failure rate; those numbers appear in Appendix B alongside the with-and-without-duplicates comparison, and we report them once rather than in both places.

F.2 WHICH RECOGNIZER, AND HOW LARGE?

The hard set provides no ground-truth recording of the target sentence, so we cannot measure how accurate a recognizer is on this domain independently of our synthesis. We instead transcribe identical audio with three recognizers. The per-utterance distribution is heavily right-skewed (skewness 7.2 for the verifier; 22 of 400 transcriptions exceed 100% CER through runaway insertion), so we report medians for the typical case and means for the tail.

Recognizer	Median	95% CI	Mean	95% CI
Whisper-small (verifier)	16.00%	[12.8, 18.9]	57.95%	[30.1, 95.7]
Whisper-large-v3 (independent probe)	8.34%	[5.0, 11.1]	30.54%	[17.1, 48.2]
Paraformer-zh (Mandarin evaluator)	3.12%	[2.0, 4.5]	7.15%	[4.9, 9.8]

The audio is transcribable: the evaluator reaches 3.1% median CER on the same files, so the verifier's 16.0% reflects the transcriber rather than synthesis quality.

Both capacity and family contribute. On paired resamples the verifier exceeds a larger model of the same family by $+7.66$ [+4.73, +11.60] points of median CER, that model exceeds the evaluator by $+5.21$ [+1.83, +7.55], and the verifier exceeds the evaluator by $+12.88$ [+10.05, +15.35]; all three differences exclude zero. On means the first comparison is not resolved ($+27.4$ [-1.7, +64.9]) because the tail dominates the estimate, which is why we draw no conclusion from overlap between mean intervals here.

Among our evaluators only Paraformer-zh can transcribe this set; HuBERT, wav2vec2-Conformer, and Paraformer-en are English-only.

G DEVELOPMENT-SPLIT REPAIR ABLATIONS AND QUALITY PROXIES

Local repair is cheaper than selective global regeneration in all four matched conditions. One of four unadjusted quality comparisons favors local repair, but none survives Holm correction; no quality dominance is claimed. Earlier development random-span estimates were $+0.112\,pp$ [-0.032, +0.257] and $+0.297\,pp$ [+0.136, +0.492]. That implementation matched summed span length rather than mask-union size, so these estimates motivated rather than adjudicated the prespecified held-out P2 ablation in Appendix C.

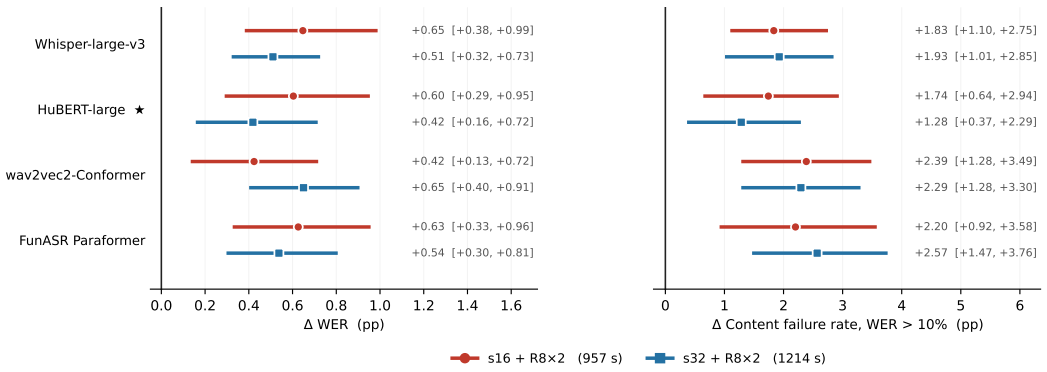


Figure 6: Development-split within-base repair effects, retained for comparison with the held-out main result in Figure 4. The x-axis convention matches the held-out rendering; numerical values are the frozen development estimates.

H Repair gain versus base failure mass

Series	Lang	Base	Base CFR (%)	Gain (pp)	95% CI
R8x1	en	s8	18.07	+2.261	[+1.762, +2.816]
R8x1	zh	s8	17.67	+2.084	[+1.799, +2.376]
R8x1	en	s16	8.72	+0.443	[+0.240, +0.702]
R8x1	en	s32	8.26	+0.414	[+0.250, +0.613]
R8x2	en	s8	18.07	+2.413	[+1.893, +2.988]
R8x2	en	s16	8.72	+0.647	[+0.386, +0.975]
R8x2	en	s32	8.26	+0.510	[+0.320, +0.727]
R8x2	zh	s32	4.65	+0.120	[+0.015, +0.219]

Fitting the English points in log–log space gives an exponent of 1.91 $[1.21, 2.83]$ for two repair rounds. The interval is wide because the fit rests on three points; a high R^2 here reflects the small number of points rather than a precisely determined exponent. We report the pattern, not the exponent, and only within the measured range of base failure rates (4.65–18.07%): extrapolating to the hard subset (22.75%) over-predicts the observed gain by roughly ten-fold.

