Evaluating Robustness of LLM-Based Ambient Scribes for SOAP Note Generation

Ehsan Latif EHSAN.LATIF@LUMS.EDU.PK
Aleema Faisal 26100031@LUMS.EDU.PK
Shaheer Hammad 26100044@LUMS.EDU.PK
Dayyan Ali Akhtar 26100007@LUMS.EDU.PK
Agha Ali Raza AGHA.ALI.RAZA@LUMS.EDU.PK
Ihsan Ayyub Qazi IHSAN.QAZI@LUMS.EDU.PK
Computer Science Department, Lahore University of Management Sciences (LUMS), PK

Abstract

Clinical documentation is a major driver of clinician workload and burnout, motivating the adoption of ambient AI scribes that transcribe clinician-patient conversations into clinical notes. Safe deployment requires both transcript-grounded fidelity and robustness to upstream Automatic Speech Recognition (ASR) noise-properties not captured by traditional ROUGE-like metrics. We propose a clinically grounded evaluation framework that decomposes notes into atomic, QNOTE-structured facts and applies a two-phase triangulated protocol: (1) align generated facts to clinician-authored gold notes to measure coverage, omission, contradiction, and candidate additions; (2) verify gold-absent generated facts against transcripts to distinguish valid elaborations from unsupported content. Across eight LLM-based note generators, we find that omissions are the primary source of contextual degradation (8.5%–24.0%), while contradictions remain relatively stable (6.2%–7.9%). A large majority of content initially flagged as "added" relative to gold is supported by the transcript (92%), highlighting the importance of transcript verification. Robustness analysis with controlled transcript-level perturbations shows that conversational redundancy often mitigates errors (38.6% recovery), whereas substitution errors (e.g., negation flips, medical homophones) are more likely to propagate when redundancy is absent. These results provide a structured approach for evaluating fidelity and robustness in clinical note generation and suggest practical considerations for safer deployment.

Data and Code Availability Code for transcript preprocessing, prompting, SOAP note generation, and all evaluation analyses are available at GitHub Repository[1], including scripts to reproduce the reported experiments, figures, and tables; this study uses the publicly available PrIMock57 dataset (Korfiatis et al., 2022), and all additional annotations created as part of our evaluation (e.g., error taxonomy labels and severity ratings) will be made available alongside the code release, subject to the dataset's original licensing terms.

Institutional Review Board (IRB) Because all datasets used in this study are publicly available, IRB approval is not required.

1. Introduction

Clinical documentation has become one of the most time-consuming responsibilities in modern healthcare. Evidence suggests that for each hour of direct patient care, physicians may spend nearly two additional hours on electronic health record (EHR) tasks, contributing to administrative overload (Kanaparthy et al., 2025). This growing documentation burden has been strongly associated with clinician burnout and diminished quality of patient interactions (Leung et al., 2025; Kanaparthy et al., 2025). Although clinical note formats, such as SOAP (Subjective, Objective, Assessment, Plan), provide a standardized structure for outpatient encounter documentation, composing a complete and clinically high-quality note still requires substantial manual effort. Prior work has introduced validated instruments for

1. https://github.com/ehsanlatif/LLM-Notes-Evaluation.git

assessing note quality, such as QNOTE (and related approaches), highlighting the importance of completeness, clarity, and accuracy in clinical documentation (Burke et al., 2014; Palm et al., 2025).

Recent advances in large language models (LLMs) have accelerated interest in automating clinical documentation, particularly through ambient AI scribes that transcribe clinician-patient conversations into structured notes (Tierney et al., 2024). Early deployments suggest reduced documentation time and improved clinician engagement (Olson et al., 2025; Kanaparthy et al., 2025), but their use in high-stakes clinical settings raises concerns regarding factual accuracy, safety, and accountability (Leung et al., 2025).

Despite the promise of LLM-powered clinical documentation, several gaps remain in the systematic understanding of model performance and safety. First, many existing evaluations are limited to a single model or a narrow clinical setting, making it difficult to compare performance across systems or generalize findings. A recent systematic review found that only a small fraction of studies evaluating LLMs in healthcare used real patient care data, and comparatively few examined summarization or clinical note-generation tasks (Bedi et al., 2025). Benchmarks, such as HealthBench, have advanced the evaluation of LLMs in healthcare, but primarily focus on question answering and short-form tasks rather than full-length clinical documentation, such as SOAP note generation (Arora et al., 2025). Second, traditional NLP similarity metrics (e.g., ROUGE, BLEU) are poorly suited to clinical note evaluation, because surface-level overlap does not reliably reflect clinical correctness or safety. Clinically meaningful evaluation must distinguish errors of addition (hallucinations), omission, and contradiction, and assess their potential downstream consequences (Asgari et al., 2025).

These evaluation gaps have important clinical implications. An AI model that hallucinates a dose or diagnosis of medication poses a fundamentally different risk profile than one that omits the relevant history or allergies, and inaccurate documentation can mislead clinicians and compromise continuity of care (Asgari et al., 2025). Although some recent studies suggest that AI-generated notes can approach clinician-written notes in certain quality dimensions, they can be more verbose and remain prone to containing hallucinated or extraneous content (Palm et al., 2025). As a result, clinicians consistently emphasize the need for careful human oversight and responsible integration into clinical workflows (Leung et al., 2025). Therefore, it is importamt to rigorously characterize the types and severity of errors produced by different LLMs, and to evaluate how robust these systems are under realistic conditions, including transcription imperfections.

In this work, we address these gaps through two research questions:

- **RQ1:** Do LLM-generated clinical notes preserve the transcript's contextual integrity?
- **RQ2:** What is the impact of introducing controlled transcription errors on the generated clinical note quality?

To answer these questions, we evaluated eight LLM-based note generators on clinician-patient conversations, using identical prompts and, when supported, deterministic decoding. Our contributions are as follows:

1. We introduce a hallucination evaluation method that performs fact extraction, comparison, and clinical categorization. Using the hallucination evaluation method, we compare eight proprietary LLMs and analyze how errors distribute across SOAP sections. We show that GPT-5 achieves the highest overall fidelity.
2. We propose a robustness analysis method that induces clinically realistic transcription errors into transcripts (Table 1) and find that transcript redundancy is a primary robustness mechanism (38.6% of induced errors recovered via alternate mentions), whereas substitution-type errors (e.g., medical homophones) are difficult to detect without redundancy; a deployment-relevant failure mode for ambient scribing systems (Bell et al., 2020).

2. Related Work

2.1. Clinical Note Generation and Hallucination Evaluation

The automation of clinical documentation has attracted significant research interest as a means to alleviate physician burden. Giorgi et al. (2023) demonstrated that fine-tuned FLAN-T5-Large could achieve strong performance on section header prediction and text generation in the MEDIQA-Chat challenge, while GPT-4 with in-context learning achieved

the highest average section scores for full SOAP note generation. This highlighted an early trade-off between conciseness and completeness in LLM-generated documentation. Biswas and Talukdar (2024) extended this work by evaluating multiple LLMs on SOAP and BIRP note generation, finding that GPT-4 consistently outperformed other models on ROUGE-based metrics. However, their qualitative analysis revealed persistent issues: models occasionally omitted clinically relevant details, introduced vague plans, and hallucinated plausible but unsupported information. Critically, the authors noted that high ROUGE scores primarily reflect lexical similarity rather than factual correctness, suggesting that aggregate metrics obscure safety-critical failure modes.

The limitations of standard evaluation metrics have received increasing scrutiny. Janiak et al. (2025) demonstrated that when evaluated by semantic judges rather than ROUGE, state-of-the-art hallucination evaluation methods saw performance drops of up to 45.9%, and that simple response length was often a more powerful indicator of hallucination than complex detection algorithms. Asgari et al. (2025) proposed a framework specifically designed to assess clinical safety and hallucination rates, distinguishing between faithfulness hallucinations (deviations from source content) and factuality hallucinations (deviations from world knowledge). They found that longer summaries contained more hallucinations and identified "Specific⇒General" errors as a major failure mode where models replaced technical clinical terms with common language. Vishwanath et al. (2024) developed a specialized framework categorizing hallucinations into five medical types, finding that expert clinician review required 91.5 minutes per note at $55 per annotation underscoring the unsustainability of manual auditing at scale.

