LexiAdapt: Auditing Contextual Speech Recognition for Unsupported Phrase Insertions

Anonymous authors
Paper under double-blind review

Abstract

Contextual automatic speech recognition (ASR) uses phrase lists from documents, contacts, or domain vocabularies to recover rare terms. Yet the recognizer can copy a listed phrase that was not spoken. Word error rate (WER) measures errors across all words. Bias-word error rate (B-WER) focuses on listed rare words, while phrase recall measures recovery. None reveals by itself whether the audio supports an accepted edit or whether a safe-looking system rejects almost every edit. We introduce LexiAdapt: it selects contextual edits without the reference and audits the transcript against it afterward. LexiAdapt tracks where each phrase came from. No-leakage lists exclude the target transcript. It also fixes the acceptance rule before evaluation. Its scorecard reports separate error rates for listed rare words and all other words. It also reports phrase precision and recall, their harmonic mean, and unsupported additions. Coverage records how often the system accepts a changed contextual transcript. Overall WER and real-time factor capture transcription quality and cost. The comparison contract covers a primary modern contextual model, a larger streaming contextual model, a Whisper anchor, and a no-context control. The design requires disjoint development and test units, frozen phrase sources and gates, matched contextual interfaces, and uncertainty at the highest independent unit. This makes the central question falsifiable: under matched conditions, does context improve bias-word and phrase recovery without increasing unsupported additions, collapsing coverage, or imposing unacceptable latency? The protocol prevents lower WER or zero observed insertions under near-total abstention from being treated as sufficient evidence.

1 Introduction

Automatic speech recognition (ASR) systems often miss the words that matter most to a user: a local name, a medical term, a project title, or a phrase from a nearby document. Contextual ASR addresses this problem by giving the recognizer a short phrase list at test time. The list can rescue a rare term that the acoustic model would otherwise miss. It can also create a new error channel by encouraging the recognizer to copy a plausible phrase that was never spoken.

Standard evaluation does not cleanly separate these outcomes. Word error rate (WER) rewards overall transcription accuracy, while phrase recall rewards recovery of listed terms. Both metrics can improve even when the context list inserts a listed phrase that the speaker did not say. A convincing evaluation must therefore answer two questions together: did context recover the intended phrase, and did the system avoid adding listed phrases without reference support?

LexiAdapt studies this question as an acceptance problem.[1] A contextual decoder may propose broadly, but an edit enters the reported transcript only if a predeclared, reference-independent rule accepts it. The protocol tracks phrase provenance and jointly evaluates recognition quality, phrase recovery, unsupported insertions, and latency. Neighbor-derived candidate phrases exclude the target utterance, but some saved ranking modes count its transcript when estimating corpus rarity. Reference-supported-phrase admission also selects the opportunity cohort; implementation controls use target phrases. None is an unrestricted no-leakage evaluation.

[1] Anonymous Open4Science code repository: `https://anonymous.4open.science/r/lexiadapt-asr-D3E4/README.md`; complete source archive: `https://anonymous.4open.science/api/repo/lexiadapt-asr-D3E4/zip`.

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.

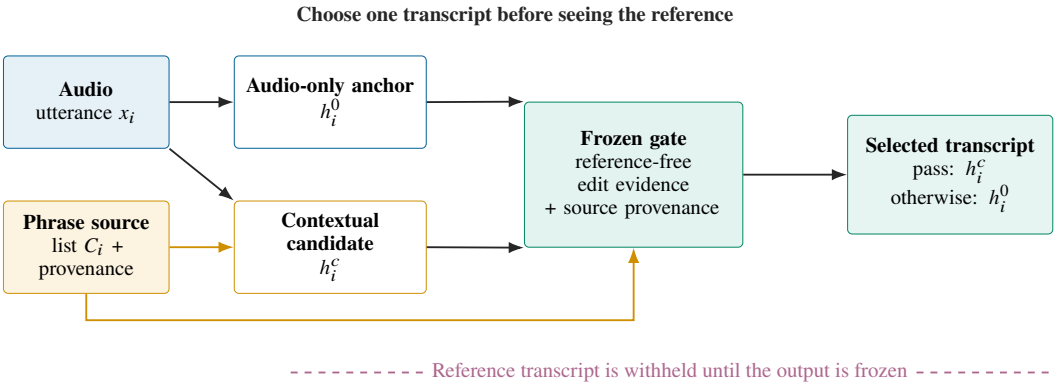

Figure 1: LexiAdapt selects one transcript without access to the reference. Audio produces an anchor and a contextual candidate, while the supplied phrase source also provides provenance features. A frozen gate accepts the contextual candidate or keeps the anchor; the reference is revealed only for the post-selection audit in Figure 2.

Figure 1 shows the central distinction. The context list supplies useful phrases and distractors that look or sound similar, the recognizer proposes alternatives, and a fixed gate either accepts the contextual transcript or keeps the baseline. A post-selection audit then checks whether the accepted edit was supported by the reference. This separation prevents a system that recovers many listed phrases from receiving credit for unsupported copying.

The evaluation contract uses a progressively harder sequence. Prompts that intentionally include reference phrases check the implementation. A no-leakage opportunity set uses neighboring utterances as the context source. Audited multilingual lists test whether filters reject similar-looking or similar-sounding distractors. A four-locale inventory tests whether a fixed transcript-distance gate behaves consistently across scripts. Finally, meeting audio paired with document context tests whether rules selected on one set of meetings transfer to disjoint meetings.

Contributions. We make three contributions. First, we define a contextual-ASR scorecard that reports released-scorer-compatible B-WER/U-WER, phrase recovery, unsupported additions, accepted-edit coverage, WER, and latency together. Second, we provide a leakage-aware protocol that separates phrase sourcing, contextual proposal, frozen acceptance, and reference-only audit. Third, we specify a common modern-model comparison in which Qwen3 is the primary contextual baseline, VibeVoice is the frontier contextual system, Whisper-large-v3-turbo is the legacy anchor, and Parakeet is a no-context control. Transcript-distance gating is the paper default, the source-backed gate is the closest evaluated baseline, and the acoustic selector remains diagnostic until genuine alternatives and calibrated scores are retained. The Whisper-family results later in the paper come from earlier runs and motivate the design, but their exact inputs, model revisions, and complete run lineage have not been jointly verified.

2 Related Work

Contextual recognition. The inspected literature spans distinct comparison axes. Le et al. (Le et al., 2021) provide the B-WER/U-WER metric and evaluation precedent; DYNAC (Sudo et al., 2025) is the closest dedicated single-speaker contextual method; the Tree-Constrained Pointer Generator studies (Sun et al., 2022; 2024) provide decoding and bias-word-error predecessors; and RECAST (Mittal et al., 2025) is the closest prompt-scaling predecessor, comparing retrieved-keyword prompts with full-list prompting at up to 4,000 keywords. CALM (Shakeel et al., 2026) is newer multi-speaker related work, but the frozen audit verified only official metadata. None of these inspected sources establishes the same fixed post-decoding acceptance plus phrase-level unsupported-addition evaluation used here; the frozen search also does not establish field-wide absence of related safety analyses.

Open speech models and reliability. The prospective comparison spans four public model roles: Qwen3-ASR-1.7B-hf as the primary modern contextual baseline (Qwen, 2026), VibeVoice-ASR-Streaming-7B as a larger streaming contextual system (Microsoft, 2026), Whisper-large-v3-turbo as a legacy anchor (OpenAI, 2024), and Parakeet-TDT-0.6B-v2 as a fast no-context control (NVIDIA, 2026). OWSM v4 has a public parent checkpoint, but the frozen audit did not verify a dedicated OWSM-Biasing route or a comparable contextual-list contract in the checked official locations (Peng et al., 2025; ESPnet, 2025). Work on efficient speech language models highlights the need to report computation and latency with accuracy (Jia et al., 2025). Inclusive ASR research further shows that aggregate error rates can conceal concentrated failures for particular languages and speaker communities (Joshi et al., 2025). LexiAdapt therefore requires utility and insertion-risk results for each evaluated population and context source.

