ACTIONABLE FEEDBACK: IMPROVING MODEL CAPABILITIES FROM USER EDIT DATA ON CLINICAL DOCUMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly being deployed in clinical documentation workflows, where clinicians review and edit their outputs before use. In principle, these clinician edits represent a scalable source of supervision for model training. In practice, not all edits represent *actionable* feedback: for instance, on data drawn from real-world deployment of an AI clinical documentation product, our automated review suggests that over 45% of edited outputs contain user-added information that is not directly supported by the inputs to the model. Naively fine-tuning on such edits risks teaching the model to hallucinate. In this paper, we develop a method to address this challenge. Given that real clinician edits on deployed documentation tools are not publicly available, we first develop a taxonomy of clinician-editing behaviors, informed by data from a deployed AI clinical documentation product. Then, we develop a method for generating semi-synthetic clinician-edit data, informed by our taxonomy, on a discharge-summary generation task based on MIMIC-III, a public dataset. Finally, we introduce simple filtering methods for identifying *actionable* edits, with accuracy evaluated by expert review. We show empirically, on our semi-synthetic dataset, that naive fine-tuning on raw edited output *increases* ungrounded content, while our approach of progressively filtering toward grounded edits enables improvement on clinician-alignment metrics while retaining faithfulness to the input context.

1 INTRODUCTION

A central challenge in clinical documentation is summarizing large volumes of patient information into concise notes that clinicians can readily use for patient care. Large language model (LLM)-based applications are increasingly used to automate this process, ranging from ambient scribes that draft patient-visit summaries from clinician-patient conversations (Tierney et al., 2024; Lukac et al., 2025; Afshar et al., 2025) to systems that generate discharge summaries from longitudinal inpatient records (Williams et al., 2025; Grolleau et al., 2026) or draft patient portal messages (Poursoltan et al., 2026). The resulting LLM outputs are closely reviewed and revised by a clinician in the loop before being finalized as documents in the electronic health record (EHR) or as messages to send to patients. In contrast to other summarization or generation tasks, where user feedback may be limited to coarse ratings of quality, these settings provide a rich environment for learning: expert edits are routinely collected during deployment and reveal the gaps between what the model produced and what a clinician would accept, providing, in theory, ideal supervision for training models.

However, there are good reasons to be cautious. First, a clinician's edit may be informed by information that was not available to the model, such as undocumented clinical observations or unrecorded conversations with the patient or other members of the care team (Figure 1, left). Training directly on such edits risks *learning to hallucinate*, encouraging the model to emit outputs that are not grounded in its inputs. This problem is not hypothetical: for instance, more than 45% of clinical notes, drafted by an ambient medical scribing product and edited by a clinician user, from a proprietary corpus of 30,000 de-identified and anonymized encounters are estimated to contain *at least one ungrounded edit* (Section 3.1, Appendix A.2). Second, edits can be heterogeneous along certain dimensions (e.g., style or format) where the desired behavior of a system should be predictable. However, studying these challenges is hindered by a lack of public datasets that contain such post-deployment data (i.e., draft $\rightarrow$ edit data), despite the growing prevalence of such settings in real-world clinical workflows.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

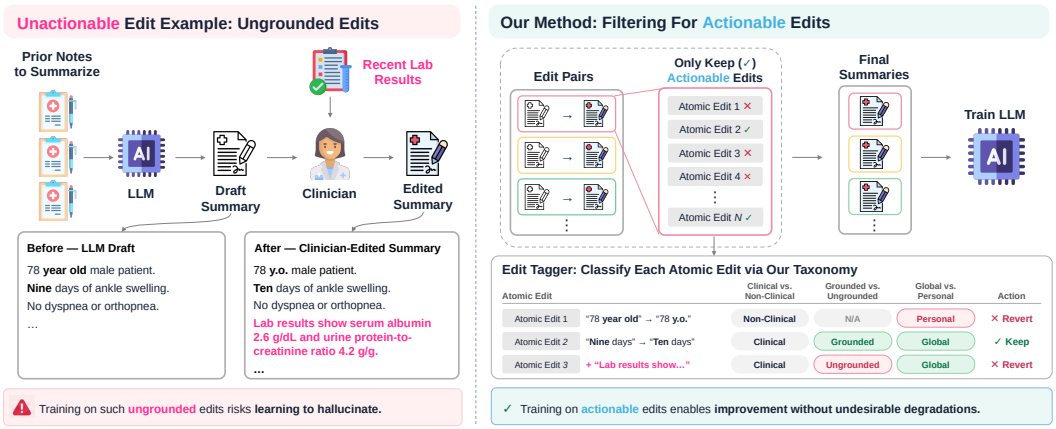

Figure 1: We propose a method for identifying and filtering for *actionable* edits, i.e., edits that reveal gaps in model capabilities and carry rich supervision signal. **(Left)** An example of an *unactionable* edit is an *ungrounded* edit, where a clinician has access to information (e.g., recent lab results) that they edit into the final document, but which is *not present* in the inputs to the LLM. Example text is illustrative. Training an LLM to generate the edited summary from prior notes alone could risk increasing hallucinations (e.g., adding plausible but made-up lab results). **(Right)** An overview of our filtering method. From each (pre-edit, post-edit) summary pair, we extract all *atomic edits* (Definition 1), categorize them based on our proposed taxonomy (Section 3.2), and revert edits that are not actionable (Section 5). The resulting *filtered* post-edit summaries are then used for model training, allowing improvements in draft quality without undesirable degradations (e.g., hallucinations).

In this paper, we explore this problem setting in the context of clinician edits to LLM-drafted discharge summaries. To address the absence of publicly available clinician-edit datasets, we construct MIMIC-III-EXT-BHC-EDIT, a semi-synthetic dataset based on MIMIC-III, with LLM-drafted discharge summaries and their *post-edit* variants (Figure 2; Section 4). The edits in our dataset are produced by grafting synthetic edits onto LLM-drafted summaries, leveraging real clinician-written discharge summaries for content. To ground the simulated edits in real clinician behavior, we design a taxonomy of edit types based on analysis of a proprietary dataset of 30,000 de-identified and anonymized encounters from real-world deployment of an ambient medical scribing product (Section 3). The taxonomy of edit types and the style of these edits inform the design of our LLM-based edit simulator. We also collect expert annotations to review the quality of the edits (Section 4.1).

Using MIMIC-III-EXT-BHC-EDIT, we demonstrate the challenges in directly training on raw edits and propose a taxonomy-based filtering approach that extracts edits that provide actionable supervision (Figure 1, right). We find that naive fine-tuning on raw post-edit summaries improves draft quality over that of the base model on several metrics that measure alignment to clinician-written reference summaries, but at the cost of introducing ungrounded and clinically unsubstantiated content (Section 7.1). In contrast, when fine-tuning on post-edit summaries after filtering out ungrounded and "personal" edits (i.e., reflecting local clinician preferences), we retain much of the performance gains observed from training on raw edit data, *while largely preserving groundedness* (Section 7.2).

Overall, our work establishes a framework for studying clinician feedback through edits, highlighting both its promise and pitfalls. Our main contributions are summarized as follows:

1. We develop a taxonomy of clinician edits to LLM-generated clinical notes, capturing how clinicians refine LLM outputs in practice (Section 3).
2. We use our edit taxonomy to propose a method for identifying and categorizing edits, which enables scalable extraction of *actionable feedback* from raw edit data (Section 5).
3. We introduce MIMIC-III-EXT-BHC-EDIT, a dataset of semi-synthetic edit pairs for the task of discharge summary generation (Section 4), and use it as a testbed for studying how to effectively leverage noisy edit data for improving model capabilities (Section 6).
4. We find that, on our dataset, (i) naively training on post-edit outputs can induce undesirable behavior, such as generating ungrounded content (Section 7.1), and that (ii) *filtering for actionable edits* is crucial to learn from edits without degradation in behavior (Section 7.2).

More broadly, our findings suggest that realizing the promise of post-deployment edit data as effective supervision signal requires methods that can reliably separate learnable signal from noise.

2 Related Works

2.1 Clinician Edits to Clinical Documentation

One widely-adopted application of LLMs in healthcare is the use of ambient medical scribes that capture the audio of patient-clinician conversations and produce a first draft of clinical notes, which a clinician then reviews and edits before signing it into the EHR (Tierney et al., 2024; Lukac et al., 2025; Afshar et al., 2025). The clinician edits naturally accumulate as a byproduct of routine documentation, motivating a nascent line of work that studies how clinicians make revisions, primarily through descriptive and qualitative analyses. Most such works focus on analyzing the *frequency* of edits across specialties and visit types, while categorizing each edit based on the *content* it modifies (e.g., medication, diagnosis, social history) with validation against human-annotated corpora (Cho et al., 2026; Guo et al., 2026a;b). Beyond edit-based analyses, other works qualitatively explore clinician perspectives on ambient AI scribes through semi-structured interviews and thematic analyses (Van Tiem et al., 2026). Collectively, these works establish that clinician edits encode systematic, clinically meaningful signal, but are limited to descriptive analyses on proprietary datasets, leaving the question of *how one should leverage edits for targeted model interventions unanswered*.

2.2 Learning from (Noisy) Edits

Gao et al. (2024) study learning from interactive user edits to LLM outputs in tasks such as news article summarization and email writing, adapting to latent user preferences inferrable from the edits. Unlike our setting, their setup assumes repeated interactions with a single user over multiple rounds, and their experiments rely on fully simulated edits that are not informed by real-world user-editing behavior. Misra et al. (2025) adopt the setting of Gao et al. (2024) and study the effectiveness and sample efficiency of different strategies for fine-tuning on edit data (e.g., supervised fine-tuning on post-edit outputs, preference learning (Rafailov et al., 2023) on (pre-edit, post-edit) output pairs). A distinct thread on learning from noisy user feedback considers noise in pairwise preference data, but models noise limited to label-flipping in pairwise comparisons (Im & Li, 2026). In the clinical domain, Yao et al. (2023) leverage token-level alignment between LLM drafts and clinician-edited summaries to improve patient-visit summary generation. However, they assume that all clinician edits are suitable as ground-truth supervision, which our findings warrant caution against.

3 Constructing a Taxonomy of Clinician Edits

3.1 Analysis of Real-World Clinician Edit Data

Given the wide adoption of ambient scribing systems in healthcare (Section 2.1), we use the edits made to outputs of an ambient medical scribing product to study how clinicians edit LLM-generated clinical documentation and inform our taxonomy of clinician edits in Section 3.2. Our qualitative exploration focuses on edits to the Assessment and Plan (A&P) section of a clinical SOAP note: a structured method of documentation used to record patient information in a chart. The A&P contains the clinician's professional analysis or diagnosis and synthesizes the subjective and objective data into a working clinical impression which includes a plan with next steps for treatment. Our dataset contains 30,000 de-identified and anonymized encounters spanning more than 50 specialties from multiple institutions across the United States between 2023 and 2026.

As textual differences between the pre-edit and post-edit drafts of a de-identified and anonymized A&P can span multiple distinct changes made by a clinician, we first extract from each (pre-edit, post-edit) pair a list of *atomic edits*, which we define as follows:

Definition 1 (Atomic Edit)**.** Given a draft and its edited version, an ***atomic edit*** is the smallest unit of textual change attributable to *a single underlying clinician intent*. All changes driven by the same reason are part of one atomic edit; changes driven by distinct reasons are separate atomic edits.