To address factual accuracy, Li et al. (2025) introduced the K-SOAP format that augments traditional SOAP sections with explicit clinical entity extraction, demonstrating that keyword guidance provided "obvious prevention of hallucinations." Miller et al. (2025) showed that dynamic few-shot prompting with retrieval of semantically similar examples achieved strong performance on clinical note section classification, reducing "structural hallucinations" where content is misplaced across sections. These works establish that explicit grounding in extracted facts can improve accuracy, a principle we adopt through fact-level decomposition using the QNOTE schema.

2.2. Robustness to Transcription Errors

Ambient AI scribe systems rely on automatic speech recognition (ASR), yet the robustness of downstream note generation to transcription errors remains understudied. Binici et al. (2025) investigated this gap by measuring error rates across ASR systems and developing an LLM-based approach to generate corrupted transcripts mirroring realistic ASR failures, including insertions, deletions, and phonetically similar substitutions. Testing on PriMock57, they found that pre-training on corrupted-clean pairs improved summary quality, but their evaluation used aggregate metrics rather than hallucination-specific measures.

Wang et al. (2025) developed a comprehensive evaluation framework for ambient digital scribing tools that systematically tested robustness to transcription noise and edge cases. Key findings revealed that masking clinically critical terms led to vague or inferred statements, models frequently retained or silently "corrected" implausible values without flagging them, and new medications posed major challenges with transcription errors leading to omissions. Prior work on ASR error characterization (Shah, 2024) classified errors into categories including dictionary errors (out-of-vocabulary medical terms), homonym errors (phonetically identical but semantically distinct words), and critical errors (negation flips, misrecognized numbers), finding that while most errors allowed plausible interpretation from context, the remaining errors posed serious clinical risks.

3. Dataset

PriMock57 (Korfiatis et al., 2022) is a clinical benchmark dataset comprising 57 mock primary-care consultations. Each consultation includes an audio recording, a manually produced utterance-level transcript, and a corresponding physician-authored clinical note. The dataset is designed to support the evaluation of automatic speech recognition (ASR) systems as well as downstream clinical note generation models. The overall dataset details are provided in Appendix A.

Initial experimentation with our hallucination evaluation revealed systematic inconsistencies between physician-authored notes and consultation transcripts, primarily in the form of omissions, contradictions, and additions. As shown in Fig. 2, omissions were most frequent (94), followed by contradictions (40) and additions (26), suggesting missing

information as the dominant source of misalignment. We then conducted a structured multi-annotator audit of all consultations, categorizing issues as resolvable or unresolvable and assigning severity ratings to unresolvable cases. Fig. 3 shows that most issues were resolvable (146, 91.2%), while a smaller subset was unresolvable (14, 8.8%) due to factors such as physician inference beyond the transcript or transcription ambiguity. Resolvable issues were corrected through targeted edits, with examples shown in Table 6. Consultations containing high-critical unresolvable errors were excluded, reducing the dataset from 57 to 51 consultations and eliminating 91% of annotated issues, resulting in a cleaner benchmark for downstream factual evaluation; however, this filtering may bias the dataset toward more consistent examples and could lead to optimistic estimates of real-world performance. Additional details are provided in Appendix C.

4. Methods

4.1. Hallucination evaluation (for RQ1)

We propose a hallucination evaluation method that operates in three stages: (1) fact extraction from generated notes, gold notes (medically vetted human expert-written notes), and transcripts; (2) fact comparison using an LLM-based agent; and (3) metric computation (Fig. 1 illustrates the overview of the hallucination evaluation method).

4.1.1. Note Generation

All models received identical prompts for SOAP note generation. It begins with a system message defining the AI's role as a clinical documentation assistant, followed by detailed style directives emphasizing conciseness, clinical abbreviations, and avoidance of hallucinations. The structure explicitly outlines the required SOAP format (Subjective, Objective, Assessment, Plan) and provides an example of the desired output style. The prompt template is provided in the Appendix B.

4.1.2. Fact Extraction

We extract structured facts from clinical text using the QNote format (Burke et al., 2014). Each fact is associated with a specific SOAP section and contains:

- A unique identifier (e.g., `hpi-003`)
- The section label (e.g., History of Present Illness)
- The fact content in natural language
- Source text from the original document

Fact extraction serves as a normalization layer between free-form clinical narratives and downstream fact-level evaluation. Rather than evaluating hallucinations at the document level, which is often ambiguous and difficult to attribute, our approach decomposes clinical notes into atomic, section-aware factual units that can be independently verified for entailment. Vladika et al. (2025) has shown that decomposing generated text into atomic factual statements enables more fine-grained and reliable factuality evaluation than document-level comparison, particularly in entailment-based settings.

The extraction process employs a structured, rule-constrained prompting strategy to guide the language model; the prompt template is provided in Appendix B. Specifically, the model is required to adhere to four constraints:

1. **Categorization:** Assign each fact to one of the 12 predefined QNOTE sections.
2. **Atomization:** Ensure each extracted fact represents a single clinically meaningful assertion (e.g., one symptom, finding, or action), decomposing compound statements when necessary.
3. **Source Attribution:** Include the exact verbatim source text from which each fact was derived.
4. **Comprehensiveness:** Extract all medically relevant facts present in the note.

Two independent annotators manually verified a subset of LLM-extracted facts for factual correctness and alignment with the source transcript. Inter-annotator agreement was quantified using Cohen's κ, yielding a value of 0.95, which corresponds to substantial agreement. These results indicate that the extraction process produces accurate and verifiable atomic facts suitable for fine-grained entailment evaluation.

Requiring explicit source attribution enables deterministic and auditable entailment checking. While this approach may introduce systematic errors related to over-decomposition, these are observable and easier to correct than implicit hallucinations embedded in free-form text.

Each extracted fact is normalized into the following JSON structure:

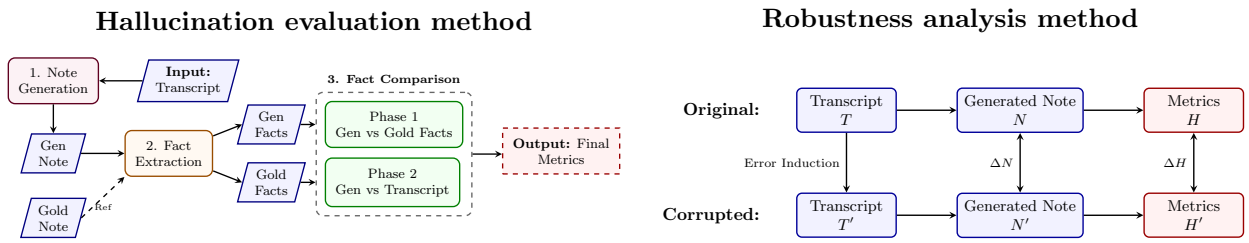

Figure 1: Overview of the proposed framework. **Left:** Fact-based evaluation method for LLM-generated notes, where transcripts are compared to gold notes via fact extraction and assessed by the LLM Comparison Agent (here we shorten the "Generated" as "Gen" to manage space). **Right:** Robustness analysis method showing the effect of corrupted transcripts on generated notes and metrics, with arrows indicating induced errors (ΔN, ΔH).

```
{
  "fact_id": "section-001",
  "content": "atomic fact string",
  "source_text": "verbatim quote"
}
```

4.1.3. Fact Comparison

Evaluating factual accuracy in AI-generated clinical notes is challenging due to clinical nuance that cannot be captured by traditional string-based metrics. In particular, it is necessary to distinguish clinically equivalent statements while separating omissions from hallucinations. To address this, we propose an LLM-based fact comparison agent operating over atomic, QNOTE-structured facts, enabling clinically meaningful evaluation of recall, precision, and hallucination.

Our method follows a *two-phase triangulated protocol*, leveraging three sources of evidence (generated notes, gold notes, and transcripts) to separate reference incompleteness from genuine model errors (shown in Fig. 1).