Scope relative to prior work. LexiAdapt is not a new acoustic model, vocabulary architecture, or language-model rescoring method. Its scope is a fixed post-decoding acceptance and evaluation protocol for supported phrase recovery and unsupported additions. The only runnable component found in the dated audit was Le et al.'s pinned scorer (Le et al., 2021); it is an evaluator, not a model baseline. Its paper, README, and pinned code disagree about the inventory used to assign insertion errors: the paper describes full-list membership, the README labels TSV column 3 as the rare-reference-word inventory and column 4 as the full supplied-list inventory, while the code reads column 3. Because this semantic discrepancy changes which metric receives an inserted distractor, we name the code-compatible and full-list measurements separately and do not present the scorer as one unambiguous canonical full-list definition. No inspected dedicated contextual-ASR system had an official implementation or checkpoint satisfying all execution gates: immutable revision, usable license, documented contextual-list interface, documented V100 inference route, and a comparable evaluation contract. DYNAC, Tree-Pointer, RECAST, and CALM are therefore literature comparators only, not executed baselines.

3 LexiAdapt Protocol

3.1 Task and audit target

For utterance i, let x_i be the audio, C_i the supplied phrase list, h_i^0 the no-context transcript, and h_i^c a context-conditioned candidate. A reference-independent gate $g(z_i)$ observes declared features z_i, such as phrase provenance, transcript distance, and an optional acoustic-score margin. It returns the contextual candidate when $g(z_i) = 1$ and the baseline otherwise:

$$\hat{h}_i = \begin{cases} h_i^c, & g(z_i) = 1, \\ h_i^0, & g(z_i) = 0. \end{cases} \tag{1}$$

The reference transcript is not part of z_i. It is revealed only for evaluation.

The post-selection audit identifies context phrases that appear in $\hat{h}_i$ but not in h_i^0. Text normalization lowercases, replaces punctuation except apostrophes with spaces, and collapses repeated whitespace. An addition is unsupported when the normalized reference does not contain it. The audit separately marks an unsupported phrase as a near-reference variant when its best normalized reference window reaches the fixed similarity threshold of 0.82. Similarity is the Python `SequenceMatcher` ratio between the phrase and reference windows whose token counts differ from the phrase's token count by at most one. This test is separate from the token-level transcript distance used by the acceptance gates. The reported strict false-positive (FP) count excludes these predeclared variants; the raw false-positive and variant counts remain in the stored score records. We report both unsupported additions relative to the no-context transcript and the total strict FP count in each output. A nonzero no-context total means that the baseline recognizer produced a listed phrase without receiving the list; it is not an insertion caused by prompting. The same matcher and threshold are applied to every system. Zero observed events describes the evaluated sample; it is not a guarantee about future audio.

3.2 Leakage-aware context construction

Each phrase carries a source identifier and a relation to the target utterance. Reference-grounded lists are used only as implementation controls. No-leakage lists exclude the target transcript and draw phrases from neighboring utterances, other entries in the same public corpus, or linked document text. The protocol also records distractors, source frequency, phrase form, locale, and script when those fields are available. These records make it possible to distinguish a useful source from one that merely contains a frequent phrase.

3.3 Proposal and acceptance

LexiAdapt separates three operations. First, a source filter and phrase-form risk score produce a bounded context list. Second, the ASR model generates a prompt-conditioned transcript or multiple ranked candidate transcripts, also called an n-best list. Third, the fixed gate decides whether the contextual output replaces the no-context anchor. The simplest gate limits transcript distance. Source-aware rules use recorded source counts for newly added phrases. Acoustic rules retain multiple transcripts and compare their decoder scores; this test is valid only when the decoder stores genuine alternatives, their ranks, and their scores.

For the reported gates, normalized anchor distance is the token-level Levenshtein distance between the no-context and contextual transcripts divided by the number of tokens in the no-context transcript, with a denominator of at least one. The four-locale gate accepts a candidate that adds a listed phrase only when this distance is at most 0.75. The AMI/QMSum source-and-distance gate requires distance at most 0.5 and a source count of at least one for the best-supported added phrase. The intended source-and-n-best extension requests eight candidate transcripts, deduplicates them, and requires each added phrase to appear in every retained unique candidate. In the saved validation outputs, candidate sampling was off and only one candidate per utterance was retained; the original raw and duplicate counts are missing. Thus the requested count does not establish distinct alternatives, and results that depend on candidate agreement remain diagnostic. These thresholds are selected on their named development or validation sets and then held fixed for the corresponding evaluation.

For the stored hybrid diagnostics, the rule first accepts a changed prompt output only if it adds a listed phrase, has anchor distance at most 0.5, has source count at least one, source confidence at least 0.75, and phrase-risk score at most 0.4. Otherwise, it accepts the stored reranker output only if that output differs from the anchor, adds no listed phrase, and has anchor distance at most 1.5; the prompt path wins ties. Prompt-supported additions and non-context reranker changes are counted separately. The eight-meeting validation hybrid is computed from feature rows after decoding (Appendix Section A.6); the four-meeting replay is a separate transfer diagnostic. Neither stored summary by itself verifies the original candidate alternatives, scores, or executed producer and input revisions, so neither establishes an acoustic-selection mechanism.

The gate is selected on validation data and then held fixed for a disjoint evaluation unit. Meeting, document, speaker, and phrase-family boundaries are kept intact when the data permit. Any rule chosen after inspecting the held-out references remains diagnostic and cannot support a transfer result.

3.4 Scorecard

For every prospective comparison, the scorecard is a required-output contract. It keeps three non-interchangeable metric lanes: released-scorer-compatible B-WER/U-WER, full-list B-WER sensitivity, and normalized multiword phrase-span contextual WER. No scientific values for these lanes are available yet. The contract also requires WER as a percentage, contextual phrase recall, precision, F1 score, raw unsupported additions, the strict phrase false-positive count as a near-reference sensitivity analysis, and real-time factor (RTF). Exact numerator and denominator counts accompany rates. Accepted-edit coverage is the number of accepted changed outputs divided by the number of changed contextual candidates; abstention is one minus this value. Edit coverage is not the fraction of scheduled runs that finished. A rule that accepts no contextual edit can have zero observed insertions without being useful. The reported scores below describe only their named finite sets.

The audited source bundle does not link the original predictions and speaker identities needed to check the hard-distractor uncertainty calculation. We therefore omit its previously reported confi-

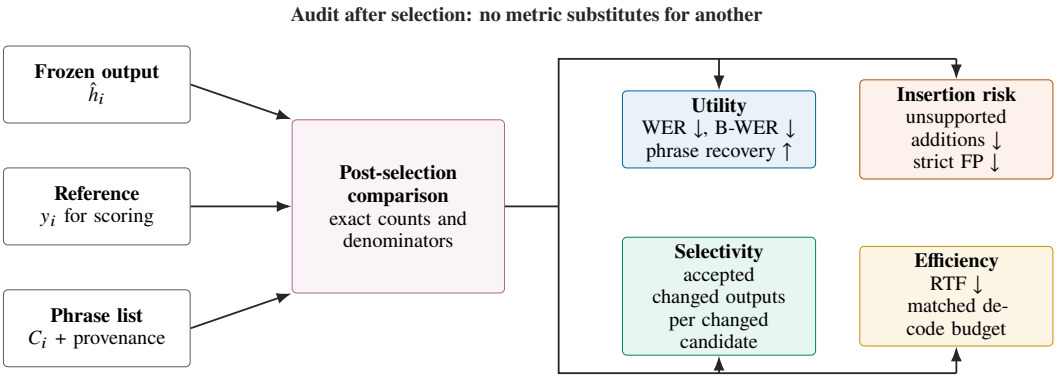

Figure 2: LexiAdapt audits a frozen output along four separate axes. Utility measures recognition and listed-phrase recovery; risk counts unsupported additions; selectivity reports how many changed contextual candidates the gate accepts; and efficiency uses matched decoding budgets and real-time factor. This is a schematic measurement contract, not an empirical result.

dence intervals. Prospective paired intervals must resample the highest independent source unit, such as a speaker or meeting.

RTF is the sum of decoding time divided by the sum of audio duration. We compare it only across runs with matched runtime conditions and decoding budgets. Every operating point fixes the context source, model checkpoint, decoding budget, and acceptance rule.

