

Evaluating Medical Terminology Changes in Clinical Post-ASR Error Correction

Fabian Schambeck^{*†} 

Michael Pucher^{*} 

^{*}Recognosco GmbH, Vienna, Austria

[†]Centre for Translation Studies, University of Vienna, Vienna, Austria

Abstract—Generative error correction for Automatic Speech Recognition (ASR) output is commonly evaluated using aggregate transcription metrics, most notably Word Error Rate (WER). In this work, we fine-tune and evaluate five pretrained language models spanning encoder-decoder and decoder-only Transformer architectures for post-ASR error correction of clinical dictation transcripts. To examine correction behavior beyond aggregate metrics, we introduce a terminology transition analysis framework that tracks whether medical terms are preserved, recovered, degraded, or remain unresolved between uncorrected and corrected transcripts. Fine-tuned models improved both overall transcription quality and medical terminology performance, with WER reductions from 4.58% to as low as 3.57% and positive Net terminology outcomes across models. However, Net gains were small, and models differed in their balance between terminology recovery and degradation. For documents containing terminology errors, WER improvements were moderately associated with Net terminology outcomes, indicating that aggregate transcription improvements are related to, but do not fully characterize, the effects of post-ASR correction on medical terminology.

Index Terms—Post-ASR Error Correction, Clinical NLP, Medical Terminology, Terminology-Aware Evaluation, Encoder-Decoder Models, Large Language Models

I. INTRODUCTION

Automatic Speech Recognition (ASR) systems are increasingly integrated into clinical documentation workflows to improve documentation efficiency and reduce manual transcription effort [1]. However, recognition errors may persist, particularly for medical terminology. Previous work has reported instances of elevated recognition error rates for such terminology, including drug names and diagnoses [2], [3]. In clinical settings, such errors are of particular concern because they may alter the intended meaning of a transcript and affect its clinical interpretation [2].

Moreover, the impact of recognition errors extends beyond errors in clinical documentation when ASR outputs are used by downstream Natural Language Processing (NLP) tasks, such as Named Entity Recognition (NER) [4], summarization [5], and automated clinical documentation generation, where language models generate clinical notes from ASR transcripts of patient encounters [6]. Errors affecting medical terminology may alter the information available to downstream systems and propagate through subsequent processing stages [7].

Recent work has demonstrated that generative post-ASR error correction can improve transcription quality through direct inference-time correction of ASR outputs [8], [9], as well as through the generation of pseudo-labels for subsequent

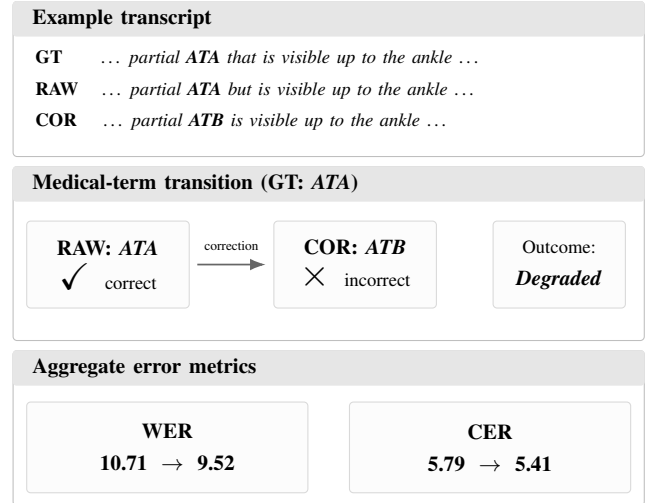


Fig. 1. Illustrative example from our corpus of a degraded medical-term transition. GT is the physician-corrected reference, RAW the uncorrected ASR output, and COR the corrected output by Qwen2.5-7B-FT. The medical term ATA (Anterior Tibial Artery), correct in RAW, is changed to ATB in COR despite improved document-level WER and CER.

model training [10]. Performance in post-ASR error correction is typically evaluated using aggregate error metrics such as Word Error Rate (WER) and Character Error Rate (CER) [11]–[13]. While these metrics assess overall transcription quality, terminology-level precision, recall, and F1 score can provide a more targeted assessment of medical terminology. Such terminology-level metrics summarize performance relative to the reference transcript, but do not characterize how the correctness of individual terminology instances changes from the ASR output to the corrected transcript. Erroneous terms provide opportunities for correction, while already-correct terms may be degraded through overcorrection [13]. Examining terminology with respect to its correctness before and after correction therefore provides a complementary view of model correction behavior. In particular, models may differ in how they balance the recovery of erroneous terms against the degradation of already-correct terms.

Figure 1 illustrates the distinction between aggregate transcription quality and terminology-specific correction behavior. Gains in document-level WER and CER may coincide with the degradation of individual medical terms, which motivates a focused analysis of terminology changes during correction.