Phase 1: Generated Note Facts vs Gold Note Facts. In Phase 1, generated and gold-standard facts are aligned bidirectionally. Each *gold* fact is assigned one of the following labels: ***Covered*** (the concept is present in the generated note, allowing semantic paraphrase), ***Contradicted*** (the generated note asserts the opposite meaning), or ***Omitted*** (the concept is absent from the generated note). Each ***generated*** fact is independently classified as: ***Supported*** (it semantically matches a gold fact), ***Contradicted*** (it conflicts with a gold fact), or ***Absent*** (it is not present in the gold reference and is passed to Phase 2).

The comparison explicitly accounts for semantic equivalence and unit normalization (e.g., "24–48 hours" ≈ "1–2 days"). We do not treat transcript silence as evidence of a negative finding; absence of evidence is considered uninformative rather than implicitly negative. The prompt template is provided in Appendix B, preventing spurious mismatches due to surface variation.

Phase 2: Generated Notes, Facts vs Transcript. Facts labeled ***Absent*** from **gold** are verified against the original transcript to distinguish legitimate elaborations from hallucinations. Verification yields one of two outcomes: ***Valid elaboration***, when the fact is supported by transcript evidence (via direct mention or clearly stated evidence), or ***Invalid elaboration***, when it is unsupported by the transcript. To assess sensitivity to evaluation assumptions, we additionally consider a stricter variant where only explicitly stated transcript evidence is accepted, excluding inferred or medically implied relationships. Hallucinations are further categorized as ***True addition*** (new information absent from the transcript, including statements that may be clinically reasonable but not explicitly stated or supported) or ***True contradiction*** (direct conflict with transcript evidence). This two-phase design ensures clinically valid elaborations are preserved while reliably identifying true hallucinations.

4.2. Robustness analysis (RQ2)

Ambient AI scribe systems are intended to reduce clinician documentation burden; however, they inherently depend on upstream automatic speech recognition (ASR) and transcription pipelines. In this setting, transcription errors are not merely cosmetic

artifacts: once recorded in an EHR note, they can persist across encounters and influence downstream clinical decision making. Patient-facing studies have shown that errors in ambulatory EHR notes are common and can impose an additional burden on patients who must repeatedly identify and request correction of incorrect medical information (Bell et al., 2020). More critically, because clinical documentation is reused for continuity of care, a single undetected error may mislead future clinicians and increase the risk of propagated misinformation across encounters, particularly in automated workflows that aim to minimize human-in-the-loop verification (Bell et al., 2020). These risks motivate the need to study robustness under realistic transcription noise, especially for safety-critical error types that can alter diagnoses, treatment plans, and medication lists.

To evaluate model robustness to transcription errors, we developed a controlled error induction method. Using the best-performing model from RQ1, we systematically introduce errors into transcripts and measure their downstream effects on generated notes.

4.2.1. Error Taxonomy

We define six categories of transcription errors, motivated by (i) commonly observed ASR failure modes in clinical speech (e.g., deletions, low-confidence span suppression, and phonetic confusions), and (ii) prior evidence that certain documentation errors (especially negation errors and medical term confusions) can directly affect clinical interpretation and management decisions. This taxonomy targets error classes that are both plausible under real clinical transcription pipelines and clinically consequential when preserved in EHR notes (Table 1 provides a complete taxonomy).

First, we include errors that remove critical clinical information: **(1) censoring of medical terms** (e.g., medications, allergies, conditions), **(2) censoring of numerical values** (e.g., dose, frequency, duration), **(3) censoring of diagnostic terms**, and **(4) omission of symptom** mentions at the utterance level. These error types model real situations where ASR systems or post-processing pipelines drop uncertain spans, where entire clinically relevant turns are missed, or where a note generator lacks the evidence needed for accurate clinical assessment. Missing information can be as harmful as incorrect information, since downstream models may fill gaps with unsupported inferences or produce incomplete plans that delay or misdirect care.

Second, we include substitution-type errors that introduce incorrect but plausible clinical facts: **(5) negation flips** and **(6) medical homophones**. Negation errors are well-known to be safety-critical in clinical NLP because they invert the presence/absence of symptoms, diagnoses, or exposures (e.g., "no fever" $\rightarrow$ "fever") and can change diagnostic reasoning and triage; automated negation detection has been specifically studied because of its frequency and importance in clinical text (Elkin et al., 2005). Similarly, medical homophone substitutions (e.g., confusing drug names) are particularly risky because the corrupted term remains linguistically fluent and may evade detection by both models and human readers. Prior work suggests documentation inaccuracies can lead to confusion, reduced trust, non-adherence, delayed follow-up, and increased length of stay (Bell et al., 2020). In automated scribing systems with minimal verification, such errors can persist in the EHR, mislead future clinicians, and propagate misinformation across future encounters.

5. Experimentation and Results

5.1. Experimental Setup

We design the experiments to mirror the proposed method (Fig. 1) and to support two research questions: (RQ1) fact-based hallucination evaluation under clean transcripts, and (RQ2) robustness under controlled transcription noise. Across both settings, we follow the same core flow: *(i) transcript $\rightarrow$ note generation, (ii) fact extraction, (iii) LLM-based fact comparison, and (iv) metric computation.* Unless otherwise stated, we keep prompts and decoding settings fixed to ensure comparability.

Models Evaluated. We evaluate eight instruction-following LLMs spanning multiple capability levels and deployment profiles, chosen to reflect a range of high-performing, commercially accessible models used in clinical documentation pipelines. We emphasize standard and lightweight variants because SOAP note generation primarily requires preserving and structuring information rather than explicit multi-step reasoning, and because these models are more cost-effective and reproducible in practice. We note that our empirical findings are limited to this model family, although the evaluation framework itself is model-agnostic and can be applied

Table 1: Taxonomy of transcription errors for robustness analysis.

ID	Error Type	Description	Example
(1)	Censor Medical Term	Remove drug names, conditions, allergies	"metformin" → [BLANK]
(2)	Censor Numerical Value	Remove doses, frequencies, durations	"aspirin, seventy five milligrams, once a day" → "aspirin, once a day"
(3)	Censor Diagnosis	Remove diagnostic terms	"You seem to have labyrinthitis" → "You seem to have"
(4)	Symptom Omission	Remove entire symptom mention dialogue	Complete sentence removal
(5)	Negation Flip	Invert yes/no responses	"no fever" → "fever"
(6)	Medical Homophone	Replace with similar-sounding medical term	"Trimethoprim" → "Triamterene"

to other architectures. Model identifiers and API settings are provided in Table 5 and Appendix A.1.

Inputs and Outputs. For each encounter transcript T, each model produces a SOAP note N using an identical note-generation prompt (Appendix B). For hallucination evaluation (RQ1), we additionally use the medically vetted gold note G. For robustness experiments (RQ2), we construct a corrupted transcript T' and generate a corresponding note N' using the best-performing model from RQ1.

Unified Prompting and Deterministic Decoding. All models receive the same prompt template for SOAP note generation, consisting of (i) a system message specifying the role as a medical scribe and formatting constraints, and (ii) a user message containing the transcript and instructions to output a structured SOAP note. To isolate model differences from sampling variance, we use deterministic decoding when supported (temperature = 0; max tokens fixed). For models with model-specific decoding defaults (e.g., reasoning-focused variants), we report the default settings used (Table 5). The fact comparison agent uses structured, few-shot prompts and emits JSON outputs to support reproducibility and auditing.

5.1.1. RQ1: Hallucination Evaluation Protocol

Given transcript T, model-generated note N, and gold note G, we compute hallucination and coverage metrics using the three-stage pipeline in Fig. 1:

1. **Fact Extraction.** We extract atomic, QNOTE-structured facts from N, G, and T (Sec. 4.1.2). Each fact includes a section label and verbatim source attribution to enable auditable verification.

2. **Two-Phase Fact Comparison.** We compare facts using the LLM-based agent. In Phase 1, N is aligned against G to measure coverage and identify candidate additions. In Phase 2, generated facts labeled *Absent* from the gold reference are verified against T to distinguish valid elaborations from true hallucinations.

3. **Metric Computation.** Using comparison outcomes, we compute the addition rate, omission rate, contradiction rate, and coverage.