I Cross-strategy comparisons and diagnostics

I.1 Held-out comparison with additional denoising

Beyond 32 steps, additional denoising yields little further improvement on the held-out evaluator panel. Only one of the four $s32$-to-$s64$ comparisons has an interval that excludes zero, and none of the four $s64$-to-$s128$ comparisons does. The direct comparison with $s128$ is less decisive than the within-base repair result: at the $s32$ operating point all four held-out WER estimates favor repair, but only the Whisper-large-v3 cell survives table-wise Holm correction ($+0.30\,pp$, [+0.13, +0.48]).

The development-split comparison favored additional denoising under three of four evaluators; the held-out comparison does not reproduce that pattern. The development comparison has one quarter of the sample, and the held-out curve shows little resolved improvement beyond 32 steps, with no evaluator resolving an $s64$-to-$s128$ gain. We report the held-out result as primary and retain the development tables as diagnostics.

This appendix reports direct comparisons with $s64$ and $s128$ and the matched-cost interpolation. In every table, Δ = single run $-$ repair, so positive values favour repair. Each table defines one multiplicity family of 16 cells (four evaluators $\times$ two operating points $\times$ two metrics). A single mark ($^{\checkmark}$) means that the unadjusted 95% interval excludes zero; a double mark ($^{\checkmark\checkmark}$) means that the cell also survives Holm correction across all 16 cells.

Table 5: Held-out direct WER comparison of s32+R8×2 with $s128$. Δ = s128 − repair, so positive values favor repair. Across the full family of 16 WER/CFR comparisons, all three Holm survivors are Whisper-large-v3 cells: WER at the $s32$ operating point and CFR at both operating points. No cell from the other three evaluators survives; the full family appears in Table 4 in Appendix C.

Evaluator	Δ WER (pp)	95% CI
Whisper-large-v3	$+0.30$	[+0.13, +0.48]
HuBERT-large	$+0.11$	[-0.13, +0.36]
wav2vec2-Conformer	$+0.21$	[-0.12, +0.55]
Paraformer-en (FunASR)	$+0.19$	[-0.03, +0.42]

Table 6: Repair versus $s64$ (Table B). The single run is 52% more expensive than the $s16$ method and 20% more expensive than the $s32$ method.

Metric	Method	Whisper-large-v3	HuBERT-large	wav2vec2-Conformer	Paraformer-en
WER	s16+R8×2	$+0.37$ [-0.11, +1.01]	$-0.54^{\checkmark}$ [-1.12, -0.02]	-0.10 [-0.68, +0.53]	$+0.13$ [-0.26, +0.55]
WER	s32+R8×2	$+0.57^{\checkmark\checkmark}$ [+0.25, +1.00]	$-0.61^{\checkmark}$ [-1.27, -0.04]	$+0.10$ [-0.58, +0.77]	-0.30 [-0.85, +0.22]
CFR	s16+R8×2	$+1.56^{\checkmark}$ [+0.18, +3.03]	-1.93 [-4.31, +0.46]	-0.18 [-2.84, +2.39]	$+0.83$ [-1.01, +2.66]
CFR	s32+R8×2	$+2.11^{\checkmark\checkmark}$ [+0.73, +3.58]	-2.20 [-4.59, +0.18]	$+0.28$ [-2.29, +2.75]	-0.18 [-2.11, +1.83]

On the primary HuBERT evaluator, both mean-WER cells are nominally negative (-0.54 and -0.61 points) but do not survive table-wise correction ($p_{\text{Holm}} = 0.52$ and 0.44). The two surviving cells favour repair under Whisper at the $s32$ operating point; no evaluator-independent dominance is supported.

Table 7: Repair versus $s128$ (Table D). The single run is 163% and 107% more expensive than the two method operating points.