Figure 2 makes the four non-interchangeable outputs explicit. Utility, insertion risk, selectivity, and efficiency are reported together, so a system that avoids additions only by accepting no contextual edits cannot appear useful through safety alone.

4 Experimental Design

Study key. The LibriSpeech and English FLEURS controls each contain 512 utterances with supplied target phrases; the 123-utterance LibriSpeech opportunity set uses neighbor-derived candidates and reference-based admission. The four-locale FLEURS development inventory has 1,280 utterances; overlap with the English control is unknown. The document-context validation set has 6,003 utterances from eight AMI meetings; the disjoint ICSI replay has 4,133 utterances from four meetings. A reference-selected 1,799-utterance set is contained in the 86,888-utterance Product set. The 5,431 repaired-input bundles are output bundles, not independent utterances or meetings. These populations and units cannot be pooled as replications.

Controlled evaluations. The supplied-target controls use LibriSpeech test-clean (Panayotov et al., 2015) and United States English from the Few-shot Learning Evaluation of Universal Representations of Speech (FLEURS) corpus (Conneau et al., 2022). The neighbor-derived LibriSpeech set has 153 supported phrases and 356 distractors. Its saved ranking mode is unverified and may count target text when estimating rarity; admission requires a reference match. The positive-sample rate is 1.0, but the screened denominator and selection flags are unresolved, so recovery does not measure opportunity prevalence or negative-example safety. The rule-audited FLEURS set has 48 target–distractor pairs, 97 target occurrences, and 6,144 distractor placements across the English control. Source and phrase-form checks plus automated reference scoring were applied, without human listening or label review.

Multilingual development evaluation. The four-locale FLEURS inventory covers English, Spanish, French, and Hindi with typed targets, script-matched distractors, source fields, and locale labels. Its human release audit remains incomplete, so it is a development set, not a released benchmark.

Document-context evaluation. Validation joins QMSum query-based summaries (Zhong et al., 2021) to individual-headset-microphone audio from the Augmented Multi-party Interaction (AMI) Meeting Corpus (Carletta, 2007). Target-utterance text is removed before prompting, but the sum-

Table 1: Controlled diagnostic results, not validated operating points. "Strict FP" is the listed-phrase false-positive sensitivity count under the frozen near-reference matcher, not a population error bound. The corresponding supplied-distractor exposure totals are 356 for LibriSpeech, 6,144 for FLEURS, and 15,021 for the four-locale set; phrase placements are descriptive exposures, not independent trials. Entries labeled "Four-locale set" use Whisper medium. A dash means that the saved decode did not retain a comparable timing measurement. The contextual results have unresolved run lineage and transfer safety.

Evaluation	System	WER (%)↓	Recall↑	Strict FP (sens.)↓	RTF↓
LibriSpeech opportunity	No context	6.28	0.6667	0	–
	Prompt	3.32	0.9804	0	–
FLEURS rule-audited	No context	6.60	0.7216	0	–
	Source+risk, beam 8	5.58	0.9794	0	0.3272
Four-locale set	No context	13.15	0.6726	0	0.1318
	Distance-gated prompt	12.69	0.9351	0	0.2662

maries were written after the meetings. This tests linked-summary stress, not context available before recognition. The meeting-disjoint ICSI replay probes transfer of a gate fixed on validation.

The ICSI replay uses linked QMSum Academic documents from three train and one validation entries in the official QMSum partition; it is not an official QMSum test split. Failed and abstaining transfers remain visible.

Models and comparison rules. Earlier evaluations used public Whisper checkpoints (Radford et al., 2023) from tiny through medium and an incompletely linked large-v3-turbo meeting replay. Conditions included no-context decoding, prompting, candidate generation, source filtering, and fixed gates. The AMI/QMSum prompt supplied two phrases; a control requested eight candidates and up to eight phrases, but the retained files do not verify candidate agreement. Checkpoint and corpus-release identities remain incomplete, so these results are diagnostic. The unexecuted prospective grid assigns Whisper-large-v3-turbo as legacy anchor, Qwen3-ASR-1.7B-hf as modern contextual baseline, VibeVoice-ASR-Streaming-7B as frontier contextual system, and Parakeet-TDT-0.6B-v2 as a no-context control. Contextual models require an empty-versus-nonempty context check on matched audio, phrases, scorer, and resources. None is eligible for comparison until revision, payload, runtime, interface, and evaluation contract pass the execution gate.

5 Diagnostic Evaluation

The numerical results below come from earlier, incompletely linked runs. Their exact input bytes, model revisions, and complete run-to-result lineage have not been jointly verified, so they cannot support the prospective comparison.

5.1 Controlled records motivate joint recognition and recovery checks

Table 1 begins with settings in which the source construction is explicit and target-transcript leakage is either restricted to a control or removed. On the no-leakage LibriSpeech opportunity set, prompting Whisper small cuts WER from 6.28% to 3.32% and raises phrase recall from 0.6667 to 0.9804, with no observed unsupported insertion. On the rule-audited FLEURS set, the source-and-risk beam-8 setting records 5.58% WER and 0.9794 recall with no observed strict FP. Its aggregate does not establish the number of distinct scored alternatives retained per utterance, so these values do not isolate a reranking effect. Within these recorded opportunity sets, the diagnostic rows show a possible recovery–insertion tradeoff under the stated matcher; no prediction-level replay or exact run-identity join has verified those values.

5.2 Acceptance rules alter the recorded operating point

The same experiments also show why raw phrase recovery is insufficient. On the rule-audited FLEURS budget study in Appendix Table 7, increasing the prompt budget from four to twelve phrases improves the Whisper-small WER from 6.17% to 5.96%. However, recall falls from 0.9691 to 0.9485 and one strict FP appears. This list was constructed using the target reference to place matching phrases first; a saved random-order control exposes the dependence of the four-phrase result on list order (Appendix Section A.3). A source-and-risk rule reaches 5.58% WER while keeping the strict count at zero. In a separate test that ranks candidates by phonetic similarity, an unguarded score inserts "Indies" when the reference says "end the innings." Requiring positive source support removes that error. The gate therefore changes which decoder setting is admissible; it is not a cosmetic analysis added after model selection.

The four-locale evaluation provides a second test (Appendix Table 9). With Whisper medium, ungated prompting reaches 12.68% WER and 0.9351 recall but introduces one strict insertion. A transcript-distance gate keeps the same recall, removes the insertion, and changes WER only to 12.69%. The distance-gated prompt path reaches this point at 0.2662 RTF, compared with 0.5938 for a four-candidate reranking path with the same aggregate utility. In this diagnostic, the wider candidate path does not improve the measured utility-risk operating point. A separate eight-candidate development summary records worse Hindi WER, recall, and unsupported-insertion count than the four-candidate setting, alongside higher latency. Its original predictions and execution identity have not been rebound, so it is an unresolved negative diagnostic, not an independent performance comparison. A WER-first leave-shard-out policy summary also records one strict insertion and a large recall loss. On a second phrase inventory over the same four-locale audio, wider raw reranking still records three unsupported Hindi insertions, while a learned acceptance rule trails a fixed distance rule (Appendix Section A.5). These development records do not establish that the fixed rule transfers safely to new audio or real user documents.

5.3 Document context exposes the transfer gap

Table 2 records the diagnostic document-context contrast. On AMI/QMSum validation, raw two-phrase prompting raises recall from 0.8974 to 0.9615, but WER degrades from 23.13% to 31.16% and the total strict false-positive count rises from one to four. Relative to the no-context transcript, the raw prompt adds five supported phrases and four unsupported phrases. One unsupported addition is excluded as a near-reference variant, leaving three new strict FPs; these join the baseline's one strict FP for a total of four. A fixed source-and-distance gate keeps all five supported additions, accepts no unsupported addition, and returns the total strict count and WER to 1 and 23.12%, respectively. It accepts only 5 of 1,545 changed contextual candidates (0.32% edit coverage; 99.68% abstention), although it retains 5 of the 9 candidates that add a supplied phrase. The source-count condition excludes no changed candidate in this aggregate, so the distance cutoff performs the additional filtering. The low observed unsupported-addition count is coupled to very low broad edit coverage.