5.1.2. RQ2: Robustness Evaluation via Controlled Error Induction

To study robustness to transcription errors, we keep the same evaluation flow but replace the clean transcript with a minimally corrupted version and analyze downstream changes (Fig. 1, right):

Error Generation ($T \rightarrow T'$). Starting from the original transcript T, we create a corrupted transcript T' by inducing *exactly one* error drawn from the taxonomy in Sec. 4.2.1. These perturbations are applied at the transcript-text level (rather than the audio or ASR decoding stage) to isolate the downstream impact of specific transcription errors. We use an LLM to propose candidate corruptions by (i) identifying clinically consequential spans across SOAP-relevant content, (ii) selecting an error type likely to change meaning, and (iii) generating the altered text. We

then filter out low-impact candidates and verify medical homophones to ensure both terms are plausible clinical entities. Each T' contains a single targeted perturbation.

Note Generation ($T' \rightarrow N'$). For each corrupted transcript T', we generate a SOAP note N' using the best-performing model from RQ1 (GPT-5). We also retain the note N generated from the original transcript T for paired comparisons.

Error Response Evaluation. We categorize the model's handling of the induced error in N' as `omitted`, `preserved`, `corrected`, or `flagged as unreasonable`. This assessment is performed with an LLM-as-a-judge using a fixed rubric, and we validated its alignment with human judgments on the complete set.

Note-to-Note Difference Analysis (ΔN). To isolate behavior changes attributable to transcription noise, we compare N' against N using an auditing rubric (LLM-as-a-judge). This captures additions, omissions, and contradictions in N' relative to N, independent of transcript support. Because this analysis is relative to N, it does not count error corrections in N' as hallucinations when they align back to the original note.

5.2. RQ1: Hallucination evaluation results

To answer **RQ1** (whether LLM-generated SOAP notes preserve the transcript's contextual integrity), we apply our two-phase fact comparison protocol (Phase 1: generated vs. gold; Phase 2: absent-from-gold vs. transcript). Table 2 reports metrics that operationalize contextual integrity as (i) *coverage* of clinically relevant transcript content, (ii) low *omission* of gold-anchored facts, (iii) low *contradiction* with transcript evidence, and (iv) low *true addition* (unsupported by both transcript and gold).

Overall contextual integrity is primarily limited by omissions. GPT-5 preserves transcript context most effectively, achieving the highest coverage (85.28% ± 0.34) and the lowest omission rate (8.48% ± 0.23), while also exhibiting the lowest contradiction rate (6.24% ± 0.11). Across models, contextual integrity varies most in *completeness*: omission rates span 8.48%-23.96%, whereas contradiction rates remain in a narrower range (6.24%–7.88). This indicates that models more often degrade context by *dropping* transcript-supported details than by introducing explicit inconsistencies.

Most non-gold content is transcript-supported, not fabricated. Many differences between generated and gold notes reflect reference incompleteness rather than loss of contextual integrity. Specifically, a large majority of content initially flagged as "added" relative to gold is verified in Phase 2 as *valid elaboration* supported by the transcript (91.92%). True additions are uncommon but model-dependent: GPT-5.2 yields the lowest true-addition rate (0.24% ± 0.06), while smaller models show higher rates (e.g., GPT-5 nano: 4.00% ± 0.16), consistent with a trade-off between conservative generation and completeness.

Context degradation is section-dependent. Section-level analysis (Fig. 4) shows that contextual integrity is not uniform across the SOAP structure. Errors concentrate in **HPI**, which contributes the largest share of omissions and contradictions, consistent with the need to preserve temporality, qualifiers, and multi-part symptom narratives. In contrast, **Plan of Care** accounts for a disproportionate share of additions (50% of all added facts), suggesting that forward-looking planning encourages inference-like generation that can diverge from transcript-grounded documentation when not explicitly stated.

Significance of model effects. A Friedman test confirms that model choice significantly affects omission ($p < 10^{-17}$) and addition rates ($p < 10^{-13}$), but not contradiction rates ($p = 0.27$). Overall, contextual integrity differences across LLMs are driven mainly by recall (coverage/omissions) and secondarily by true additions identified via transcript verification, while contradictions appear comparatively stable across systems.

5.3. RQ2: Robustness Analysis

To answer **RQ2** (the impact of controlled transcription errors on generated SOAP note quality), we performed robustness analysis by inducing a single clinically realistic transcription error into each transcript ($T \rightarrow T'$) and generating SOAP notes from the corrupted input ($T' \rightarrow N'$). We then assess how the induced error affects the resulting note by analyzing (i) whether the erroneous span is *preserved*, *corrected*, *omitted*, or *flagged as unreasonable*, and (ii) which recovery mechanisms drive downstream robustness (Fig. 1, right).

Table 2: Overall model performance on SOAP note generation. Best values are **bolded**, second-best are underlined. Values represent mean ± standard deviation across all consultations.

Model	Coverage ↑	Omission ↓	Contradiction ↓	Addition ↓	Elaboration ↓
GPT-4.1	81.84 ± 0.70	10.38 ± 0.55	7.78 ± 0.16	0.61 ± 0.03	20.85 ± 0.19
GPT-4.1 mini	79.29 ± 1.26	13.60 ± 1.48	7.12 ± 0.22	1.14 ± 0.24	19.93 ± 0.52
GPT-4o	69.14 ± 0.22	23.96 ± 0.02	6.89 ± 0.19	0.39 ± 0.04	18.91 ± 1.20
GPT-5.2	83.93 ± 1.15	9.20 ± 0.23	6.88 ± 0.91	**0.24 ± 0.06**	22.27 ± 0.13
GPT-5	**85.28 ± 0.34**	**8.48 ± 0.23**	**6.24 ± 0.11**	0.41 ± 0.16	22.70 ± 0.39
GPT-5 mini	78.94 ± 0.60	13.18 ± 0.69	7.88 ± 0.09	2.02 ± 0.39	**18.54 ± 2.47**
GPT-5 nano	77.63 ± 1.38	14.82 ± 0.41	7.54 ± 0.97	4.00 ± 0.16	21.98 ± 0.45
o3	83.52 ± 0.11	9.92 ± 1.24	6.56 ± 1.13	3.50 ± 0.82	20.37 ± 0.13

Error recovery / handling pattern	Count	% of 44
Mentioned elsewhere (redundancy)	13	29.5%
Mentioned elsewhere – doctor repeated and confirmed afterwards	2	4.5%
Mentioned elsewhere (doctor summarized at the end)	2	4.5%
Inferred from surrounding	4	9.1%
Contradicted with another mention but recovered (nonsense omitted)	4	9.1%
Assumed default value (correct)	1	2.3%
No hints in surrounding	1	2.3%
Recovered via redundancy (aggregate)	17	38.6%

Table 3: Distribution of model behaviors under controlled transcript corruption (RQ2). Percentages are computed over 44 induced transcript errors.

Many transcription errors do not degrade note quality due to redundancy. The largest fraction of induced errors has a limited downstream impact because transcripts often contain repeated clinical facts (e.g., confirmation questions or end-of-visit summaries). In 38.6% of cases (17/44), the generated note recovered the correct fact using alternate mentions in the dialogue (Table 3), indicating that conversational redundancy acts as an implicit robustness mechanism even without explicit error detection.

When redundancy is absent, errors can trigger transcript-ungrounded inference. In the absence of redundant evidence, models sometimes fill missing information by inferring from the surrounding context. We observe such inference-driven recovery in 9.1% of cases (4/44). While these corrections may appear clinically plausible, they reduce transcript-grounded note quality by introducing content not explicitly supported by the input and therefore represent hallucinations under our evaluation definition.

Substitution-type errors are most likely to propagate into the note. For substitution errors such as medical homophones, models frequently preserve fluent but contextually incongruent terms, directly propagating transcription noise into the generated SOAP note. This indicates limited sensitivity to local semantic mismatch when no contradicting mention exists elsewhere. In contrast, in 9.1% of cases (4/44), models avoided degradation by leveraging another transcript mention that contradicted the corrupted span, effectively omitting the nonsense term and aligning with the consistent evidence (Table 3). Overall, these findings show that the impact of transcription errors on note quality is highly conditional: redundancy substantially mitigates downstream degradation, whereas single-span substitutions can persist and alter clinical meaning when redundancy is absent.

