Discharge Instructions are not One Task: Grounding Differences Between Surgical and Non-Surgical Admissions

Mayank Jobanputra[1] **Justin Xu**[2] **Samarth H. Oza**[3] **Hulma Naseer**[1]
Yifan Wang[1] **Blerta Veseli**[1] **Chandralekha Kona**[3] **Haochen Cui**[2]
David W. Eyre[2] **Vera Demberg**[1]

[1]Saarland University [2]University of Oxford [3]Independent Researcher

Abstract

Discharge instructions are patient-facing, safety-critical documents that guide medication use, follow-up care, and recovery after hospitalization. Because they must synthesize information across the clinical record and often include post-discharge guidance not stated verbatim in the EHR, they are a difficult target for clinical text generation. In this work, we study discharge instructions in MIMIC-IV through a grounding-first lens. Using two LLMs, we decompose each discharge instruction into medically relevant statements and verify them against the Electronic Health Record (EHR). We find that discharge instructions for Surgical admissions are much longer, averaging roughly 24–25 statements per admission versus 11–12 in Non-Surgical cases, while supported content remains similar in absolute count. The additional Surgical content is dominated by statements that are not directly stated in the record or require clinically plausible extrapolation. Through this analysis, we advocate for better grounding and completeness evaluations at a fine-grained level, establishing a foundational step toward safer and more reliable discharge-instruction generation.

1 Introduction

Discharge instruction generation has become a central clinical text generation problem because it sits directly at the interface between hospital care and patient self-management. At the center of this progress is the BioNLP 2024 shared task *Discharge Me!*, which established a public benchmark and made discharge documentation a focal task for the community (Xu et al., 2024).

At a high level, discharge instructions should answer three questions for the patient: what happened during this admission, what they should do after leaving the hospital, and what signs should prompt them to seek further medical care. Yet, Discharge instructions can vary a lot depending on the type of admission. Some admissions, especially surgical admissions, contain substantially more content for post-procedural counseling and clinician-authored guidance that may not be directly inferable from the EHR alone. Such cases pose a more challenging grounding problem than shorter, more routine non-surgical admissions. Figure 1 provides an overview of these differences per admission type.

This difference in complexity is easy to overlook in aggregate evaluation. Xu et al. (2024) reports high correctness scores for top-performing systems during clinician review, but it does not provide explicit quantitative detail about the length or procedural complexity of the reviewed discharge instructions. Without that information, it is difficult to know how well current systems handle the most complex and weakly grounded cases. We therefore argue that before asking whether systems can generate discharge instructions well, we should first ask a more foundational question related to grounding and completeness:

> *What content in a discharge instruction can be derived from the EHR?*

In this work, we systematically study the clinician-written discharge instructions themselves to understand which parts of human-written discharge instructions are explicitly grounded in the EHR. Carrying out this study manually is not feasible due to the required number of clinician hours. Hence, we use Large Language Models (LLMs) to analyze the MIMIC-IV dataset, similar to the VeriFact (Chung et al., 2025) framework. For the systematicity, we segment the data into Surgical and Non-Surgical subsets to answer:

1. What are the quantitative differences in the discharge instructions between Surgical and Non-Surgical cases?
2. Which category relies more heavily on ungrounded clinical extrapolation?

