

When Does User-Scoped Glossary Injection Recover ASR-Corrupted Domain Terms in Meeting Summarization? Separating Term Recovery from Summary Utility

Gihwan Kim¹, Hyoungjun Lim¹, Yoonseo Chung¹, Heeyong Park¹,
Changhyeon Jeong¹, Seunghee Ma¹, Joohye Park¹, Junghwan Kim¹

¹LOTTE INNOVATE

{gihwan.kim, hyoungjun.lim, yoonseo.chung, huiyong.bak,
Changhyeon.Jeong, seunghee.ma, joohyepark, jhwan_kim}@lotte.net

Abstract

Prompt-level glossary injection can recover ASR-corrupted domain terms, but its effectiveness depends sharply on glossary scope. In enterprise meeting summarization, a black-box STT system passes only a 1-best transcript to the summarizer; its errors on loanwords, acronyms, and domain terms can replace the canonical forms needed downstream. Motivated by the keyword-input UX of a deployed Korean meeting summarization service, we evaluate training-free prompt-level injection of an annotation-derived meeting-scoped glossary proxy without transcript rewriting. On 30 public Korean meetings processed offline by operational STT, 27% of annotated terms are corrupted under normalized matching, and a no-glossary summarizer recovers only 1.3% of these corrupted terms. Meeting-scoped injection raises term recovery to 53.8%, with 89.2% of output keywords grounded in the transcript. The effect collapses under broader scopes: on the same 45 terms, meeting-level recovery is 66.7% versus 6.7% for a 2,669-term full-dictionary proxy. Coverage-controlled size sweeps show this collapse on Qwen3-8B, and three EXAONE-3.5 points reproduce it, indicating prompt-level over-biasing. A four-annotator human evaluation does not establish improved summary utility; we therefore treat domain-term recovery as an operational dimension, not a summary-preference proxy. We contribute a public-data evaluation protocol and deployment analysis of domain-term recovery under black-box STT constraints.

1 Introduction

Automatic meeting summarization is increasingly deployed as a pipeline that feeds the 1-best speech-to-text (STT) transcript into an LLM summarizer. In enterprise settings, the STT component may be an external API or a separately operated non-public model; downstream summarization cannot access decoder states, word lattices, or alternative hypothe-

ses. It can only process the received text and control prompt construction.

This black-box setting hides a failure mode that is weakly reflected by overall transcription quality. Summaries often need the canonical forms of terms that anchor tasks, decisions, and follow-up actions: transliterated loanwords, Korean renderings of acronyms such as GIS, NPO, and TF, and administrative or technical terminology. When such forms disappear, users must inspect the transcript or recording even if WER is low. Prior work similarly shows that NER on spontaneous-speech transcripts can fail even on correctly transcribed entity spans and degrades further when ASR inserts, substitutes, or deletes tokens within those spans (Szymański et al., 2023), and that transcript noise can propagate to summarization (Shapira et al., 2025; Koay et al., 2020).

Existing responses either correct entities or ASR errors at the transcript stage (Pusateri et al., 2025; Ghosh et al., 2025), or train summarizers to tolerate synthetic ASR errors (Binici et al., 2025). These approaches are valuable, but they add correction outputs that must be validated operationally or require retraining as domains and base models change. We instead study a training-free deployment pattern that attempts recovery at summarization time, without modifying the STT output or model parameters.

The pattern comes from the operational UX of a deployed enterprise meeting summarization service for Korean meetings. The service exposes two keyword-entry paths: account-level keywords for organization- or work-specific terms and meeting keywords for terms tied to a particular meeting. Service logic merges these inputs into a user-provided keyword signal, which can serve as a meeting-scoped source of candidate terminology. We formalize the controlled abstraction of this signal as *user-scoped glossary injection*: a glossary narrowed to the user or meeting context is provided to the summarizer while the transcript remains un-

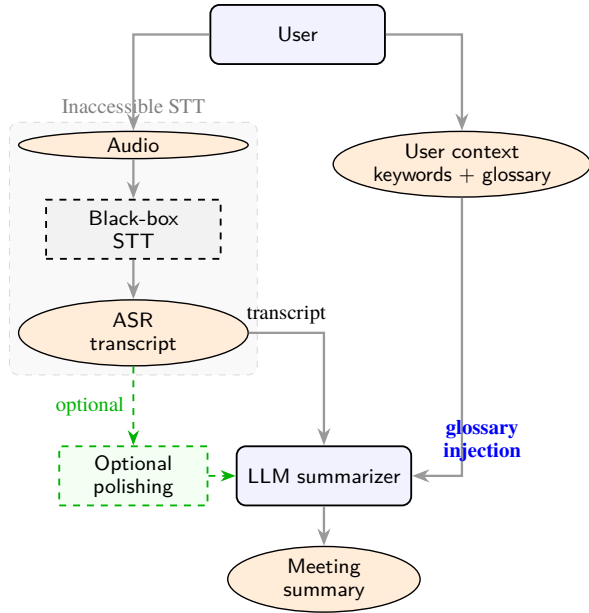


Figure 1: Simplified view of the deployed meeting summarization system. The 1-best transcript from a black-box STT system and user-scoped glossary context are injected at the summarization stage.

changed. To avoid using operational logs, we evaluate the same constraint and intervention point by applying an operational Korean STT system offline to public meeting audio with term annotations.

We do not claim that glossary injection improves overall summary preference. We ask when prompt-level injection recovers ASR-corrupted domain terms and when it fails. Our contributions are: (1) quantifying failed spontaneous recovery under black-box STT, where 27% of annotated Korean terms are corrupted under normalized matching and only 1.3% are recovered without a glossary; (2) isolating a training-free intervention that improves recovery in both an English neologism stress test (0.0%→53.3%) and the Korean public-audio setting (1.3%→53.8%); (3) showing a prompt-level size–accuracy trade-off: meeting-scoped glossaries work, but organization-scale lists collapse toward the no-glossary baseline even when they contain the relevant terms; and (4) separating domain-term recovery from overall summary preference, arguing that term surfacing should be monitored as a distinct operational dimension rather than treated as a proxy for general summary quality.