6. Discussion

Our results emphasize that trustworthy evaluation (and deployment) of ambient scribe systems requires disentangling three sources of deviation; importantly, our contribution is in evaluation methodology rather than directly mitigating hallucinations: (i) *reference incompleteness* in clinician-authored notes, (ii) *model generation errors* that degrade transcript-grounded contextual integrity (omissions, contradictions, true additions), and (iii) *upstream transcription noise*. In PriMock57, transcript-to-note inconsistencies were common, motivating our multi-annotator audit and removal of highly critical unresolvable cases to en-

able reliable factual benchmarking (BN et al., 2025). This directly supports our methodological contribution: the two-phase triangulation protocol (Phase 1: Gen vs Gold; Phase 2: Absent vs Transcript) reduces false hallucination flags when gold notes omit transcript-supported details, while still isolating *true* unsupported content; a key challenge in clinical factuality evaluation (Vladika et al., 2025).

For **RQ1**, we find that contextual integrity is primarily limited by *incompleteness* rather than fabrication: omission rates vary substantially across models, whereas contradiction rates remain comparatively stable. Clinically, this shifts the main safety concern from rare hallucinatory additions to missing or partially preserved encounter context. Errors are also section-dependent: HPI concentrates on omissions/contradictions, while Plan of Care accounts for a disproportionate share of additions, reflecting a tendency for forward-looking sections to invite inference-like generations. These findings align with evidence that documentation inaccuracies can meaningfully affect downstream care and impose burden when errors persist across encounters (Bell et al., 2020), motivating *section-aware* review strategies and interfaces that highlight evidence for high-risk facts.

For **RQ2**, controlled transcription errors degrade note quality in a conditional manner: many errors do not propagate when alternate mentions exist, making transcript redundancy the primary robustness mechanism (Lukac et al., 2025). However, when redundancy is absent, models sometimes reconstruct missing information via contextual inference (e.g., censored diagnoses), which may appear clinically plausible but reduces transcript-grounded note quality by introducing unsupported content. Substitution-type perturbations (e.g., medical homophones) are particularly likely to propagate because they remain fluent yet semantically incongruent, and models do not reliably flag them without contradicting evidence elsewhere. Given the growing real-world adoption of ambient scribing systems (Olson et al., 2025), these findings suggest practical deployment considerations: emphasizing explicit transcript evidence (or multiple mentions) for safety-critical categories (diagnoses, medications, allergies, negations), surfacing uncertainty when generation depends on inference, and applying targeted checks for substitution-like ASR failures that can silently alter clinical meaning (Stults et al., 2025).

7. Limitations

Our experiments are conducted on PriMock57 (Korfiatis et al., 2022), a mock primary-care dataset that enables controlled benchmarking but may not reflect real-world clinical audio conditions (e.g., background noise, code-switching, demographic variation, and specialty-specific complexity). In practice, ASR systems may exhibit error distributions that differ from and are more severe than those in curated benchmarks, which could alter both hallucination rates and robustness outcomes (Vishwanath et al., 2024).

In addition, our evaluation pipeline relies on LLMs for fact extraction, comparison, and judge-based robustness labeling. Despite structured prompting, JSON constraints, and manual audit checks, LLM-based evaluators can be biased or brittle on safety-critical edge cases (e.g., subtle negation, temporal qualifiers, and numerical reasoning) (BN et al., 2025). While we observe strong agreement with manual verification on sampled subsets, residual sensitivity to judge behavior remains a limitation of this approach. Our robustness analysis further induces a single controlled error per transcript; real ASR pipelines may produce multiple interacting errors that compound downstream effects (Croxford et al., 2025). Finally, we evaluate SOAP generation under standardized prompts and deterministic decoding (Stults et al., 2025); results may be sensitive to prompt design and instruction style.We also do not include clinician end-user evaluation, which limits assessment of perceived clinical risk, usability, and workflow impact in real-world settings (Olson et al., 2025).

8. Conclusion

We propose a clinically grounded evaluation framework for measuring hallucination and robustness in LLM-based medical note generation, rather than a method for directly reducing hallucinations. Our two-phase triangulated protocol reduces over-penalization due to incomplete gold notes by verifying absent-from-gold facts against transcripts, enabling more deployment-relevant benchmarking. Across eight LLMs, omissions are the dominant failure mode and the primary driver of model differences, while robustness experiments show that conversational redundancy mitigates many transcription errors but that contextual inference can introduce transcript-ungrounded content. These findings support evaluation beyond gold-note matching and motivate

section-aware oversight and safety guardrails for high-risk categories (diagnoses, medications, and negations) in ambient scribing deployments.

Acknowledgment

During the preparation of this work, the authors used ChatGPT in order to check grammar and polish the wording. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1): 274, 2025.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. Testing and evaluation of health care applications of large language models: a systematic review. *Jama*, 2025.

Sigall K Bell, Tom Delbanco, Joann G Elmore, Patricia S Fitzgerald, Alan Fossa, Kendall Harcourt, Suzanne G Leveille, Thomas H Payne, Rebecca A Stametz, Jan Walker, et al. Frequency and types of patient-reported errors in electronic health record ambulatory care notes. *JAMA network open*, 3(6): e205867–e205867, 2020.

Kuluhan Binici, Abhinav Ramesh Kashyap, Viktor Schlegel, Andy T Liu, Vijay Prakash Dwivedi, Thanh-Tung Nguyen, Xiaoxue Gao, Nancy F Chen, and Stefan Winkler. Medsage: Enhancing robustness of medical dialogue summarization to asr errors with llm-generated synthetic dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23496–23504, 2025.

Anjanava Biswas and Wrick Talukdar. Intelligent clinical documentation: Harnessing generative ai for patient-centric clinical note generation. *arXiv preprint arXiv:2405.18346*, 2024.

Suhas BN, Han-Chin Shing, Lei Xu, Mitch Strong, Jon Burnsky, Jessica Ofor, Jordan R Mason, Susan Chen, Sundararajan Srinivasan, Chaitanya Shivade, et al. Fact-controlled diagnosis of hallucinations in medical text summarization. *arXiv preprint arXiv:2506.00448*, 2025.

Harry B Burke, Albert Hoang, Dorothy Becher, Paul Fontelo, Fang Liu, Mark Stephens, Louis N Pangaro, Laura L Sessums, Patrick O'Malley, Nancy S Baxi, et al. Qnote: an instrument for measuring the quality of ehr clinical notes. *Journal of the American Medical Informatics Association*, 21(5): 910–916, 2014.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, et al. Evaluating clinical ai summaries with large language models as judges. *npj Digital Medicine*, 8(1):640, 2025.

Peter L Elkin, Steven H Brown, Brent A Bauer, Casey S Husser, William Carruth, Larry R Bergstrom, and Dietlind L Wahner-Roedler. A controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making*, 5(1):13, 2005.

John Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin An, Grace Zheng, and Bo Wang. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 323–334, 2023.

Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Jan Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34716–34733, 2025.

Naga Sasidhar Kanaparthy, Yenny Villuendas-Rey, Tolulope Bakare, Zihan Diao, Mark Iscoe, Andrew Loza, Donald Wright, Conrad Safranek, Isaac V Faustino, Alexandria Brackett, et al. Real-world

evidence synthesis of digital scribes using ambient listening and generative artificial intelligence for clinician documentation workflows: rapid review. *JMIR AI*, 4:e76743, 2025.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. Primock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 588–598, 2022.

Tiffany I Leung, Andrew J Coristine, and Arriel Benis. Ai scribes in health care: balancing transformative potential with responsible integration. *JMIR Medical Informatics*, 13(1):e80898, 2025.

Yizhan Li, Sifan Wu, Christopher Smith, Thomas Lo, and Bang Liu. Improving clinical note generation from complex doctor-patient conversation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 209–221. Springer, 2025.

Paul J Lukac, William Turner, Sitaram Vangala, Aaron T Chin, Joshua Khalili, Ya-Chen Tina Shih, Catherine Sarkisian, Eric M Cheng, and John N Mafi. A randomized-clinical trial of two ambient artificial intelligence scribes: Measuring documentation efficiency and physician burnout. *medRxiv: the preprint server for health sciences*, 2025.