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 426–434

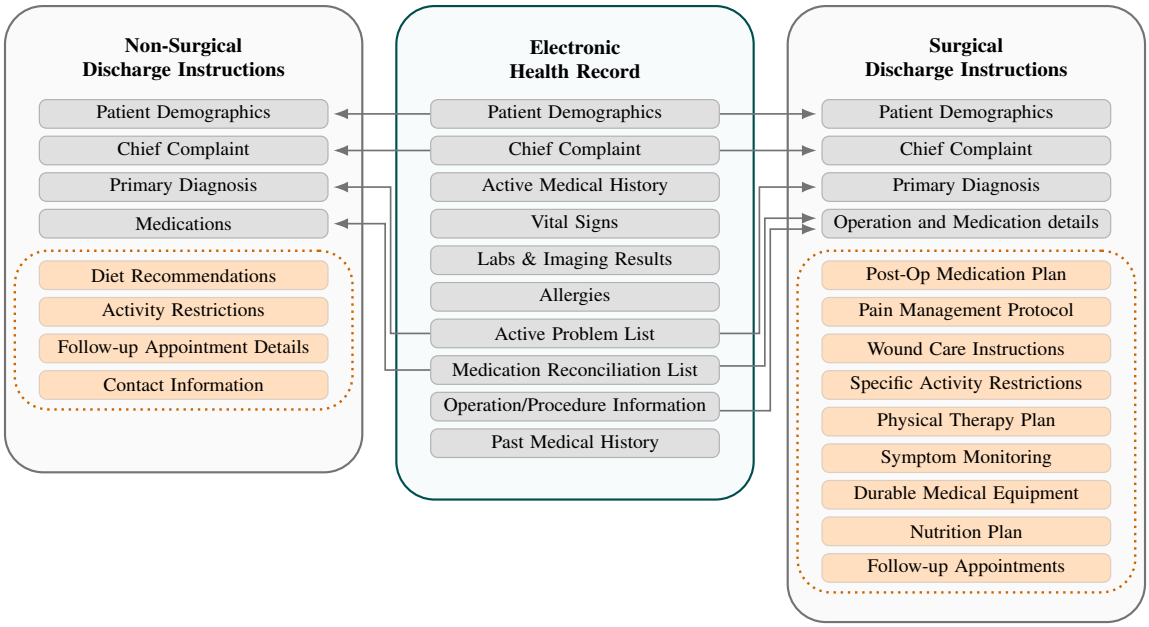

Figure 1: **Surgical and Non-Surgical discharge instructions differ not only in length but in verifiability.** Surgical discharge instructions contain substantially more medically important post-discharge care statements per admission than Non-Surgical instructions. However, the increase is driven primarily by statements that are not directly verifiable from the Electronic Health Records.

Our choice to classify cohorts using the clinical narrative is driven by the structural limitations of structured Electronic Health Record (EHR) fields. While ICD-10 diagnosis codes indicate clinical conditions, they do not reliably confirm whether an operative procedure was performed during a specific hospitalization. Conversely, while procedure codes provide a more direct signal, individual admissions frequently contain multiple overlapping codes. Constructing a rule-based mapping table to resolve these cases would require extensive exception logic and sensitivity analyses to handle procedural ambiguities. Utilizing the text-based clinical narrative allows the pipeline to dynamically determine the primary treatment pathway based on the comprehensive hospital course.

We view this work as a stepping stone toward creating a fine-grained benchmark for evaluating the faithfulness and completeness of generated discharge instructions.

2 Related Work

Early work on discharge-instruction generation used more structured formulations. J Kurisinkel and Chen (2019) cast the task as set-to-ordered-text generation from ICD codes, showing that discharge instructions often require inference rather than direct copying from the input. Re^{3}Writer moves closer to record-grounded generation with a retrieve-reason-refine framework designed to improve faithfulness in patient-instruction generation (Liu et al., 2022). Recent work has focused on benchmarking and improving discharge-section generation at scale. The BioNLP 2024 shared task *Discharge Me!* established a public benchmark for generating *Brief Hospital Course* and *Discharge Instructions*, and representative system papers explored retrieval, section selection, and prompt- or fine-tuning-based methods (Xu et al., 2024; Koontz et al., 2024; Lyu et al., 2024; Damm et al., 2024).

Building upon these evaluations, our work takes a step further by studying the fine-grained clinical differences within the dataset. Specifically, we explore the distinction between Surgical and Non-Surgical admissions to evaluate how grounding complexity varies across different types of patient care.

3 Method

In this section, we provide an overview of each stage of our pipeline for classifying the parts of the discharge instructions.

Dataset. We analyze the *Discharge Me!* Phase II Test split for all our experiments to reduce the risk that the evaluated discharge instructions were seen by the LLMs during instruction tuning or post-training. As the final held-out test split, it provides

a cleaner setting for studying record-grounded verification than training or validation data. This split contains discharge instructions for 10,962 Hospital admissions.