A cross-meeting source-swap control rotates document sources across the eight validation meetings while holding the target audio, model, sampling seed, requested candidate count, and decoding settings fixed (Appendix Table 11). The recorded rule with a retained-candidate phrase-support threshold accepts three contextual additions with the original sources and none with swapped sources; neither condition accepts an unsupported addition or raises the strict count. A rule without that threshold, but with the same source and distance checks, accepts six swapped-source additions, four of them unsupported, and raises the strict count by one. A targeted prediction check found eight distinct scored transcripts in two of the three original-source acceptances and one in the third; all six swapped-source ablation acceptances had multiple distinct scored transcripts. The full candidate census and executed run identity remain unverified. A support fraction of one over a single transcript cannot test agreement across alternatives. Thus the recorded decisions depend on the swapped source and the additional threshold, but this contrast does not isolate an acoustic-consensus mechanism or establish useful transfer.

A separate eight-meeting validation diagnostic tests reranking without decoder prompting and a selective combination with the prompt gate (Appendix Section A.6). The raw reranker has an adverse recorded WER, while the combination applies most accepted reranker edits to transcripts without a new listed phrase. These distinct outcomes cannot establish contextual recovery by acoustic selec-

Table 2: Document-context diagnostic results, not validated transfer results. "Added S/U" is the primary raw count of supported/unsupported phrase additions relative to the no-context transcript. "Strict FP" is a sensitivity count after excluding near-reference variants at the fixed 0.82 similarity threshold; AMI counts use 73,186 supplied-distractor exposures. Edit coverage is accepted changed outputs divided by changed contextual candidates. The AMI raw prompt has 1,545/1,545 coverage; its fixed gate has 5/1,545. The raw prompt adds four unsupported phrases, one of which is a near-reference variant, so its strict-FP total increases from one to four. The four-meeting ICSI replay has unavailable edit coverage because the candidate pool is incomplete. Original prediction-to-display and run identities remain unresolved.

Evaluation	System	WER (%)↓	Recall↑	Strict FP (sens.)↓	Added S/U S↑/U↓	Edit coverage (%)↑
AMI/QMSum validation	No context	23.13	0.8974	1	0/0	–
	Raw prompt	31.16	0.9615	4	5/4	100.00
	Source+distance gate	23.12	0.9615	1	5/0	0.32
Frozen ICSI internal replay	No context	62.65	0.7143	6	0/0	–
	Fixed hybrid	64.94	0.7143	5	0/0	–

tion. Their original prediction, evaluator, and execution identities are not jointly verified, so they are not used as held-out evidence.

Transfer is harder in the earlier evaluations. Raw prompting proposes nine contextual phrase additions on validation. In a 129-meeting empty-context replay, the prompt and gate values were computed over all 86,888 inspected samples (Appendix Table 11). Raw prompting proposes 385 additions, of which 276 (71.69%) are unsupported. The validation-selected permissive gate accepts 132 edits, including 47 unsupported additions, and increases the strict false-positive count by 26; the stricter source-and-distance rule accepts two edits, neither unsupported, and preserves the baseline strict count. This analysis covers all 129 meetings, but its values remain diagnostic because the exact source bytes, checkpoint revision, and run lineage are unresolved; they are not held-out evidence. A separate source-plus-acoustic packet has 1,668 of 5,431 bundles complete (26,688 of 86,888 sample outputs), with 3,763 bundles missing and zero context entries. Its held-out input lists were empty despite nonempty validation lists; this is an input-contract failure, not a context-bearing test of an acoustic selector. Its governing generation and denominator remain unresolved, so it stays fail-closed. The separate 1,799-sample context-bearing replay has 14 context-addition entries and no recorded WER change; it does not resolve either incomplete analysis. On a separate frozen-policy, meeting-disjoint ICSI replay with 4,133 utterances, Whisper medium stays at recall 0.7143 to 0.7143 and its hybrid accepts no supported contextual addition; its stored acoustic reranker output differs from the anchor on 1,305 outputs without adding listed phrases, worsening WER from 62.65% to 64.94%. Whisper large-v3-turbo starts at recall 0.6429 and reaches 0.7143 in the hybrid, with one supported prompt-context addition, but its hybrid still fails the overall gate and changes 1,109 outputs. These results do not establish that document relevance supplies useful utterance-level acoustic support.

An integrity check found that the earlier real-document prediction files retained only one transcript per utterance. Those files can evaluate source and transcript gates, but they cannot test an acoustic comparison among alternatives. A separately discovered repaired-input run has paired evaluator and prediction filenames for all 5,431 held-out bundles. In a read-only check of five spaced bundles, all 72 stored predictions have a nonempty context and report prompt application, but each retains only one candidate after requesting eight. This sample does not establish raw generation diversity, an acoustic margin, or the run's aggregate scores. The repaired run is thus visible as an unresolved result family, not as evidence that an acoustic selector succeeds on held-out meetings.

The 1,799-utterance context-bearing replay selects examples with at least one phrase supported by the reference transcript. Every selected utterance also occurs in the larger 86,888-utterance Product set. Across matched feature records, the no-context error count differs for 15 utterances and the prompted error count for 199. The corresponding stored transcripts also differ for all 15 and 199 cases, respectively. For the 199 prompted disagreements, the named Product input shards have empty phrase lists, while their stored predictions mark prompting as applied and record 7–16 available context phrases. This count does not state how many phrases entered the prompt. Neither prediction set records an input-file or producer digest. The named shards therefore do not establish the exe-

cuted inputs or explain the transcript differences. This reference-conditioned subset is neither an independent test population nor a paired estimate of the effect of adding context. Its scores remain diagnostic pending executed-input, prediction, and run-lineage reconciliation. A separate large-v3-turbo hybrid summary also names a 1,799-example context-bearing set; it has not been joined to the 14-addition source/acoustic bridge above. That summary reports 17.32% WER for its prompt-only rule and 17.46% for the rule that also accepts non-context reranker edits, an absolute WER difference of +0.14% (hybrid minus prompt-only). Its acceptance check fails both the required sample count and majority-meeting improvement condition; Appendix Section A.8 gives the remaining scope and provenance limits. This is an adverse recorded contrast, not a verified held-out effect of acoustic selection.

5.4 What the diagnostic records suggest

Across the diagnostic sets, one contrast recurs. Contextual prompting is associated with recorded recovery gains in controlled opportunity sets when the list contains a relevant phrase and the acceptance rule separates it from confusable alternatives. In the document-linked evaluations, source frequency remains high even for phrases that were not spoken, and the accepted set either becomes unsafe or has very low coverage. The exact inputs, model revisions, and run lineage remain unverified, so these results motivate an evaluation that reports this failure separately from WER.

6 Discussion

Evaluation implication. A contextual-ASR result should identify the source of its phrase list, prevent target leakage, freeze the decision rule before held-out scoring, and report supported recovery and unsupported additions together. WER alone cannot reveal whether context improved the intended phrase or introduced a different listed term.

Modeling implication. The remaining technical problem is utterance-level evidence. Source relevance and phrase-form risk help construct a safer list, while transcript distance limits large edits. Neither signal proves that a listed phrase matches the audio. Candidate-acoustic comparison is therefore promising only when genuine alternatives and calibrated scores are retained and evaluated on disjoint units.

Scope. The largest recorded positive changes occur in controlled opportunity sets and the 512-utterance English FLEURS diagnostic. These results come from earlier runs; their exact inputs, model revisions, and complete run lineage have not been jointly verified. The multilingual inventory is still undergoing release review, and the meeting evaluations show negative or abstaining transfer. Zero observed strict FPs are finite-sample counts under one matcher, not population guarantees. The reported runs cover Whisper-family models and do not establish cross-model generalization. The study uses automated reference-based audits and does not include human listening judgments.

Ethics. User context can contain private names, contacts, documents, or community vocabulary. A contextual recognizer can leak this information or amplify bias in the supplied list. Public release files exclude private sources and must document provenance and consent while distinguishing unsupported insertions from near-reference spelling or punctuation variants. Evaluation by language and locale is necessary because an aggregate score can hide concentrated failures.