We implement the atomic-edit extraction step with GPT-4o (Hurst et al., 2024), which also outputs structured tags (e.g., factual error, omission) and descriptions of the clinician intent for each edit. Then, we obtain text descriptions of salient edit patterns via (i) topic modeling and (ii) sparse-autoencoder (SAE; Huben et al., 2024)-based feature discovery. For (i), we apply the BERTopic

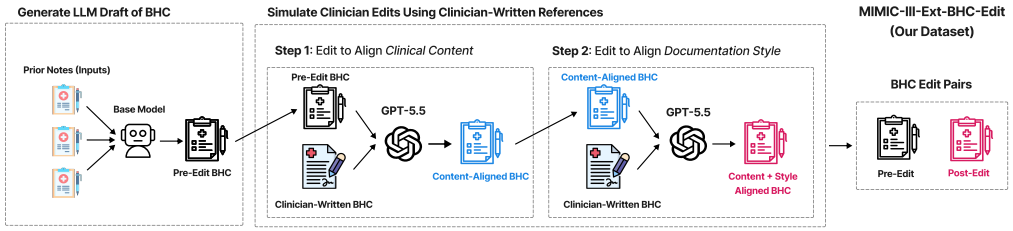

Figure 2: Overview of our generation process for MIMIC-III-EXT-BHC-EDIT, a semi-synthetic dataset of edit pairs for the task of generating brief hospital course (BHC) summaries (Section 4).

pipeline (Grootendorst, 2022) to sentence embeddings of clinician-intent descriptions to identify clusters of atomic edits that reflect similar intentions. For (ii), we train an SAE on sentence embeddings of the clinician-intent descriptions and use the top-activating examples to characterize each learned feature. Anecdotally, we observe that using the clinician-intent descriptions instead of the raw pre/post-edit notes is more effective in isolating the associated semantic change from other surface-level changes within each edit pair. We prompt GPT-5-mini (OpenAI, 2025) to obtain text descriptions of the atomic edit clusters from (i) and the learned SAE features from (ii). We incorporate our high-level findings from this process into constructing our taxonomy in Section 3.2.

3.2 TAXONOMY OF CLINICIAN EDITS TO CLINICAL DOCUMENTATION

We construct a taxonomy that characterizes clinician edits to LLM-generated clinical documentation along three axes, incorporating insights from (i) our exploration of clinician edit data from an ambient medical scribing vendor (Section 3.1) and (ii) prior studies of clinician-editing behavior in the literature (Guo et al., 2026a; Cho et al., 2026). We design our taxonomy to capture the heterogeneity of edits observed throughout deployment and provide a principled basis for identifying edits suitable as training signal. The three axes in our taxonomy each capture: what content is being edited (*Clinical vs. Non-Clinical*), whether the edit is supported by the input context (*Grounded vs. Ungrounded*), and whether the edit reflects local preferences (*Global vs. Personal*).

Clinical vs. Non-Clinical Edits. We distinguish between *clinical* and *non-clinical* edits based on whether they modify clinical content in the pre-edit draft. This distinction matters because the two types of edits carry fundamentally different implications for model improvement: clinical edits often encode generalizable signal about what the model should produce, whereas non-clinical edits may mainly reflect personal preferences. We further categorize each clinical edit as one of five subtypes, based on the underlying clinician intent: factual error correction, omission correction, condensation, specificity adjustment, and confidence-level adjustment (definitions in Appendix A.1). Non-clinical edits, by contrast, leave clinical content intact and instead modify formatting, organization, or stylistic properties (e.g., bullet points vs. free-form prose, abbreviation style of clinical terminology).

Grounded vs. Ungrounded Edits. For *clinical* edits, we further distinguish between edits whose resulting changes in clinical content are supported by the context provided to the model (e.g., prior notes) and others. A clinical edit is *grounded* if the associated change can be explained by information either present or absent from the supporting evidence. For example, an omission correction is grounded when the added information appears in the input context (e.g., prior notes) but was missed by the model, and a factual error correction is grounded when the edited version aligns with—and the original conflicted with—clinical details present in the input context. A clinical edit is *ungrounded* when no such evidence-based explanation exists: the edit reflects information, judgment, or preference exogenous to the model's inputs. These edits are not necessarily inaccurate, as they may reflect privileged knowledge or context the model cannot access, and are prevalent in real clinician-edit data (Appendix A.2). However, treating them as supervision targets risks training the model to confabulate beyond its evidence, since the "correct" output is not derivable from the model's inputs.

Global vs. Personal Edits. This third axis introduces another level of actionability. Among edits suitable as learnable signal, some reflect the population as a whole while others reflect feedback from particular subgroups or individuals. Personal edits reflect stylistic preferences that vary across clinicians: e.g., bullet points or free-form prose, expanded or contracted abbreviations for medical terminology, and how to order and group problems within the assessment. Such edits are most often non-clinical, but the categories are not coextensive: confidence-level adjustments and condensation patterns, for example, could reflect personal practice style despite affecting clinical content.

4 MIMIC-III-EXT-BHC-EDIT: A SEMI-SYNTHETIC CLINICIAN EDIT DATASET FOR BRIEF HOSPITAL COURSE GENERATION

We introduce MIMIC-III-EXT-BHC-EDIT, a semi-synthetic dataset based on MIMIC-III consisting of LLM-drafted discharge summaries (i.e., the *pre-edit* notes) and their *post-edit* versions, simulated based on the clinician-written summary available for each admission (Figure 2). Following prior works on discharge summary generation (Adams et al., 2021; Xu, 2024; Chung et al., 2026), we focus on the BHC section, which details the key clinical events, diagnoses, and treatments during a patient's stay in the hospital. As inputs for the task, we provide all prior notes available for the patient up to discharge time (e.g., nursing charts, chest X-ray reports). We extract a total of 5,821 admissions from MIMIC-III, using the cohort selection criteria by Chung et al. (2026) (Appendix B.1).

Generating the pre-edit BHCs. To generate the initial BHC draft for each admission, we zero-shot prompt three LLMs with the pre-discharge EHR notes as context (Figure 2 (left)): Gemma-3-4B-IT (Kamath et al., 2025), Qwen-2.5-7B-Instruct (Yang et al., 2025), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) (see Appendix E.2 for the prompting details). We select these models as they support long-context inputs (up to 128K), cover various model architectures and sizes, and are open-weight models that can be fine-tuned. We simulate the clinician edits on top of the zero-shot outputs, which provide the supervision signal for fine-tuning each model for BHC generation.

Generating the post-edit BHCs. From the pre-edit BHCs, we generate their *post-edit* versions, using the clinician-written reference BHC for each admission as a proxy for how a clinician would have edited the LLM draft (Figure 2, middle). We perform this in two steps: **Step 1** (*clinical content* alignment) updates the clinical content of the pre-edit BHC to match that of the reference; and **Step 2** (*documentation style* alignment) updates the style (e.g., problem-based organization, use of shorthands) of the content-aligned BHC to match that of the reference. We implement each step with GPT-5.5 (OpenAI, 2026) in compliance with the PhysioNet Credentialed Health Data License.

In **Step 1**, we provide an LLM with the pre-edit and reference BHCs for each admission and prompt it to add or remove clinical details such that the two BHCs match in clinical content, while preserving the documentation style (e.g., paragraph structure, prose style). A *clinical detail* is a discrete piece of clinical information, such as a single medication with its dosage or a single diagnosis with its supporting rationale. As these edits are all *clinical* based on our taxonomy (Section 3.2), we simulate each edit as one of six possible operations each aligned with a clinical-edit subtype: factual error correction, omission correction, condensation, specificity adjustment, confidence level adjustment, and retain. We provide full details on per-operation rules in Appendix B.3. As the pre-edit drafts are generally much shorter than the clinician-written references, frequent omission corrections (around 23–32% of edit operations; Tables B1–B3) tend to make the post-edit BHCs longer (Table B4).

In **Step 2**, we provide an LLM with the content-aligned BHC from **Step 1** and the reference BHC and prompt it to revise the former to match the structural and stylistic conventions of the latter, *without altering its clinical content*, in five aspects: organization by problem vs. chronology; bullet points vs. paragraphs; shorthands (e.g., "year old" vs. "y.o."); fragmented vs. complete sentences; and the ordering and grouping of problems. Since clinicians often vary in documentation preferences, **Step 2** can be viewed as *personalizing* the content-aligned BHC to the clinician who wrote the reference.

4.1 EXPERT VALIDATION OF THE SIMULATED EDITS

To ensure that the simulated edits result in BHCs that (1) read as coherent and professionally written documentation; and (2) are well-aligned with clinician-written references in clinical content, we conduct a small expert annotation study with a pool of 11 nurse practitioners (NPs) with hospitalist backgrounds to assess each aspect. Starting from the pool of (pre-edit, **Step 2**) BHC pairs generated for all three base LLMs, we randomly sample 25 edit pairs for assessing both (1) and (2). From each edit pair, we extract a list of all *atomic edits* (Definition 1) and randomly select one for annotation. Alongside each edit being annotated, we provide the NPs with the corresponding clinician-written reference BHC, and the pre- and post-edit BHCs to provide relevant context. For each example, we obtain annotations from three NPs (75 total judgments over 22.5 hours) and acquire a single consensus label via majority voting. We include the annotation guide used in our study in Appendix B.7.

Task 1: Coherence & Real Documentation Practice. To assess whether each edit produces coherent and professionally written changes that practitioners would reasonably make, we ask each annotator to provide one of three ratings: *Reasonable*, *Unreasonable*, and *Unsure* (detailed defini-

Table 1: Expert validation of the simulated edits in MIMIC-III-EXT-BHC-EDIT (Section 4.1).

Annotation Task	Expert Rating	Examples		Inter-Annotator Agreement	
		%	n	Unanimous	Majority
Task 1: Coherence & Documentation Practice	**Reasonable**	**100%**	25	22	3
	Unreasonable	0%	0	0	0
	Unsure	0%	0	0	0
	All	100%	25	**88%** (22/25)	**12%** (3/25)
Task 2: Clinical Content Alignment to Reference	**Aligned**	**84%**	21	19	2
	Not Aligned	0%	0	0	0
	Unsure	0%	0	0	0
	N/A	16%	4	1	3
	All	100%	25	**80%** (20/25)	**20%** (5/25)

tions of the ratings in Appendix B.7). All 25 examples were labeled as *Reasonable*, with unanimous agreement on 88% (22/25) and majority agreement on 12% (3/25) of cases (Table 1, top row).

Task 2: Clinical Content Alignment to the Reference. To assess whether each edit only involves changes that agree with the reference in clinical content, we ask each annotator to provide one of four ratings: *Aligned*, *Not Aligned*, *Unsure*, or *N/A*. 84% (21/25) of examples were labeled as *Aligned*, while 16% (4/25) were labeled as *N/A* (i.e., edit does not change clinical content), with unanimous agreement on 80% (20/25) and majority agreement on 20% (5/25) of cases (Table 1, bottom row).

Overall, we observe that experts generally find our simulated edits reasonable and well-aligned with clinician-written references. Meanwhile, we do not intend to claim that our simulated edits reproduce how clinicians edit LLM drafts under real-world workflow constraints. Rather, as every added or corrected detail in the post-edit BHCs comes from a real clinician-written BHC, ungrounded content reflects information clinicians actually document rather than arbitrary insertions, yielding a controlled testbed in which grounded, ungrounded, global, and personal edits can be isolated. Given the high costs of expert validation, we leave further large-scale dataset validation to future work.

5 IDENTIFYING AND ISOLATING ACTIONABLE EDITS FOR TRAINING

To identify the *actionable* edits to use as supervision signal, we develop a GPT-5.5-based *edit tagger* that takes an edit pair as input and produces structured per-edit annotations aligned with our taxonomy. Given (i) the pre-edit BHC, (ii) the post-edit BHC, and (iii) all pre-discharge EHR notes for a given admission, the edit tagger extracts all atomic edits (Definition 1) and classifies each along the three axes of our taxonomy (full details in Appendix C.1). Table C1 shows the distribution of edit types with the **Step 1** and **Step 2** outputs as the post-edit BHC, as determined by the edit tagger.