2 Related Work

Downstream propagation of ASR errors. Prior work shows that NER errors occur even on

correctly transcribed entity spans and increase further when ASR insertions, substitutions, or deletions occur within those spans (Szymański et al., 2023), that transcript noise propagates to downstream tasks including summarization, with named-entity repair among the more effective targeted transcript-cleaning strategies for transcript-level tasks (Shapira et al., 2025), and that domain terminology can substantially affect meeting-summarization performance in an extractive ICSI setting (Koay et al., 2020). These studies establish that the problem is real. Our work differs by measuring term corruption directly in an industry-constrained setting: we apply an operational STT system to public meeting audio with term-span annotations, and then evaluate a recovery intervention in the downstream summary output.

Recovery at the ASR stage. Retrieval-augmented entity correction (Pusateri et al., 2025) and generative ASR error correction with synthetic data and retrieval augmentation (Ghosh et al., 2025) modify the transcript itself. Pusateri et al. start from a possibly errorful textual ASR hypothesis, retrieve candidate entities, and use an adapted LLM for correction; they evaluate the corrected transcript primarily by WER. Their method therefore does not require decoder states, lattices, or input audio. Our distinction is the intervention and evaluation point: we leave the received 1-best transcript unchanged, add no transcript-correction stage, and measure canonical-term recovery in downstream summary fields.

Robustness at the summarization stage. MED-SAGE (Binici et al., 2025) uses LLM-generated ASR-error dialogues to retrain a summarizer, making it the closest prior work in problem setting: dialogue summarization under ASR errors with domain entities. In contrast, the deployment pattern studied here operates only through inference-time prompt injection. It therefore requires no synthetic training set, retraining cycle, or modification of the underlying summarization model.

Relation to contextual biasing. Contextual biasing provides ASR systems with dynamic context phrases or biasing lists to improve recognition of rare entities and user- or domain-specific terms. CLAS learns context n-gram embeddings jointly with ASR (Pundak et al., 2018), and recent work injects text context into pretrained LLM-based ASR models to learn contextualization (Lakomkin et al.,

2024). These methods, however, require access to the decoder, a biasing module, or the ASR training process. They are therefore difficult to apply when a deployed summarization system only receives the 1-best transcript from a commercial or internally operated black-box STT system. We move the intuition that “a term dictionary can guide recognition” from the ASR internals to the summarization prompt. Le et al. (2021) combine trie-based deep biasing and shallow fusion in streaming RNN-T systems and show that rare-word recognition is sensitive to biasing-list size when reference rare words are retained and 100–2,000 distractors are added; our scoping experiments test an analogous empirical sensitivity at the summarization-prompt level, without positing a shared mechanism.

3 Deployed System

The work starts from a deployed meeting summarization service implemented as a workflow graph for preprocessing, semantic segmentation, chunk-level summarization, global merging, structured-field extraction, and post-processing. Figure 1 is a simplified view of the two paths that matter here: the STT 1-best output and glossary injection. The central deployment constraint is that STT is a separate non-public black-box pipeline, so the summarizer receives only 1-best text and cannot access decoder states or alternative hypotheses.

The deployed UI provides two keyword-entry paths: account-level keywords for recurring organization- or work-specific terms, and meeting-specific keywords for terms to emphasize in a given meeting. Service logic merges them into one user-provided keyword signal, which can function as a meeting-scoped source of candidate terminology for the summarizer. In the controlled evaluation, we instantiate this signal as an annotation-derived meeting-level glossary proxy and inject it through the canonical-form domain-glossary block described in Appendix C, isolating term-form recovery rather than general topical focus. The resulting intervention is *user-scoped glossary injection*: the block acts only at summarization time and does not rewrite the transcript.

The deployed system also has an optional polishing module for readability, but it targets fluency rather than domain-term recovery. We exclude it from the controlled evaluation and leave its interaction with glossary injection to future work.

The experiments below do not evaluate opera-

tional traffic logs. Instead, they apply the operational STT system offline to raw audio from public meeting corpora, preserving the deployment constraint while fixing the data and annotations to public resources.

4 Problem Characterization: ASR Error Propagation

Before evaluating glossary injection, we characterize STT-error propagation by comparing a gold-transcript, no-glossary condition (C0) with an STT-transcript, no-glossary condition (C1). Full resources, summarizer, and statistics are described later; here we introduce the C0–C1 diagnostic. English experiments use audio from the ICSI meeting corpus (Janin et al., 2003), a public English meeting-speech dataset, with general-query reference summaries from QMSum (Zhong et al., 2021); Korean experiments use Meeting Speech Recognition Data by Major Domains (AI Hub 464) (NIA AI Hub, 2021), a public Korean meeting-speech corpus with domain-term annotations. The English axis measures field-level drops in term recovery, while the Korean axis measures domain-term corruption and spontaneous recovery failure under operational STT on public meeting audio. For English, $\text{TRR}_{\text{fix}}(y) = |\{t \in D : t \text{ appears in } y \text{ in canonical form}\}|/|D|$ uses a fixed denominator D , the union of glossary terms appearing in the reference summary or any condition output.

Quantitative analysis in English. We applied Whisper-large-v3 (Radford et al., 2023) to 59 ICSI meetings and summarized C0 and C1 with the same model. TRR_{fix} drops from 0.645 to 0.448, a 31% relative decrease over 57 valid meetings. Field-level drops are 0.031 for the abstractive summary, 0.089 for decisions, 0.175 for action items, and 0.362 for speaker notes (Table A2), showing that propagation is larger in fields that rely on named entities and domain terms.