In this work, we investigate generative post-ASR error correction for clinical dictation using a proprietary corpus of real-world British English speech documents. Because privacy, latency, and infrastructure constraints can make local deployment desirable in clinical settings, we compare established, locally deployable language models spanning encoder-decoder and decoder-only Transformer architectures, model scales, and training configurations. We evaluate correction performance at both the overall transcription and medical-terminology levels and focus on terminology changes during correction. Specifically, our contributions are as follows:

- We adapt and evaluate five small- to medium-scale Transformer language models for generative post-ASR error correction of clinical dictation transcripts, spanning encoder-decoder and decoder-only architectures.
- We investigate identity sampling as a mechanism for controlling correction behavior and optimize the identity sampling ratio for each model through hyperparameter search using a combined WER and CER objective.
- We introduce a terminology transition analysis framework that distinguishes preserved, recovered, degraded, and unresolved terms based on their correctness before and after correction.
- We analyze post-ASR error redistribution at both the transcript and medical-terminology levels and compare correction patterns across models.
- We show that WER improvements are related to, but do not fully characterize, the balance between recovered and degraded medical terms, as measured by their Net difference.

II. BACKGROUND

Modern end-to-end ASR systems achieve high transcription accuracy, yet residual recognition errors remain, particularly under challenging acoustic conditions and distribution shifts [14]. Post-ASR error correction has emerged as an approach to addressing such errors, but its evaluation in specialized domains such as healthcare requires consideration of medical terminology in addition to aggregate transcription quality.

A. Generative Post-ASR Error Correction

Generative post-ASR error correction has recently been explored using a range of pretrained and instruction-tuned Transformer-based language models and adaptation strategies [9], [11], [12], [15]. Encoder-decoder models have commonly been employed for supervised text transformation tasks due to their explicit sequence-to-sequence formulation [8], [12], [16], whereas decoder-only Large Language Models (LLMs) have more recently been applied to such tasks in instruction-following settings [9], [11]. In both settings, pretrained language models are adapted to ASR error correction using task-specific supervision derived from paired erroneous and reference transcripts.

Traditional post-ASR correction operates on single transcript hypotheses as a text-only post-processing step [8], [17]. Alternative approaches exploit N-best ASR hypotheses

through reranking or generative refinement [9], [15], or incorporate speech-derived information alongside textual ASR output [18]. In this work, we focus on text-only correction of single ASR hypotheses.

B. Medical Terminology and Evaluation Practices

Evaluation of post-ASR error correction systems commonly relies on transcript-level aggregate metrics, particularly WER [8], [11], [12], [15]. Clinical and biomedical ASR research has employed specific metrics for evaluating medical terminology and entities. These include Entity Error Rate (EER) [16], medical-specific WER and CER alongside medical-entity recall [3], and entity-level precision, recall, and F1 scores for biomedical entity extraction from ASR transcripts [4]. Related work has also examined the relationship between ASR and NER errors and highlighted limitations of F1 for evaluating NER on conversational transcripts [19]. To the best of our knowledge, prior work has not explicitly tracked individual medical-terminology instances relative to a reference transcript to examine changes in correctness from the ASR output to the corrected transcript.

III. EXPERIMENTS

A. Dataset

We conduct our experiments on a proprietary corpus of real-world British English clinical physician dictations collected from 2,319 speakers under varying channel and noise conditions. Each sample consists of an ASR transcript produced by a single DNN-HMM-based production ASR system, hereafter referred to as the *raw transcript* (RAW), and a physician-corrected reference transcript, referred to as the *ground truth* (GT). GT transcripts correspond to the final physician-corrected documentation and were not specifically constructed as independently audio-verified references for this study. Therefore, differences between RAW and GT may also reflect general physician revisions in addition to recognition errors. Given a correction model m , the RAW transcript is transformed into a model-specific *corrected transcript*, denoted COR_m . For evaluation, GT is treated as the reference transcript against which both RAW, serving as the uncorrected ASR baseline, and COR_m are compared.

The corpus comprises 70,081 dictations spanning a broad range of medical specialties and clinical domains, including radiology, pathology, general practice, and various other areas, with radiology accounting for the majority of samples. Across all splits, RAW transcripts contain an average of 86.7 ± 43.8 words per dictation and approximately 6.08 million words in total. The dataset was partitioned at the document level into training, validation, and test sets of 67,081, 1,500, and 1,500 dictations, respectively. Speaker-disjoint partitions and deduplication of recurring report templates were not enforced. All clinical data were collected and processed in accordance with applicable data-protection requirements, including the General Data Protection Regulation (GDPR). Due to the proprietary and sensitive nature of the clinical data, the corpus and fine-tuned model checkpoints are not publicly available.

TABLE I
PRETRAINED LANGUAGE MODELS USED FOR MODEL ADAPTATION.

Architecture	Model	Parameters	Type	Adaptation
Encoder-Decoder	T5-base	220M	Base	Full FT
	T5-large	770M	Base	Full FT
	FLAN-T5-base	250M	Instruction	Full FT
Decoder-Only	Qwen2.5-3B-Instruct	3B	Instruction	LoRA
	Qwen2.5-7B-Instruct	7B	Instruction	LoRA

B. Experimental Setup

We compare five Transformer-based models for post-ASR error correction. T5-base, T5-large [20], and FLAN-T5-base [21] were fully fine-tuned on the training set, whereas Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct [22] were adapted using Low-Rank Adaptation (LoRA) [23]. The resulting model configurations are summarized in Table I. For the instruction-tuned FLAN-T5 and Qwen2.5 variants, zero-shot performance was evaluated prior to adaptation.