3.1 DI extraction and cleaning

We first transform the discharge notes into a structured sentence-level format using a constrained extraction pipeline. We use GPT-OSS-120B (OpenAI, 2025) to: (i) identify the *Discharge Instructions* section, (ii) copy them verbatim, (iii) remove formatting artifacts and rhetorical questions, (iv) group related clauses into meaningful declarative sentence units, and (v) predict a binary label, SURGICAL or NON-SURGICAL, for the given datapoint. The exact prompts are available in the appendix D.

We observed that GPT-OSS-120B classified most discharge instructions as Non-Surgical. To mitigate this issue, we introduced a secondary deterministic validation step that is applied to the raw clinical narrative. Any admission initially classified as Non-Surgical is re-categorized as Surgical if the discharge instruction documents a major invasive procedure or matches a predefined lexicon of surgical keywords (e.g., *post-op*, *incision*, *sutures*).

3.2 Sentence categorization with LLMs

We then evaluate each discharge-instruction sentence using two LLMs: Llama-3.3-70B-Instruct (Grattafiori et al., 2024) and MediX-R1 (Mullappilly et al., 2026). We choose a Llama-family model because VeriFact uses Llama 3.1 70B for EHR-grounded verification, and we include MediX-R1 as a recent medically specialized LLM with strong reported benchmark performance. Each LLM decomposes a larger declarative sentence unit into smaller medically important statements, retrieves supporting or refuting evidence from the diagnoses and hospital course, and assigns a verification label. The exact prompts are available in the appendix D. We use three main verification labels: *Supported*, *Not Supported*, and *Not Addressed*.

Supported statements are explicitly backed by the clinical context. *Not Supported* statements are addressed by the record but only partially supported, lack sufficient detail, or are inconsistent with it. *Not Addressed* statements are not directly verifiable from the available evidence. We also track two finer subcategories. Within *Not Supported*, we separate *Contradiction* from statements that lack sufficient detail. Within *Not Addressed*, we separate *Clinically Plausible Extrapolation* from statements that are completely not addressed by the record. Since, we are evaluating clinician written summaries, we expect most of the statements to be either *Supported* or *Clinically Plausible Extrapolation*. Other cases may reveal the limitations of LLMs or genuine errors by clinicians while writing Discharge Instructions.

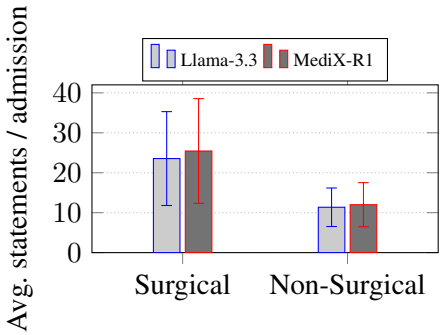

Figure 2: Average number of medically important statements per admission, with standard deviation as error bars. Surgical discharge instructions are approximately twice as long as Non-Surgical ones for both LLMs.

Code. Our code is available on GitHub[1].

4 Results

We start by analyzing the lengths of discharge instructions. We find that the discharge instructions for Surgical admissions are much longer than Non-Surgical admissions. Figure 2 shows that this difference is large and consistent across both LLMs. This suggests that the grounding problem is not uniform across the dataset. Surgical discharge instructions place a substantially larger informational burden on any generation or verification system.

Figure 3 shows that the additional content in Surgical discharge instructions is concentrated in statements that are not directly verifiable from the available EHR evidence (see Table 4 in the Appendix for exact details). For both LLMs, the number of *Supported* statements is similar in Surgical and Non-Surgical admissions, but Surgical admissions contain many more statements in the broad *Not Addressed* category. This means that Surgical discharge instructions are longer not because they contain more directly supported information, but because they include more content that is missing from the diagnoses and Brief Hospital Course or requires inference. Surprisingly, within the *Not*