7 Conclusion

LexiAdapt evaluates accepted contextual edits using phrase provenance, a reference-independent rule, rare-word and phrase recovery, unsupported additions, coverage, WER, and latency. The matched modern-model comparison remains unexecuted; the Academic/ICSI route lacks verified corpus, audio, checkpoint, and preflight identities. Earlier run summaries lack complete input and run lineage. They cannot establish model gains, benchmark readiness, or deployment safety. A useful contextual edit needs source and audio support, not merely lower aggregate WER.

AI Use Statement

Yes. AI tools were used to aid or polish the writing and for retrieval and discovery, including finding related work. They also assisted in planning and checking the evaluation, developing and testing analysis code, and interpreting recorded results. Details of these uses are described in the paper.

Reproducibility Statement

Sections 3 and 4 specify the protocol and study populations, and Appendix A gives additional diagnostics. The MIT-licensed supplementary source archive, `lexiadapt-source.zip`, contains the implementation, tests, and a synthetic example. It does not reproduce empirical scores, which require audio, model checkpoints, and run outputs that are not supplied.

REFERENCES

Jean Carletta. Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007. doi: 10.1007/s10579-007-9040-x. URL https://doi.org/10.1007/s10579-007-9040-x.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*, 2022. URL https://arxiv.org/abs/2205.12446.

ESPnet. espnet/owsm_v4_base_102M, 2025. URL https://huggingface.co/espnet/owsm_v4_base_102M. Exact inspected revision 1652139ae2605c13b50e1c17d31e0be85aa05c7c, accessed 2026-09-19.

Junteng Jia, Gil Keren, Wei Zhou, Egor Lakomkin, Xiaohui Zhang, Chunyang Wu, Frank Seide, Jay Mahadeokar, and Ozlem Kalinli. Efficient streaming LLM for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890853. URL https://doi.org/10.1109/ICASSP49660.2025.10890853.

Sakshi Joshi, Eldho Ittan George, Tahir Javed, Kaushal Bhogale, Nikhil Narasimhan, and Mitesh M. Khapra. Recognizing every voice: Towards inclusive ASR for rural bhojpuri women. In *Interspeech 2025*, pp. 4243–4247, 2025. doi: 10.21437/Interspeech.2025-2734. URL https://arxiv.org/abs/2506.09653.

Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, Yangyang Shi, Jay Mahadeokar, Julian Chan, Yuan Shangguan, Christian Fuegen, Ozlem Kalinli, Yatharth Saraf, and Michael L. Seltzer. Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion. In *Interspeech 2021*, pp. 1772–1776, 2021. doi: 10.21437/Interspeech.2021-1566. URL https://www.isca-archive.org/interspeech_2021/le21_interspeech.html.

Microsoft. microsoft/VibeVoice-ASR-Streaming-7B, 2026. URL https://huggingface.co/microsoft/VibeVoice-ASR-Streaming-7B. Exact inspected revision 60d858b518b4e19d404af3737f848fc185b30177, accessed 2026-09-19.

Ashish Mittal, Sunita Sarawagi, and Preethi Jyothi. RECAST: Retrieval-augmented contextual ASR via decoder-state keyword spotting, 2025. URL https://aclanthology.org/2025.findings-emnlp.203/.

NVIDIA. nvidia/parakeet-tdt-0.6b-v2, 2026. URL https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2. Exact inspected revision ae9ad07059c7c739ffaf932226a8fe64ae2620b0, accessed 2026-09-19.

OpenAI. openai/whisper-large-v3-turbo, 2024. URL https://huggingface.co/openai/whisper-large-v3-turbo. Exact inspected revision 41f01f3fe87f28c78e2fbf8b568835947dd65ed9, accessed 2026-09-19.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://doi.org/10.1109/ICASSP.2015.7178964.

Yifan Peng, Muhammad Shakeel, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin, and Shinji Watanabe. OWSM v4: Improving open whisper-style speech models via data scaling and cleaning. In *Interspeech 2025*, pp. 2225–2229, 2025. doi: 10.21437/Interspeech.2025-1062. URL https://arxiv.org/abs/2506.00338.

Qwen. Qwen/Qwen3-ASR-1.7B-hf, 2026. URL https://huggingface.co/Qwen/Qwen3-ASR-1.7B-hf. Exact inspected revision bcd2b5b7f32b480ab5790554cfa8347f246a14f3, accessed 2026-09-19.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Muhammad Shakeel, Yosuke Fukumoto, Chikara Maeda, et al. CALM: Joint contextual acoustic-linguistic modeling for personalization of multi-speaker ASR, 2026. URL https://ieeexplore.ieee.org/document/11463102/.

Yui Sudo, Yosuke Fukumoto, Muhammad Shakeel, Yifan Peng, Chyi-Jiunn Lin, and Shinji Watanabe. DYNAC: Dynamic vocabulary-based non-autoregressive contextualization for speech recognition. In *Interspeech 2025*, pp. 2215–2219, 2025. doi: 10.21437/Interspeech.2025-129. URL https://arxiv.org/abs/2506.00422.

Guangzhi Sun, Chao Zhang, and Philip C. Woodland. Minimising biasing word errors for contextual ASR with the tree-constrained pointer generator, 2022. URL https://arxiv.org/abs/2205.09058v2.

Guangzhi Sun, Chao Zhang, and Philip C. Woodland. Graph neural networks for contextual ASR with the tree-constrained pointer generator, 2024. URL https://arxiv.org/abs/2305.18824v1.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5905–5921, 2021. doi: 10.18653/v1/2021.naacl-main.472. URL https://aclanthology.org/2021.naacl-main.472/.

A Extended Evaluation Results

This appendix examines the saved outputs from the same earlier runs as the main paper. The exact input bytes, sample selections, commands, seeds, model revisions, and complete run-to-result lineage are not yet jointly verified. These scores describe only their named sets and cannot support a model-effect estimate.

A.1 Reference-grounded controls

Table 3 records lower WER with prompting than without it for the tested Whisper checkpoints on both datasets. The audited source bundle does not link these rows to their original predictions and exact run identity, so this pattern does not independently verify the executed implementation. Because target transcripts supply positive phrases in this control, it also does not establish no-leakage generalization. Table 4 records no copied distractor under prompting for the tested models in a 1,024-utterance distractor-only control; its original predictions have not been joined to the displayed values. A separate saved text-repair summary on the named distractor-only lists records two false-positive entities for each tested Whisper checkpoint on FLEURS and three or four on LibriSpeech. The repair rule modifies a stored transcript after decoding and was not timed. These counts expose a different failure path from the prompted rows in Table 4; without original input, prediction, and scorer joins, they do not establish a paired safety difference or a mechanism for the insertions.

Table 3: Scaled real-audio contextual ASR diagnostic from earlier runs. Each setting aggregates 16 32-utterance shards for 512 utterances per dataset. Phrase metrics use phrases extracted from each reference transcript and supplied distractors, so this table records a prompt signal but does not independently verify the executed pipeline or manually check every phrase. Distractor FP rate divides false-positive listed phrases by supplied distractor instances. Repair is text-level post-processing over the no-context hypothesis. Its additional runtime was not timed, so a dash marks unavailable RTF and does not denote zero cost.