Metric	Method	Whisper-large-v3	HuBERT-large	wav2vec2-Conformer	Paraformer-en
WER	s16+R8×2	$+0.20$ [-0.22, +0.59]	-0.47 [-1.00, +0.08]	-0.60 [-1.24, +0.09]	-0.20 [-0.62, +0.22]
WER	s32+R8×2	$+0.40^{\checkmark}$ [+0.09, +0.74]	-0.55 [-1.24, +0.08]	-0.40 [-1.05, +0.23]	$-0.63^{\checkmark\checkmark}$ [-1.14, -0.19]
CFR	s16+R8×2	$+0.37$ [-1.38, +2.20]	-1.74 [-4.22, +0.83]	-2.11 [-4.68, +0.37]	-0.37 [-2.29, +1.56]
CFR	s32+R8×2	$+0.92$ [-0.28, +2.20]	-2.02 [-4.40, +0.28]	-1.65 [-4.22, +0.83]	-1.38 [-3.21, +0.37]

Within the development split, the point estimates have the same sign at both operating points: Whisper favours repair and the other three evaluators favour additional denoising. This development-only pattern is retained as a diagnostic and is not used for held-out claims. Under the 16-cell Holm family, only the Paraformer-en $s32$ WER cell survives; the Whisper $s32$ WER cell is nominally significant but does not survive ($p_{\text{Holm}} = 0.16$).

Only four of the sixteen interpolated cells survive table-wise correction, and they are split across evaluators and operating points. These comparisons are therefore secondary to direct within-base repair effects.

On the development split, UTMOS changes by 0.0019 [-0.0121, +0.0160] from $s32$ to $s128$. Repair changes it by 0.0056 [+0.0009, +0.0105]. These measurements do not isolate verifier–evaluator coupling from differences in output properties.

Table 8: Repair versus a matched-cost interpolated single run (Table C). Unlike Tables B and D, the comparator is not directly measured.

Metric	Method	Whisper-large-v3	HuBERT-large	wav2vec2-Conformer	Paraformer-en
WER	s16+R8×2	$+0.31$ [-0.17, +0.83]	$+0.40$ [-0.07, +0.90]	$+0.40$ [-0.29, +1.12]	$+0.90^{\checkmark\checkmark}$ [+0.43, +1.40]
WER	s32+R8×2	$+0.55^{\checkmark\checkmark}$ [+0.31, +0.87]	-0.23 [-0.67, +0.17]	$+0.31$ [-0.16, +0.74]	$+0.01$ [-0.37, +0.37]
CFR	s16+R8×2	$+1.39$ [-0.56, +3.30]	$+1.25$ [-0.89, +3.37]	$+1.66$ [-1.00, +4.36]	$+3.33^{\checkmark\checkmark}$ [+1.43, +5.29]
CFR	s32+R8×2	$+2.04^{\checkmark\checkmark}$ [+1.08, +3.10]	-0.91 [-2.52, +0.68]	$+1.02$ [-0.72, +2.73]	$+0.84$ [-0.51, +2.21]

J ORDERING SEARCH AND ORACLE CONTROLS

Our question is downstream of the unmasking order, but a reader may reasonably ask whether a better order exists that we failed to find. We report three controls, each closing a different form of that question.

Is there headroom in the order itself? We compare best-of-16 under two regimes: varying the unmasking order, and holding the order fixed while varying only the sampling seed. If order mattered, the first should beat the second.

	8 steps		16 steps	
	single	best-of-16	single	best-of-16
Random order ×16	5.233	0.935	2.800	0.838
Fixed order, seed ×16	5.140	0.742	2.648	0.854

Search recovers a large amount in both regimes (4.30 and 4.40 percentage points at 8 steps), but almost none of it is attributable to the order: the order-specific headroom is -0.10 points at 8 steps and $+0.17$ at 16. Comparing the two best-of-16 results directly gives -0.19 [-0.44, +0.01] and $+0.02$ [-0.13, +0.17]; neither excludes zero.

Do the proposed heuristics help? We evaluate five commitment rules at a matched step budget against the confidence-with-Gumbel-noise default.

Rule	WER_8	Δ	WER_{16}	Δ
Confidence + Gumbel (default)	4.893	—	2.708	—
Confidence, greedy	5.531	-0.64	2.523	$+0.19$
Random	5.214	-0.32	2.840	-0.13
Conditional-logit confidence	5.282	-0.39	2.867	-0.16
Margin	9.001	-4.11	2.708	-0.00
Entropy	9.649	-4.76	3.094	-0.39
Left-to-right	108.9	-104	74.6	-71.8

Note. This ordering comparison is a separate generation run from the detection experiments; the same confidence + Gumbel rule gives 4.893 WER here versus 5.321 in the detection/recovery run at $s8$. We therefore interpret the deltas only within this ordering run.

Negative values are worse than the default. Random ordering is not significantly worse than confidence-first at either budget. Margin and entropy, which help in text, roughly double the error rate at 8 steps. Left-to-right decoding collapses into runaway insertion. Order selection can therefore matter a great deal, but in our setting only in the harmful direction.

Would knowing the correct order help? An oracle that selects, at each step, the position whose prediction matches the ground-truth token recovers 0.27–0.33% of tokens at 8 steps against 0.13–0.16% for confidence-first and 0.12–0.18% for random ordering. The absolute level is what matters here: the oracle criterion is itself close to uninformative.