Model inputs were formatted according to the respective model families. T5-base and T5-large directly mapped RAW transcripts to physician-corrected GT transcripts using the sequence-to-sequence formulation. For FLAN-T5-base, a task instruction was prepended to the RAW transcript; Qwen2.5 models used a chat-style prompt comprising a fixed system message specifying correction behavior and a user message containing the RAW transcript:

FLAN-T5-base prompt:

Input: Correct any ASR transcription errors in the following English medical report. Preserve all numbers, abbreviations, and medical terms. Output only the corrected text.

<RAW>

Output: <COR_m>

Qwen2.5 prompt:

System Input: You are a medical transcription editor. Each user turn is an ASR transcript of an English medical report. Correct transcription errors only.

Rules:

- Fix words the ASR got wrong (wrong word, missing word, spurious word)
- If the text is already correct, copy it exactly as written
- Do not paraphrase, reword, or change meaning
- Preserve all numbers, abbreviations, and medical terms unless clearly wrong
- Output only the corrected transcript, nothing else

User Input: <RAW>

Output: <COR_m>

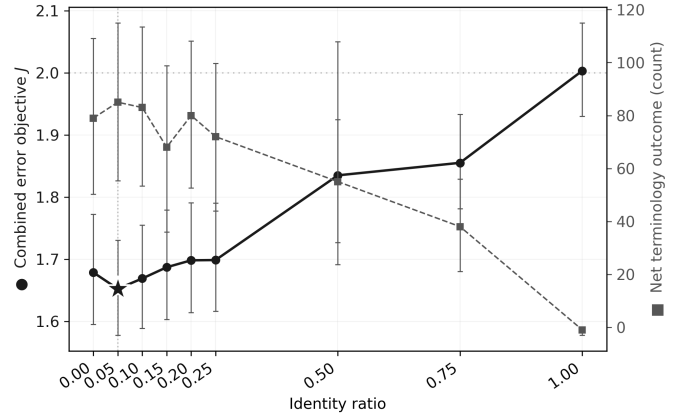


Fig. 2. Effect of the identity sampling ratio on T5-large. Shown are the normalized combined error objective J and the Net terminology outcome (Recovered – Degraded), with 95% bootstrap confidence intervals. The horizontal reference line marks the RAW baseline ($J = 2$), and the star marks the lowest observed J at an identity ratio of 0.05.

Model adaptation used AdamW with linear warmup and learning-rate decay, bf16 mixed precision, gradient checkpointing, and global gradient clipping at 1.0. Encoder-decoder models were fully fine-tuned for up to five epochs with a sequence length of 512, whereas Qwen models were adapted using LoRA for up to three epochs with a sequence length of 2048 and gradient accumulation to an effective batch size of 16. Validation was performed after each epoch, with early stopping at a patience of three epochs for encoder-decoder models.

Hyperparameters were optimized separately for each model using Optuna over 20 trials with a TPE sampler. During hyperparameter optimization and final checkpoint selection, validation performance was evaluated using the normalized combined error objective

$$J = \frac{\text{WER}_{\text{COR}}}{\text{WER}_{\text{RAW}}} + \frac{\text{CER}_{\text{COR}}}{\text{CER}_{\text{RAW}}}, \quad (1)$$

where lower values of J indicate better performance.

Identity sampling was incorporated into training, with the sampling ratio treated as a hyperparameter. A proportion of correction examples (RAW→GT) was replaced with identity mappings (GT→GT), leaving the total training set size unchanged. The identity sampling ratio was optimized within [0.0, 0.5]. Figure 2 illustrates its effect for T5-large in a separate identity-ratio sweep, with the lowest observed value of J at a ratio of 0.05.

The remaining optimized hyperparameters were learning rate, per-device batch size, warmup ratio, and weight decay. Learning rates were sampled log-uniformly from model-specific ranges and batch sizes from model-specific categorical sets. Warmup ratio and weight decay were sampled uniformly from [0.03, 0.15] and [0.0, 0.1], respectively. For the Qwen2.5 models, LoRA rank, scaling factor, and dropout were sampled from {8, 16, 32, 64}, {16, 32, 64, 128}, and uniformly from [0.0, 0.15], respectively.

During inference, all models used greedy decoding (beam size = 1), with maximum generation lengths of 512 and 1024 tokens for encoder-decoder and decoder-only models, respectively. All experiments were conducted on a single NVIDIA RTX A6000 GPU with 48 GB VRAM.