Kurt Miller, Steven Bedrick, Qiuhao Lu, Andrew Wen, William Hersh, Kirk Roberts, and Hongfang Liu. Dynamic few-shot prompting for clinical note section classification using lightweight, open-source large language models. *Journal of the American Medical Informatics Association*, 32(7):1164–1173, 2025.

Kristine D Olson, Daniella Meeker, Matt Troup, Timothy D Barker, Vinh H Nguyen, Jennifer B Manders, Cheryl D Stults, Veena G Jones, Sachin D Shah, Tina Shah, et al. Use of ambient ai scribes to reduce administrative burden and professional burnout. *JAMA Network Open*, 8(10):e2534976–e2534976, 2025.

Erin Palm, Astrit Manikantan, Herprit Mahal, Srikanth Subramanya Belwadi, and Mark E Pepin. Assessing the quality of ai-generated clinical notes: validated evaluation of a large language model ambient scribe. *Frontiers in Artificial Intelligence*, 8: 1691499, 2025.

Savyasachi V Shah. Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. *JAMA Network Open*, 7(8):e2425953–e2425953, 2024.

Cheryl D Stults, Sien Deng, Meghan C Martinez, Joseph Wilcox, Nina Szwerinski, Kevin H Chen, Stephanie Driscoll, Joanna Washburn, and Veena G Jones. Evaluation of an ambient artificial intelligence documentation platform for clinicians. *JAMA Network Open*, 8(5):e258614–e258614, 2025.

Aaron A Tierney, Gregg Gayre, Brian Hoberman, Britt Mattern, Manuel Ballesca, Patricia Kipnis, Vincent Liu, and Kristine Lee. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery*, 5(3):CAT–23, 2024.

Prathiksha Rumale Vishwanath, Simran Tiwari, Tejas Ganesh Naik, Sahil Gupta, Dung Ngoc Thai, Wenlong Zhao, SUNJAE KWON, Victor Ardulov, Karim Tarabishy, Andrew McCallum, et al. Faithfulness hallucination detection in healthcare ai. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*, 2024.

Juraj Vladika, Annika Domres, Mai Nguyen, Rebecca Moser, Jana Nano, Felix Busch, Lisa C. Adams, Keno K. Bressem, Denise Bernhardt, Stephanie E. Combs, Kai J. Borm, Florian Matthes, and Jan C. Peeken. Improving reliability and explainability of medical question answering through atomic fact checking in retrieval-augmented llms. *arXiv preprint arXiv:2505.24830*, 2025.

Haoyuan Wang, Rui Yang, Mahmoud Alwakeel, Ankit Kayastha, Anand Chowdhury, Joshua M Biro, Anthony D Sorrentino, Jessica L Handley, Sarah Hantzmon, Sophia Bessias, et al. An evaluation framework for ambient digital scribing tools in clinical applications. *npj Digital Medicine*, 8(1): 358, 2025.

Appendix A. Dataset Details

A.1. Models Evaluated

Here, we provided all the details and rationale behind the selection of each model.

Table 4: Dataset statistics for Primock57_cleaned.

Statistic	Value
Number of consultations	51
Average transcript length (words)	1632.0 ± 431.0
Average gold note length (words)	198.2 ± 50.0
Sections covered	Chief Complaint, HPI, PMH, Medications, Allergies, Family Hx, Social Hx, Assessment, Plan, Follow-up, Physical Exam, Review of Systems

- **GPT-4.1 & GPT-4.1 mini:** The then-latest iteration of the GPT-4 series, emphasizing advanced instruction-following and reliability. GPT-4.1 mini is its cost-efficient counterpart.
- **GPT-4o:** A multimodal model optimized for balanced speed, cost, and capability across text and vision tasks.
- **GPT-5 nano, GPT-5 mini, & GPT-5:** The first generation of the GPT-5 series, representing a significant architectural advance. The nano and mini variants offer scaled-down, efficient versions of the full GPT-5 model.
- **o3:** A reasoning-focused model designed for complex problem-solving, featuring configurable reasoning effort levels (minimal, medium, high).

Appendix B. Prompts

B.1. SOAP Note Generation Prompt

Listing 1: System prompt for SOAP note generation

```
You are a clinical documentation
    assistant who converts raw
    clinician-patient
transcripts into accurate and terse
    SOAP consult notes.

KEY STYLE DIRECTIVES (Follow Strictly):
- Conciseness is critical: Use bullet
    points, sentence fragments, and
  clinical shorthand. The goal is a
      dense summary.
- Omit if empty: If information for a
    section or point is not in the
  transcript, omit it completely.
- Use abbreviations: Use common,
    unambiguous clinical abbreviations
  (e.g., LLQ, PMH, SOB, ADLs, etOH).
- Pertinent positives and negatives: If
     the clinician explicitly asks
  about a symptom or function, always
      include the patient's response.
- Avoid elaboration: Do not infer or
    paraphrase unless explicitly stated
  verbatim in the transcript.

STRUCTURE (SOAP):
- Header: Patient, MRN, DOB, etc. (omit
     fields not present).
- S (Subjective): CC, HPI, relevant PMH
    /PSH, meds, allergies, family/
    social
  history. Use bullet points.
- O (Objective): Vitals, physical exam
    findings, diagnostics.
- A (Assessment): Numbered list of
    problems with brief summaries.
- P (Plan): Bulleted list of actions
    for each problem.

EXAMPLE OUTPUT STYLE:
Subjective:
. 3-day history of watery diarrhea (~6/
    day); no blood in stool
. LLQ crampy, intermittent abdominal
    pain
. PMH: Asthma
. DH: Inhalers
. SH: Accountant; nil smoking/etOH
    history

HALLUCINATION POLICY:
- Do not add or infer clinical facts
    beyond the transcript.
- Never create values (e.g., vitals,
    lab numbers).
- If a medication is mentioned without
    dose/route/frequency, record the
  name only.

FORMATTING:
```

Table 5: Model specifications and configuration parameters.

Model	API Version	Temperature	Max Tokens	Access Date
GPT-4.1	gpt-4.1-preview	0	4096	Jan 2026
GPT-4.1 mini	gpt-4.1-mini-preview	0	4096	Jan 2026
GPT-4o	gpt-4o-2024-08-06	0	4096	Jan 2026
GPT-5	gpt-5-001	0	4096	Jan 2026
GPT-5 mini	gpt-5-mini-preview	0	4096	Jan 2026
GPT-5.2	gpt-5-mini-preview	0	4096	Jan 2026
GPT-5 nano	gpt-5-nano-preview	0	4096	Jan 2026
o3	o3-2025-12-19	1 (Default)	N/A	Jan 2026

```
- Output in markdown.
- Use `.` for bullet points.
```

B.2. Fact Extraction Prompt

We use a two-part prompt consisting of a system instruction defining strict behavioral constraints and a user instruction specifying the extraction task and schema.

Listing 2: System prompt for QNOTE fact extraction

```
FACT_EXTRACT_SYSTEM =
You are a specialized Clinical Data
    Extraction Engine.
Your ONLY function is to convert
    clinical text into strict QNOTE-
    schema JSON.

CRITICAL BEHAVIORAL CONSTRAINTS:
1. NO CONVERSATION: Do not output any
    introductory text.
2. NO HALLUCINATION: Extract only what
    is explicitly stated.
3. STRICT SCHEMA: Use only the 12
    allowed QNOTE section keys.
4. FORMATTING: Output raw JSON only. No
     markdown.
```

Listing 3: User prompt defining the QNOTE extraction task

```
FACT_EXTRACT_USER =
The Universal Prompt for QNOTE Fact
    Extraction
1. ROLE AND GOAL:
You are an expert clinical data
    extraction AI. Your task is to
    analyze a single, unstructured
    clinical note and extract all
    medically relevant facts into a
    structured JSON object based on the
     QNOTE schema.

2. INPUT:
You will be given a single block of
    text representing a clinical note.

3. CORE RULES OF EXTRACTION:
Rule 1 (Categorize): Categorize
    information into one of the 12
    QNOTE sections.
Rule 2 (Atomize): Each extracted fact
    should be a single, concise, and
    atomic piece of information.
Rule 3 (Cite Your Source): Every fact
    MUST include the  source_text  key
    with the exact, verbatim substring
    from the note.
Rule 4 (Be Comprehensive): Extract all
    relevant facts.

4. OUTPUT FORMAT (THE QNOTE SCHEMA):
Your output must be a single, clean
    JSON object.
Keys: Chief_Complaint,
    History_of_Present_Illness,
    Past_Medical_History, Medications,
    Adverse_Drug_Reactions_and_Allergies
    , Family_History,
    Social_and_Family_History,
    Assessment, Plan_of_Care,
    Follow_up_Information,
    Physical_Findings,
    Review_of_Systems.

Fact Object Structure:
{{
   fact_id :  section-001,
```

```
   content :  atomic fact string ,
   source_text :  verbatim quote
}}

5. TASK:
Process the following clinical note and
     generate the QNOTE-structured JSON
     object.
Here is the note:
{note_text}
```