The reason is that exact token recovery is not an achievable target in this representation. Ground-truth semantic tokens are almost all distinct within an utterance (for example 118 distinct values across 119 frames). Even on the model's own training task at low masking ratios, top-1 accuracy stays in the single digits:

Masked	top-1	top-10	top-50	top-200
10%	10.6	40.2	73.6	94.1
20%	8.3	35.9	67.3	90.4
50%	6.7	28.8	57.9	83.1
75%	3.1	16.7	38.2	64.9

With 90% of the ground-truth context supplied, the model still places the exact token first only 10.6% of the time, while placing it in the top 50 73.6% of the time. Incorrect predictions are not random: their embeddings have cosine similarity 0.322 to the ground-truth token against 0.054 for random pairs. The model is in the right neighbourhood and rarely on the exact point.

This is why we do not adopt the exact-match formulation used in text-domain planning work. An objective built on $\mathbb{1}[\hat{x} = x]$ has almost no signal in this token space, independently of the order in which positions are committed.

K FAILURE VARIANCE AND SIGNAL-AXIS DIAGNOSTICS

All analyses in this section are post-hoc descriptive analyses on the development split. They are used to diagnose the structure behind the held-out verification gap, not as confirmatory evidence.

Failure ICC estimator. For binary failure, we estimate the between-utterance variance of the latent failure probability with a moment estimator and apply the finite-sample correction $\mathrm{Var}(\pi) = \mathrm{Var}(\hat{\pi}) - \mathbb{E}[\hat{\pi}(1-\hat{\pi})]/(n-1)$ at $n = 16$. The correction removes an 11% upward bias in the $s8$ ICC (0.30 before correction to 0.27 after correction). Continuous signals use ANOVA $\mathrm{ICC}(1,1)$, which we also apply to failure for comparability (0.271 versus the moment estimate 0.270). Intervals are from 4000 utterance-level bootstrap resamples.

How often does the same utterance fail? The coarse "fails in 1–8 runs" summary hides substantial structure among affected utterances. The table resolves the 16-seed failure count and, for $s8$, reports the probability that best-of-2 or best-of-4 sampling recovers at least one successful realization.

Failures in 16 runs	$s8$	$s16$	$s8$ best-2	$s8$ best-4
0	53 (24.3%)	143 (65.6%)	—	—
1	55 (25.2%)	31 (14.2%)	0.996	1.000
2	26 (11.9%)	13 (6.0%)	0.984	1.000
3–4	44 (20.2%)	11 (5.0%)	0.958	0.998
5–8	26 (11.9%)	8 (3.7%)	0.853	0.976
9–12	7 (3.2%)	5 (2.3%)	0.547	0.788
13–15	3 (1.4%)	2 (0.9%)	0.197	0.352
16	4 (1.8%)	5 (2.3%)	0.000	0.000

At $s8$, 81 of 165 affected utterances (49%) fail only once or twice and are recovered with high probability by two samples; the 5–8 band requires four. Utterances that fail in all sixteen runs are not recovered at any budget, and their number does not fall when the step count is doubled.

Observable correlates of the between-utterance component. Observable utterance attributes explain little of the between-utterance component ($R^2 = 0.10$ at $s8$, 0.05 at $s16$). At $s8$, only word count is resolved and its coefficient is negative. At $s16$, word count and the uppercase-initial-word proxy have negative intervals; neither low-R^2 model supports a simple observable difficulty account for the between-utterance component. No development utterance contains a numeric token or an abbreviation under the feature definitions used here ($0/218$ for each), so those two indicators are excluded from the regression rather than interpreted as zero-effect covariates.

Signal ICC and within-utterance centering. Confidence, entropy, and degeneracy have ICCs of 0.74–0.75 at $s8$ and 0.81–0.87 at $s16$, whereas round-trip transcription has ICCs of 0.25 and 0.46, respectively. To remove between-utterance differences, we subtract each utterance's mean score across its sixteen sampling seeds from each seed-level score. At $s8$, this changes entropy AUC from 0.605 to 0.637 and degeneracy AUC from 0.607 to 0.645; at $s16$, the corresponding changes are 0.510 to 0.566 and 0.501 to 0.566. Round-trip AUC decreases from 0.862 to 0.703 at $s8$ and from 0.852 to 0.619 at $s16$. Because centering requires multiple realizations of the same utterance, we use it only as a diagnostic.