[1] https://github.com/mayankjobanputra/dischage_differences

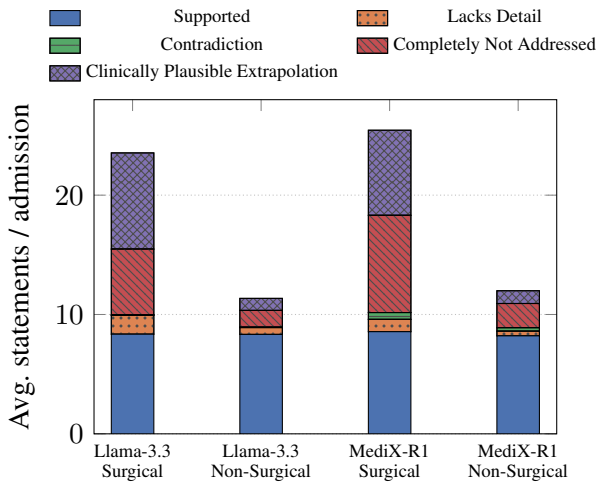

Figure 3: Fine-grained classification results for Surgical and Non-Surgical hospital admissions.

Category	Surgical	Non-Surgical
Broad labels (%)	79.79	85.16
Broad-label κ	0.6366	0.6645
Fine labels (%)	66.02	81.02
Fine-label κ	0.4845	0.5955

Table 1: Agreement stats for Llama-3.3 and MediX-R1. Broad labels refer to *Supported*, *Not Supported*, and *Not Addressed*; fine labels use the full five-label scheme.

Addressed category, the increase in Surgical admissions is distributed across both *Completely Not Addressed* statements and *Clinically Plausible Extrapolation*. We discuss our findings from the manual error analysis on this in Section 5.

Next, we measure the agreement between the two LLMs. Overall, we observe over 82% agreement (κ = 0.66) between Llama-3.3 and MediX-R1 at the broad categories level, but at the fine-grained level, the agreement drops to 72.41% (κ = 0.53). We provide a more detailed analysis in Table 1.

Based on the results from Table 1, we can also see that the agreement is lower in the Surgical admissions than in the Non-Surgical admissions at both the broad level and the fine level. While the agreement levels are encouraging, the overall results need to be verified by clinicians before they can be taken as evidence that the LLM labels are reliable on their own and also to decide which LLM performs better on this classification task.

5 Human Evaluation

To test whether the LLM labels are clinically meaningful, we conducted a targeted human study on 24 admissions comprising 544 medically important statements. The study was stratified across the three subsets introduced above: *Cross-LLM Consensus*, *Cross-LLM Disagreement*, and *Clinically Plausible*. One clinician annotated all 24 admissions. A second clinician independently annotated a subset of 6 admissions, yielding 146 aligned statements for inter-rater agreement analysis.

Comparison	Agreement (%)	κ
Human–human (broad)	96.58	0.8874
Human–human (fine)	91.78	0.7785
Human–Llama-3.3 (broad)	81.2	0.4819
Human–Llama-3.3 (fine)	45.8	0.2333
Human–MediX-R1 (broad)	71.1	0.3591
Human–MediX-R1 (fine)	45.6	0.1918

Table 2: Human-evaluation results. Human–human agreement is computed on the double-annotated subset of 146 aligned statements. Human–Llama-3.3 and Human–MediX-R1 agreements are computed on the 544 aligned statements.

Table 2 shows high human–human agreement at both the broad level ($\kappa = 0.8874$) and the fine level ($\kappa = 0.7785$), indicating that the labeling scheme can be applied consistently by clinicians.

At the broad level, Llama-3.3 agrees with the human annotations more than MediX-R1 (81.2% vs. 71.1%; $\kappa = 0.4819$ vs. 0.3591). At the fine level, however, both LLMs perform similarly poorly (45.8% vs. 45.6%), with low κ values in both cases. This shows that neither LLM is reliable enough to replace clinician judgment, especially for closely related categories such as *Not Addressed* and *Clinically Plausible Extrapolation*.