B.3. Fact Comparison Prompt

Listing 4: System prompt for Phase 1 fact comparison

```
PHASE1_SYSTEM =
You are an expert Clinical Auditor.
Your task is to compare  Generated
    Facts  against  Gold Facts  (Ground
     Truth) to evaluate accuracy.

You must handle Semantic Equivalence
    intelligently:
1. Time/Units: Treat  24-48 hours  as
    equal to 1-2 days . Treat  bid  as
      twice daily .
2. Implied Negatives: If Gold says  No
    spread , and Generated says  Rash
    localized to chest  (implying no
    spread), mark as COVERED.
3. Elaboration: If the Generated fact
    adds detail that is logically
    consistent with the Gold fact (e.g
    ., Gold:  Pain , Generated:
    Throbbing Pain ), check if it
    contradicts. If it implies the same
     clinical reality, it is SUPPORTED.

Constraint: - You must distinguish
    between a FACT MISSING (Omission)
    and a FACT WRONG (Contradiction).
```

Listing 5: User prompt for Phase 1 evaluation

```
PHASE1_USER =
GOLD FACTS:
{gold_facts}

GENERATED FACTS:
{gen_facts}

STEP 1: ASSESS GOLD FACTS (Recall &
    Accuracy)
For every GOLD Fact, determine its
    status in the Generated Facts:
- COVERED: The clinical concept is
    present (even if phrased
    differently).
- MISMATCH: The Generated facts do not
    match Gold fact (e.g., Gold:  No
    fever , Gen:  Fever ).
- OMITTED: The concept is completely
    absent.

STEP 2: ASSESS GENERATED FACTS (
    Precision)
For every GENERATED Fact, determine its
     relationship to the Gold Facts,
    STRICTLY following the categories
    below. DO NOT invent new categories
    :
- SUPPORTED: Matches a Gold fact (
    semantically).
- MISMATCH: Conflicts with a Gold fact.
- NOT_IN_GOLD: The fact is NOT present
    in the Gold Facts.

OUTPUT JSON:
{{
   gold_assessment : [
    {{
       fact_id :  hpi-001 ,
       status :  COVERED ,
       reasoning :  Gen fact 'hpi-x'
          mentions '1-2 days' which
          matches Gold '24-48 hrs'.
    }},
    {{
       fact_id :  hpi-002 ,
       status :  MISMATCH ,
       reasoning :  Gold says 'No blood
          ', Gen says 'Blood in stool
          '.
    }}
  ],
   gen_assessment : [
    {{
       fact_id :  gen-001 ,
       status :  NOT_IN_GOLD ,
       reasoning :  Mentions working
          from home. Not found in Gold
          summary .
    }}
  ]
}}
```

Listing 6: Prompt for Phase 2 evaluation

```
PHASE2_SYSTEM =
You are an expert Clinical Fact Checker
    . Your task is to verify  Extra
    Facts  that were found in an AI-
    generated note but were missing or
    mismatched from the human Gold
    Summary.
You must determine if these facts are
    valid details found in the
    Transcript or if they are
    hallucinations.

You will be given:
1. The Original Transcript (The
    absolute truth).
2. A list of Extracted Facts (The
    Extra  details to verify).

For EACH fact, you must classify it
    into one of these strict categories
    :

-  VALID_ELABORATION : The fact is
    supported by the transcript. It may
     be a direct quote, a clear
    clinical inference, or a correct
    statement that something was
    negative/not discussed. (This is a
    GOOD extra detail).
-  TRUE_ADDITION : The fact introduces
    information that is NOT present in
    the transcript. The model
    hallucinated specific values, dates
    , or events. (This is a BAD
    hallucination).
-  CONTRADICTION : The fact directly
    conflicts with the transcript. (e.g
    ., Transcript says  No fever , Fact
     says  Fever ).

Crucial Evaluation Rules:
1. Implicit Negatives: If the fact
    states something was  not discussed
      or  unremarkable , and the
    transcript is silent on it, mark as
     VALID_ELABORATION.
2. Clinical Synonyms: Treat medical
    synonyms (Tylenol = Acetaminophen)
    as matches.
3. Inference: If the fact is a logical
    clinical conclusion from the
    transcript (e.g.,  Blue inhaler  ->
      SABA ), mark as VALID_ELABORATION
    .

Output Schema:
You must output a single valid JSON
    object containing a  verdict  list.
     Each object in the list must use
    EXACTLY these keys:
{{
   verdict : [
    {{
       fact_id :  The exact ID provided
           in the input ,
       status :  MUST be one of:
          VALID_ELABORATION,
          TRUE_ADDITION, CONTRADICTION
           ,
       reasoning :  A concise
          explanation quoting the
          transcript if possible
    }}
  ]
}}
```

B.4. Error Generation Prompt

Listing 7: Error Generation Prompt

```
ERRORS_PROMPT =
You are a clinical NLP expert. You are
    given all transcript facts in the
    following JSON structure:

{facts_json}

Each fact includes:
- fact_id
- content
- source_text
- and is grouped under a section key (e
    .g., HPI, ROS, PMH).

Sections can be categorized into one of
     SOAP categories with these Mapping
    :**
- **S (Subjective):** Chief_Complaint,
    History_of_Present_Illness,
    Past_Medical_History, Medications,
    Allergies
- **O (Objective):** Exam, Labs, Vital
    Signs, Imaging
- **A (Assessment):** Assessment,
    Diagnosis
- **P (Plan):** Plan, Medications,
    Procedures
```

```
Your tasks:
1. Review ALL facts and their
    source_text.
2. Select **10 total facts** that are
    highly suitable for one of the
    following clinically important
    error types. Ensure that there are
    facts from **each SOAP section**.
    Error types:
- Omission
- Censor Medical Term
- Censor Value
- Replace with Similar Medical Term
- Negation Flip
- Diagnosis Censor

Suitability rules (must match
    source_text):
- **Omission**: Applies to any fact.
- **Censor Medical Term**: Only if the
    text contains specific medical
    terms, drug names, or disease names
    .
- **Censor Value**: Only if numbers,
    dosages, or measurements occur.
- **Replace with Similar Medical Term
    **: Only if a specific disease,
    symptom, test, or drug name appears
     that can be plausibly swapped.
- **Negation Flip**: Only if the text
    clearly includes negation ( denies
     ,  no ,  not ) or a positive
    assertion that can be flipped.
- **Diagnosis Censor**: Only if a
    diagnosis is explicitly mentioned (
    e.g.,  you have gastroenteritis ).

3. For each selected fact:
- Assign exactly **one** error_type.
- Produce an **altered_source_text**
    with the chosen error realistically
     applied according to these rules:
- **Omission**: Set altered_source_text
     = null.
- **Censor Medical Term**: Delete key
    medical terms, drug names, disease
    names or diagnostic terms.
- **Censor Value**: Remove key
    numerical values, lab/test results,
     dosages or measurements.
 - **Replace with Similar Medical Term
     **: Swap a disease, symptom, test
     or drug name with a real, similar-
     sounding medical term (plausible
     but incorrect).
- **Negation Flip**: Flip the negation
    status (e.g.,  no fever  ->  fever
     ).
- **Diagnosis Censor**: EITHER remove
    the diagnosis name (e.g.,  you have
     [BLANK] ) OR remove the entire
    string (empty string). Randomize
    between these two options.

Return **only JSON**, with 10 items
    total, each of the form:
[
 {{
         fact_id :  ... ,
         fact_section :  ... ,
         error_type :  ... ,
         original_source_text :  ... ,
         altered_source_text :  ...
 }}
]
Do not include any explanation outside
    the JSON.
```