Predictive degeneracy is not a source of stochastic failure. We tested whether utterances with greater predictive degeneracy also show larger within-utterance variability in failure and find no association. At $s8$ the correlation between utterance-mean degeneracy and $\hat{\pi}(1-\hat{\pi})$ is $+0.0141$ [-0.1288, +0.1574]; all four tested $s8$ contrasts are unresolved. The corresponding $s16$ correlation is -0.0566 [-0.2059, +0.1059], and all three recorded $s16$ contrasts are unresolved. This negative control is consistent with the decomposition above: degeneracy is itself predominantly an utterance-level quantity (ICC 0.74) and does not track within-utterance variation.

L Reproducibility details

Software stacks. Generation and evaluation run in separate environments; we report both because they differ in every version that could affect numerics.

	Generation	Evaluation	Plotting
Python	3.10	3.10	3.10
PyTorch	2.0.1+cu117	2.4.1+cu121	—
torchaudio	2.0.2+cu117	2.4.1+cu121	—
transformers	4.41.2	4.46.3	—
NumPy	1.26.0	2.2.6	1.26.0
FunASR	—	1.3.29	—

Hardware and execution. All reported runs use a single NVIDIA A100 (80 GB) on one partition. Generation runs at fp32 with batch size 1; Whisper-based evaluation runs at fp16 with batch size 8; FunASR runs at batch size 1. Every job is a cold start. Model loading (approximately 25 s per pass, independent of the number of utterances) is excluded from throughput accounting.

We pinned the partition after a preliminary check found that identical seeds and inputs produce different outputs on different GPU architectures ($+0.152$ [+0.027, +0.320] percentage points between two architectures). Five seeds do not separate an architecture effect from realization variance, so we treat this as a reason to pin hardware rather than as a finding.

Checkpoints. Amphion commit `26f68831`. MaskGCT component checkpoints (MD5, first 16 hex digits) and evaluator revisions:

Component	Hash or revision
semantic codec	`35a659a565907fd1`
acoustic codec	`53f56e9aa933eee5`
acoustic codec (layer 1)	`3b20b10298edea5f`
T2S model	`b132a6d1387c0ef4`
S2A model (full)	`7782c0fcbd779b55`
S2A model (single layer)	`b504d7c039fe0f15`
HuBERT-large (primary evaluator)	`ece5fabbf034c107`
Whisper-large-v3 (evaluator)	`06f233fe06e71032`
wav2vec2-Conformer (evaluator)	`6b36ef01c6443c67`
wav2vec2-base (circularity control, not in panel)	`22aad52d435eb6db`
Whisper-small (verifier)	`973afd24965f72e3`
Paraformer-en (English evaluator)	`bf1a82a3d5a7624b`
Paraformer-zh (Mandarin evaluation, SeACo variant)	`4a8d75afdee06a54`
w2v-BERT 2.0 (semantic front end)	`da985ba0987f70aa`
WavLM-Large + ECAPA-TDNN (speaker similarity)	`efe934a9578bcb5f`
UTMOS (naturalness proxy)	`85ba046a0f88cae0`

Hugging Face entries are repository revisions; MaskGCT components, the ModelScope models, and the two quality proxies are MD5 digests of the checkpoint files. The two ModelScope models are pinned by name only (`master`), so we record file digests instead of a revision.

Speaker similarity uses the WavLM-Large upstream with an ECAPA-TDNN verification head, the configuration specified by the benchmark's reference pipeline. It is not the Hugging Face `wavlm-base-plus-sv` model.

The four evaluators reported in the main text are Whisper-large-v3, HuBERT-large, wav2vec2-Conformer, and Paraformer-en. Two further recognizers appear only in controls and are not part of that panel: wav2vec2-base (circularity control) and Paraformer-zh (Mandarin evaluation). We name the two language-specific models throughout, because they use different checkpoints and architecture variants—the cross-lineage control in Section E depends on that distinction.

S2A schedule. Under the default MaskGCT S2A schedule, codebooks 0 and 1 use 25 and 10 iterative steps; codebooks 2–11 are single-pass. S2A is held at this schedule in every condition.

Cost model and its validation. Throughput is modelled as $N\,(c_0 + c_1 s)$ for $N = 218$ utterances and s denoising steps, with $c_0 = 1.801$ s and $c_1 = 0.0761$ s. We compare the model against measured first-pass job times.

Steps	Model (s)	Measured (s)	Rel. error	Source
8	525	576	8.8%	median of 5
16	658	701	6.1%	uncontended seed
32	923	957	3.5%	median of 5
64	1454	1485	2.1%	median of 5
128	2516	2532	0.6%	median of 5

The model is accurate at the budgets that carry the cost comparisons and degrades at the smallest ones: the relative error is 0.6% at 128 steps, 2.1% at 64, and 3.5% at 32, but 6.1% at 16 and 8.8% at 8. The model therefore under-predicts wall-clock time at low step counts; we state agreement within 5% only for budgets of 32 steps and above.