We also assess the accuracy of edit tagger judgments along all three axes via expert annotations (full details provided in Appendix C.3). For the *Clinical vs. Non-Clinical* and *Global vs. Personal* axes, we compare the edit tagger outputs to expert annotations that we collect from a pool of 10 NPs with hospitalist backgrounds, on 25 randomly sampled atomic edits. For each edit, we obtain annotations from three NPs (75 expert judgments over 22.5 hours). For the *Grounded vs. Ungrounded* axis, we compare the edit tagger outputs for 150 randomly sampled atomic edits to the clinician annotations from VERIFACT-BHC (Chung et al., 2026), which assess the groundedness of each "claim" from a reference BHC. As our atomic edits and the VERIFACT-BHC claims do not completely align, we only review edits whose text spans align with at least one claim to obtain the clinician ground-truth.

Our edit tagger agrees most with experts on the *Clinical vs. Non-Clinical* (75% = 18/24, excluding one case labeled *Unsure*) and *Grounded vs. Ungrounded* (80% = 120/150) axes, where experts agree unanimously on 84% (21/25) and 63.1% (149/236[1]) of cases, respectively. Agreement between the edit tagger and experts is lowest on *Global vs. Personal* (52% = 13/25), which is also more subjective for experts than *Clinical vs. Non-Clinical*: experts agree unanimously on only 64% (16/25) of the same edits. We thus treat *Global vs. Personal* tags as noisier signal than those on the other two axes.

Learning from *actionable* edits. After extracting and tagging the atomic edits from each (pre-edit, post-edit) BHC pair, we construct various *filtered* training sets, where we selectively *revert* edits of specific types. Given each pair and filter, we update the post-edit BHC by undoing every atomic edit whose tags are excluded by the filter: reverting insertions, restoring deletions, and substituting the

[1]Refers to the 236 VERIFACT-BHC claims that overlap with the 150 atomic edits reviewed.

original draft span back in for substitutions (Appendix D). This yields a family of training targets that share the same input but differ in which edits they treat as supervision signal. We define two filters that progressively restrict the supervision to increasingly conservative subsets of edits:

- ***Grounded* edits**: Removes *clinical* edits that are *ungrounded* in the inputs, restricting supervision to other edits explicitly substantiated by the input evidence available to the model.
- ***Grounded + Global* edits**: Additionally removes *personal* edits, restricting to edits that improve general model capabilities rather than reflecting individual clinician preferences.

We focus on removing ungrounded and personal edits, as they are least likely to provide generalizable supervision: the former are unsupported by the input context; the latter reflect local preferences.

6 EXPERIMENTS

We validate our method for extracting and learning from actionable clinician edits (Section 5) using MIMIC-III-EXT-BHC-EDIT (Section 4). In particular, we focus on improving the clinical utility of the generated BHC summary *without losing groundedness to the prior notes provided as input*. We also assume a setting where clinicians edit LLM drafts to complete documentation tasks, and we thus do not have access to the "oracle" summaries that clinicians would have written on their own.

Starting from Gemma-3-4B-IT, Qwen-2.5-7B-Instruct, and Llama-3.1-8B-Instruct—the three models whose zero-shot BHC generations serve as the *pre-edit* BHCs in our MIMIC-III-EXT-BHC-EDIT dataset (Section 4)—we perform supervised fine-tuning (SFT) with low-rank adapters (LoRA; Hu et al., 2022), using each model's simulated *post-edit* BHCs as the ground-truth. We compare between SFT on the *Grounded* and *Grounded + Global* edit pairs resulting from our filtering method (Section 5) vs. SFT on the raw, unfiltered post-edit BHCs. We also compare with using the original clinician-written BHCs for SFT (where one assumes access to fully human-authored notes), as well as on intermediate outputs from post-edit simulation (**Step 1**; Section 4) (see Appendix E).

We evaluate on two test sets—one based on the 100 admissions that overlap with VERIFACT-BHC (Chung et al., 2026), and another obtained from taking a 90–5–5 train, validation, and test split by patient on the remaining 5,721 admissions from our dataset. We curate the former as a *grounded* test set, using the human annotations from VERIFACT-BHC of whether each atomic claim from the reference (clinician-written) BHC is substantiated by the prior notes (Appendix E.3). The *grounded* test set is considered to avoid rewarding models for generating ungrounded clinical information.

Evaluation metrics. We assess the quality of model-generated BHCs with the following metrics: (i) *clinical claim/term-matching* metrics: DocLens-Precision/Recall (Xie et al., 2024) and MEDCON (Yim et al., 2023); (ii) *clinical factuality* metrics: MedScore (Huang et al., 2026), DocLens-Attribution (Xie et al., 2024), and PDSQI-9 (Croxford et al., 2025); and (iii) natural language generation (NLG) metrics: BERTScore (Zhang et al., 2020), ROUGE-L (Lin, 2004), and BLEU-2 (Papineni et al., 2002). The metrics in (i) assess whether the clinical claims/terms in each model-generated BHC match those in the reference BHC, measuring alignment to what a clinician would document. The metrics in (ii) assess whether the claims in each model-generated BHC are supported by the input prior notes, measuring faithfulness to the input. DocLens-Attribution is an adaptation of the "Citation Recall" metric in Xie et al. (2024), checking whether each claim from the model-generated BHC is entailed by the prior notes instead of a model-generated citation. For PDSQI-9, we focus on its "accurate" score (1–5), which measures factuality at the summary level. As other metrics range between 0 and 1, we shift and scale PDSQI-9 to the same range. We compute the reference-based metrics in (i) and (iii) on the *grounded* test set (VERIFACT-BHC), and all other metrics on examples from both the VERIFACT-BHC and MIMIC-III-EXT-BHC-EDIT test sets.

7 RESULTS

We present the main findings from our experiments (Section 6), focusing on the *clinical claim/term-matching* and *clinical factuality* metrics. Overall, we find that naive SFT on raw edit data improves BHC generation but degrades groundedness in the input prior notes (Section 7.1), while filtering out ungrounded edits beforehand enables performance gains without compromising groundedness (Section 7.2). We include the results from NLG-based evaluation and additional analyses in Appendix G, where we show that similar trends hold for SFT on the original clinician-written references available from MIMIC-III ("oracle" supervision) or the *intermediate* post-edit BHCs (**Step 1**; Section 4).

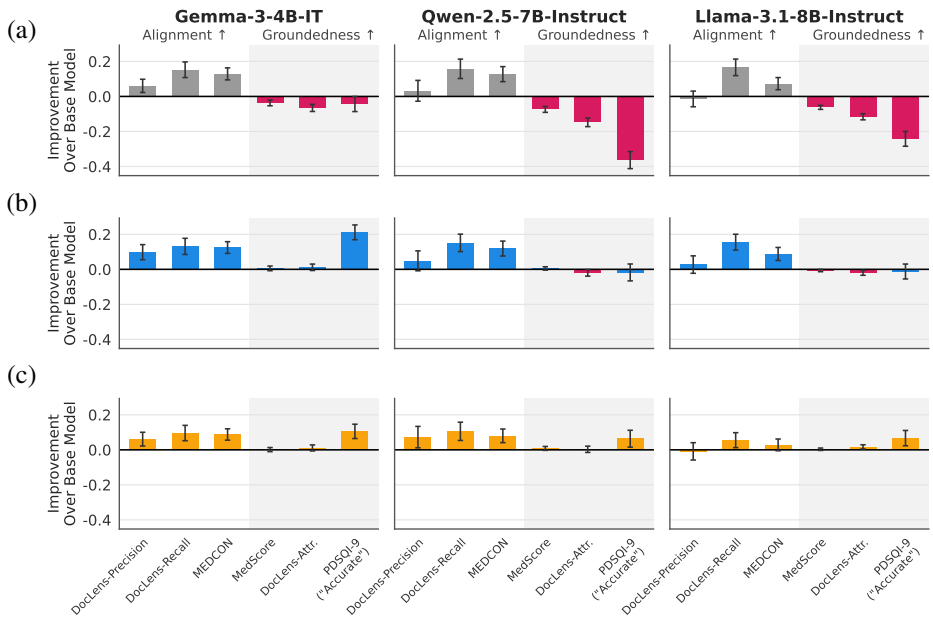

Figure 3: Naive fine-tuning on raw post-edit BHCs degrades groundedness in the input prior notes (Section 7.1), while filtering for actionable edits beforehand improves alignment to clinician-written references, *while largely preserving groundedness* (Section 7.2). We report improvements over the base model in the clinical claim/term-matching metrics and the groundedness metrics, after SFT on the (a) raw, (b) *Grounded*, and (c) *Grounded + Global* post-edit BHCs. In each panel, the left three metrics measure alignment and the shaded right three metrics measure groundedness. Note that the metrics are not necessarily comparable on the same scale, but the point is in their directionality. Error bars indicate the 95% bootstrap CIs, computed by resampling the test set with replacement 10,000 times, and red bars indicate statistically significant degradations relative to the base model.

7.1 Naive SFT on Raw Edits Improves Performance But Degrades Factuality

Across all three models, directly fine-tuning on the raw edit data generally improves alignment to clinician-written reference BHCs, suggesting that edits can provide effective supervision for model training (Figure 3(a), left half of each panel). For instance, the average MEDCON score on the *grounded* test set (Section 6) increases by +0.130 for Gemma-3-4B-IT (0.229 → 0.359), +0.127 for Qwen-2.5-7B-Instruct (0.253 → 0.380), and +0.073 for Llama-3.1-8B-Instruct (0.290 → 0.363), all of which are statistically significant improvements (i.e., 95% bootstrap confidence intervals (CIs) in average improvement lie above 0). DocLens-Recall scores also increase significantly, indicating that the BHCs generated by the fine-tuned models provide better coverage of clinical details in the clinician-written reference, while maintaining similar DocLens-Precision scores as the base model.

However, these improvements come at the cost of groundedness in the input prior notes (Figure 3(a), shaded right half of each panel). For example, the average PDSQI-9 "accurate" score (on the original scale of 1–5) drops by -0.18 for Gemma-3-4B-IT (2.32 → 2.14), -1.45 for Qwen-2.5-7B-Instruct (3.39 → 1.94), and -0.97 for Llama-3.1-8B-Instruct (2.92 → 1.95). In fact, all models show statistically significant degradations on MedScore, DocLens-Attribution, and PDSQI-9 (shown in the red bars). Notably, such degradation in groundedness is a direct consequence of treating the raw post-edit BHCs with high levels of ungroundedness as the ground-truth (Figure G2, middle). Our results suggest that while training on raw edit data can help improve the general summarization capabilities of the model, it can also induce the model to generate content that is not faithful to the input context.

These results also suggest that relying on reference-based metrics for evaluating model-generated clinical summaries warrants caution. Even after removing any ungrounded details from the reference (e.g., *grounded* test set), such metrics only indicate whether a summary sufficiently captures the key details that a clinician would highlight, and may fail to penalize cases where the generated details are consistent with the reference but not supported by the input prior notes. Therefore, it is crucial to assess reference-based agreement *subject to the constraint that groundedness does not degrade*.

7.2 SFT ON FILTERED EDITS IMPROVES PERFORMANCE WHILE PRESERVING FACTUALITY

After applying our *Grounded* and *Grounded + Global* filters to the BHC edit pairs (Section 5), SFT on the post-edit BHCs improves alignment to clinician-written references *while generally preserving groundedness* (Figure 3(b,c)). For instance, fine-tuning Gemma-3-4B-IT with the filtered data results in significant improvements on *all* reference-based metrics (i.e., 95% bootstrap CIs in average improvement lie above 0), with similar or better groundedness scores (notably, PDSQI-9) than those of the base model. Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct also generally show improvements on DocLens-Recall and MEDCON, especially in the *Grounded* filter setting, with little to no degradation in groundedness. These trends arise from treating BHCs with high groundedness scores as the ground-truth, as enabled by our edit filters (Figure G2, right). Varying the fraction of ungrounded edits removed shows that groundedness after SFT improves steadily as more such edits are filtered out, while reference-alignment gains remain largely unchanged (Appendix G.3).