Dataset	System	WER (%)↓	Phrase recall↑	Phrase F1↑	False pos.↓	Distractor FP rate↓	RTF↓
FLEURS en_us	Whisper tiny, none	13.45	0.6943	0.8196	0	0.0000	0.0173
FLEURS en_us	Whisper tiny, prompt	9.30	0.8955	0.9449	0	0.0000	0.0173
FLEURS en_us	Whisper tiny, repair	10.91	0.9414	0.9688	2	0.0002	–
FLEURS en_us	Whisper base, none	8.14	0.7432	0.8527	0	0.0000	0.0201
FLEURS en_us	Whisper base, prompt	6.34	0.9160	0.9562	0	0.0000	0.0207
FLEURS en_us	Whisper base, repair	6.08	0.9678	0.9826	2	0.0002	–
FLEURS en_us	Whisper small, none	6.60	0.7988	0.8882	0	0.0000	0.0341
FLEURS en_us	Whisper small, prompt	5.03	0.9258	0.9615	0	0.0000	0.0348
FLEURS en_us	Whisper small, repair	4.86	0.9766	0.9872	2	0.0002	–
LibriSpeech clean	Whisper tiny, none	6.56	0.7857	0.8800	0	0.0000	0.0192
LibriSpeech clean	Whisper tiny, prompt	4.52	0.9378	0.9679	0	0.0000	0.0201
LibriSpeech clean	Whisper tiny, repair	5.12	0.9347	0.9642	4	0.0005	–
LibriSpeech clean	Whisper base, none	5.18	0.8153	0.8983	0	0.0000	0.0245
LibriSpeech clean	Whisper base, prompt	3.41	0.9561	0.9776	0	0.0000	0.0261
LibriSpeech clean	Whisper base, repair	3.80	0.9571	0.9766	3	0.0004	–
LibriSpeech clean	Whisper small, none	4.01	0.8418	0.9141	0	0.0000	0.0427
LibriSpeech clean	Whisper small, prompt	2.68	0.9561	0.9776	0	0.0000	0.0424
LibriSpeech clean	Whisper small, repair	2.78	0.9643	0.9803	3	0.0004	–

Table 4: Distractor-only prompt stress diagnostic from earlier runs. Context lists contain only phrases not supported by the utterance transcript. Distractor FP rate divides copied distractor phrases by supplied distractor phrase instances. The recorded prompted outputs contain no copied distractor phrase for the tested Whisper models on this 1,024-utterance set. A separate text-repair summary on the named lists records nonzero false-positive entities; neither arm's original predictions and run lineage have been jointly checked.

Dataset	System	WER (%)↓	False pos.↓	Distractor FP rate↓	RTF↓
FLEURS en_us	Whisper tiny	11.91	0	0.0000	0.0164
FLEURS en_us	Whisper base	8.16	0	0.0000	0.0200
FLEURS en_us	Whisper small	6.45	0	0.0000	0.0319
LibriSpeech clean	Whisper tiny	6.26	0	0.0000	0.0196
LibriSpeech clean	Whisper base	5.00	0	0.0000	0.0257
LibriSpeech clean	Whisper small	3.66	0	0.0000	0.0432

A.2 No-leakage retrieval and distractor stress

Table 5 gives the recorded neighbor-derived LibriSpeech comparison across Whisper sizes, conditional on a reference-supported phrase entering the opportunity set. Table 6 replaces broad distractors with orthographically confusable phrases. Its recorded point estimates show lower WER and higher recall under prompting for all three checkpoints in this selected set.

Text-only phrase repair substitutes a listed phrase for a similar transcript span after decoding. Its saved Whisper-small reports on the 123-utterance list record 3.53% WER, 0.9542 recall, and six false-positive entities, versus 3.32%, 0.9804, and zero in the prompt entry of Table 5. On the distinct 195-utterance list, repair records 3.08% WER, 0.9467 recall, and 14 false-positive entities; the prompt entry in Table 6 records 3.15% WER, 0.9877 recall, and zero exact false positives. The repair reports use a 0.85 similarity threshold, but their original inputs and scorer are not jointly verified with the prompt rows. Their false-positive field is not reconciled to the strict hard-distractor matcher. These aggregates do not establish a paired effect or a uniformly favorable tradeoff.

The same named 195-utterance hard-distractor list has a saved transcript-distance calibration for the Whisper-small prompt-and-rerank candidate. Its selected threshold of 0.25 accepts one changed transcript out of 195; the report records 3.15% baseline WER, 3.06% gated WER, and no false-positive entities. This tiny coverage is a different outcome from the ungated prompt comparison in Table 6, even though the rounded baseline WER matches. The threshold was selected on this list, and original predictions, scorer, and run identity have not been jointly verified. Neither the small WER difference nor the zero observed count establishes a useful held-out safety operating point.

Table 5: Neighbor-derived LibriSpeech opportunity set. Candidate phrases come from same-chapter neighbors and need two source occurrences. Admission requires a reference-supported phrase, so all 123 selected utterances have an opportunity; the list contains 153 supported phrases and 356 distractors. The executed ranking mode and result lineage are unresolved.

Model	Context	WER (%)↓	Phrase recall↑	Phrase F1↑	False pos.↓
Whisper tiny	none	9.56	0.5882	0.7407	0
Whisper tiny	prompt	6.35	0.8954	0.9448	0
Whisper base	none	7.81	0.6340	0.7760	0
Whisper base	prompt	5.21	0.9346	0.9662	0
Whisper small	none	6.28	0.6667	0.8000	0
Whisper small	prompt	3.32	0.9804	0.9901	0

Table 6: Hard-distractor retrieval stress test on LibriSpeech clean. The list contains 195 utterances, 244 reference-supported phrases, and 968 similar-looking non-reference distractors drawn from the same neighboring-utterance phrase pool. Exact FP counts listed phrases found in the output but absent from the reference; strict hard FP counts those hits among hard distractors after near-reference variants are removed.

Model	Context	WER (%)↓	Recall↑	F1↑	Exact FP↓	Strict hard FP↓
Whisper tiny	none	8.27	0.6721	0.8039	0	0
Whisper tiny	prompt-8	5.45	0.9590	0.9770	1	0
Whisper base	none	6.64	0.7090	0.8297	0	0
Whisper base	prompt-8	4.18	0.9713	0.9854	0	0
Whisper small	none	5.25	0.7377	0.8491	0	0
Whisper small	prompt-8	3.15	0.9877	0.9938	0	0

The previously reported paired intervals are withheld: the audited source bundle does not link the original prediction records to this comparison, and the asserted utterance-level resampling does not resolve the unknown speaker grouping. These point estimates remain diagnostic until their prediction-level lineage is reconstructed.

A.3 Rule-audited FLEURS controls

Table 7 shows that a larger prompt is not uniformly safer or more effective in this diagnostic. Table 8 shows the coverage cost associated with increasing the source-count threshold. Moving from one source occurrence to two leaves the reported strict-error count at zero but reduces retained supported phrases from 85 to 41 and recall from 0.8557 to 0.4227. These results do not establish the safety of either threshold outside the tested set, and the missing run lineage prevents selecting a principal operating point from this comparison.

The audited list was built by finding lexicon phrases in each target reference and placing those matches before distractors. Its first four positions cover all 97 supported phrases in the saved list summary. This is a reference-aided construction, not a test-time source selector. A seeded random-order list over the same named 512-utterance FLEURS cohort puts only 19 of those phrases in the first four positions. In retained Whisper-small grouped reports, the four-phrase prompt records 6.17% WER and 0.9691 recall with the audited order, versus 6.51% and 0.7629 with random order. At twelve phrases, the random-order report records 6.03% WER, 0.8969 recall, and one strict unsupported insertion; its four-phrase counterpart has no strict insertion under the saved audit. The earlier list builder also used Python's process-dependent string hash in its distractor shuffle, so its numeric seed alone cannot recreate the old phrase lists. These are sensitivity diagnostics, not evidence that random ordering is a deployable selector or that order alone caused the score difference: exact input identities, predictions, scorer, and run lineage have not been jointly verified. The prompt-budget result depends on reference-aided list construction and phrase position, not only on budget size.

A separate saved Whisper-medium report on this named 512-utterance set records 5.24% WER under three-phrase prompting. Two-candidate reranking records 5.01% WER and one listed false positive; its source-guarded counterpart keeps 5.01% WER and records zero. A leave-shard-out policy summary chooses that guard in all 16 shards, changing 41 transcripts and accepting one phrase addition, versus 42 changed transcripts and two additions under ungated reranking. Four-candidate source-guarded reranking records 4.83% WER, but its saved real-time factor is 0.4179, versus 0.2327 for the two-candidate guard. These report-level comparisons do not verify matched executed inputs, timing, scorer, or policy-selection independence. They expose a possible safety and cost tradeoff, not a validated acoustic operating point.

A.4 Multilingual gate and transfer checks

Table 9 compares the no-context anchor, ungated contextual candidate, and fixed transcript-distance gates on the four-locale development inventory. Table 10 tests the same feature family across independently constructed inventories. In the reverse direction, open-context acceptance produces four unsupported non-target phrases; restricting acceptance to target phrases removes them but also lowers recall. This asymmetry cautions against treating one threshold as universally portable.