Refitting the two constants on the same measurements gives $c_0 = 2.019$ s and $c_1 = 0.07492$ s and fits all five budgets within 1%. We report the original constants because every cost figure in the paper uses them, and we verified that the substantive conclusion is insensitive to the choice: recomputing the probability that the repair condition is cheaper than a CFR-equivalent single run changes it by at most 0.01 in all four evaluator families.

One caveat on the measurements themselves: four of the five 16-step jobs ran in a batch that was approximately 19% slower than the fifth (838–839 s versus 701 s) under otherwise identical settings. We use the uncontended measurement, which also agrees with the refitted model.

Text normalization. We use the benchmark reference pipeline unchanged. For English, hypotheses and references are lower-cased and stripped of punctuation, and comparison is at the word level. For Mandarin, punctuation (including full-width forms) is stripped, text is converted to simplified characters, and comparison is at the character level. The same normalization is applied to all four evaluators and to the verifier, since all of them are scored by the same routine; no evaluator-specific rule is used. No number or abbreviation expansion is applied.

Held-out run. The held-out evaluation was run once, after the split and all development-split selections were frozen. It comprises 870 English utterances at nine conditions, 1616 Mandarin and 320 hard utterances at two conditions each, and 870 utterances for the acoustic-stage reproduction, every condition at five seeds—the same seed count as the development split, so the two are directly comparable. Total generation and evaluation took approximately 37 hours of wall-clock time across four concurrent A100 partitions. The cost model above is calibrated on the development split and applied unchanged; it is a per-utterance model, so the larger held-out set does not require recalibration.

Round-trip AUCs use midrank tie handling (Mann–Whitney). An earlier internal computation used ordinal ranks, which overstated round-trip AUC by roughly 0.005–0.18 where verifier-missed failures tie at zero score. Decoder-derived continuous signals are unaffected.

Confirmatory and descriptive status. Not all analyses have the same inferential status:

Analysis	Status
English T2S detector panel	Frozen confirmatory
Within-base repair (16 cells)	Frozen confirmatory
$s64$/$s128$ comparison families	Frozen confirmatory
S2A held-out evaluation	Prespecified condition; executed post-unblinding; secondary
Review-budget trade-off	Post-hoc descriptive; grid $\{5, 10, 20, 30, 50\}\%$ fixed 2026-08-01 after unblinding; all grid points reported (Appendix C); tie-interpolated per the 2026-09-14 tie-handling decision
Held-out $s32$ detection and label-family control	Post-hoc; executed after unblinding; frozen detector definitions; no re-selection
Verifier-family control (data2vec-audio-base-960h)	Post-hoc; executed after unblinding; frozen definitions; no re-selection
Verifier-size control (Whisper-medium)	Post-hoc; executed after unblinding; frozen definitions; no re-selection
$s16$ posterior-likelihood re-scoring and 20-feature probe	Post-hoc; executed after unblinding; frozen definitions; no re-selection
Held-out verifier miss share	Post-hoc; executed after unblinding; frozen definitions; no re-selection
15-feature matched control	Post-hoc robustness
Hard-set verifier-error strata	Post-hoc mechanism analysis
Flag-rate fallback retrospective	Post-hoc illustration
P2 localization ablation	Prespecified (SHA-256 `8e08ee3595cec350`); executed after unblinding of primary results; primary cell not met
Failure variance decomposition ($s8$, $s16$)	Post-hoc descriptive, development split
Signal ICC and within-utterance centring	Post-hoc descriptive, development split
Ordering and oracle controls	Development split, reported for scope

Freeze chronology and post-unblinding changes. Development-split selections, including the policy, detector definitions, aggregation choices, recalibrator, and probe specification, were frozen before held-out generation and evaluation. The held-out results were examined only after these choices had been fixed.

After unblinding, thresholds, condition lists, the selected aggregation pair, recalibrator, probe specification, bootstrap seed, and evaluator panel were not changed. Reporting-only changes included

the posterior-likelihood row merge, split-specific captions, execution of the previously omitted S2A held-out evaluation, and regeneration of `paper_data.md`.

The review-budget analysis (`review_response.json`) was later recomputed with the tie-interpolation rule adopted for the canonical analysis; this changed the 20% round-trip cell from 64.3% to 64.1% and precision from 58.7% to 58.6%. Ordinal-legacy values are retained in the file.

Statistical procedure. All intervals are 95% percentile bootstrap intervals over 4000 resamples (seed 20260729) with utterances as the resampling unit, so that the five seeds of an utterance are always drawn together. Where a difference is reported, the same resample is used for both conditions.