Korean: operational STT on public audio. For Korean, we transfer the deployed STT condition to public data. From AI Hub 464, we select the 30 highest term-density meetings and run the audio through the operational Korean STT system offline. Among 824 annotated terms, 27% (225 terms) are corrupted under normalized substring matching; strict surface-form matching gives 53%. Corrupted terms include loanwords and acronyms

such as turnkey, GIS, NPO, and TF, plus administrative and technical terms such as heat meters, overseas business travel expenses, local finance grants, budget committees, and amendment motions.

Corruption may be invisible on public benchmarks. On a separate public Korean meeting-speech corpus, Meeting Speech (AI Hub 132 (NIA AI Hub, 2020), 10 meetings), we compare the operational STT system with a public general-purpose STT system (Whisper-large-v3, Korean setting). This is not an STT ranking; it shows that term corruption depends on the STT–vocabulary combination. Strong public-benchmark STT systems may therefore miss deployment conditions involving domain-specific terminology.

Summarization does not repair the errors by itself. Given only the corrupted transcript, the no-glossary summarizer (C1) outputs the canonical form for only 1.3% of corrupted terms (3/225). The following experiments therefore test whether glossary injection can improve on this low spontaneous-recovery baseline.

5 User-Scoped Glossary Injection

The method takes an STT transcript x and a user- or meeting-scoped glossary $G = \{(c_i, V_i)\}_{i=1}^k$, where c_i is the canonical output form and V_i is an optional set of aliases or misrecognized forms. The key property is scope: G is not an organization-wide dictionary by default, but a candidate set relevant to the meeting. The ablations widen this scope.

At inference time, we do not rewrite the transcript. We add G to the prompt as a separate *Domain glossary* block with a conditional instruction: if a term is mentioned in canonical or likely misrecognized form, the output should use c_i . This standardizes transcript-grounded mentions at generation time, without transcript correction or ASR re-decoding.

Keyword extraction field. The output schema contains a keywords field for domain terms and proper nouns, in addition to the abstractive summary, decisions, action items, and speaker notes. This field matches the deployed keyword feature and is necessary for evaluation, because sparse domain terms can be omitted from the summary even when correctly recovered.

Handling long meetings. When transcript plus glossary exceeds the context budget, the system

uses map–merge summarization: the glossary is injected at both chunk and merge stages, and its length is included in the chunking threshold.

Comparison conditions. C0 uses the gold transcript without a glossary as a clean-transcript reference. C1 uses the STT transcript without a glossary as the baseline. C2 uses the STT transcript plus G . All three conditions keep the summarizer, output schema, and map–merge procedure fixed; only C2 adds the glossary block. Thus, C0–C1 estimates propagation from STT corruption, while C2–C1 estimates glossary-based recovery. The design is training-free and leaves both transcript and model parameters unchanged.

6 Experimental Setup

Evaluation resources. English ICSI/QM-Sum (Janin et al., 2003; Zhong et al., 2021) is a reproducible auxiliary axis: we apply Whisper-large-v3 to 59 meetings to examine field-level propagation and quality side effects. The main Korean axis uses AI Hub 464, a public meeting-speech corpus with term-span annotations: from its validation split, we select the 30 highest term-density meetings (824 annotated terms, 25.3 hours) and apply an operational Korean STT system offline to public audio. We intentionally use the highest term-density meetings as a stress-test design: the goal is not to estimate average session-level summary quality over all meetings, but to obtain enough annotated domain-term events to measure corruption, recovery, and over-biasing under a black-box STT constraint. AI Hub 132 (10 meetings) is used to compare corruption patterns between operational and public STT systems and to calibrate the detector-specific thresholds of the deterministic baseline (Section 7.5).

Summarizer and decoding. The main experiments use Qwen3-8B (Qwen Team, 2025), served with vLLM (Kwon et al., 2023) in non-thinking mode (temperature 0.7, top- p 0.8, top- k 20, 32,768-token context). We repeat the three key size-sweep points with EXAONE-3.5-7.8B-Instruct (LG AI Research, 2024) using greedy decoding. All conditions keep the summarizer, output schema, and map–merge procedure fixed.

Glossary scope. The meeting-level glossary is built from each evaluation meeting’s term annotations: we repeatedly strip a fixed list of trailing particle and ending strings, discard forms shorter than

two characters, and deduplicate in first-occurrence order, yielding 16–113 canonical entries. The AI Hub 464 main-axis list contains no observed STT-error variants, and ASR output is not used to select entries or construct their forms. This is a clean proxy for the merged keyword signal (the account-level and meeting keywords combined in Section 3); because it is derived from each meeting’s gold annotations, it is an oracle-like candidate set that pre-solves which terms are relevant. The main experiment therefore evaluates the summarizer’s response to a given candidate set, not end-to-end glossary selection; we discuss this proxy gap and the unevaluated selection step in the Limitations. Domain-level glossaries contain annotation terms appearing in at least two training meetings within the same domain (140–386 terms across the five evaluated domains), while the full dictionary contains terms appearing in at least two training meetings across all domains (2,669 terms). Neither construction uses evaluation-split term annotations; they approximate department-level and organization-wide bulk injection, respectively.

Metrics and statistics. The main Korean metrics are corrupted-term recovery and keyword grounding. The recovery denominator is 225 (meeting, annotated-term-string) pairs whose canonical form does not occur as a substring of the meeting’s full STT transcript after normalization (lowercasing and removing characters other than alphanumerics and Hangul); recovery is scored by testing whether the canonical form occurs as a substring of the five output fields serialized and normalized the same way (regardless of field, context, or count). Grounding is the fraction of output keywords present in the gold transcript. English experiments use TRR_{fix} , whose denominator is the union of glossary terms appearing in the reference summary or any condition output. The primary uncertainty estimate is meeting-level paired bootstrap resampling (10,000 samples, 95% CI); term-level McNemar tests are auxiliary discordance statistics. The Korean small-scale human evaluation follows a blind pairwise protocol (Appendix D).