C. Medical Terminology Identification

To enable terminology-specific evaluation, we constructed a fixed medical terminology inventory independent of the correction models. The inventory was derived from a proprietary British English lexical dictionary used in a production clinical setting and separate from the model training data. We obtained fastText word embeddings [24] and clustered the vocabulary using a Gaussian Mixture Model (GMM). The Bayesian Information Criterion (BIC) selected 28 clusters from a candidate range of 10 to 30. All clusters were manually inspected and assigned semantic labels based on their predominant lexical content. Nine clusters primarily containing anatomical structures, diagnoses, medications, procedures, abbreviations, and clinical descriptors were designated as medical. The union of their single-word vocabulary items constituted the final inventory of 12,781 terms.

D. Evaluation Metrics

Overall transcription quality was evaluated using document-level WER and CER, macro-averaged across documents. We additionally report improvements relative to the uncorrected RAW baseline, defined as $\Delta\text{WER} = \text{WER}_{\text{RAW}} - \text{WER}_{\text{COR}}$ and analogously for CER, such that positive values indicate improvement. Medical terminology quality was evaluated using GT-anchored, micro-averaged precision, recall, and F1 derived from word-level alignments between GT and the evaluated transcripts. These token-aligned metrics are position-sensitive and account for substitutions, deletions, and insertions within GT terminology spans. Because terminology introduced outside these spans does not contribute false positives, we additionally report mention-level terminology precision, recall, and F1 using position-independent surface-form multiset matching after lowercasing and whitespace normalization, under which unmatched terminology mentions in the evaluated transcript contribute false positives.

To analyze terminology-specific correction behavior, individual GT terminology occurrences were tracked by position in the word-level alignment, and the corresponding aligned term in RAW and COR was identified. Repeated occurrences of the same term were therefore tracked separately by position. A terminology instance was considered correct if the aligned term matched the GT term after lowercasing and removing non-alphanumeric characters; otherwise, it was considered incorrect. The test set contains 23,772 GT medical-term occurrences across 1,498 of the 1,500 test documents, with a mean of 15.87 ± 10.65 occurrences per document; the remaining two documents were excluded from GT-anchored terminology analyses.

Each terminology instance was assigned to one of four transition categories: *preserved* (correct in RAW and COR),

		COR	
		Correct	Incorrect
RAW	Correct	Preserved	Degraded
	Incorrect	Recovered	Unresolved

Fig. 3. Terminology transition matrix. Rows indicate whether RAW matches GT, and columns indicate whether COR matches GT. Terms are classified as preserved, recovered, degraded, or unresolved.

recovered (incorrect in RAW, correct in COR), *degraded* (correct in RAW, incorrect in COR), or *unresolved* (incorrect in both RAW and COR). Figure 3 shows the resulting transition matrix.

Because the terminology transition analysis is anchored to medical terms occurring in GT, medical terms introduced by RAW or COR outside GT terminology spans are not captured by the transition categories. Recovery rates were computed relative to terminology instances that were incorrect in RAW, whereas degradation rates were computed relative to terminology instances that were correct in RAW. Net rates were computed relative to all GT terminology occurrences.

The transition categories complement terminology-level precision, recall, and F1 by distinguishing the recovery of erroneous terminology from the degradation of previously correct terminology. Terminology transition outcome was additionally summarized as

$$\text{Net} = \text{Recovered} - \text{Degraded}, \quad (2)$$

where positive values indicate that more terminology instances were recovered than degraded. Net therefore characterizes the balance between terminology recovery and degradation rather than overall terminology accuracy.

To investigate the relationship between global transcription improvements and terminology transition outcomes, we computed Spearman rank correlations between per-document WER improvements and per-document Net scores. This analysis was restricted to the 419 documents containing at least one GT terminology occurrence that was incorrectly recognized in RAW, ensuring that each included document provided an opportunity for terminology recovery.

To analyze error redistribution, we compared corpus-level error counts between RAW and COR. At the word level, total errors comprise substitutions, deletions, and insertions; at the terminology level, only substitutions and deletions within GT terminology spans were included, consistent with the GT-anchored analysis. Relative changes were calculated with respect to RAW as $100 \times (E_{\text{COR}} - E_{\text{RAW}}) / E_{\text{RAW}}$, such that negative values indicate fewer errors.

Unless otherwise stated, 95% confidence intervals were estimated using non-parametric percentile bootstrap resampling

TABLE II

OVERALL TRANSCRIPTION AND TOKEN-ALIGNED MEDICAL TERMINOLOGY PERFORMANCE. WER AND CER ARE MACRO-AVERAGED ACROSS DOCUMENTS; Δ WER AND Δ CER DENOTE CHANGES RELATIVE TO RAW, WITH POSITIVE VALUES INDICATING IMPROVEMENT. TERMINOLOGY P/R/F1 ARE MICRO-AVERAGED. VALUES ARE POINT ESTIMATES $\pm$ 95% BOOTSTRAP CI HALF-WIDTHS. ALL FINE-TUNED MODELS SIGNIFICANTLY IMPROVE WER AND CER RELATIVE TO RAW, WHEREAS ALL ZERO-SHOT MODELS SIGNIFICANTLY DEGRADE BOTH METRICS ($p < 0.001$).