Notably, the models trained with our filtered edit data retain groundedness *while producing longer and more detailed BHCs* than the original base model, a setting where factuality is more challenging to achieve (Table G2). For instance, for Qwen-2.5-7B-Instruct, the average length of BHCs from the base model is only 109.3 words, while the average lengths of BHCs after SFT on the *Grounded* and *Grounded + Global* post-edit BHCs are 219.9 and 177.3 words, respectively. We observe similar trends for Gemma-3-4B-IT and Llama-3.1-8B-Instruct (Table G2, top and bottom rows). These results demonstrate the effectiveness of our edit tagger and filtering strategies (Section 5) in translating noisy edit data into effective supervision signal for model training.

Meanwhile, although improvements are evident along certain axes, no single training setup dominates across all metrics. For example, the groundedness scores for Qwen-2.5-7B-Instruct and Llama-3.1-8B-Instruct are maximized by our most aggressive filter (*Grounded + Global*; Figure 3(c)), while the reference-based metrics (MEDCON, DocLens) increase more when only filtering out ungrounded details (*Grounded*; Figure 3(b)), suggesting that some edits tagged as "personal" may still provide useful supervision (Section 5). These results suggest that more sophisticated filtering strategies may be necessary to yield further gains and generalize across different edit types.

7.3 EXPERT VALIDATION OF MODEL-GENERATED BHC GROUNDEDNESS

For further validation of our findings in Sections 7.1–7.2, we obtain expert annotations from a pool of 8 NPs with hospitalist backgrounds on 10 admissions randomly selected from the test sets. For each admission, we ask each annotator to compare the three BHCs generated by the fine-tuned models (raw, *Grounded*, *Grounded + Global*) against the zero-shot-generated BHC from the base model in terms of their groundedness. For each pairwise comparison, the annotators are asked to provide one of three ratings: $A > B$, $A < B$, $A \approx B$, where A and B each denote either the BHC generated by one of the three fine-tuned models or the zero-shot-generated BHC, shown in random order. For every comparison, we obtain 3 annotations, totaling $10 \times 3 \times 3 = 90$ judgments collected over 30 hours.

Comparing SFT (raw) to the base model, we find that the base-model BHCs are considered strictly better grounded in 80% of admissions. In contrast, the BHCs generated by SFT (*Grounded*) and SFT (*Grounded + Global*) are both similar or better than the base-model BHCs in 90% (Better: 30%, Equal: 60%) and 60% (Better: 50%, Equal: 10%) of admissions, respectively (Table F1). These results corroborate our finding that our filtering strategies help mitigate degradation in groundedness. We provide the remaining details (e.g., inter-annotator agreement) from this study in Appendix F.

8 DISCUSSION AND CONCLUSION

In this paper, we studied how to leverage clinician-edit data for model training. Informed by real-world clinician-edit data, we constructed a taxonomy of clinician edits, distinguishing edits that are actionable (e.g., grounded, global) from others that are not (e.g., ungrounded, personal). We then introduced MIMIC-III-EXT-BHC-EDIT, a semi-synthetic edit dataset based on MIMIC-III, given the lack of public edit datasets. On this dataset, we found that naively training on raw edits can degrade generation quality, and that targeted removal of edits with undesirable properties (e.g., ungrounded) can help translate edit data into effective supervision without clear degradations. More broadly, our findings suggest that realizing the potential of user-edit data as effective supervision signal warrants care and requires methods that can reliably separate learnable signal from noise.

Limitations. Due to space constraints, we include a discussion of our limitations in Appendix H.

AI USE STATEMENT

In this work, we used generative AI tools for analyzing the de-identified and anonymized data from an ambient medical scribing vendor (Section 3.1), generating our semi-synthetic MIMIC-III-EXT-BHC-EDIT dataset (Section 4), implementing the edit tagger (Section 5), implementing the pipeline for LLM fine-tuning (Section 6), and formatting the data used for expert annotation studies (Sections 4.1, 5, and 7.3). We have not used generative AI tools for developing conceptual frameworks, proposing or refining hypotheses, or interpreting our main results. Other required disclosure tasks such as developing theoretical models, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, or assisting with translation are not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images, summarizing or analyzing relevant literature, and editing parts of writing for better readability. We have reviewed all AI-assisted work. All LLM-generated code was verified and tested for correctness by 2 authors of this paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. What's in a Summary? Laying the Groundwork for Advances in Hospital-Course Summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4794–4811, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.382. URL https://aclanthology.org/2021.naacl-main.382/.

Majid Afshar, Mary Ryan Baumann, Felice Resnik, Josie Hintzke, Anne Gravel Sullivan, Graham Wills, Kayla Lemmon, Jason Dambach, Leigh Ann Mrotek, Mariah Quinn, Kirsten Abramson, Peter Kleinschmidt, Thomas B. Brazelton, Margaret A. Leaf, Heidi Twedt, David Kunstman, Brian Patterson, Frank Liao, Stacy Rasmussen, Elizabeth S. Burnside, Cherodeep Goswami, and Joel Gordon. A Pragmatic Randomized Controlled Trial of Ambient Artificial Intelligence to Improve Health Practitioner Well-Being. NEJM AI, 2, 11 2025. doi: 10.1056/aioa2500945. URL https://doi.org/10.1056/aioa2500945.

Anthropic. Introducing Claude Opus 4.8. https://www.anthropic.com/news/claude-opus-4-8, May 2026.

Ha Na Cho, Yawen Guo, Sairam Sutari, Emilie Chow, Steven Tam, Danielle Perret, Deepti Pandita, and Kai Zheng. Consumer-to-Clinical Language Shifts in Ambient AI Draft Notes and Clinician-Finalized Documentation: A Multi-level Analysis. arXiv preprint arXiv:2603.18327, 2026.

Philip Chung, Akshay Swaminathan, Alex J. Goodell, Yeasul Kim, S. Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, David Seong, Andrew A. Lee, Caitlin E. Coombes, Brad Bradshaw, Mahir A. Sufian, Hyo Jung Hong, Teresa P. Nguyen, Mohammad R. Rasouli, Komal Kamra, Mark A. Burbridge, James C. McAvoy, Roya Saffary, Stephen P. Ma, Dev Dash, James Xie, Ellen Y. Wang, Clifford A. Schmiesing, Nigam Shah, and Nima Aghaeepour. Verifying Facts in Patient Care Documents Generated by Large Language Models Using Electronic Health Records. NEJM AI, 3(1): AIdbp2500418, 2026. doi: 10.1056/AIdbp2500418. URL https://ai.nejm.org/doi/full/10.1056/AIdbp2500418.

Emma Croxford, Yanjun Gao, Nicholas Pellegrino, Karen K. Wong, Graham Wills, Elliot First, Miranda Schnier, Kyle Burton, Cris G. Ebby, Jillian Gorskic, Matthew Kalscheur, Samy Khalil, Marie Pisani, Tyler Rubeor, Peter Stetson, Frank Liao, Cherodeep Goswami, Brian Patterson, and Majid Afshar. Development and Validation of the Provider Documentation Summarization Quality Instrument for Large Language Models. Journal of the American Medical Informatics Association, 32(6):1050–1060, 2025.

Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. Aligning LLM Agents by Learning Latent Preference from User Edits. In Conference on Neural Information Processing Systems (NeurIPS), 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu,

Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent,

Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783, 2024.

François Grolleau, April S. Liang, Timothy Keyes, Stephen P. Ma, Thomas Lew, Tridu R. Huynh, Natasha Steele, Philip Chung, Paige Qin, Gowri Chandra, Stephanie F. Wang, Evan Mullen, Lauren Carpenter, Mita Hoppenfeld, Matthew Morrin, Baffour A. Kyerematen, Nerissa Ambers, Nikesh Kotecha, Emily Alsentzer, Jason Hom, Nigam H. Shah, Kevin Schulman, and Jonathan H. Chen. Physician-Reported Safety Outcomes of AI-Generated Hospital Course Summaries. JAMA Network Open, 9(5):e2616556–e2616556, 05 2026. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2026.16556. URL https://doi.org/10.1001/jamanetworkopen.2026.16556.

Maarten Grootendorst. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. arXiv preprint arXiv:2203.05794, 2022.

Yawen Guo, Di Hu, Ziqi Yang, Seungjun Kim, Brian Tran, Jamie Lee, Sitha Vallabhaneni, Rachael Zehrung, Sairam Sutari, Steven Tam, Emilie Chow, Danielle Perret, Deepti Pandita, and Kai Zheng. What Do Clinicians Edit in Ambient AI-Drafted Clinical Documentation? A Qualitative Content Analysis. Journal of the American Medical Informatics Association, 33(8):1457–1465, 08 2026a. ISSN 1527-974X. doi: 10.1093/jamia/ocag073. URL https://doi.org/10.1093/jamia/ocag073.

Yawen Guo, Yiliang Zhou, Di Hu, Sairam Sutari, Emilie Chow, Steven Tam, Danielle Perret, Deepti Pandita, and Kai Zheng. Understanding Clinician Edits to Ambient AI Draft Notes: A Feasibility Analysis Using Large Language Models. medRxiv, pp. 2026–02, 2026b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In International Conference on Learning Representations (ICLR), 2022.

Heyuan Huang, Alexandra DeLucia, Vijay Murari Tiyyala, and Mark Dredze. MedScore: Generalizable Factuality Evaluation of Open-ended Long-form Medical Answers by Domain-adapted Claim Decomposition and Verification. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), Findings of the Association for Computational Linguistics: ACL 2026, pp. 14149–14180, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.693. URL https://aclanthology.org/2026.findings-acl.693/.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models. In International Conference on Learning Representations (ICLR), 2024.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb,

Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya

Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. GPT-4o System Card. arXiv preprint arXiv:2410.21276, 2024.

Shawn Im and Yixuan Li. How Well Can Preference Optimization Generalize Under Noisy Feedback? Transactions on Machine Learning Research (TMLR), 2026.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo A. Celi, and Roger G. Mark. MIMIC-III, a Freely Accessible Critical Care Database. Scientific Data, 3(160035), 2016.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 Technical Report. arXiv:2503.19786, 2025.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. International Conference on Learning Representations (ICLR), 2019.

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. Ambient AI Scribes in Clinical Practice: A Randomized Trial. NEJM AI, 2, 11 2025. doi: 10.1056/aioa2501000. URL https://doi.org/10.1056/aioa2501000.

Dipendra Misra, Aldo Pacchiano, Ta-Chung Chi, and Ge Gao. Principled Fine-tuning of LLMs from User-Edits: A Medley of Preference, Supervision, and Reward. Conference on Neural Information Processing Systems (NeurIPS), 2025.

OpenAI. Introducing GPT-5, August 2025. URL https://openai.com/index/introducing-gpt-5/.

OpenAI. Introducing GPT-5.5, April 2026. URL https://openai.com/index/introducing-gpt-5-5/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pp. 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Lily Poursoltan, Jie Cao, Wei Chen, Eileen Kim, Tejpreet K. Nakai, Aaron Boussina, Andrew Chua, Jeffrey Pan, Marlene Millen, Ming Tai-Seale, Christopher A. Longhurst, Kevin Zhu, and Karandeep Singh. Physician Edits to AI-Drafted Patient Messages and Their Impact on Clinical Workload. NEJM AI, 3(9):AIoa2501034, 2026. doi: 10.1056/AIoa2501034. URL https://ai.nejm.org/doi/full/10.1056/AIoa2501034.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In Conference on Neural Information Processing Systems (NeurIPS), 2023.

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, 5, 2 2024. doi: 10.1056/cat.23.0404. URL http://dx.doi.org/10.1056/cat.23.0404.

Jen Van Tiem, Elizabeth Cramer, Christopher Iverson, Korey Kennelty, Noah Andrys, Julie Lee, Lindsey Knake, Jason Misurac, James Blum, and Heather Schacht Reisinger. Listening to the Note: Clinician Perspectives on Ambient Artificial Intelligence Scribes in Medical Documentation. Journal of the American Medical Informatics Association, 33(2):255–262, 2026.