Error analysis The most frequent error is confusion between *Not Addressed* and *Clinically Plausible Extrapolation*. Routine post-operative advice, such as lifting restrictions, driving restrictions while taking pain medication, light exercise, and incision care, is often labeled *Not Addressed* by the LLMs even when the clinician labels it *Clinically Plausible Extrapolation*.

A second recurring error is over-attribution of support to discharge boilerplate. MediX-R1 more often labels service-level instructions as *Supported*, while the clinician more often labels them *Clinically Plausible Extrapolation* or *Not Addressed*. Examples include refill instructions, brace use, diet advice, and resumption of home medications.

MediX-R1 appears relatively stronger on procedure-specific *Clinically Plausible Extrapolation*, while Llama-3.3 is stronger on explicitly

Supported statements and achieves higher overall agreement with human labels. We provide example statements along with human and LLM labels in Appendix C.

6 Conclusion

In this work, we show that Surgical and Non-Surgical admissions differ substantially in both length and verifiability: Surgical discharge instructions contain many more medically important statements, and the increase is driven mainly by content that is not directly supported by the EHR text. This suggests that future work on discharge-instruction generation should report results separately for Surgical and Non-Surgical cohorts and move toward more fine-grained evaluation of grounding. Our human study shows that LLMs can help surface likely ungrounded content and identify difficult regions of the dataset, but they are not reliable enough for fine-grained labeling without clinician validation. Based on our findings, we suggest that reliable evaluation of generated discharge instructions likely requires resources beyond the EHR alone. A potential future direction can be an agentic setup that can use additional clinical tools, protocols, and knowledge sources in a Biomni-like framework (Huang et al., 2025) adapted to discharge-specific needs.

Limitations

Our study has some limitations. We use only open-weight LLMs in our experiments and do not compare against strong closed-source models, which may differ in both decomposition quality and verification behavior. We also report results from a single run per model, so we do not measure variability across random seeds or decoding configurations. Our analysis is based on a single prompt configuration per stage and does not explore a broader prompt sensitivity study. Different prompt wording, output constraints, or evidence-formatting choices could affect both the extracted statements and the assigned verification labels. More generally, we evaluate only two LLMs, so the observed agreement patterns may not fully characterize the behavior of other model families. We also do not provide any analysis on how exactly this affects recent LLMs's capability of generating factual and clinically complete discharge instructions.

Ethics Statement

This work studies de-identified clinical text and proposes auditing methods for safety-critical patient-facing generation. We emphasize that unsupported or weakly supported discharge instructions can create direct patient harm, and that automated systems should not be deployed without clinician oversight, calibration, and site-specific validation. We also do not use any closed-source LLMs such as GPT-5, Gemini to avoid any kind of sensitive data leakage. All clinicians participating in the study are board-certified physicians and are co-authors of this work.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback and insightful comments. MJ and VD acknowledge the funding by Deutsche Forschungsgemeinschaft (DFG) grant 389792660 as part of TRR 248 – CPEC, see https://perspicuous-computing.science. YW and BV were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2853/1 "Neuroexplicit Models of Language, Vision, and Action" - project number 471607914. JX gratefully acknowledges joint support from Canadian Institutes of Health Research (CIHR) Project ID 202410BCB-535721-77482 (Bioinformatics and Computational Biology), Nuffield Department of Medicine (NDM), and Oxford University Press (OUP). HC is supported by the Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (CIFMS) Grant Number 2024-I2M-2-001-1.

References

Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, and 1 others. 2025. Verifying facts in patient care documents generated by large language models using electronic health records. *NEJM AI*, 3(1):AIdbp2500418.

Hendrik Damm, Tabea Margareta Grace Pakull, Bahadır Eryılmaz, Helmut Becker, Ahmad Idrissi-Yaghir, Henning Schäfer, Sergej Schultenkämper, and Christoph M. Friedrich. 2024. WisPerMed at "discharge me!": Advancing text generation in healthcare with large language models, dynamic expert selection, and priming techniques on MIMIC-IV. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 105–121, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, and 1 others. 2025. Biomni: A general-purpose biomedical ai agent. *biorxiv*.