Model	Document-level transcript quality				Token-aligned terminology quality		
	WER	CER	Δ WER	Δ CER	P	R	F1
RAW	4.58 $\pm$ 0.20	2.63 $\pm$ 0.13	–	–	98.05 $\pm$ 0.23	97.54 $\pm$ 0.26	97.79 $\pm$ 0.24
T5-base-FT	3.83 $\pm$ 0.22	2.30 $\pm$ 0.16	+0.75 $\pm$ 0.14	+0.33 $\pm$ 0.11	98.35 $\pm$ 0.20	97.75 $\pm$ 0.25	98.05 $\pm$ 0.22
T5-large-FT	3.75 $\pm$ 0.22	2.25 $\pm$ 0.15	+0.83 $\pm$ 0.14	+0.38 $\pm$ 0.09	98.44 $\pm$ 0.20	97.89$\pm$0.24	98.17$\pm$0.21
FLAN-T5-base-FT	3.77 $\pm$ 0.21	2.26 $\pm$ 0.13	+0.81 $\pm$ 0.11	+0.37 $\pm$ 0.06	98.45$\pm$0.19	97.85 $\pm$ 0.23	98.15 $\pm$ 0.20
Qwen2.5-3B-FT	3.70 $\pm$ 0.20	2.23 $\pm$ 0.13	+0.88 $\pm$ 0.12	+0.40 $\pm$ 0.07	98.43 $\pm$ 0.20	97.79 $\pm$ 0.23	98.11 $\pm$ 0.21
Qwen2.5-7B-FT	3.57$\pm$0.21	2.16$\pm$0.14	+1.01$\pm$0.12	+0.47$\pm$0.07	98.45$\pm$0.19	97.85 $\pm$ 0.23	98.15 $\pm$ 0.20
FLAN-T5-base-ZS	32.73 $\pm$ 2.71	30.01 $\pm$ 2.52	-28.15 $\pm$ 2.74	-27.38 $\pm$ 2.55	92.48 $\pm$ 1.73	64.08 $\pm$ 2.95	75.71 $\pm$ 2.39
Qwen2.5-3B-ZS	6.26 $\pm$ 0.49	3.99 $\pm$ 0.46	-1.68 $\pm$ 0.46	-1.36 $\pm$ 0.45	97.92 $\pm$ 0.23	96.56 $\pm$ 0.47	97.24 $\pm$ 0.31
Qwen2.5-7B-ZS	6.48 $\pm$ 0.28	3.51 $\pm$ 0.18	-1.90 $\pm$ 0.21	-0.88 $\pm$ 0.13	96.63 $\pm$ 0.30	95.57 $\pm$ 0.34	96.10 $\pm$ 0.31

TABLE III

GT-ANCHORED MEDICAL TERMINOLOGY CORRECTION OUTCOMES. RECOVERED AND DEGRADED ARE REPORTED AS COUNTS, WITH RATES (%) RELATIVE TO THE 585 RAW-INCORRECT AND 23,187 RAW-CORRECT TERMINOLOGY OCCURRENCES IN PARENTHESES, RESPECTIVELY. NET DENOTES RECOVERED – DEGRADED, WITH THE CORRESPONDING RATE (%) RELATIVE TO ALL 23,772 GT TERMINOLOGY OCCURRENCES IN PARENTHESES. VALUES FOLLOWING $\pm$ ARE 95% BOOTSTRAP CI HALF-WIDTHS.

Model	Recovered	Degraded	Net
T5-base-FT	118 (20.17 $\pm$ 4.77)	67 (0.29 $\pm$ 0.10)	+51 (+0.21 $\pm$ 0.17)
T5-large-FT	118 (20.17 $\pm$ 4.55)	34 (0.15$\pm$0.06)	+84 (+0.35$\pm$0.14)
FLAN-T5-base-FT	125 (21.37 $\pm$ 4.67)	51 (0.22 $\pm$ 0.08)	+74 (+0.31 $\pm$ 0.16)
Qwen2.5-3B-FT	131 (22.39 $\pm$ 4.38)	72 (0.31 $\pm$ 0.08)	+59 (+0.25 $\pm$ 0.15)
Qwen2.5-7B-FT	149 (25.47$\pm$4.82)	74 (0.32 $\pm$ 0.09)	+75 (+0.32 $\pm$ 0.18)
FLAN-T5-base-ZS	9 (1.54 $\pm$ 1.18)	7,962 (34.34 $\pm$ 3.05)	-7,953 (-33.46 $\pm$ 3.00)
Qwen2.5-3B-ZS	89 (15.21 $\pm$ 3.34)	321 (1.38 $\pm$ 0.42)	-232 (-0.98 $\pm$ 0.43)
Qwen2.5-7B-ZS	108 (18.46 $\pm$ 3.73)	576 (2.48 $\pm$ 0.25)	-468 (-1.97 $\pm$ 0.27)

with 10,000 replicates, with speakers resampled as clusters to account for repeated documents from the same speaker. Differences relative to RAW were assessed using paired speaker-clustered bootstrap tests. Confidence intervals for Spearman correlations were estimated using speaker-clustered bootstrap resampling, with significance assessed using size-stratified whole-speaker block permutation tests that preserve all documents from a speaker within the same block.