B.5. Robustness Analysis Prompts

Listing 8: System Prompt for Note to Note Comparison in Robustness Evaluation

```
RQ2_DIFF_SYSTEM =
You are a  Stability Analyst  for
    Clinical AI.
Your Goal: Compare two sets of clinical
     facts generated by the SAME model
    but from slightly different inputs
    (Input A vs. Input B).

You must identify stability issues:
1. Did the model drop information
    present in the clinical note? (
    OMISSION)
2. Did the model change details
    regarding the same event? (
    CONTRADICTION)
3. Did the model add new things not in
    the initial clean set? (ADDITION)

Definitions:
- REFERENCE = Facts from the  Clean
    Transcript run.
- CANDIDATE = Facts from the  Noisy/
    Modified  Transcript run.
```

Listing 9: User Prompt for Note to Note Comparison in Robustness Evaluation

```
RQ2_DIFF_USER =
We ran a clinical model on a  Clean
    Transcript  (Reference) and then
    again on a  Noisy Transcript  (
    Candidate).
Compare the extracted facts to see how
    the noise affected the output.

REFERENCE FACTS (Clean Baseline):
{clean_facts}

CANDIDATE FACTS (Noisy Run):
{noisy_facts}

INSTRUCTIONS:
1. **Map Reference to Candidate:** For
    every fact in the Reference, check
    if it survived in the Candidate.
   - **PRESERVED:** The fact exists in
       the Candidate (semantically
       equivalent).
   - **OMITTED:** The fact is
       completely missing in the
       Candidate.
   - **CONTRADICTED:** The Candidate
       contains a conflicting version (
       e.g.,  Seroxat  vs  Cerazette ,
        Left side  vs  Right side ).

2. **Check for Ripple Effects (
    Additions):** Check if the
    Candidate contains *new* facts not
    present in the Reference.
   - **NEW_ADDITION:** Information
       found in Candidate but NOT in
       Reference. (This suggests the
       noise triggered a hallucination)
       .

OUTPUT JSON FORMAT:
{{
   stability_analysis : [
    {{
       ref_fact_id :  clean-hpi-01 ,
       status :  PRESERVED ,
       candidate_match_id :  noisy-hpi
          -01 ,
       reasoning :  Both state patient
          has headache .
    }},
    {{
       ref_fact_id :  clean-hpi-02 ,
       status :  OMITTED ,
       reasoning :  Reference mentions
          'Diabetes', but Candidate
          completely ignores it .
    }}
  ],
   noise_induced_hallucinations : [
    {{
       candidate_fact_id :  noisy-plan
          -04 ,
       content :  Patient referred to
          Cardiology. ,
       reasoning :  This referral was
          NOT in the Clean run. The
          noise might have confused the
           model into adding it .
    }}
  ]
}}
```

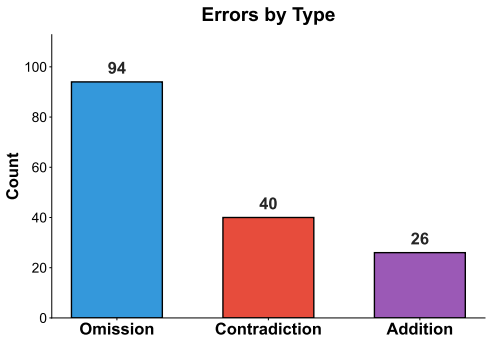

Figure 2: Distribution of annotated error types in the dataset.

Appendix C. Additional Results

We summarize the outcomes of the manual dataset annotation conducted on PriMock57. Across all consultations, a total of 162 factual issues were identified. Of these, 148 (91.4%) were classified as resolvable through targeted corrections to transcripts or clinical notes, while 14 (8.6%) were deemed unresolvable.

Fig. 2 presents the distribution of identified error types, with omissions constituting the majority, followed by contradictions and additions. Fig. 3 illustrates the proportion of resolvable versus unresolvable issues. These results reflect the final dataset composition used for evaluation.

Table 6: Examples of Resolvable and Unresolvable Annotation Issues

Category	Type	Clinical Note Says	Transcript Says
Resolvable Issues — Corrected in Dataset			
Resolvable	Omission	(No allergy documented)	"I'm allergic to penicillin"
Resolvable	Contradiction	"No known drug allergies"	"I'm allergic to Clindamycin"
Resolvable	Contradiction	"Blood in stool"	"Blood in urine"
Resolvable	Contradiction	"Lives with parents"	"I live with my husband and his family"
Resolvable	Addition	"Plan: Paracetamol"	(No mention of paracetamol)
Unresolvable Issues — Consultation Excluded from Dataset			
Unresolvable	Contradiction	"No dysphagia"	"Something stuck in my throat while drinking" [1]
Unresolvable	Contradiction	"Not short of breath"	"I've been quite short of breath lately" [1]
Unresolvable	Contradiction	"DH: on Cerazette"	"Seroxat" [2]

Limitation	Phase	Mitigation
Contradictions misclassified as omissions	1	Explicit CONTRADICTED label with required justification
Duplicate reference facts	1	Fact-level deduplication via QNOTE schema
Ambiguous elaboration	2	Transcript verification with VALID_ELABORATION label
Numerical or unit variation	1	Temporal and unit equivalence normalization
Synonymy and paraphrasing	Both	Clinical synonym handling and inference rules

Table 7: Limitations of the fact comparison agent and corresponding mitigation strategies.

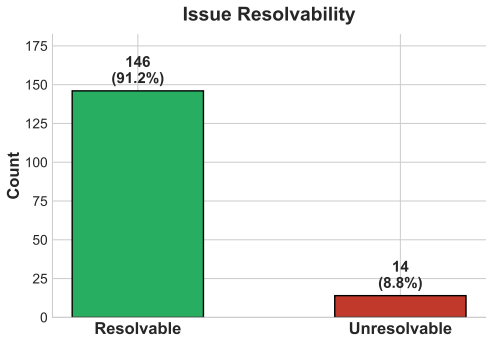

Figure 3: Distribution of annotated error types in the dataset.

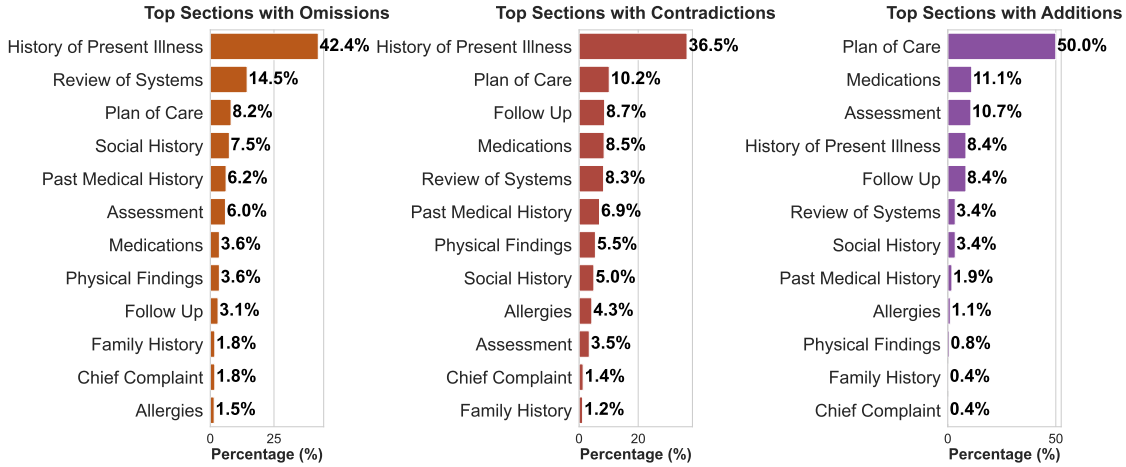

Figure 4: Distribution of errors across clinical note sections. The horizontal bars represent the percentage of total errors (Omissions, Contradictions, and Additions) contributed by each section. History of Present Illness accounts for the largest share of errors. Plan of Care is notably prone to hallucinations, accounting for 50% of all additions generated by the models.