Litton J Kurisinkel and Nancy Chen. 2019. Set to ordered text: Generating discharge instructions from medical billing codes. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6165–6175, Hong Kong, China. Association for Computational Linguistics.

Jordan C. Koontz, Maite Oronoz, and Alicia Pérez. 2024. Ixa-med at discharge me! retrieval-assisted generation for streamlining discharge documentation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 658–663, Bangkok, Thailand. Association for Computational Linguistics.

Fenglin Liu, Bang Yang, Chenyu You, Xian Wu, Shen Ge, Zhangdaihong Liu, Xu Sun, Yang Yang, and David A. Clifton. 2022. Retrieve, reason, and refine: Generating accurate and faithful patient instructions. In *Advances in Neural Information Processing Systems*, volume 35.

Mengxian Lyu, Cheng Peng, Daniel Paredes, Ziyi Chen, Aokun Chen, Jiang Bian, and Yonghui Wu. 2024. UF-HOBI at "discharge me!": A hybrid solution for discharge summary generation through prompt-based tuning of GatorTronGPT models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 685–695, Bangkok, Thailand. Association for Computational Linguistics.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Omair Mohamed, Mohamed Zidan, Fahad Khan, Salman Khan, Rao Anwer, and Hisham Cholakkal. 2026. Medix-r1: Open ended medical reinforcement learning. *Preprint*, arXiv:2602.23363.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98, Bangkok, Thailand. Association for Computational Linguistics.

A Dataset statistics

The summary of corpus and label is provided in Table 3.

B Detailed Distribution Statistics

Table 4 provides the numerical distribution of the broad and fine-grained verification categories for both Llama-3.3 and MediX-R1.

C Qualitative Error Analysis

The examples in Table 5 summarizes the main qualitative error patterns observed in the human evaluation. These errors are grouped by the subset from which they were sampled.

D Prompt Templates

D.1 DI extraction and cleaning: system prompt

```
You are an expert clinical data extraction
↪  assistant. Your primary directive is to
↪  analyze complex hospital admission
↪  records and extract specific,
↪  structured information into a strictly
↪  formatted JSON object.
```

D.2 DI extraction and cleaning: task prompt

```
**TASKS:**

**Task-1: Verbatim Copy of Discharge
↪  Instructions**
- Locate the "Discharge Instructions"
↪  section within the clinical note.
- Extract and copy the entire text of this
↪  section verbatim, exactly as it appears
↪  in the text.

**Task-2: Split and Group into Meaningful
↪  Sentences**
- Take the verbatim discharge instructions
↪  and split them into declarative
↪  sentences.
- Remove all rhetorical or conversational
↪  questions (e.g., "Questions?").
- Remove all splitting markers or visual
↪  separators like "========" or "****".
- **IMPORTANT:** You must group sentences
↪  that talk about the exact same topic or
↪  refer to each other into a single
↪  string item.
  - *Example of Grouping:* If the text says:
  ↪  "We have also prescribed a nebulizer
  ↪  machine that you can use at home. We
  ↪  only recommend using this if you are
  ↪  too short of breath to use your
  ↪  albuterol inhaler effectively."
    *Do NOT split this. Output it as a
    ↪  single array item:* ["We have also
    ↪  prescribed a nebulizer machine that
    ↪  you can use at home. We only
    ↪  recommend using this if you are too
    ↪  short of breath to use your
    ↪  albuterol inhaler effectively."]

**Task-3: Identify the Treatment Service
↪  Type**
```

	Llama-3.3	MediX-R1
Corpus statistics		
Admissions with outputs from both models	10,958	
Total sentences	74,576	77,281
Sentences / admission (avg.)	6.81	7.05
Sentences / admission (max.)	53	49
Total Medical statements	159,280	168,649
Medical statements / admission (avg.)	14.54	15.39
Medical statements / admission (max.)	102	115
Verification label counts		
Supported	94,669	93,503
Not Supported	8,571	5,793
Lacks Detail	8,060	1,936
Contradiction	511	3,857
Not Addressed	55,529	65,496
Completely Not Addressed	26,144	38,627
Clinically Plausible Extrapolation	29,385	26,869

Table 3: Summary of corpus and label statistics.