IV. RESULTS AND DISCUSSION

Results are reported on the 1,500-document test set; GT-anchored terminology analyses exclude the two documents without medical reference terms. We examine overall and terminology performance, terminology transitions, and error redistribution.

A. Correction and Terminology Performance

Table II summarizes overall transcription and token-aligned medical terminology performance. All fine-tuned models significantly improved WER and CER relative to RAW ($p < 0.001$), despite the comparatively low baseline WER of 4.58%, whereas all zero-shot models produced significantly higher

error rates. Qwen2.5-7B-FT achieved the lowest WER and CER, at 3.57% and 2.16%, respectively.

Medical terminology performance followed the same broad pattern, but the model ranking differed. All fine-tuned models improved token-aligned terminology F1 over RAW. T5-large-FT obtained the highest F1, whereas Qwen2.5-7B-FT achieved the lowest WER and CER. This difference suggests that improvements in overall transcription quality and medical terminology are related but need not favor the same correction model.

Because the token-aligned metrics are anchored to GT terminology spans, they do not penalize medical terminology introduced elsewhere in the transcript. We therefore additionally evaluated position-independent terminology mentions, for which unmatched COR mentions count as false positives. Although this does not incorporate introduced terminology into the transition analysis, it provides a complementary measure of unmatched terminology mentions. The mention-level analysis showed the same overall improvement pattern. The fine-tuned models produced 466–490 unmatched terminology mentions in COR, compared with 527 in RAW, and also improved mention-level precision, recall, and F1. Thus, the improvements observed within GT terminology spans were not accompanied by an increase in unmatched COR terminology mentions.

B. Terminology Transition Analysis

Of the 23,772 GT terminology instances, 23,187 were already correct in RAW and were mostly preserved after correction, while the majority of the 585 initially incorrect instances remained unresolved. Table III shows that the fine-tuned models recovered 118–149 of the 585 terminology instances incorrect in RAW. Of the 23,187 initially correct instances, only 34–74 were degraded. All fine-tuned models produced positive Net outcomes, ranging from +51 to +84 across 23,772 GT terminology occurrences.

The model comparison reveals different recovery–degradation balances. Qwen2.5-7B-FT recovered the most erroneous terminology occurrences (149) but also produced

TABLE IV

ASSOCIATION BETWEEN DOCUMENT-LEVEL WER IMPROVEMENT AND TERMINOLOGY TRANSITIONS FOR FINE-TUNED MODELS. SPEARMAN ρ IS COMPUTED OVER DOCUMENTS CONTAINING AT LEAST ONE TERMINOLOGY INSTANCE INCORRECT IN RAW ($n=419$; 27.9% OF THE TEST SET). FOR DOCUMENTS WITH IMPROVED WER, NET OUTCOMES ARE CATEGORIZED AS NET LOSS (NET < 0), NO NET CHANGE (NET = 0), OR NET GAIN (NET > 0). BRACKETS DENOTE 95% SPEAKER-CLUSTERED BOOTSTRAP CIs; ALL CORRELATIONS ARE SIGNIFICANT AT $p < 0.001$ USING SIZE-STRATIFIED WHOLE-SPEAKER BLOCK PERMUTATION TESTS.

Model	Correlation $\rho(\Delta\text{WER}, \text{Net})$ [95% CI]	n	Among WER-improved documents		
			Loss (%)	No net change (%)	Gain (%)
T5-base-FT	0.43 [0.34, 0.51]	225	0.9	64.0	35.1
T5-large-FT	0.47 [0.38, 0.54]	225	0.0	62.2	37.8
FLAN-T5-base-FT	0.50 [0.43, 0.57]	219	0.0	60.3	39.7
Qwen2.5-3B-FT	0.51 [0.43, 0.58]	230	1.3	57.8	40.9
Qwen2.5-7B-FT	0.55 [0.46, 0.62]	240	1.2	56.2	42.5

the most degradations (74), whereas T5-large-FT recovered fewer occurrences (118) but produced substantially fewer degradations (34), yielding the highest Net outcome (+84). Therefore, comparable overall terminology scores may mask differing correction patterns, depending on the trade-off between fixing pre-existing errors and maintaining terminology that was already correct.

C. WER and Terminology Outcomes

To assess whether improvements in overall transcription quality correspond to favorable terminology changes, we examined the relationship between document-level WER improvement and Net terminology outcome. Among the 419 documents containing at least one terminology instance incorrect in RAW, WER improvement was moderately positively correlated with Net terminology outcome across all fine-tuned models ($\rho = 0.43\text{--}0.55$, all $p < 0.001$; Table IV). WER and terminology improvement are thus connected, but the modest correlations suggest that these two metrics cannot be used interchangeably.