Christopher Y. K. Williams, Charumathi Raghu Subramanian, Syed Salman Ali, Michael Apolinario, Elisabeth Askin, Peter Barish, Monica Cheng, W. James Deardorff, Nisha Donthi, Smitha Ganeshan, Owen Huang, Molly A. Kantor, Andrew R. Lai, Ashley Manchanda, Kendra A. Moore, Anoop N. Muniyappa, Geethu Nair, Prashant P. Patel, Lekshmi Santhosh, Susan Schneider, Shawn Torres, Michi Yukawa, Colin C. Hubbard, and Benjamin I. Rosner. Physician- and Large Language Model–Generated Hospital Discharge Summaries. JAMA Internal Medicine, 185(7):818–825, 07 2025. ISSN 2168-6106. doi: 10.1001/jamainternmed.2025.0821. URL https://doi.org/10.1001/jamainternmed.2025.0821.

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. DocLens: Multi-Aspect Fine-Grained Evaluation for Medical Text Generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 649–679, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.39. URL https://aclanthology.org/2024.acl-long.39/.

Justin Xu. Discharge Me: BioNLP ACL'24 Shared Task on Streamlining Discharge Documentation. PhysioNet, April 2024. doi: 10.13026/0zf5-fx50. URL https://doi.org/10.13026/0zf5-fx50. Version 1.3.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. arXiv preprint arXiv:2412.15115, 2025.

Zonghai Yao, Benjamin Schloss, and Sai Selvaraj. Improving Summarization with Human Edits. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 2604–2620, 2023.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. ACI-Bench: A Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. Scientific Data, 10(586), 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. International Conference on Learning Representations (ICLR), 2020.

A Additional Details on the Edit Taxonomy

A.1 Clinical Edit Subtypes

We further categorize *clinical* edits (Section 3.2) as one of five subtypes and provide the definition for each below:

1. *Factual error correction*: Clinician revises information in the draft that conflicts with the underlying encounter.
2. *Omission correction*: Clinician adds clinically relevant information absent from the draft.
3. *Condensation*: Clinician removes information that, while not incorrect, is determined to be extraneous.
4. *Specificity adjustment*: Clinician rewrites lay or imprecise language (e.g., a patient's verbatim symptom description) into standardized clinical terminology.
5. *Confidence-level adjustment*: Clinician perturbs the amount of certainty or hedging.

A.2 Prevalence of Ungrounded Edits in Real Clinician Edit Data (Section 3.2)

Following our definition of *grounded* vs. *ungrounded* edits in Section 3.2, we estimate that more than 45% of clinical notes from our dataset of 30,000 de-identified and anonymized encounters contain one or more ungrounded edits. We estimate this prevalence through a two-stage process. First, we use our LLM-based edit tagger (Section 5) on the anonymized corpus of 30,000 encounters. 57% of the notes are flagged by the tagger as containing ungrounded edits. Following manual review of a random subset of 20 de-identified and anonymized encounters by one of the authors with expertise in assessing clinical documentation, we find that 80% are true positive cases, which leads to an estimated prevalence of $0.57 \times 0.80 \times 100 = 45.6\%$.

B Additional Details on MIMIC-III-Ext-BHC-Edit

Here, we provide additional details on how we construct MIMIC-III-Ext-BHC-Edit (Section 4).

B.1 MIMIC-III Cohort Selection

As discussed in Section 4, we follow the cohort selection criteria proposed by Chung et al. (2026) in the "Details Involving the Creation of VeriFact-BHC" section of their Supplementary Appendix, which we replicate below. We extract all hospital admissions in the MIMIC-III Clinical Database v1.4 (Johnson et al., 2016) belonging to patients that satisfy the following conditions:

1. Patients whose most recent hospital admission contains a discharge summary with Brief Hospital Course (BHC) section extractable by regular-expression pattern matching;
2. Patients had at least two additional physician notes in the patient's EHR;
3. Patients had at least ten notes preceding the discharge summary, including those from prior hospital admissions and encounters.

While Chung et al. (2026) randomly subsample 100 patients satisfying the above criteria for their analysis, we extract all such patients and construct our MIMIC-III-Ext-BHC-Edit dataset using the most recent hospital admissions corresponding to 5,821 patients.

B.2 Pre-Edit BHC Generation: Handling Admissions with Input Context Exceeding Context Window Sizes

Among the 5,821 admissions extracted as in Appendix B.1, some admissions contain a significantly large number of input prior notes, resulting in cases where the concatenation of all input prior notes does not fit into the context windows of the base models after tokenization. As such, for each of the three base models, we pre-tokenize the concatenation of all prior notes available for each admission, and exclude the admissions with inputs that exceed the context window of the model. Thus, we only use the remaining admissions that fit into the context window of each model for post-edit BHC

Table B1: Frequency of edit operations from **Step 1** (clinical content alignment) of our edit simulation process (Section 4) for Gemma-3-4B-IT, across all admissions in MIMIC-III-EXT-BHC-EDIT. Coverage denotes the % of admissions in which the operation appears at least once.

	Per operation		Per admission	
Edit operation	Count	% of total	Mean count	Coverage (%)
Specificity Adjustment	17,298	23.7	3.50	97.3
Omission Correction	20,455	28.1	4.14	98.2
Retain	12,123	16.6	2.46	92.1
Factual Error Correction	12,383	17.0	2.51	93.8
Condensation	9,744	13.4	1.97	88.7
Confidence Level Adjustment	831	1.1	0.17	16.2
Total	72,834	100.0	14.76	—

Table B2: Frequency of edit operations from **Step 1** (clinical content alignment) of our edit simulation process (Section 4) for Qwen-2.5-7B-Instruct, across all admissions in MIMIC-III-EXT-BHC-EDIT. Coverage denotes the % of admissions in which the operation appears at least once.

	Per operation		Per admission	
Edit operation	Count	% of total	Mean count	Coverage (%)
Specificity Adjustment	11,698	25.0	3.37	98.2
Omission Correction	14,755	31.5	4.25	98.6
Retain	8,758	18.7	2.52	94.7
Factual Error Correction	6,890	14.7	1.99	88.9
Condensation	4,261	9.1	1.23	69.6
Confidence Level Adjustment	446	1.0	0.13	12.3
Total	46,808	100.0	13.49	—

generation, and in turn, the training and evaluation of models (as detailed in Section 6). See Table B4 for the final number of admissions included for each base model.

B.3 POST-EDIT BHC GENERATION: DETAILED DESCRIPTION OF STEP 1 (CLINICAL CONTENT ALIGNMENT)

As discussed in Section 4, we provide GPT-5.5 with the pre-edit (i.e., zero-shot generations) and clinician-written BHCs for each admission and prompt it to add or remove clinical details from the pre-edit BHC to match the content in the clinician-written BHC, while preserving the overall documentation style (e.g., paragraph structure, sentence order, prose style). In alignment with our taxonomy (Section 3.2), each edit operation takes one of the following forms: factual error correction, omission correction, condensation, specificity adjustment, confidence level adjustment, and retainment. If a clinical detail (a discrete piece of clinical information, such as a single medication with its dosage or a single diagnosis with its supporting rationale) is *consistent* between the two BHCs, we either keep the original detail (*retainment*) or update it to match the specificity of the corresponding detail in the clinician-written BHC (e.g., "prescribed antibiotics" → "prescribed vancomycin") (*specificity adjustment* or *confidence adjustment*). If a clinical detail is *inconsistent* between the two BHCs, we either remove the original detail or substitute it with correct information to match what the clinician-written BHC states (*factual error correction*). If a clinical detail is only discussed in the pre-edit BHC, then we remove it (*condensation*), assuming that the clinician-written BHC omits it for a specific reason (e.g., conciseness). Clinical details only discussed in the clinician-written BHC are always added to the pre-edit BHC (*omission correction*). Tables B1–B3 show the distribution of edit operations resulting from **Step 1** of the post-edit generation process, each corresponding to Gemma-3-4B-IT, Qwen-2.5-7B-Instruct, and Llama-3.1-8B-Instruct, respectively.

B.4 POST-PROCESSING OF ZERO-SHOT BASE MODEL GENERATIONS

In practice, we observe that the zero-shot BHC generations from the base models may: (i) not be valid BHCs (e.g., do not read as summaries that a clinician would write, regurgitates the input prior notes); (ii) generate unnecessary content (e.g., conversational text like "`Okay, here is`

Table B3: Frequency of edit operations from **Step 1** (clinical content alignment) of our edit simulation process (Section 4) for Llama-3.1-8B-Instruct, across all admissions in MIMIC-III-EXT-BHC-EDIT. Coverage denotes the % of admissions in which the operation appears at least once.

	Per operation		Per admission	
Edit operation	Count	% of total	Mean count	Coverage (%)
Specificity Adjustment	20,800	27.1	4.13	99.2
Omission Correction	17,814	23.2	3.54	96.5
Retain	14,972	19.5	2.97	96.7
Factual Error Correction	12,806	16.7	2.54	93.4
Condensation	9,421	12.3	1.87	84.0
Confidence Level Adjustment	960	1.3	0.19	18.0
Total	76,773	100.0	15.24	—

`the brief hospital course section for the patient:"` or section headers like `"**BHC:**"`); or (iii) repeat the same content until consuming the entire output token budget. Simulating edits on such generations can result in cases where either the edit itself is not reflective of the types of post-deployment edits that clinicians would make in practice, or the post-edit version reads as a complete rewrite of the original LLM-generated draft.

To avoid these issues, we always post-process the zero-shot generated draft by applying the following steps:

1. **Section header removal:** We strip leading section headers (e.g., `Brief Hospital Course:`, `**BHC:**`, `## Brief Hospital Course (BHC)`) as well as inline header prefixes (e.g., `BHC: <content>`) by matching a set of regular expressions anchored to the start of the text, applied iteratively (up to three passes) to handle stacked headers.
2. **EHR-regurgitation removal:** We truncate the output at the earliest line-anchored occurrence of any marker indicating that the model has begun echoing the raw EHR export, such as note-index dividers (e.g., `## Note X/Y`), long underscore separators, `----Protected Section----` banners, field labels like `Admitting Diagnosis:`, or radiology placeholder tokens (e.g., `[**Clip Number (Radiology) ...**]`).
3. **Repetition removal:** We extract the trailing suffix of the output at each of several fixed lengths (200, 150, 100, 80, 60 characters), and for each suffix we search for additional exact occurrences earlier in the text; if a given suffix recurs at least twice with gaps between consecutive occurrences that are all within 15% of the mean gap (signaling a periodic loop rather than coincidental repetition), we record a candidate cut at the end of its first occurrence, and truncate the output at the earliest such cut across all suffix lengths.
4. **Truncation removal:** If the remaining output ends without a sentence-ending punctuation, we truncate everything after the last position matching `[.!?]` (optionally followed by a closing quote or bracket and then whitespace or end-of-text), except when it would remove more than 30% of the text (heuristic safeguard against cases where the output is not actually truncated).

The post-processed zero-shot BHC generations then become the pre-edit BHCs upon which **Step 1** and **Step 2** of our edit simulation pipeline are subsequently applied.

B.5 POST-EDIT BHC GENERATION: PROMPTS

Below, we include the prompts that we use for implementing **Step 1** and **Step 2** of the post-edit generation process. For both cases, we always generate the output with greedy decoding.

Step 1 - Prompt

```
System Prompt
You are a medical expert in discharge summaries, specifically the
```

Brief Hospital Course (BHC) section. You are simulating realistic clinician edits to LLM-generated BHCs.

User Prompt Template

You will be given:
1. An LLM-generated BHC (the "pre-edit BHC draft" that a clinician is editing)
2. A clinician-written BHC for the same admission (the gold-standard BHC)

DEFINITION: A "clinical detail" is a discrete piece of clinical information in the BHC (e.g., a single diagnosis with its supporting rationale, a single medication with its dose and indication, a single procedure with its outcome). Clinical details are the units of comparison between the two BHCs: each detail may be present in one BHC, both BHCs, or neither, and may be expressed with different levels of specificity, certainty, or terminology across the two. A clinical detail can span multiple atomic facts when those facts are tightly coupled and would naturally be edited together (e.g., a medication and its dose).