	Surgical Admissions		**Non-Surgical Admissions**	
Category	**Llama-3.3**	**MediX-R1**	**Llama-3.3**	**MediX-R1**
Broad Labels				
Supported	8.35 ± 5.58	8.56 ± 6.08	8.33 ± 3.96	8.21 ± 4.06
Not Supported	1.64 ± 2.32	1.59 ± 2.13	0.61 ± 0.93	0.69 ± 1.02
Not Addressed	13.55 ± 9.61	15.29 ± 10.43	2.41 ± 2.27	3.10 ± 2.89
Fine-Grained Subcategories				
Lacks Detail	1.57 ± 2.28	1.03 ± 1.70	0.57 ± 0.88	0.39 ± 0.78
Contradiction	0.07 ± 0.33	0.57 ± 1.24	0.04 ± 0.25	0.29 ± 0.61
Completely Not Addressed	5.49 ± 7.00	8.15 ± 8.16	1.40 ± 1.57	2.03 ± 1.77
Clinically Plausible Extrapolation	8.06 ± 8.60	7.13 ± 8.11	1.01 ± 1.74	1.07 ± 2.16

Table 4: Mean and standard deviation (Mean $\pm$ SD) of medically important statements per admission assigned by Llama-3.3 and MediX-R1 across all evaluation categories.

```
- Analyze the hospital course and diagnoses.
- Choose `SURGICAL` if the patient underwent
↪  a major therapeutic surgical procedure
↪  in an operating room during this stay
↪  and the "Service" is "SURGERY".
- Choose `NON-SURGICAL` if the patient was
↪  managed medically.
*(Note: Try your best to classify this
↪  based on the clinical narrative. The
↪  system will double-check your work).*

**Task-4: Extract Major Surgical
↪  Procedures**
- If the service is `SURGERY`, list the
↪  names of the "Major Surgical or
↪  Invasive Procedure" performed.
- If no major surgical procedure is found or
↪  the service is `MEDICINE`, return
↪  ["NA"].

**Task-5: Anonymization**
- Assign a fake name to the patient
↪  `Noname`.
- Ensure any remaining anonymization
↪  markers (like `___` or `[**...**]`) are
↪  replaced with generic placeholder
↪  `[masked info]`.

**REQUIRED JSON OUTPUT FORMAT:**
{
  "patient_name": "<Output from Task 5>",
  "verbatim_discharge_instructions":
  ↪  "<Output from Task 1>",
  "discharge_instructions_sentences":
  ↪  ["<Output from Task 2>"],
  "clinical_pathway": "<Output from Task
  ↪  3>",
  "procedures": ["<Output from Task 4>"]
}

**PATIENT DATA:**
Subject ID: {{ subject_id }}
Hospital Admission ID: {{ hadm_id }}

--- DIAGNOSES ---
{{ diagnoses_text }}

--- RAW CLINICAL NOTE ---
{{ raw_note_text }}
```

Subset	Claim	Human	Llama	Medix
Cross-LLM Consensus	"Do not drive until you have stopped taking pain medicine and feel you could respond in an emergency"	CPE	NA	NSup.
Cross-LLM Consensus	"Don't lift more than [masked info] lbs for 4 weeks"	CPE	NA	NA
Cross-LLM Consensus	"You may start some light exercise when you feel comfortable"	CPE	NA	NA
Cross-LLM Consensus	"You will need to stay out of bathtubs or swimming pools for a time while your incision is healing"	CPE	NA	NA
Cross-LLM Consensus	"You may feel weak or washed out for a couple of weeks"	CPE	NA	NA
Cross-LLM Disagreement	"You may climb stairs"	CPE	NA	Sup.
Cross-LLM Disagreement	"Eat a normal healthy diet."	CPE	NSup.	Sup.
Cross-LLM Disagreement	"You do not need a brace."	CPE	NA	Sup.
Cross-LLM Disagreement	"You should resume taking your normal home medications."	CPE	NSup.	Sup.
Cross-LLM Disagreement	"Please allow 72 hours for refill of narcotic prescriptions, so please plan ahead."	CPE	NA	Sup.
Cross-LLM Disagreement	"We are not allowed to call in narcotic prescriptions to the pharmacy."	NA	NA	Sup.
Clinically Plausible	"You may have a sore throat because of a tube that was in your throat during surgery"	CPE	NA	CPE
Clinically Plausible	"You may shower at this time but keep your incision dry."	CPE	NA	CPE
Clinically Plausible	"It is best to keep your incision open to air but it is ok to cover it when outside."	CPE	NA	CPE
Clinically Plausible	"Avoid heavy lifting, running, climbing, or other strenuous exercise until your follow-up appointment."	CPE	NA	CPE
Clinically Plausible	"You may take leisurely walks and slowly increase your activity at your own pace once you are symptom free at rest."	CPE	NSup.	CPE
Clinically Plausible	"No contact sports until cleared by your neurosurgeon."	CPE	NA	CPE