Table 7: Whisper-small safety controls on the 512-utterance rule-audited FLEURS set. A larger prompt improves WER but loses phrase recovery and introduces one strict unsupported insertion; selection by source support and phrase-form risk improves both utility measures without an observed strict insertion.

System	WER (%)↓	Recall↑	Precision↑	F1↑	Strict↓
No context	6.60	0.7216	1.0000	0.8383	0
Prompt, 4 phrases	6.17	0.9691	1.0000	0.9843	0
Prompt, 12 phrases	5.96	0.9485	0.9684	0.9583	1
Source+risk, beam 8	5.58	0.9794	1.0000	0.9896	0

Table 8: Source-support threshold on the rule-audited FLEURS set. Higher thresholds retain fewer supported phrases even when the observed strict unsupported-insertion count remains zero.

Support threshold	Decoder	WER (%)↓	Recall↑	Strict errors↓	Kept↑	RTF↓
≥ 1 source	Prompt	4.91	0.8557	0	85	0.0861
≥ 1 source	Prompt-rerank	4.91	0.8557	0	85	0.3494
≥ 2 sources	Prompt	5.03	0.4227	0	41	0.0852
≥ 2 sources	Prompt-rerank	5.03	0.4227	0	41	0.3489

Table 9: Whisper-medium acceptance frontier on the four-locale development inventory. These results are diagnostic, with unresolved run lineage and transfer safety, and do not establish a validated operating point.

Operating point	WER (%)↓	Recall↑	F1↑	Strict errors↓	RTF↓
No-context anchor	13.15	0.6726	0.8042	0	0.1318
Ungated prompt	12.68	0.9351	0.9650	1	0.1344
Distance-gated prompt	12.69	0.9351	0.9665	0	0.2662
Beam-4 distance gate	12.69	0.9351	0.9665	0	0.5938

Table 10: Cross-inventory transfer of transcript distance. S matches each distractor's writing system to its target; R independently ranks distractors by source evidence. Open-context rules may accept any listed phrase, while target-only rules may accept only designated targets. Transfer is asymmetric.

Threshold fit → test	Acceptance scope	WER (%)↓	Recall↑	F1↑	Strict errors↓	Non-target↓
S → R	Open context	12.76	0.9469	0.9727	0	0
S → R	Target only	12.91	0.8407	0.9135	0	0
R → S	Open context	35.36	0.7802	0.8751	4	4
R → S	Target only	35.96	0.5892	0.7415	0	0

A.5 Second phrase-inventory diagnostic

On the separate script-matched, four-locale development inventory, a saved fixed-threshold sweep exposes a particularly severe Hindi tradeoff. For Whisper base, the ungated prompt reports 140.56% WER, 0.1750 phrase recall, and eight listed false positives; the distance-and-repetition rule at threshold 0.75 records 155.46% WER, zero recall, and zero false positives. For Whisper tiny, the same reports give 214.97% WER, 0.0500 recall, and 32 false positives without the gate, versus 180.88% WER, zero recall, and zero false positives with it. Whisper small retains 0.4125 recall under the gated rule, down from 0.4375 with the ungated prompt, while its listed false positives fall from three to zero. Thus the zero-count entries for base and tiny also recover none of the named phrases. These are repeated conditions on the same inventory, not independent locale replications. Their original input, prediction, scorer, and executed-rule lineage remains incomplete; the above-100% WER values require insertion and normalization replay before any model-quality interpretation. The sweep cannot establish a safe threshold for new Hindi speech.

The same script-matched inventory has separate saved Whisper-medium four- and eight-candidate reranking reports. For its 256 Hindi utterances, the reported WER rises from 40.97% with four candidates to 41.51% with eight; phrase recall falls from 0.8375 to 0.8250, while strict unsupported insertions rise from three to four. Reported real-time factor rises from 0.8080 to 2.0672. The four- and eight-candidate summary files name the same source-list group, but their original predictions, scorer, checkpoint revision, and executed-input identity have not been joined. This cost-and-error pattern is an adverse development diagnostic, not a verified paired effect of candidate width or an operating-point recommendation.

The separately constructed pending-source phrase lists use the same 1,280 four-locale FLEURS utterances, not an independent audio population. In their retained Whisper-medium summaries, raw four-candidate reranking raises Hindi WER from 40.73% without context to 61.07% and records three strict unsupported insertions. Eight-candidate reranking records 65.75% Hindi WER and the same three insertions; its sample-weighted real-time factor is 1.2304 across locales, compared with 0.1827 for eight-candidate prompting without reranking. These values do not isolate an acoustic-selection benefit. For four-candidate prompting on this inventory, the fixed transcript-distance rule records 12.74% aggregate WER, 0.9440 recall, 0.9712 F1, and no strict insertion. A risk-capped learned acceptance rule records 12.78% WER, 0.9410 recall, 0.9696 F1, and no strict insertion. The learned rule therefore does not improve the fixed rule on the recorded metrics. The acoustic-score-only search leaves one strict insertion; a Hindi-shard search including transcript distance selects distance in all eight held-out shards. These are development diagnostics with incomplete original prediction and executed-run lineage, not a prospective test or a demonstrated safety guarantee. All 15,360 phrase occurrences in this inventory have source-type metadata, but none has the source-count field required by the source-count filter. Consequently, these lists do not test that filter under its stated input contract.

A.6 Eight-meeting reranking and selective hybrid diagnostic

This validation analysis asks whether acoustic-only reranking recovers document-supported terms without sending broad document phrases to the decoder. The saved Whisper-medium, eight-candidate rerank summary covers 6,003 utterances in eight AMI/QMSum meetings and records 114.35% word error rate (WER), 0.8974 listed-phrase recall, three false-positive entities, and 3.6088 real-time factor. The no-context summary for the named inventory records 23.13% WER and the same recall. The saved feature counts reproduce the rerank summary's 376 shard WER values: 138 utterances with at least 100 recorded edit errors contribute 80.23% of its total, and 107 of those references contain one word. These high-error records occur in all eight meetings. The original rerank prediction and evaluator files were not found in the checked project storage, so this arithmetic replay cannot determine whether the long outputs came from decoding, mismatched inputs, or another producer failure. The raw rerank is therefore an adverse score from the saved run, not a verified paired effect or a successful contextual-recovery path.

A separate post-decoding calculation selects among the stored anchor, guarded prompt output, and reranker output using the rules in Section 3.3. Its 6,003 paired feature entries match in shared population identity and anchor counts and reproduce the saved summary. The anchor and guarded-prompt WERs are 23.13% and 23.12%; a gate accepting 791 reranker changes that add no listed phrase

records 22.29%, and the combined gate records 22.28% WER with five supported contextual additions. All four rows retain the anchor's one strict false positive; the combined rule adds none. The combined rule is a selective calculation, not the raw 114.35% rerank arm. Its five contextual additions occur in four meetings, and its WER is worse than the anchor on one of eight meetings (17.82% versus 18.10% on ES2011d). The saved records do not bind producer and checkpoint identity, executed input bytes, the full distinct-candidate population, or group-aware uncertainty to these scores. Consequently, the lower recorded error under selective non-context edits is a diagnostic lead, not evidence of a transferable acoustic contextual-ASR benefit.

A.7 Document-context diagnostics

Table 11 joins the source-swap control with the 129-meeting transfer. The recorded rule with a retained-candidate phrase-support threshold abstains after a swap, while the rule without that threshold accepts four unsupported additions. A targeted check found multiple distinct scored alternatives for two original-source acceptances and all six swapped-source ablation acceptances, but a singleton for the third original-source acceptance. The full run lineage remains unverified, so this contrast does not establish an acoustic-selection effect. On disjoint meetings, raw proposals are mostly unsupported; the permissive gate remains unsafe, while the strict gate accepts only two edits.

Table 11: Document-context transfer audit. Raw rows show proposals; gate rows show accepted edits; dashes are inapplicable. The three 129-meeting prompt/gate rows were computed over all 86,888 inspected samples. Unsupported edits precede the 0.82 near-reference exclusion. Strict-FP increase is relative to each no-context baseline, so zero does not imply zero total.