YOUR TASK: Simulate a clinician editing the LLM-generated BHC draft so that its clinical content aligns with the clinician-written BHC. Decompose each BHC into its constituent clinical details, then for each detail in the LLM-generated BHC, cross-reference it against the corresponding detail (if any) in the clinician-written BHC to decide what edit should be made. The resulting post-edit BHC should look like a **minimally edited version of the LLM draft**, not a rewrite.

EDITING RULES:
- Preserve the LLM draft's overall paragraph structure, sentence order, and prose style. Do NOT reorganize into problem-oriented format, add headers, or convert between prose and bullets unless the LLM draft already uses that format.

- For clinical details discussed in both BHCs with *equivalent* content: Keep the LLM's phrasing as-is. Use the action "retain".

- For clinical details discussed in both BHCs where the clinician version contains a *specific clinical fact* that the LLM version generalizes or omits (e.g., a specific drug name, dose, route, frequency, lab value, anatomical location, device specification, or precise timing): Replace the vague phrasing with the specific fact, preserving the surrounding sentence structure. Use the action "specificity_adjustment". Do NOT use this action for paraphrasings or wording variations that don't add a concrete clinical fact. If the LLM and clinician versions convey the same clinical information using different words, use "retain" -- wording differences alone are not specificity adjustments.

- For clinical details discussed in both BHCs where the clinician version expresses a different *level of clinical certainty* (e.g., the LLM-generated BHC asserts "the patient had sepsis", while the clinician-written BHC says "the patient was treated empirically for possible sepsis"): Adjust the LLM version's hedging to match the clinician-written version. Use the action "confidence_level_adjustment".

- For clinical details discussed in both BHCs with *conflicting* content: Replace the LLM version with the clinician-written version. Use the action "factual_error_correction".

- For clinical details only discussed in the LLM-generated BHC: Decide whether to remove or retain based on the *type* of content. Use the action "condensation" ONLY when the LLM-only content is one of the following: (a) redundant content that restates information already covered elsewhere in the LLM draft; (b) low-value procedural or process detail (e.g., generic statements like "the patient was monitored", "the patient tolerated the procedure well") that doesn't advance clinical understanding; or (c) verbose narration of a fact already conveyed more concisely. For LLM-only content that does NOT fit the above (e.g., a specific clinical finding, a meaningful diagnostic or therapeutic step, a notable event during the admission): Use the action "retain". Many LLM-only details are clinically valid -- clinicians vary in what they choose to include in a BHC, and not every omission by the clinician means the content should be removed.

- For clinical details only discussed in the clinician-written BHC: Add them to the LLM draft where they fit naturally without breaking coherence or contradicting existing content. Keep the additions concise and consistent with the LLM's writing style. Use the action "omission_correction".

- The final post-edit note should be approximately similar in length to the LLM draft. Small increases are fine for necessary additions, but avoid overinflating the length.

PROCESS: First, produce an edit plan listing each change (or explicit retain) from start to end of the document. Then, produce the simulated post-edit note.

LLM-Generated BHC (Pre-Edit Draft)
{llm_bhc}

Clinician-Written BHC (Gold-Standard BHC)
{human_bhc}

Step 2 - Prompt

System Prompt
You are a medical expert who reformats clinical notes to match a clinician's preferred documentation style.

User Prompt Template
You will be given:
1. A *clinician-edited* BHC, where an LLM-generated BHC was edited by a clinician to ensure that the clinical content is correct
2. A *clinician-written* BHC showing the clinician's preferred documentation style

YOUR TASK: Reformat the edited BHC to match the clinician-written BHC's structural and stylistic conventions. Do NOT change any clinical content; only change formatting, organization, and writing style.

Aspects to align:
- Section headers / problem-oriented structure vs. free-form prose
- Bullet points vs. flowing paragraphs
- Abbreviation style (e.g., "year old" vs "y. o. ", "status post" vs "s/p", "follow-up" vs "f/u")
- Sentence fragments vs. complete sentences

```
- Problem ordering and grouping

Produce ONLY the reformatted note.  Do not add or remove clinical
content.

Edited BHC (Content-Correct, Needs Reformatting)
{step_1_output}

Clinician-Written BHC (Reflects Preferred Documentation Style)
{human_bhc}

Reformatted Note
```

B.6 Final Dataset Statistics

Table B4: Statistics for the model-specific edit-pair datasets. For each base model, we report sample counts and BHC length distributions for the train and validation sets for their corresponding edit pairs. Pre-edit BHCs are the zero-shot generations from each model (after post-processing); post-edit BHCs are the final outputs of our two-stage simulated edit pipeline (Section 4). Post-edit BHCs are not generated for two test sets introduced in Section 6.

			Pre-Edit BHC (words)		Post-Edit BHC (words)	
Base Model	Split	N	Mean (SD)	Median	Mean (SD)	Median
Gemma-3-4B-IT	Train	4,680	191.6 (98.7)	180	368.8 (162.6)	346
	Validation	256	184.1 (58.6)	174	354.1 (161.1)	334
Qwen-2.5-7B-Instruct	Train	3,302	107.0 (42.3)	98	305.3 (138.1)	282
	Validation	167	104.6 (40.8)	96	290.8 (139.0)	267
Llama-3.1-8B-Instruct	Train	4,775	252.4 (142.9)	229	378.5 (170.4)	353
	Validation	262	254.0 (106.2)	244	381.2 (184.8)	350

The final summary statistics for our edit pair dataset are shown in Table B4. The edit datasets for each model follow from: (i) filtering out the admissions whose input context exceeds the context window size of the given model; (ii) zero-shot generating the BHCs; (iii) applying the post-processing pipeline (Appendix B.4); (iv) running **Step 1**; and finally, (v) running **Step 2**. Note that we only generate the post-edit BHCs for the train and validation sets of MIMIC-III-Ext-BHC-Edit, since we only need the clinician-written BHCs as references for evaluation.

B.7 Details and Annotation Guide for Expert Validation of Simulated Edits

Below, we include the definitions of ratings for our expert annotation tasks described in Section 4.1:

Task 1: Coherence & Real Documentation Practice.

- **Reasonable**: Resulting text is coherent with other parts of the post-edit draft and written in professional language, organization, and style appropriate for clinical documentation.
- **Unreasonable**: Resulting text has at least one clear issue in terms of coherence with other parts of the post-edit draft or the language, organization, or style used.
- **Unsure**: Unclear whether the coherence and language used are reasonable or unreasonable.

Task 2: Clinical Content Alignment to the Reference.

- **Aligned**: The resulting changes only (i) substitute/add in clinical details substantiated by the reference BHC and (ii) remove clinical details that are not consistent with the reference BHC.

- **Not Aligned**: The resulting changes either (i) substitute/add in clinical details that are not substantiated by the reference BHC or (ii) remove clinical details that are consistent with the reference BHC and integral to the summarization.
- **Unsure**: It is unclear whether the given edit is aligned or not.
- **N/A**: The edit does not substantively modify the clinical content present in the BHC.

We also include the full annotation guide provided to the NP annotators:

Annotation Guide

```
We are evaluating the quality of edits made to an LLM-generated
discharge summary, specifically to its brief hospital course (BHC)
section.  The BHC section details key clinical events, diagnoses,
and treatments for a patient during their hospitalization.  The
pre-edit draft was obtained by providing an LLM with all of
the EHR notes available for a patient prior to their discharge
from the hospital.  For each hospitalization, there is also a
clinician-written reference BHC, which serves as a source of
ground-truth.

The goal of this annotation task is to check if each individual
edit to the BHC draft is "reasonable", considering:
- Whether the resulting changes are coherent with other parts
of the final BHC draft and written in professional language,
organization, and style appropriate for clinical documentation.
- Whether they only (i) substitute/add in clinical details
substantiated by the reference BHC and (ii) remove clinical details
that are not consistent with the reference BHC.

What You Will See
Each annotation example will include the following items:
- A clinician-written reference BHC
- An LLM-generated draft of the BHC (pre-edit BHC)
- The final BHC draft (post-edit BHC)
- One of the edits made to the pre-edit BHC

In the pre-edit and post-edit BHCs, the text added/removed via the
edit are also demarcated with "{{ }}".

What To Do
For each edit, determine whether it reads as a "reasonable" edit,
considering the description above.  Specifically, provide the
following information:

1-1. Coherence & Documentation Practice
- Reasonable:  The resulting text is coherent with other parts
of the post-edit draft and written in professional language,
organization, and style appropriate for clinical documentation.
- Unreasonable:  The resulting changes have at least one clear
issue in terms of coherence with other parts of the post-edit draft
or the language, organization, or style used.
- Unsure:  You are unsure whether the coherence and language used
are reasonable or unreasonable.

1-2. Free-Text Comment
Please briefly explain your reasoning.

2-1. Clinical Content Alignment to Reference
- Aligned:  The resulting changes only (i) substitute/add in
clinical details substantiated by the reference BHC and (ii) remove
clinical details that are not consistent with the reference BHC.
- Not Aligned:  The resulting changes either (i) substitute/add in
clinical details that are not substantiated by the reference BHC or
```

(ii) remove clinical details that are consistent with the reference BHC and integral to the summarization.
- Unsure: You are unsure whether the given edit is aligned or not.
- N/A: The edit does not substantively modify the clinical content present in the BHC.

2-2. Free-Text Comment
Please briefly explain your reasoning.

Important Notes (Please Read)
- "Reasonable" does not mean "you would personally make this change". The focus should be on whether an edit clearly degrades the clinical quality or utility of the BHC.
- Because all BHCs have undergone a de-identification process, the pre-/post-edit BHCs may contain redactions (e.g., "_____", "**First Name4**", "**Hospital1**", "NamePattern2") or imputed dates (e.g., "[**2122-6-20**]") that render the BHCs awkward or grammatically incorrect. Please ignore such de-identification artifacts when determining the "reasonability" of edits.

C Additional Details on the Edit Tagger

C.1 Prompt for Tagging Edits

Edit Tagging

System Prompt
You are a medical expert reviewing clinician edits made to AI-generated discharge summaries (in particular, the Brief Hospital Course (BHC) section). The discharge summary generation model receives the patient's pre-discharge EHR notes as input, and produces a draft discharge summary. A clinician then reviews the draft and makes edits to produce the final discharge summary. While clinicians generally make edits to improve upon the original AI-generated BHC, keep in mind that they may make edits that introduce new errors, add information unsubstantiated by the pre-discharge notes, or serve purely stylistic purposes.

User Prompt Template
You will be given:
1. An LLM-generated BHC draft (the "draft" that a clinician is editing)
2. A final clinician-edited BHC for the same admission
3. All pre-discharge EHR notes available for the patient

YOUR TASK: Identify all observable atomic edits between the LLM-generated BHC draft and the final edited BHC, then characterize each edit.