Availability. AI Hub 464/132 are available directly from AI Hub under its access procedures and data-use policy. That policy restricts third-party provision of the original data without approval; interpreting the scope of redistribution rights conservatively, we do not redistribute the AI Hub source data or its dataset-derived artifacts (operational-

Gold transcript (translated)	
...budget for the NPO Support Center ...	
Operational STT output	<i>ASR Error</i>
...budget for the MPO Support Center ...	
C1 output: no glossary	<i>Error Propagated</i>
keywords: [MPO Support Center , ...]	
C2 output: user-scoped glossary	<i>Term Restored</i>
keywords: [NPO Support Center , ...]	

Figure 2: Real Korean meeting example, translated for readability. The transliteration error is rendered as NPO → MPO: without glossary injection it propagates to keywords, while user-scoped glossary injection restores the canonical term.

STT transcripts, glossary and distractor files, filled runtime prompts, and model outputs). Implementation code is also not publicly released. The prompt template format is documented in Appendix C. Exact end-to-end reproduction of the reported generations is therefore unavailable, while the documented procedures (prompt format, decoding configuration, preprocessing, and scoring) are reproducible on independently obtained data.

7 Controlled Offline Evaluation and Results

7.1 Controlled Test: Neologism Stress Test

To reduce the confound that public-corpus terms may already be known to the LLM, we replace 15 ICSI glossary candidates with artificial neologisms. The selected items are uppercase acronym or proper-name-like terms appearing in the QM-Sum reference summary, clean transcript, and STT transcript. C0 receives canonical neologisms in the clean transcript; C1/C2 receive matched ASR-like corrupted forms in the STT transcript, with the remaining transcript text kept as the actual Whisper output. C1 recovers none of the neologisms (0.000), while C2 recovers 0.533 (C0: 0.667; Appendix Table A3). C0 is below 1.0 because clean-input summaries need not surface every term.

7.2 Offline Evaluation with Operational STT on Public Audio

Table 1 shows the main result. C2 raises corrupted-term recovery from 0.013 to 0.538 (+0.524, 95% CI [0.405, 0.663]; auxiliary McNemar $\chi^2 = 114.1$, $p < 0.001$).¹ Recovered terms include NPO, TF,

¹In a matched rerun, recovery is robust to the target definition: on a strict 190-target set (after removing within-meeting

Condition	Corrupted-term recovery (225)	Keyword grounding
C0 gold transcript (no glossary)	0.253 (57/225)	0.847
C1 STT (baseline)	0.013 (3/225)	0.692
C2 STT + meeting-level glossary	0.538 (121/225)	0.892

Table 1: Main Korean experiment over the 225 annotated terms corrupted by operational STT. C0–C1 measures propagation, and C2–C1 measures glossary-based recovery. For C2–C1, primary inference uses meeting-level paired bootstrap (+0.524, 95% CI [0.405, 0.663]); term-level McNemar is auxiliary ($\chi^2 = 114.1$, $p < 0.001$).

Budget Committee, and the Enforcement Decree of the Local Finance Act (Figure 2). C2 exceeding C0 does not mean better STT input: the metric measures whether canonical terms are surfaced in the output schema, and C0 is a clean-input reference rather than a term-surfacing upper bound. Likewise, grounding is an output-keyword precision measure rather than an input-quality upper bound; C2 maintains keyword grounding of 0.892, so broad glossary copying is not observed under meeting-relevant scope.

7.3 Scoping Ablation: Glossaries Must Be Narrow

Table 2 shows that scope is decisive. Domain-level and full-dictionary glossaries have lower coverage, but even for covered terms they lower recovery and grounding. On the matched 45-term subset covered by both glossaries, meeting-level recovery is 0.667 versus 0.067 for the full dictionary; one-sided recoveries concentrate on meeting-level scope (29 vs. 2 terms; 15 vs. 1 non-tied meetings).

To confirm that this contrast is not an artifact of the processing path, we re-compare after forcing both conditions through map–merge on all 30 meetings, equalizing per-meeting chunk counts (in the original path the full dictionary is chunked on all 30 meetings while the meeting-scoped glossary is a single call on 19). Meeting-level recovery is 112/225 and whole-dictionary recovery is 4/225 (meeting-level paired bootstrap 95% CI [0.357, 0.616]), so the contrast holds. This equalizes the discrete processing path, chunk boundaries,

duplicates and subsuming items), C2 recovers 106/190 versus 3/190 for C1 (meeting-level paired bootstrap 95% CI [0.409, 0.673]), and all six target definitions agree. Fuzzy-similarity and boundary audits flag at most 7/225 potential false positives and 40/599 potential false negatives, respectively; these are candidates for inspection, not confirmed errors.

Scope	Terms	Covered	Recall	Ground.
Meeting-level	16–113	225	0.538	0.892
Domain-level	140–386	17	0.176	0.525
Full dictionary	2,669	45	0.067	0.391

Note. All results use Qwen3-8B. **Covered** is the number of corrupted terms included in each glossary; **Recall** uses those terms as its denominator, which therefore differs across rows. **Ground.** denotes keyword grounding. The meeting-level glossary covers all corrupted terms by construction, whereas the broader glossaries use only training-split terms to avoid leakage from evaluation annotations. On the 45 terms covered by both the meeting-level and full-dictionary conditions, recall is 0.667 and 0.067, respectively.

Table 2: C2 glossary-scope ablation. The annotation-derived meeting-level condition approximates user- or meeting-scoped injection; domain-level and full-dictionary conditions approximate department- and organization-wide bulk injection, respectively.

and glossary-presentation counts; glossary-size-dependent prompt and attention load remains part of the treatment. Per-size processing-path statistics are in Appendix Table A4.