This distinction is also apparent at the document level. Among documents with improved WER, 56.2–64.0% showed no net terminology change, while 35.1–42.5% showed a positive Net outcome. Negative Net outcomes were rare (0–1.3%). Thus, although medical terminology benefited in some cases, more than half of the documents with improved WER showed no corresponding Net terminology gain. This suggests that WER improvements can reflect corrections elsewhere in the transcript without necessarily improving medical terminology.

D. Error Redistribution

Finally, we examined how fine-tuning redistributed errors relative to RAW. Table V summarizes the resulting changes at the word and medical-terminology levels. At the word level, all fine-tuned models reduced substitutions and insertions while increasing deletions, resulting in overall error reductions of 16.8–21.9%. Qwen2.5-7B-FT achieved the largest reduction (21.9%), consistent with its lowest WER.

Medical terminology showed a similar redistribution, with fewer substitutions but more deletions after correction. Total terminology errors nevertheless decreased for every fine-tuned model. Unlike at the word level, T5-large-FT achieved

TABLE V

RELATIVE CHANGES IN ERROR COUNTS COMPARED WITH RAW. WORD-LEVEL TOTALS COMPRISE SUBSTITUTIONS, DELETIONS, AND INSERTIONS; MEDICAL TERMINOLOGY TOTALS COMPRISE SUBSTITUTIONS AND DELETIONS WITHIN GT TERMINOLOGY SPANS. NEGATIVE VALUES INDICATE FEWER ERRORS.

Model	Word-level				Medical terminology		
	$\Delta\text{Subs.}$	$\Delta\text{Del.}$	$\Delta\text{Ins.}$	ΔTotal	$\Delta\text{Subs.}$	$\Delta\text{Del.}$	ΔTotal
T5-base-FT	-14.8	+33.3	-40.4	-17.1	-15.6	+17.1	-8.7
T5-large-FT	-15.5	+19.2	-35.4	-17.7	-20.3	+8.1	-14.4
FLAN-T5-base-FT	-14.1	+32.1	-39.8	-16.8	-20.8	+17.9	-12.6
Qwen2.5-3B-FT	-16.1	+23.2	-39.8	-19.1	-19.7	+26.0	-10.1
Qwen2.5-7B-FT	-19.1	+17.4	-41.2	-21.9	-20.8	+17.1	-12.8
FLAN-T5-base-ZS	+139.3	+4139.3	+167.7	+803.6	+168.2	+5834.1	+1359.5
Qwen2.5-3B-ZS	+24.0	+161.6	-3.3	+35.6	+5.4	+168.3	+39.7
Qwen2.5-7B-ZS	+53.4	+114.5	+2.4	+43.1	+71.4	+112.2	+80.0

the largest reduction in terminology errors (14.4%), whereas Qwen2.5-7B-FT achieved the largest reduction in word-level errors (21.9%). Together, these results show that improvements in overall transcription quality are related to, but do not fully characterize, medical terminology transition outcomes.

V. CONCLUSIONS

We investigated generative post-ASR error correction for clinical dictation transcripts, focusing on changes in medical terminology. Fine-tuned models improved both overall transcription quality and medical terminology performance. The transition analysis showed that all fine-tuned models recovered more erroneous terminology than they degraded previously correct terminology, although the resulting Net gains were modest. WER improvements were moderately associated with Net terminology outcomes, but models differed in their balance between recovery and degradation and in the reduction of terminology errors. These findings suggest that aggregate transcription metrics do not fully characterize the effects of post-ASR correction on medical terminology.

Our study has several limitations. The experiments use a single clinical dictation corpus and ASR setting, and speaker and template overlap across data splits limits evaluation of generalization to unseen speakers and documentation patterns. The terminology inventory is restricted to single-word terms. Introduced medical terminology is captured as false positives in the mention-level evaluation but is not represented in the GT-anchored terminology transition analysis. In addition, Net assigns equal weight to all terminology instances and does not reflect differences in their potential clinical significance. Because the evaluated models also differ in architecture, scale, and adaptation strategy, differences between models cannot be attributed to any one of these factors in isolation.

Future work could extend the transition analysis to introduced terminology and evaluate correction across additional ASR settings and speaker- and template-independent splits. The differences observed between recovery and degradation also motivate approaches that directly control this trade-off rather than optimizing aggregate transcription error alone.