DEFINITION:
- An atomic edit is the smallest unit of change between the draft and final BHC that corresponds to a single reason for modification. It should correspond to one specific reason why the clinician modified the BHC. All text changes driven by that reason belong to the same edit; changes driven by different reasons should be considered as separate edits. For example, if a clinician changes "the patient had a fever and cough" to "the patient had a fever", this would be one atomic edit corresponding to the reason of correcting an unsupported symptom (cough).
- Split edits if they differ in:
(a) type of content being edited: clinical vs non-clinical

```
(b) grounding (supported vs. unsupported by EHR)
(c) the clinician's intent (e.g., correcting a factual error vs.
correcting an omission vs. making a stylistic change)

For each edit, provide the following information:
(1) a concise description of the textual changes resulting from the
atomic edit
(1a) the surface edit operation: addition, substitution, or
deletion
(1b) the exact verbatim text span from the Original BHC
corresponding to this atomic edit
- copy the text exactly as it appears in the Original BHC
- if the atomic edit is an addition with no relevant Original-BHC
span, return an empty string
- prefer the smallest contiguous text span that uniquely identifies
the edit
(1c) the exact verbatim text span from the Final Edited BHC
corresponding to this atomic edit
- copy the text exactly as it appears in the Final Edited BHC
- if the atomic edit is a deletion with no relevant
Final-Edited-BHC span, return an empty string
- prefer the smallest contiguous text span that uniquely identifies
the edit

(2) whether the edit modifies clinical content in the LLM-generated
BHC
- clinical edits are edits that modify the clinical content of the
BHC, such as changes to diagnoses, treatments, symptoms, or other
medical information.
- non-clinical edits are edits that leave the clinical content
unchanged and modifies formatting, organization, grammar fixes,
or stylistic properties.

(3) if the edit is a clinical edit, determine whether the edit is
clearly supported by the pre-discharge EHR notes
- grounded: Supported by EHR notes (e.g., a diagnosis mentioned in
the BHC is also mentioned in the EHR notes)
- ungrounded: Not supported by EHR notes (e.g., a diagnosis
mentioned in the BHC is not mentioned in any of the EHR notes)

(4) reasoning for whether the edit is or isn't substantiated by the
pre-discharge EHR notes

(5) if the edit is ungrounded, select one of the following
categories that best describes the edit's relationship to the
pre-discharge EHR notes:
- contradictory: conflicts with information in the pre-discharge
notes
- unsupported: not explicitly present and not inferable from EHR
- inferred but uncertain: not explicitly stated but could be
inferred from the available information, but there is doubt as to
whether or not the inference is correct

(6) if the edit is a clinical edit, select one of the following
categories that best describes the edit type:
- factual_error_correction: fixes incorrect or unsupported content
in draft
- omission_correction: adds missing clinically relevant
information supported by EHR
- refinement: modifies specificity, detail (e.g., adding more
specific terms or taking out details), or certainty of clinical
content without correcting a factual error or adding omitted
content
```

```
- other (if the edit does not fit into any of the above categories)

(7) if the edit is a clinical refinement edit, select one of the
following categories that best describes the edit type:
- condensation of content:  when the clinician takes out
information that is not conflicting with the EHR data
- specificity adjustment:  when the clinician adds more specific
information that is supported by the EHR data (e.g., changing
"antibiotics" to "vancomycin and cefepime")
- hedging/certainty change:  when the clinician changes the
certainty level of a statement (e.g., changing "the patient had a
fever" to "the patient may have had a fever")

(8) if the edit is a nonclinical edit, select one of the following
categories that best describes the edit type:
- formatting:  changes related to the structure, layout, or
organization of the text (e.g., adding bullet points, changing
section headers)
- grammar:  changes that fix grammatical errors, punctuation, or
spelling
- wording:  changes that adjust phrasing or terminology without
altering clinical content
- other:  any other non-clinical edits that do not fit into the
above categories

(9) a description of the clinician's intent underlying the edit,
e.g.:
- Example 1:  The clinician corrected an imprecise admission
diagnosis in the AI-generated note by substituting in the specific
clinical diagnosis established during the hospitalization.
- Example 2:  The clinician corrected an incomplete description of
the patient's antibiotic course by substituting in the specific
agents used and the de-escalation that occurred during the
hospitalization.
- Example 3:  The clinician added a consultation and its
recommendations that occurred during the hospitalization but were
omitted from the AI-generated note.
- Example 4:  The clinician removed speculative differential
diagnosis content that is not appropriate for the summary nature
of the brief hospital course.

(10) based on the clinician's intent and the nature of the edit,
your judgment if this is a global vs.  personal edit:
- global edits are edits that you think most clinicians would make
given the same LLM-generated BHC draft and the same pre-discharge
EHR notes.  These edits are driven by reasons that are likely to be
shared across different clinicians.
- personal edits are edits that may be driven by individual
clinician preferences, such as stylistic choices or different
thresholds for including certain information in the BHC. These
edits are less likely to be shared across different clinicians.

IMPORTANT:
- Do not merge changes driven by different reasons into a single
edit.
- The `pre_span_text` and `post_span_text` fields must be copied
verbatim from the corresponding BHC text, not paraphrased.
- If incorrect clinical content is replaced with correct clinical
content in the edit, this should be treated as correcting a factual
error, not an omission of the added content.  In this sense,
factual-error corrections and omission corrections are mutually
exclusive for a given atomic edit.
- For each item, keep your response concise and clear, around 1-2
```

```
sentences.
```

C.2 Edit Distribution

We report the distribution of edits in our **Step 1** and **Step 2** post-edit BHC generations in Table C1.

Table C1: Distribution of edit types across models and admissions included. **Step 1** and **Step 2** show the distributions after each step of the post-edit simulation process (Section 4), with atomic edits extracted with respect to the pre-edit BHCs. "N/A" groundedness corresponds to non-clinical edits, since it is only assessed for clinical edits. Mean denotes the average number of edits per note.

		Count		Operation (%)			Clinical (%)		Groundedness (%)			Scope (%)	
Model	Setting	Total	Mean	Add	Sub	Del.	Clin.	Non-C.	Grnd.	Ungr.	N/A	Global	Pers.
Gemma-3-4B-IT	**Step 1**	99,960	20.25	53.15	28.28	18.57	97.17	2.83	64.97	30.93	2.83	76.65	23.35
	Step 2	122,516	24.82	55.71	27.31	16.99	89.15	10.85	61.30	26.34	10.85	70.73	29.27
Qwen-2.5-7B-Instruct	**Step 1**	64,377	18.56	59.58	27.85	12.57	98.59	1.41	61.27	36.14	1.41	73.41	26.59
	Step 2	80,455	23.19	60.22	28.66	11.13	89.42	10.58	57.06	31.05	10.58	67.90	32.10
Llama-3.1-8B-Instruct	**Step 1**	103,812	20.61	51.34	32.73	15.94	98.14	1.86	65.12	31.89	1.86	74.59	25.41
	Step 2	126,503	25.11	53.37	31.60	15.02	88.72	11.28	60.28	27.09	11.28	67.25	32.75

C.3 Details on Edit Tagger Evaluation

Clinical vs. Non-Clinical & Global vs. Personal Edits. As discussed in Section 5, we assess the edit tagger judgments along the *Clinical vs. Non-Clinical* and *Global vs. Personal* axes by collecting expert annotations from nurse practitioners (NPs) with hospitalist backgrounds. In order to achieve uniform coverage over different base LLMs—i.e., Gemma-3-4B-IT, Qwen-2.5-7B-Instruct, and Llama-3.1-8B-Instruct—and all types of edits (as determined by the edit tagger), we use a stratified sampling approach for selecting the 25 atomic edits to annotate, which all come from the training and validation sets of MIMIC-III-Ext-BHC-Edit. Given that we annotate atomic edit along both *Clinical vs. Non-Clinical* and *Global vs. Personal* axes simultaneously, we first allocate roughly the same number of examples to sample from each axis combination: 7 clinical & global edits, 6 clinical & personal edits, 6 non-clinical & global edits, and 6 non-clinical & personal edits. Then, for each combination, we sample roughly the same number of edits from each base LLM. Through this process, we obtain an annotation set that consists of 13 clinical vs. 12 non-clinical edits, 13 global vs. 12 personal edits, with 9 edits from Gemma-3-4B-IT, 8 edits from Qwen-2.5-7B-Instruct, and 8 edits from Llama-3.1-8B-Instruct. We note that our sampling procedure upsamples certain combinations of the two axes that occur less frequently in practice (e.g., non-clinical edits are 89% personal in MIMIC-III-Ext-BHC-Edit, according to the edit tagger outputs), and thus our annotation set may not be reflective of the true distribution of edit types in MIMIC-III-Ext-BHC-Edit.

Below, we include the full annotation guide provided to the NP annotators:

Annotation Guide

```
We are evaluating the accuracy of an edit tagger, which extracts
and categorizes all atomic edits observable between the pre-edit
and post-edit versions of a discharge summary.  We specifically
focus on the brief hospital course (BHC) section of each discharge
summary, which details the key clinical events, diagnoses, and
treatments for a patient.  The pre-edit draft was obtained by
providing an LLM with all of the EHR notes available for a patient
prior to their discharge from the hospital.

The edit tagger categorizes each edit along each the following
axes:
```

1. Clinical vs. Non-Clinical
- Clinical edits are changes that alter the clinical content in the original draft (e.g., factual error correction, omission correction, condensation, specificity adjustment, and confidence-level adjustment), impacting the characterization of the patient's past and current clinical states. For instance, specificity adjustment such as "recent headache" → "headache since 5 days ago" or factual error correction such as "He was transferred to the CICU" → "He was transferred to the SICU" are examples of Clinical edits that change the clinical characterization of patients.
- Non-clinical edits are changes that leave clinical content intact and instead only modify the formatting, organization, or other stylistic properties of the BHC. For instance, a change from free-form prose to a bullet-pointed list of problems or a change from contracted abbreviations to expanded medical terms would be examples of Non-Clinical edits, assuming that they do not meaningfully change the documented clinical content.

2. Global vs. Personal
- Global edits are changes that the general population of hospitalists/practitioners would make to improve the clinical accuracy and quality of the BHC. For instance, an edit that alters core clinical details directly relevant to the key problems and events during a patient's hospitalization is likely a correction that most hospitalists/practitioners would make.
- Personal edits are changes that reflect local preferences, which can vary across hospitalists/practitioners. For instance, when discussing the same problem exhibited by a patient, one hospitalist/practitioner may prefer to provide extensive context including details that are not integral to its discussion, while another hospitalist/practitioner may prefer to document in a focused manner. As another example, a hospitalist/practitioner may prefer to use different terms or language to describe the same clinical content (e.g., hedging language).

The goal of this annotation task is to annotate whether a given edit is Clinical vs. Non-Clinical and Global vs. Personal, which will serve as labels for validating the edit tagger outputs.

What You Will See
Each annotation example will include the following:
- An LLM-generated draft of a BHC section (pre-edit BHC)
- The final BHC draft (post-edit BHC)
- One of the edits made to the pre-edit BHC

In the pre-edit and post-edit BHCs, the text added/removed via the edit are also demarcated with "{{ }}".

What To Do
For each annotation example, please provide the following information:

1-1. Clinical vs. Non-Clinical Label
- Clinical: The edit corresponds to a change that alters the clinical content in the original draft (e.g., factual error correction, omission correction, condensation, specificity adjustment, and confidence-level adjustment), impacting the characterization of the patient's past and current clinical states.
- Non-Clinical: The edit corresponds to a change that leaves clinical content intact and instead only modifies the formatting, organization, or other stylistic properties of the BHC.
- Unsure: You should select this label if you find the choice

```
ambiguous.

1-2. Free-Text Comment
Please briefly explain your reasoning for the Clinical vs.
Non-Clinical annotation task.

2-1. Global vs. Personal Label
- Global:  The edit corresponds to a change that
hospitalists/practitioners would generally make to improve the
clinical accuracy and quality of the draft.
- Personal:  The edit reflects local preferences that vary across
hospitalists/practitioners.
- Unsure:  You should select this label if you find the choice
ambiguous.

2-2. Free-Text Comment
Please briefly explain your reasoning for the Global vs.  Personal
annotation task.

Important Notes (Please Read)
- Because all BHCs have undergone a de-identification process, the
pre-/post-edit BHCs may contain redactions (e.g., "_____", "**First
Name4**", "**Hospital1**", "NamePattern2") or imputed dates (e.g.,
"[**2122-6-20**]") that render the BHCs awkward or grammatically
incorrect.  Please ignore such de-identification artifacts when
annotating the edits.
```

Grounded vs. Ungrounded Edits. We evaluate the edit tagger's groundedness judgments against VERIFACT-BHC (Chung et al., 2026), which provides clinician annotations of whether claims in BHC text are supported by prior EHR notes. Since VERIFACT-BHC annotates groundedness at a different granularity from atomic edits, there is no exact one-to-one correspondence between the units it annotates and the units our tagger produces. We therefore construct a comparable subset as follows: we restrict to clinical edits whose text spans align to at least one annotated proposition (via lexical similarity to VERIFACT-BHC chunks), then collapse the matched proposition verdicts into a single edit-level label–treating *Supported* as grounded and *Not Supported* or *Not Addressed* as ungrounded. Edits that match no propositions are excluded from evaluation. To ensure the validity of our evaluation, we conduct manual review of this procedure to mainly verify that text spans are aligned reasonably and resolve edits whose matched propositions yield mixed verdicts. We conduct manual evaluation on a randomly sampled (n=150) subset. On this subset, our tagger achieves an overall agreement rate of 78.0%, 80.0%, 75.3% with the clinician annotations for Gemma-3-4B-IT, Qwen-2.5-7B-Instruct, and Llama-3.1-8B-Instruct, respectively.