Coverage-controlled size sweeps give the same conclusion, within the evaluated pipeline (Figure 3). On the common size-sweep subset, adding irrelevant terms while keeping the relevant corrupted terms included drops Qwen3-8B recovery from 0.507 to 0.024 and grounding from 0.876 to 0.52; EXAONE-3.5 shows the same pattern. For Qwen, total keywords increase from 743 to 1,622, while ungrounded keywords rise from 92 to 778. Thus, over-biasing appears as list-copying rather than omission, consistent with contextual-biasing size–accuracy trade-offs (Le et al., 2021). This size sweep is the main deployment lesson: relevant terms being present in the prompt is not sufficient. Prompt-level glossary injection also requires relevance-based scoping; otherwise, the model shifts from recovering corrupted terms to copying or overusing unrelated glossary entries.

7.4 Robustness to User-Input Proxies

Because annotation-derived meeting glossaries are cleaner and more complete than real user-entered keywords, we use coverage and distractor perturbations as a proxy-gap check rather than as a new upper-bound experiment. This test asks whether the recovery gain survives when the scoped glossary is incomplete or contains irrelevant user-provided terms. With 50% coverage, overall recovery drops to 0.369, but recovery over injected terms remains 0.618. The 1x and 3x distractor settings occupy the

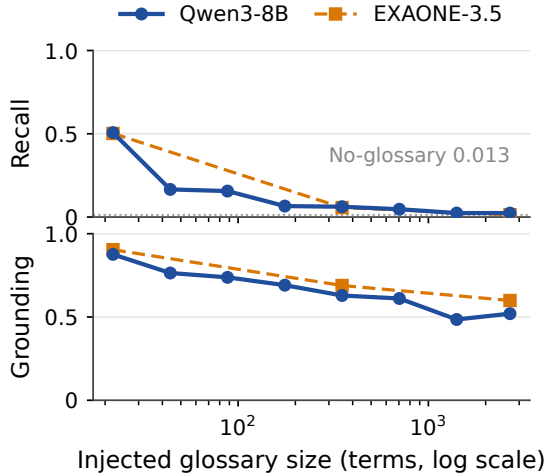


Figure 3: Glossary size–recovery trade-off. Top: corrupted-term recall; bottom: keyword grounding. Related terms are always included; only distractors increase. Solid circles=Qwen3-8B, dashed squares=EXAONE-3.5; recall drops 0.507→0.024 and 0.502→0.014.

left side of the size sweep: recovery is diluted, but grounding remains above 0.74.

7.5 Measurement and Quality

Scoring only the abstractive summary makes the C2–C1 difference nearly invisible (+0.017), because sparse domain terms may be omitted even when recovered. An extraction slot such as keywords is therefore needed to measure this behavior. Post-hoc jamo fuzzy matching of C1 keywords raises recovery only to 0.022, as relevant terms are often not emitted.

A generation-free string matcher also locates about half of the corrupted terms in the transcript (character matcher: 121/225; jamo matcher: 112/225). Detection, however, is a different task from placement: of the 112 terms detected by the jamo matcher, the no-glossary C1 emits the canonical form for just 3. Detector calibration, the detection–recovery cross-tabulation, and per-type examples are in Appendix B.

We check whether the measured effect is confined to the structured keyword output in two ways—removing the keyword field entirely, and scoring narrative fields only—and neither establishes narrative-level recovery (Appendix B); the effect is concentrated in the structured keyword output.

C2 shows no clear degradation on automatic QM-Sum quality metrics, including ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) (Appendix

Figure A1).

In a four-annotator blind pairwise evaluation on 10 Korean meetings (40 judgments; Appendix D), task usefulness favors C2 in 24 of 36 non-tied cases, but inter-annotator agreement is very low (Fleiss $\kappa = 0.074$) and judgments cluster by meeting and rater, so we draw no confirmatory conclusion in either direction. Clearly inappropriate terms or content appear in neither output in 29 of 40 judgments (Appendix D). Term errors are most visible in entity-dense structured fields such as action items and speaker notes (Table A2), while overall usefulness judgments can be driven by readability and content organization, making per-summary preference a weak instrument for term recovery. We therefore treat term recovery as a separate operational dimension and claim recovery and scoping conditions, not preference improvement.

8 Lessons for Deployment

First, scoping is the load-bearing design choice: systems should narrow glossaries by user or meeting context instead of injecting an organization-wide dictionary directly. Because large dictionaries can reduce both recovery and grounding even when relevant terms are present, candidates should be selected before injection. Second, because recovery may not track overall preference, the output schema should retain an explicit term-surfacing field such as keywords, making recovery visible and monitorable. Third, teams should regression-test recovery and grounding on their own STT, domain terms, and output schema, not only on generic benchmarks.

9 Conclusion

Scoping is the key condition for prompt-level glossary injection under black-box STT. Meeting-level injection recovers 53.8% of corrupted terms with 89.2% grounding, but broader lists collapse even when relevant terms remain in the prompt. These results make domain-term recovery a separate operational dimension, not a proxy for summary preference, and motivate merged keyword signals for training-free scoping.

Limitations

This study diagnoses domain-term recovery and glossary scoping under black-box STT, not general summary improvement. The annotation-derived meeting-level glossary is an oracle-like proxy for

the user-entered keyword signal (Section 3): we neither observe real user-keyword distributions nor evaluate end-to-end candidate selection. Coverage and distractor tests only partially bound this gap; retrieval-based or detect-then-supply selection remains future work.

The English and Korean-132 auxiliary glossaries are ASR-informed: 429 of 482 English entries (89.0%) were included only when corrupted, while the Korean-132 entries derive from clean-STT differences. They may therefore over-represent corruption and are treated only as diagnostics; neither contributes to the headline 0.538 result, whose AI Hub 464 glossary is STT-independent.

Model robustness is limited: Qwen3-8B provides the main results and full sweep, while EXAONE-3.5 repeats only three points. Glossary order and irrelevant-term sampling were not varied, and seeds were repeated only for the main condition. Absolute rates and list-copying behavior may therefore vary; the observed size-accuracy trade-off is a deployment diagnosis, not a universal law.