Table 5: Representative qualitative disagreement cases from the human evaluation. CPE = *Clinically Plausible Extrapolation*, NA = *Not Addressed*, NSup. = *Not Supported*, Sup. = *Supported*.

D.3 Sentence categorization: system prompt

```
You are an expert attending physician and clinical auditor. Your task is to verify the factual accuracy of hospital discharge instructions against the patient's ground-truth clinical record.
```

D.4 Sentence categorization: evaluation prompt

```
**TASKS:**

**Task-1: Atomic Extraction**
- For each sentence in the "TARGET SENTENCES TO EVALUATE" list, break it down into independent, verifiable atomic claims.
- Example: "Take 5mg of Lisinopril daily and follow up in 2 weeks." -> Claim 1: "Take 5mg of Lisinopril daily.", Claim 2: "Follow up in 2 weeks."

**Task-2: Evidence Retrieval and Reasoning**
- Search the provided CLINICAL CONTEXT for evidence that supports or refutes each atomic claim.
- Extract the direct quotes from the context (`evidence_quoted`). Write "None" if the claim is not addressed.
- Explain your clinical logic step-by-step (`thought_process`) *before* assigning a verdict.
- **CRITICAL JSON RULE:** Do NOT use double quotes (") inside your `thought_process`, `atomic_claim`, or `evidence_quoted` fields. If you must quote something from the text, use single quotes ('). Unescaped double quotes will corrupt the JSON object.

**Task-3: Verdict Classification**
- Classify each atomic claim into exactly one of the following categories (`verdict`):
  - `Supported`: The claim is explicitly backed by the context.
  - `Clinically Plausible Extrapolation`: The claim isn't explicitly written, but is standard, safe medical practice based on the diagnoses (e.g., advising a heart failure patient to limit salt).
  - `Contradiction`: The claim directly conflicts with the context (e.g., wrong medication, wrong dosage, wrong follow-up timeline).
  - `Not Addressed`: The claim cannot be verified or safely extrapolated from the context.
  - `Not Supported`: If the claim is addressed and not fully supported by the reference context. This includes cases where the text is only partially supported by the reference context.
```

```
**REQUIRED JSON OUTPUT FORMAT:**
{
  "evaluations": [
    {
      "original_sentence": "<The full target
      ↪  sentence being evaluated>",
      "atomic_evaluations": [
        {
          "atomic_claim": "<The extracted
          ↪  atomic claim>",
          "thought_process": "<Your
          ↪  step-by-step reasoning>",
          "evidence_quoted": "<Direct quote
          ↪  from context or 'None'>",
         "verdict": "<Supported | Clinically
          ↪  Plausible Extrapolation |
          ↪  Contradiction | Not Addressed
          ↪  | Not Supported>"
        }
      ]
    }
  ]
}

**CLINICAL CONTEXT:**
--- Diagnoses ---
{{ diagnoses_text }}

--- Hospital Course ---
{{ hospital_course_text }}

**TARGET SENTENCES TO EVALUATE:**
{{ sentences_json_list }}

Return the result strictly as a structured
↪  JSON object matching the requested
↪  schema.
```