Holm correction is applied within, and only within, families that we defined before looking at the corrected results:

Family	Size	Outcome
Local repair versus selective global regeneration	4	$0/4$ survive
Probe versus other decoder-derived readouts	7	$1/7$ survives
Held-out repair versus its own base	16	$16/16$ survive
Held-out cost comparison, $s64$	16	$1/16$ survives
Held-out cost comparison, $s128$	16	$3/16$ survives
P2 localization: A versus B, C, and D2	24	$11/16$ A-versus-B/C cells favor A; $3/8$ A-versus-D2 cells favor D2 after Holm correction. D1 is outside the family.

Comparisons reported as a grid are not corrected; we instead report how many cells exclude zero. Where a main-text claim rests on individual cells of such a grid, we apply Holm correction across the full table and report the corrected outcome. Content failure rate across thresholds, quality proxies, and hard-subset thresholds remain descriptive grids, and we refrain from citing an isolated crossing as a finding.

The frozen protocol is recorded in `frozen_policy.yaml` with SHA-256 prefix `7db96aa5bb8e9fee` and freeze date 2026-07-31. The held-out exclusion list was frozen before the held-out evaluation (Appendix B).

Post-unblinding canonical regeneration. One canonical file carries a timestamp after unblinding: `detection_auc_bootstrap.npz` was regenerated at 2026-08-01 18:57 to reflect the posterior-likelihood label merge in the reporting table. A row-wise `canon_diff.py` comparison confirms no numeric change: all 17 rows are identical to machine precision ($0.00\text{e}{+00}$ maximum difference), and only one label differs. The underlying development analysis was completed at 2026-07-31 03:16:42, before the held-out run began.

M CANONICAL RESULT FILES

Reported numeric results are traced to the canonical outputs listed below or to the three auxiliary measurements identified in the audit paragraph; figure and table scripts do not recompute statistics. Associated preregistration and generator artifacts are listed alongside them for provenance. Point estimates and confidence bounds are stored in the outputs, and bootstrap analyses archive the full resample distribution in companion NPZ files with seed 20260729.

Automated numeric cross-check. Of $1{,}372$ numbers extracted from the source, $1{,}320$ match a canonical entry exactly. The remaining 52 are derived quantities—differences, ratios, and percentages computed from canonical values—or come from three identified auxiliary measurements: the evaluator reference rates, the verifier-as-evaluator control, and the third-evaluator control. The audit script is released with the canonical files, so this count can be reproduced; no number in the source fails to trace to either a canonical entry or one of these three measurements.

File	Generator	Contents
detection_auc_bootstrap.npz	canonical_detection.py	17 detectors × 4000
repair_effect_bootstrap.npz	canonical_repair.py	64 contrasts × 4000
equivalent_step_bootstrap.npz	canonical_equivstep.py	8 rows × 4000
dose_response_bootstrap.npz	canonical_doseresponse.py	8 points, 4 fits
guardrail_ci.json	guardrail_ci.py	6 proxies, 5 seeds
hard_multi_asr_fp_s32.json	hard_multi_asr.py	3 recognizers, $n = 400$
sealed_exclude_manifest.json	make_sealed_exclude.py	frozen exclusion list
detection_test_excl.npz	canonical_test.py	7 T2S detectors × 4000
detection_s2a_test.npz	canonical_s2a_test.py	2 S2A detectors × 4000
repair_test_excl.npz	canonical_test.py	48 contrasts × 4000
test_full.json	canonical_test_full.py	operating points, CFR grid, cost
zh_hard_test.json	canonical_test_full.py	Mandarin and hard, with/without
utmos_wer_mediation.json	utmos_wer_mediation.py	UTMOS–WER decomposition
review_response.json	review_response_20260801.py	probe, rates, hard strata, review-budget curves
hard_paired.json	canonical_hard_paired.py	5 hard-set thresholds
hard_paired_bootstrap.npz	canonical_hard_paired.py	5 thresholds ×10000
ordering_ablation.json	canonical_ordering_ablation.py	7 ordering rules × 2 T2S budgets
wallclock_validation.json	canonical_wallclock.py	cost model versus measured wall-clock
variance_decomposition.json	variance_decomposition.py	$s8$ failure/signal variance and diagnostics
variance_decomposition_bootstrap.npz	variance_decomposition.py	$s8$ bootstrap distributions ×4000
variance_decomposition_s16.json	variance_decomposition.py	$s16$ failure/signal variance and diagnostics
variance_decomposition_s16_bootstrap.npz	variance_decomposition.py	$s16$ bootstrap distributions ×4000
p2_localization.json	p2_analysis.py	P2 estimates, operations, recovery
p2_localization_bootstrap.npz	p2_analysis.py	32 P2 contrasts × 4000
p2_preregistration.yaml	—	P2 family, primary cell, decision rule
dev_selections_frozen.json	canonical_detection.py	selections frozen before the held-out run
detection_verifier_crossfam.npz	canonical_detection.py	round-trip AUC with data2vec-audio-base-960h verifier, $s8/s16/s32$ under Whisper/HuBERT labels, paired vs. probe
detection_verifier_size.npz	canonical_detection.py	round-trip AUC with Whisper-medium verifier, same layout
s16_posterior_test.npz	canonical_detection.py	held-out $s16$ posterior likelihood and 20-feature probe
selector_miss_test.json	canonical_detection.py	held-out verifier miss share, 5-seed typing, W/H labels
review_budget_test.npz	canonical_review_budget.py	review-budget grid $\{5, 10, 20, 30, 50\}\%$, tie-interpolated
review_budget_test.json	canonical_review_budget.py	same, point estimates