REFERENCES

- [1] F. R. Goss, S. V. Blackley, C. A. Ortega, L. T. Kowalski, A. B. Landman, C. T. Lin, M. Meter, S. Bakes, S. C. Gradwohl, D. W. Bates, and L. Zhou, "A clinician survey of using speech recognition for clinical documentation in the electronic health record," *Int. J. Med. Inform.*, vol. 130, p. 103938, Oct. 2019.
- [2] L. Zhou, S. V. Blackley, L. Kowalski, R. Doan, W. W. Acker, A. B. Landman, E. Kontrient, D. Mack, M. Meter, D. W. Bates, and F. R. Goss, "Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists," *JAMA Netw. Open*, vol. 1, no. 3, p. e180530, Jul. 2018.
- [3] T. Afonja, T. Olatunji, S. Ogun, N. Etori, A. Owodunni, and M. Yekini, "Performant ASR models for medical entities in accented speech," in *Proc. Interspeech*, Kos Island, Greece, Sep. 2024, pp. 2315–2319.
- [4] N. Ebadi, K. Morgan, A. Tan, B. Linares, S. Osborn, E. Majors, J. Davis, and A. Rios, "Extracting biomedical entities from noisy audio transcripts," in *Proc. Joint Int. Conf. Comput. Linguistics, Lang. Resour. Eval. (LREC-COLING)*, Torino, Italy, May 2024, pp. 7023–7034.
- [5] K. Binici, A. R. Kashyap, V. Schlegel, A. T. Liu, V. P. Dwivedi, T.-T. Nguyen, X. Gao, N. F. Chen, and S. Winkler, "MEDSAGE: Enhancing robustness of medical dialogue summarization to ASR errors with LLM-generated synthetic dialogues," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Philadelphia, PA, USA, 2025, pp. 23496–23504.
- [6] M. Kadkhodaie Elyaderani, J. Glover, and T. Schaaf, "Reference-free estimation of the quality of clinical notes generated from doctor-patient conversations," in *Proc. Interspeech*, Kos Island, Greece, Sep. 2024, pp. 1425–1429.
- [7] O. Shapira, S. Chazan, and A. D. N. Cohen, "Measuring the effect of transcription noise on downstream language understanding tasks," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Vienna, Austria, Jul. 2025, pp. 29978–30004.
- [8] A. Mani, S. Palaskar, N. V. Meripo, S. Konam, and F. Metze, "ASR error correction and domain adaptation using machine translation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 6344–6348.
- [9] C.-H. H. Yang, T. Park, Y. Gong, Y. Li, Z. Chen, Y.-T. Lin, C. Chen, Y. Hu, K. Dhawan, P. Zelasko, C. Zhang, Y.-N. Chen, Y. Tsao, J. Balam, B. Ginsburg, S. M. Siniscalchi, E. S. Chng, P. Bell, C. Lai, S. Watanabe, and A. Stolcke, "Large language model-based generative error correction: A challenge and baselines for speech recognition, speaker tagging, and emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Macao, China, Dec. 2024, pp. 371–378.
- [10] J. Prakash, B. Kumar, K. Hacioglu, B. Sharma, S. Gopalan, W. Chetlur, S. Venkatesan, and A. Stolcke, "Better pseudo-labeling with multi-ASR fusion and error correction by SpeechLLM," in *Proc. Interspeech*, Rotterdam, The Netherlands, Aug. 2025, pp. 579–583.
- [11] C.-H. H. Yang, Y. Gu, Y.-C. Liu, S. Ghosh, I. Bulyko, and A. Stolcke, "Generative speech recognition error correction with large language models and task-activating prompting," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop (ASRU)*, Taipei, Taiwan, Dec. 2023, pp. 1–8.
- [12] M. La Quatra, V. M. Salerno, Y. Tsao, and S. M. Siniscalchi, "FlanEC: Exploring Flan-T5 for post-ASR error correction," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Macao, China, Dec. 2024, pp. 608–615.
- [13] T. Udagawa, M. Suzuki, M. Muraoka, and G. Kurata, "Robust ASR error correction with conservative data filtering," in *Proc. Conf. Empir. Methods Natural Lang. Process. (EMNLP Industry Track)*, Miami, FL, USA, Nov. 2024, pp. 256–266.
- [14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Honolulu, HI, USA, Jul. 2023, pp. 28492–28518.
- [15] R. Ma, M. J. F. Gales, K. M. Knill, and M. Qian, "N-best T5: Robust ASR error correction using multiple input hypotheses and constrained decoding space," in *Proc. Interspeech*, Dublin, Ireland, Aug. 2023, pp. 3267–3271.
- [16] Y. Jiang and C. Poellabauer, "A sequence-to-sequence based error correction model for medical automatic speech recognition," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Houston, TX, USA, Dec. 2021, pp. 3029–3035.
- [17] A. Mani, S. Palaskar, and S. Konam, "Towards understanding ASR error correction for medical conversations," in *Proc. Workshop Natural Lang. Process. Med. Conversations (NLP/MC)*, Jul. 2020, pp. 7–11.
- [18] Y. Li, P. Chen, P. Bell, and C. Lai, "Crossmodal ASR error correction with discrete speech units," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Macao, China, Dec. 2024, pp. 431–438.
- [19] P. Szymański, L. Augustyniak, M. Morzy, A. Szymczak, K. Surdyk, and P. Zelasko, "Why aren't we NER yet? Artifacts of ASR errors in named entity recognition in spontaneous speech transcripts," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Toronto, ON, Canada, Jul. 2023, pp. 1746–1761.
- [20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, Jan. 2020.
- [21] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," *J. Mach. Learn. Res.*, vol. 25, no. 70, pp. 1–53, Jan. 2024.
- [22] A. Yang et al., "Qwen2.5 Technical Report," arXiv:2412.15115, 2025.
- [23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Vienna, Austria, Apr. 2022.
- [24] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguistics*, vol. 5, 2017, pp. 135–146.