Evaluation	System	Proposed edits (raw)	Accepted edits (gate)	Unsupported edits↓	Strict FP rise↓
Validation, 8 meetings	Raw prompt	9	–	4	3
Validation, original source	Source+distance+support	–	3	0	0
Validation, swapped source	Source+distance+support	–	0	0	0
Validation, swapped source	Source+distance only	–	6	4	1
129-meeting prompt/gate	Raw prompt (earlier diagnostic)	385	–	276	246
	Validation-selected gate	–	132	47	26
	Strict source+distance gate	–	2	0	0

The three 129-meeting entries form one earlier empty-context replay whose prompt/gate values use all 86,888 inspected samples; they do not describe the current source-plus-acoustic route. The separate source-plus-acoustic packet has 1,668 of 5,431 bundles complete, corresponding to 26,688 of 86,888 sample outputs, with 3,763 bundles missing and zero context entries. Its governing generation remains unresolved across conflicting summaries, so its gate remains incomplete and fail-closed. A later repaired-input run has paired output filenames for all 5,431 bundles but no verified sample-level score reconstruction. In a five-bundle sample, each of 72 prompted predictions stores one candidate after deduplication despite requesting eight return sequences; raw generation and duplicate counts are unavailable. The stored predictions do not test an acoustic comparison among alternatives. A separate 1,799-example context-bearing sibling replay has 14 context-addition cases; it does not complete or widen either unresolved result. Table 12 reports the frozen-policy, meeting-disjoint ICSI replay for two checkpoints; it is disjoint from rule-selection meetings but is not an official QMSum test split.

Another attempted 129-meeting comparison built context phrases from neighboring meeting documents. Its retained partial Whisper-medium analysis pairs only 41 of 5,431 planned input bundles (656 utterances). On this selected subset, prompting raises WER by an absolute 12.97% and adds four strict unsupported insertions relative to no-context decoding. All four recorded insertion cases contain the same recurring meeting-role phrase. A reference-aided analysis of these 656 utterances marks 25 helpful prompt changes without a gained context-phrase hit, but all four harmful context-

gain cases add a strict unsupported insertion. This split helps explain the tested guard's behavior: it accepts none of the prompted outputs, avoiding those recorded insertions without showing useful contextual recovery. The analysis groups outcomes using references after decoding; these groups are not test-time labels.

The retained guard sweep on the same 656 utterances exposes the cost of relaxing that decision. The no-context anchor records 18.54% WER. A rule accepting one changed prompt records 18.66% WER and one additional strict insertion; accepting three records 19.12% WER and three additional strict insertions. Removing the context-hit requirement accepts 591 changed prompts with 20.70% WER and one additional strict insertion; a looser boilerplate-only rule accepts 624 with 29.50% WER and three additional strict insertions. These exploratory settings trade abstention for recorded errors on this partial set; they do not establish a selected operating point for new meetings.

The source files and original predictions are not currently bound to the analysis, the executed checkpoint and policy are unknown, and the subset may overlap the other meeting analyses. The absence of matching files in the current result directory does not erase the older report or verify it. A full input-to-prediction and executed-run join is needed before these values can support a population-level conclusion.

A.8 Context-bearing sibling diagnostic

The context-bearing subset contains all 1,799 Product utterances with at least one reference-supported phrase under its selection rule; the source-corpus split labels within it do not establish a separately held-out test set. Its retained feature records disagree with the Product records on no-context errors for 15 matched utterances and prompted errors for 199. Each disagreement has a corresponding pair of different stored transcripts. The named input shards agree on recorded audio URI, reference, speaker, audio ID, and split for the 199 prompted cases, but the Product shards have empty phrase lists while the Product predictions mark prompting as applied and record 7–16 available context phrases. The number actually passed to the prompt is not recorded by this field. This metadata conflict leaves the executed input unknown. The sibling therefore cannot replace the Product result or isolate a context effect without the complete input, prediction, and run-lineage joins.

Another large-v3-turbo hybrid summary names a 1,799-example context-bearing set. Its feature exports join to Product IDs, but the summary has not been joined by sample ID and executed policy to the separate 14-addition source/acoustic analysis. It reports 17.87% WER for the no-context anchor, 17.32% for the prompt-only rule, 18.02% for the non-context reranker rule, and 17.46% for the hybrid. The hybrid accepts 121 prompt-context additions and 210 non-context reranker changes; the prompt-only rule accepts the same 121 additions. Both report zero strict FPs in the saved summary. Relative to prompt-only, the hybrid has an absolute WER difference of +0.14% and lower phrase recall (0.9446 versus 0.9493). The recorded acceptance check fails its minimum of 80,000 examples and its requirement for WER improvement in at least half the meetings (55 of 127). The four summary entries can be recomputed from the retained prompt and reranker feature exports, but the original input, checkpoint, prediction, and executed-policy identities remain unbound. Neither an "audit-ready" filename-pair count nor this derived-feature replay validates those underlying runs. These numbers therefore describe an adverse diagnostic within a reference-selected subset, not a prospectively held-out test or a verified acoustic-selection effect. They do not resolve the larger Product or incomplete source-plus-acoustic studies.

Table 12: Frozen-policy, meeting-disjoint ICSI replay (4,133 utterances; all expected prompt and reranker run pairs present). WER triples are anchor/prompt/hybrid: the prompt uses the source-and-distance gate, while the hybrid adds acoustic-only non-context edits. Recall and strict-FP pairs are anchor/hybrid, with strict FPs counted after the 0.82 exclusion. Edit pairs are supported contextual/non-context reranker edits. Both transfer gates fail.

Model	WER (%)↓ A/P/H	Recall↑ A/H	Strict FP A/H	Edits context/rerank	Gate
Whisper medium	62.65/62.65/64.94	0.7143/0.7143	6/5	0/1,305	Fail
Whisper large-v3-turbo	60.09/60.09/64.88	0.6429/0.7143	5/5	1/1,109	Fail

Whisper medium stays at recall 0.7143 to 0.7143 and accepts no supported contextual addition; large-v3-turbo moves from 0.6429 to 0.7143 and accepts one supported prompt-context addition. The 1,305 and 1,109 non-context reranker edits worsen WER, so both hybrids still fail the overall gate. Passing requires lower WER, higher recall and F1, supported contextual and acoustic-only edits, no unsupported contextual addition, and no strict-FP increase. Thus source and audio alignment do not identify a useful utterance-level edit here.

B Audit Definitions and Additional Scope

Phrase matching. "Exact FP" counts listed phrases found in the output but absent from the normalized reference. A hit is near-reference when its maximum Python `SequenceMatcher` ratio over normalized reference windows of phrase length ±1 reaches 0.82. The strict count removes these hits; stored score records retain raw and variant counts. This fixed test differs from token-level Levenshtein gate distance.

Table 13 makes the scorecard denominators explicit. In particular, edit coverage measures selector behavior; it must not be confused with the fraction of scheduled runs that completed.

Table 13: Mapping from contextual edits to reported audit quantities. All counts are accumulated before rates are computed.

Quantity	Event counted	Denominator
Supported addition	Added supplied phrase present in the reference	Accepted additions
Raw unsupported addition	Added supplied phrase absent from the reference	Accepted additions
Near-reference variant	Raw unsupported addition with similarity at least 0.82	Raw unsupported additions
Strict unsupported addition	Raw unsupported addition excluding near-reference variants	Accepted additions
Total strict FP	Strict listed phrase in the final output, including any inherited anchor event	Evaluated outputs
Accepted-edit coverage	Accepted output that differs from its audio-only anchor	Changed contextual candidates

Selection and release. Neighbor-derived candidates exclude the target utterance, but the earlier rarity and balanced ranking modes included it in corpus-frequency counts. The builder now excludes the target from those counts; without the executed settings and input lists, the change does not validate earlier results. Reference-supported-phrase admission selects the opportunity cohort, while references score the frozen outputs and cannot enter the acceptance rule. The reference-grounded control supplies target phrases and tests prompting, not reference-independent list construction. Transfer requires complete coverage, a serialized validation rule, and disjoint evaluation; later tuning is diagnostic. The multilingual inventory remains developmental pending final source and audio checks. Public release files exclude private sources and review records.