Canonical artifact provenance. The newly supplied canonical results and their generator sources were checked byte-for-byte before packaging. The variance outputs identify `src/eval/variance_decomposition.py` as their generator; the archived source is stored at that same path. SHA-256 digests are:

File	SHA-256
ordering_ablation.json	6a1b9b031262a26937233d62c35230eb9f017e3ac1501400 2063a1dbf1eec6b9
wallclock_validation.json	26d2755b65a9f756f6e913687aa7eb3705e0b899fc301749 4151930ff07449d1
hard_paired.json	6d546770d6dea6d0aa5a0c010e0eb12465b64a2fc5ff0e17 2ea743b0e2a258a9
hard_paired_bootstrap.npz	15a3dcd99d31b9ef56969ff4f27d2dd11512304954bbab94 e5f4f379a3eb9b3c
variance_decomposition.json	5bc719be210738552a734849ecfff40d9aa5a88f973e7e03 b83f7d8c35e74c44
variance_decomposition_bootstrap.npz	49b78240cd3fdcc7afbe1bae1310669d1ae5b81f396f3c52 dfe6fbdab53deb9f
variance_decomposition_s16.json	6c93beaa0a4e75ef5dc89ece7bfbbe98fd016eb2951f461b 07229e1ef32ea553
variance_decomposition_s16_bootstrap.npz	345b796ec1309d1e6c805281ccfee9e52eeda615cd8a7057 907edc9cafb90df5
src/eval/canonical_ordering_ablation.py	516ba02a8c392b629730e25d0c9e282dc5dda55c992158f5 0e85c0ffc88c261c
src/eval/canonical_wallclock.py	5498fd184e5def1d53e581ebb8aed60be12925cf16696a77 932276f616d831c4
src/eval/canonical_hard_paired.py	2260cc492715da7bbcda4dc07e48b33c09cd1ed86db69710 603fb244cebe1d3d
src/eval/variance_decomposition.py	dde6a54b9da4b677ec0660d0121c90f54b04b734df4f3f64 6506abd2e50b3f22
src/eval/tex_number_audit.py	46ccc5aa07174bc1f93e98e131d6f465c7db5dd2d753683a 2d39f5e0a575e4d4
detection_verifier_crossfam.npz	1e7b817d861d400f290a7dbf48e82828ab89461a544cd7cc e3a1a3d3820c6977
detection_verifier_size.npz	c6540a4b9bf47f32b26faf96ede3553d7d6a01e49eaff13a 52c8d48a6911bdce
s16_posterior_test.npz	b8ee8e15a2375b5245370de2f2e8e8aa77db3829a5ac6133 f39856a4bac8b440
selector_miss_test.json	75703c4c0c1e41d0d08fffa8873a0900d7fe57cc3ef4a557 78010161a20acf03
review_budget_test.npz	865a80ce81432649af89ddcecf832038735d347a624497b7 d8382ffeffb62723
review_budget_test.json	f65fce0c2285d81eb8350620558a053251023ee079d0a32f c4933d9e7eed68e8
review_response.json	f2ff35e4fd9a235b903c230acaeb8c324a117558b2babca5 ad9824a129fb429f

P2 provenance. The P2 artifacts are archived with the following SHA-256 hashes:

File	SHA-256
p2_localization.json	262b219eed1eac8660169dc7498a2486c75cad9b2b84b73a3134fbec97e70a12
p2_preregistration.yaml	8e08ee3595cec350763804513f0ee70b721aefe14ac220be2b6285ba134a9a94
src/eval/p2_analysis.py	890976c6679a6a18e3249749307dda31b17b38900d2cdcf50c92f99b76abbf51

The preregistration defines the Holm family as A-versus-B, A-versus-C, and A-versus-D2 across four evaluators and two metrics; the canonical JSON records that $11/16$ B/C cells favor A and survive correction, while $3/8$ D2 cells favor D2 and survive correction.