The operational STT evaluation covers Korean only; the English auxiliary path uses public Whisper-large-v3. Here, STT agnosticism means consuming the 1-best text without ASR internals, not demonstrated robustness across languages or recognizers; cross-language and cross-recognizer evaluation remains future work. The non-public Korean STT output distribution is not independently reproducible under the availability constraints in Section 6. We did not run a controlled correction-then-summarization comparison; it remains future work, and post-hoc canonicalization is diagnostic rather than a substitute.

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Leading Generative AI Human Resources Development (IITP-2026-RS-2026-25544188) grant funded by the Korea government (MSIT)

References

Kuluhan Binici, Abhinav Ramesh Kashyap, Viktor Schlegel, Andy T Liu, Vijay Prakash Dwivedi, Thanh-Tung Nguyen, Xiaoxue Gao, Nancy F Chen, and Stefan Winkler. 2025. MEDSAGE: Enhancing robustness of medical dialogue summarization to ASR errors with LLM-generated synthetic dialogues. *Pro-*

ceedings of the AAAI Conference on Artificial Intelligence, 39(22):23496–23504.

Sreyan Ghosh, Mohammad Sadegh Rasooli, Michael Levit, Peidong Wang, Jian Xue, Dinesh Manocha, and Jinyu Li. 2025. Failing forward: Improving generative error correction for ASR with synthetic data and retrieval augmentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2466–2482.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *Proceedings of ICASSP*.

Jia Jin Koay, Alexander Roustai, Xiaojin Dai, Dillon Burns, Alec Kerrigan, and Fei Liu. 2020. How domain terminology affects meeting summarization performance. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5689–5695.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, New York, NY, USA. Association for Computing Machinery.

Egor Lakomkin, Chunyang Wu, Yassir Fathullah, Ozlem Kalinli, Michael L. Seltzer, and Christian Fuegen. 2024. [End-to-end speech recognition contextualization with large language models](#). In *Proceedings of ICASSP 2024*, pages 12406–12410.

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. 2021. [Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion](#). In *Proc. Interspeech 2021*, pages 1772–1776.

LG AI Research. 2024. [EXAONE 3.5: Series of large language models for real-world use cases](#). *Preprint*, arXiv:2412.04862.

Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

NIA AI Hub. 2020. Meeting speech (AI hub dataset 132). <https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=132>.

NIA AI Hub. 2021. Meeting speech recognition data by major domains (AI hub dataset 464). <https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=464>.

- Golan Pundak, Tara N. Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao. 2018. [Deep context: End-to-end contextual speech recognition](#). In *Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 418–425.
- Ernest Pusateri, Anmol Walia, Anirudh Kashi, Bortik Bandyopadhyay, Nadia Hyder, Sayantan Mahinder, Raviteja Anantha, Daben Liu, and Sashank Gondala. 2025. Retrieval augmented correction of named entity speech recognition errors. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [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*, pages 28492–28518. PMLR.
- Ori Shapira, Shlomo E Chazan, and Amir David Nissan Cohen. 2025. Measuring the effect of transcription noise on downstream language understanding tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29978–30004.
- Piotr Szymański, Lukasz Augustyniak, Mikolaj Morzy, Adrian Szymczak, Krzysztof Surdyk, and Piotr Żelasko. 2023. [Why aren't we NER yet? artifacts of ASR errors in named entity recognition in spontaneous speech transcripts](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1746–1761.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations (ICLR)*.
- Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. [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 (NAACL)*.

A Additional Analyses and Examples

This appendix lists supporting materials referenced in the paper: resource details (Table A1), field-level propagation (Table A2), the neologism control (Table A3), per-size processing-path statistics (Table A4), and automatic quality checks (Figure A1).

ICSI/QMSum (English)	
Meetings	59 (57 valid for TRR)
Gold resources	ICSI transcripts + QMSum academic summaries
STT condition	Whisper-large-v3
Glossary source	STT-gold differences (diagnosis), neologism substitution (control)
Main role	Problem characterization and neologism control
AI Hub 464 (Korean, main experiment)	
Meetings	30 (highest term density among 95 validation meetings)
Gold resources	Term-span annotations (824 terms)
STT condition	Operational Korean STT applied offline to public audio
Glossary source	Meeting annotations (main experiment), training-split terms (scope ablation)
Main role	Main experiment and scope ablation

Table A1: Evaluation resources. AI Hub 464 provides utterance-level term-span annotations, allowing a gold denominator that is independent of model outputs. Separately, 10 AI Hub 132 meetings are used as an auxiliary set to compare term-corruption patterns between the operational STT system and a public STT system.

Field	C0	C1	Δ
Summary	0.529	0.498	0.031
Decisions	0.599	0.510	0.089
Action items	0.625	0.450	0.175
Speaker notes	0.786	0.424	0.362
Overall	0.645	0.448	0.197

Table A2: Effect of STT input on field-level term recovery across 57 English ICSI/QMSum meetings. Values are TRR_{fix} ; C0 uses gold transcripts, C1 uses Whisper-large-v3 transcripts, and Δ is C0–C1. The largest drop appears in the entity-dense `speaker_notes` field.

Condition	TRR_{fix}
C0 clean transcript (canonical forms)	0.667
C1 damaged STT (no glossary)	0.000
C2 damaged STT + glossary	0.533

Table A3: Neologism stress test on English ICSI (6 meetings, 15 substituted items). Uppercase acronym- or proper-name-like glossary candidates that appear in the QMSum reference summary, clean transcript, and STT transcript are replaced with artificial neologisms. C0 uses canonical neologism forms in the clean transcript, whereas C1/C2 use synthetic ASR-corrupted forms in the STT transcript, created through phonological confusions and word splitting; the rest of the transcript remains the actual Whisper output. Recovery in this table was measured using the earlier four-field output schema, before the keywords extraction field used for the main Korean experiments was introduced; these values are therefore not directly comparable to results obtained under the current five-field schema.

	Inj. size (entries)	Single-ctx (of 30)	Map chunks	Gloss. inj.	Chars/ chunk	$\approx$ Tok/ chunk
	22	19/30	1.60	2.60	16,863	14,053
	44	16/30	1.97	2.97	14,912	12,427
	88	15/30	2.03	3.03	14,466	12,055
	176	12/30	2.40	3.40	12,558	10,465
	352	10/30	2.63	3.63	11,358	9,465
	704	9/30	2.73	3.73	10,804	9,003
	1,408	5/30	3.03	4.03	9,333	7,777
	2,684	0/30	3.50	4.50	7,666	6,388
Domain (249)	11/30	2.50	3.50	11,981	9,984	
Whole dict. (2,669)	0/30	3.50	4.50	7,666	6,388	

Table A4: Per-size processing-path statistics (`chunk_size=40`). Injection size is the meeting median; the remaining columns are averages over the 30 meetings, with single-context meetings counted as 0 map chunks. Approximate token counts assume 1.2 characters per token; they are not exact tokenizer outputs. As the glossary grows, single-context meetings fall from 19/30 to 0/30 while map-chunk and glossary-presentation counts rise and transcript characters per chunk fall. The first eight rows are the coverage-controlled size sweep (Figure 3); the last two are the domain-level and whole-dictionary conditions of Table 2. This glossary-size-dependent load is part of the treatment (Section 7.3).

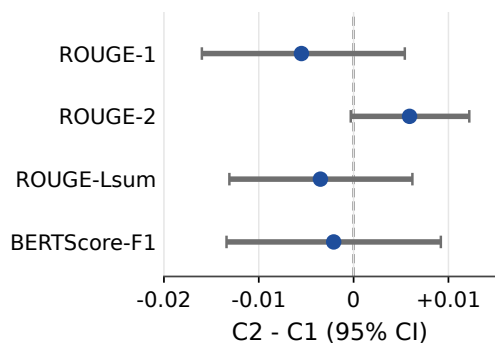


Figure A1: Automatic QMSum quality changes on 57 English meetings. Points are C2–C1 means; horizontal bars are meeting-level paired-bootstrap 95% confidence intervals. All four metrics include 0 (dotted line). ROUGE uses `rouge-score` with stemming, and BERTScore uses the English default `roberta-large` model with baseline rescaling.

B Error-Type Breakdown of Recovery and Mis-emission

The deterministic detection baseline summarized in Section 7.5 is detailed here. We fuzzy-match each glossary entry against the full operational-STT transcript, with detector-specific thresholds selected by held-out F1 on AI Hub 132 and applied unchanged. Of the 225 corrupted terms, a character matcher (threshold 0.72) detects 121 and a jamo matcher (0.84) detects 112. For the jamo matcher, strict precision is 1.0 on the meeting-scoped list only because no distractor can be emitted, and it falls to 0.038 as the list grows to a whole-dictionary scale—a model-independent echo of the scope collapse in Section 7.3. In the matched rerun used for cross-tabulation, 63 of the 120 terms C2 recovers (versus 121/225 in the main run) fall outside the matcher’s calibrated operating point (0.84), while of the 112 terms the matcher detects, the no-glossary C1 emits the canonical form for just 3.

We cross-tabulate detection against recovery over these 225 terms (Table B1). The examples below illustrate selected cells rather than an exhaustive coding under mutually exclusive type rules.

Not detected, recovered (63). This cell contains targets recovered by C2 but missed by the matcher at its calibrated operating point. Only 1 of the 63 has zero maximum jamo similarity and 57 are at or above 0.6, so this cell is not composed solely of cases with no string-level transcript evidence. Representative examples include transliter-

ated acronyms (NPO, TF, GIS, O2B) and numerals (year 2005, agenda item three), non-acronym loanwords (chutzpah, supply chain), and corrupted Korean administrative and technical terms (flexible-work system, the Citizens’ Public-Interest Activity Committee, investment- or contribution-funded institutions, non-tax revenue, polymerase). Lowering the threshold can detect some of these cases, but only at a cost in precision.

Detected, not recovered (55). Representative examples in this cell are particle-suffixed label fragments (annotation labels retaining Korean endings) and long descriptive labels. The matcher locates a similar transcript span, but under exact-match scoring an output is counted as not recovered unless it reproduces the full annotated surface form. For some of these cases, particles or endings retained in annotation labels contribute to exact-match failure, so the entire cell should not be read uniformly as model failure. Representative examples in the not-detected/not-recovered cell (50) are short two-to-three-syllable acronyms (PC, TPP), missed by the matcher and absent from the C2 output.

Mis-emission. A separate exploratory control examined cases where an irrelevant term is emitted in canonical form. When a same-sized mixed list containing meeting terms and irrelevant candidates is injected, 42 of the 121 retained corrupted terms are recovered while 22 of the 402 terms absent from the transcript are emitted (all in the keyword field), and 12 of those 22 are concentrated in a single science-topic meeting. These 12 comprise five multi-word science terms (carbon nanotube, Faraday’s law, steady-state cosmology, Cepheid variable, gamma-linolenic acid), four committee-type labels (one retaining a particle, plus the urban-planning, resident-autonomy, and administrative-autonomy committees), one multi-word political term (mixed-member proportional representation), one other short label, and one transliterated acronym (CPTPP). This control does not determine whether topical compatibility contributed to these emissions. Moreover, this mixed-list control leaves 5 normalized duplicate groups and 8 substring chains and does not implement a preregistered exclusion rule, so we read it only as exploratory; the robust over-biasing evidence is the size sweep in Section 7.3 (ungrounded keywords rising from 92 to 778).

Concentration in structured output. We detail the two checks from Section 7.5. First, after removing the keyword field and its merge instruction entirely and forcing both conditions through the same map–merge path, no reserved keyword key appears in any final merged raw output.² Recovery is then 5/225 with the glossary and 2/225 without it (95% CI [0.000, 0.029], exact McNemar $p = 0.25$), a difference that is not statistically significant. Second, under strict narrative scoring that also excludes the reserved keyword arrays the model leaks into speaker notes (in 2 of 30 meetings for C2), C2 recovers 7/225 versus C1’s 2/225: the meeting-level bootstrap interval is positive (CI [0.006, 0.040]), whereas the exact McNemar test does not reach 0.05 ($p = 0.0625$), so we treat this as weak, non-confirmatory evidence. Neither diagnostic establishes narrative-level recovery, indicating that the measured effect is concentrated in the structured keyword output.

	Recovered	Not rec.	Total
Detected	57	55	112
Not detected	63	50	113
Total	120	105	225

Table B1: Detection (rows: deterministic jamo matcher at the held-out threshold 0.84) $\times$ recovery (columns: C2 output) over the same 225 corrupted terms, in the matched rerun used for cross-tabulation (C2 recovers 120 here vs. 121 in the main run). The bold cell (63) holds terms C2 recovers but the matcher misses at its operating point; only 1 has zero maximum similarity and 57/63 are at or above 0.6. Of the 112 detected terms, no-glossary C1 emits the canonical form for only 3.

C Prompt Template

The Korean and English experiments used the same summarization template. The prompt asks the model to write the output in the language of the transcript. C0 and C1 omit the *Domain glossary* block below; C2 inserts the meeting-scoped glossary at that position.

Summarization prompt template

System. You are a careful meeting summarization assistant. Output exactly one valid JSON object. Do not duplicate the same content across fields. Do not infer or invent items not explicitly stated in the transcript.

User. [Meeting transcript]
{transcript}

²The archived artifacts for this ablation do not preserve the actual request prompts or chunk-level partial outputs.

[Task]

Read the entire meeting transcript and summarize it. Avoid omitting key information, but do not include content not present in the transcript. Write the output in the same language as the meeting transcript, preserving proper nouns and domain terms verbatim. Populate keywords with every domain-specific term, proper noun, and acronym that appears in the meeting, using the canonical spelling from the glossary whenever a term matches one.

[Domain glossary; C2 only]

The transcript may contain mis-transcribed forms of the terms below. When you mention such a term in the output, write the exact canonical spelling on the left of each entry, never the mis-transcribed form.

– {canonical term} (may be mis-transcribed as: {variant forms})

[Output JSON schema]

```
{
  "summary_items": [...],
  "decisions": [...],
  "action_items": [{"owner": "", "task": "",
    "due": ""}],
  "speaker_notes": {"participant": [...]},
  "keywords": [...]
}
```

[Required constraints]

Separate discussion summaries from confirmed decisions. Keep action items to follow-up work after the meeting. If a field has no supported content, return an empty list. Output JSON only, with no Markdown fences.

When a meeting exceeds the context budget, the template is applied separately to each transcript chunk. The merge prompt then asks the model to deduplicate per-chunk JSON summaries into one consolidated meeting summary. The glossary block is injected at both the chunk and merge stages.

D Korean Small-Scale Human Evaluation Protocol

Purpose. The Korean small-scale human evaluation is not intended to replace automatic term-recovery metrics. It serves as an auxiliary check for task usefulness and clearly inappropriate terms or content in the setting where a user reads the summary. In particular, Q2 is not a fine-grained groundedness judgment against the gold transcript; keyword grounding in the main text remains the primary safety metric.

Sample and blinding. The evaluated outputs are C1 (STT, no glossary) and C2 (STT with a meeting-level glossary) from the main Korean experiment. We select 10 meetings with many corrupted terms, but do not show the corrupted-term list to the annotators. This avoids contaminating human Q2 with the same target list that C2 is designed to recover, which would make the human judgment largely du-

plicate automatic corrupted-term recall. For each meeting, the two outputs are randomly assigned to A/B, and the mapping from A/B to C1/C2 is hidden until aggregation.

Questions. Each annotator answers two questions for each meeting. Q1 asks which summary is more useful for understanding the meeting and tracking follow-up work; the choices are A, B, and tie. Q2 asks which output contains terms, keywords, or content that is clearly inconsistent with the meeting domain or with the internal context of the summary; the choices are A, B, both, and none. The instructions tell annotators to choose *none* when an issue is ambiguous or appears to be only a spelling variation.

Aggregation and interpretation. For Q1, we recover the A/B key to C1/C2 and report C2 preference, C1 preference, ties, and the C2 preference rate excluding ties. For Q2, we report whether clearly inappropriate terms or content appear only in C1, only in C2, in both outputs, or in neither output. The result is interpreted as auxiliary evidence about practical usefulness and clearly inappropriate terms or content, not as an independent validation of the automatic recovery gain from C2.

Result. Four annotators judged the 10 meetings (40 judgments in total). After recovering the A/B key, Q1 task usefulness favors C2 in 24 cases, C1 in 12, with 4 ties, so the C2 preference rate excluding ties is $24/36 = 0.667$. However, inter-annotator agreement is very low (Fleiss $\kappa = 0.074$; only 2 of the 10 meetings are unanimous), and the raters split two-to-two (two favor C2, two favor C1). A pooled, tie-excluded sign test gives $p \approx 0.065$, but this value is descriptive only because judgments are clustered by meeting and rater. For Q2, clearly inappropriate terms or content appear only in C1 in 3 cases, only in C2 in 5 cases, in both outputs in 3 cases, and in neither output in 29 cases. Because judgments cluster by meeting and rater, we do not use the pooled counts as confirmatory evidence, and we read this result as a limited-sample observation that does not support a confirmatory conclusion in either direction.