D ADDITIONAL DETAILS ON EDIT FILTERING

D.1 DETAILS ON EDIT FILTERING PROCESS

For each pair and filter, after we produce a target by reverting every atomic edit whose tags are excluded by the filter, as illustrated in Section 5, we apply a "smoothing" step based on GPT-5.5 that repairs local fluency artifacts (e.g., broken sentences, dangling references, inconsistent tense). To minimize unwanted content changes from smoothing, we implement it with explicit instructions not to alter the propositional content of any retained edit (see prompt in Appendix D.2).

Smoothing ablations. Here, we show that our findings in Section 7 are not dependent on having the smoothing step within the edit filtering process. To demonstrate this point, we conduct additional ablations with Qwen-2.5-7B-Instruct, where we compare between three settings:

1. **Default Smoothing:** The default approach based on the prompt in Appendix D.2.

Table D1: The impact of different smoothing step implementations in edit filtering on SFT results, in the *Grounded*-filter setting. We show the results for Qwen-2.5-7B-Instruct.

	Alignment to Clinician-Written Reference			Groundedness		
Setting	DocLens-P	DocLens-R	MEDCON	MedScore	DocLens-Attr.	PDSQI-9
Base	0.271	0.174	0.253	0.948	0.900	0.599
SFT (**Step 2**)	0.303	0.331	0.380	0.874	0.753	0.236
SFT (*Grounded*, Default Smoothing)	0.320	0.325	0.371	0.953	0.880	0.580
SFT (*Grounded*, Minimal Smoothing)	0.304	0.310	0.368	0.948	0.866	0.577
SFT (*Grounded*, No Smoothing)	**0.334**	**0.339**	**0.383**	**0.954**	**0.861**	**0.574**

2. **Minimal Smoothing:** We update the LLM prompt for smoothing to explicitly restrict any repairs to be at the level of punctuation, spacing, duplicated conjunctions, dangling fragments, or sentence boundaries that are directly broken by the edit reversal.
3. **No Smoothing:** Edits are reverted with regex, and no repairs are made to fix broken or incoherent sentences.

In Table D1, we see that the choice of smoothing has little impact on the observed trends—i.e., improvements in alignment to reference, less degradation in groundedness compared to when fine-tuning on the raw post-edit BHCs (**Step 2**). Notably, even with no LLM-based smoothing in the filtering step (No Smoothing, in bold), we observe that filtering mitigates degradation in groundedness after SFT, while still allowing improvements over the base model.

D.2 Prompt for Filtering Post-Edit BHCs

Filtering Prompt.

```
System Prompt
You are a medical expert revising a clinician-edited brief hospital
course (BHC). Some edits in the clinician-edited BHC have already
been identified as edits that should be reversed under a predefined
filter policy.  Your job is to remove or revert only those specific
edits, while preserving all remaining content.

You are not deciding which edits should be reversed.  That decision
has already been made for you.  You must execute the requested
changes and keep the note coherent.

Rules:
- Apply only the listed edit reversals.
- Do not remove content beyond the listed edits.
- Do not add any new clinical facts.
- Keep changes minimal and local.
- It is allowed to make small wording, punctuation, or
sentence-flow repairs that are necessary after removing/reverting
the listed content.
- Preserve the overall structure and style of the clinician-edited
BHC as much as possible.
- Output only the final filtered BHC text.

User Prompt Template
You will be given:
1.  The original AI-generated BHC draft
2.  The clinician-edited BHC
3.  A list of already-identified edits that must be reversed

How to interpret the listed edits:
- If the edit operation is `addition`, remove the listed
`post_span_text` content.
- If the edit operation is `substitution`, replace the listed
```

```
'post_span_text' with the listed 'pre_span_text'.
- If the edit operation is 'deletion', restore the listed
'pre_span_text' content into the note where it naturally belongs.

You may make only small local wording or punctuation repairs
that are necessary after applying those changes.  Do not add new
clinical facts.  Do not remove anything beyond what is needed to
reverse the listed edits.

Original AI-Generated BHC
{pre_edit_bhc}

Clinician-Edited BHC
{post_edit_bhc}

Targeted Edits To Reverse
{targeted_edits}
```

E Additional Details on Experiments

E.1 Model Training

For SFT, we train all models with LoRA adapters for 5 epochs with rank $r = 16$, batch size $B = 64$, and weight decay $\lambda = 0.01$, using the AdamW optimizer (Loshchilov & Hutter, 2019) with cosine annealing (peak learning rate $\eta =$ 1e-4) and a linear warmup ratio of 0.03. We add the LoRA adapters to all of the self-attention projection weights and the multi-layer perceptron (MLP) layer weights. We use the validation cross-entropy loss for early stopping and use the resulting model for final evaluation on the test set.

E.2 Model Inference and Prompts

To generate the BHCs from each model (including both the base model and the fine-tuned models), we use the following prompt shown below and sample with greedy decoding. We only include the system prompt for zero-shot generations using the base model. For the SFT runs, we exclude the system prompt in training and evaluation.

BHC Generation Prompt

```
System Prompt
You are a medical expert on documenting discharge summaries, in
particular, its brief hospital course (BHC) section.  Given all
of the pre-discharge EHR notes available for a patient, your
job is to generate the BHC section.  The BHC section should be
concise and to the point, and all of the clinical details in the
BHC section should be clearly substantiated by the pre-discharge
notes.  Do not output anything other than the BHC section.  No
headers, explanations, or text corresponding to other sections of
the discharge summary.

User Prompt Template
Generate the brief hospital course (BHC) section of the discharge
summary for this patient.

## Note 1/N: [{CATEGORY}] [{DESCRIPTION}] [{CHARTDATE}]
⟨Prior Note⟩

⋮
```

```
## Note N/N: [{CATEGORY}] [{DESCRIPTION}] [{CHARTDATE}]
⟨Prior Note⟩

**YOUR RESPONSE:**
```

As noted in Section B.4, we find that models may fail to generate a valid BHC or contain degenerate outputs (e.g., repetitions until consuming the full output token budget), especially when zero-shot prompting a base model. To mitigate the impact of such artifacts in evaluation, we always apply the same post-processing pipeline to all models.

E.3 Construction of the Grounded Test Set (VeriFact-BHC)

To construct the *grounded* test set based on VeriFact-BHC (Section 6), where we remove any ungrounded clinical details from the clinician-written reference BHCs, we use the clinician annotations provided by Chung et al. (2026). In particular, for each clinician-written reference BHC, we retrieve the atomic claim-level *clinician consensus* annotations on whether the given claim is deemed as "Supported" (fully supported by the input prior notes), "Not Supported" (partially supported at best), or "Not Addressed" (not discussed at all) and remove all claims that are labeled as the latter two. To implement this approach, we provide GPT-5.5 with the following prompt below, where "flagged_claims_block" is a list of atomic claims from VeriFact-BHC with "Not Supported" or "Not Addressed" labels. We emphasize here that the claims removed are solely determined by the clinician annotations from VeriFact-BHC and that GPT-5.5 only carries out the removal.

Grounded Test Set - Prompt

```
System Prompt
You are a clinical documentation expert.  You are removing
specified clinical details from a Brief Hospital Course (BHC)
section of a discharge summary while preserving the rest of the
document verbatim and the document's coherence.

User Prompt Template
You will be given:
1.  A human-written BHC (the document to edit)
2.  A list of atomic claims that have been flagged for removal,
each with a verdict label
- "Not Supported":  the claim is addressed in the prior EHR notes
but is not fully supported by them
- "Not Addressed":  the claim is not addressed at all in the prior
EHR notes

YOUR TASK: Produce a stripped version of the BHC in which the
flagged claims have been removed.  The goal is to obtain a version
of the BHC that is entirely grounded in the prior EHR notes --
every clinical detail in the stripped BHC should correspond to
content that is supported by, or at least addressed in, those
notes.

EDITING RULES:
- Preserve all clinical content NOT in the flagged-claims list
verbatim.  Do not paraphrase, reorganize, or compress supported
content.
- Preserve the original paragraph structure, section headers (e.g.,
"# AMS"), bullets vs.  prose, abbreviations, and writing style.
The stripped BHC should look like a minimally edited version of the
original -- not a rewrite.
- For each flagged claim, locate the corresponding clinical detail
in the BHC and remove it.  The flagged claim text is an atomic-fact
summary, not a verbatim substring; identify the BHC text that
*expresses* the claim and remove that text.
```

```
- When a flagged claim is fused with supported content in the
same sentence (e.g., "The patient was treated with vancomycin
for sepsis" where only "for sepsis" is unsupported), remove only
the unsupported portion and adjust grammar/punctuation minimally
so the remaining sentence reads naturally.  Mark these cases as
"partially_removed".
- If removing a flagged claim would leave a dangling reference, an
empty section, or a grammatically broken sentence, prefer removing
the surrounding scaffolding (e.g., the introductory clause or the
now-empty header) rather than retaining the unsupported claim.
Only mark a claim as "retained" if removing it is genuinely not
possible without rewriting supported content.
- Section-header cleanup:  if removing flagged claims leaves a
problem-oriented section header (e.g., "# DM2", "# URINARY TRACT
INFECTION") with no remaining clinical content beneath it, remove
that header as well.  Do not leave empty or near-empty sections
behind.
- Do NOT introduce any new clinical content, hedging language, or
transitional phrases.  The stripped BHC must be a strict subset of
the information in the original BHC.
- Do NOT remove content that was not flagged.  Supported claims and
unflagged content must remain.

PROCESS: First, for each flagged claim (in the order given),
produce an entry in the edit plan recording how it was handled.
Then, produce the stripped BHC.

Original Human-Written BHC
{bhc_text}

Flagged Claims
{flagged_claims_block}
```

F Additional Details on Expert Validation of BHC Groundedness

Here, we provide the remaining details for the expert validation study on the groundedness of model-generated BHCs, discussed in Section 7.3. We randomly sample 10 admissions to annotate, with 5 admissions coming from the *grounded* test set overlapping with VeriFact-BHC and the remaining 5 admissions coming from the MIMIC-III-Ext-BHC-Edit test set.

For each admission, we ask each annotator to compare the BHCs generated after SFT on the (i) raw post-edit BHC, (ii) *Grounded* filtered post-edit BHC, and (iii) *Grounded* + *Global* filtered post-edit BHC against the zero-shot-generated BHC, providing one of the following ratings:

- **A > B:** Clinical details in draft A are better grounded in the prior EHR notes.
- **A < B:** Clinical details in draft B are better grounded in the prior EHR notes.
- **A $\approx$ B:** The two drafts exhibit similar levels of groundedness. They are either *similarly grounded* or *similarly ungrounded* in the prior EHR notes, but one draft is not significantly worse than the other in terms of groundedness.

A and B refer to either the zero-shot-generated BHC or one of the three BHCs generated after SFT on raw or filtered post-edit BHCs, which are provided to the annotators in fully randomized order.

Below, we include the full annotation guide provided to the annotators:

Annotation Guide

```
We are assessing the groundedness / faithfulness of LLM-drafted
discharge summaries (specifically its brief hospital course (BHC)
section) to the input context provided to the LLM. The BHC section
```

details key clinical events, diagnoses, and treatments for a patient during their hospitalization. Since an LLM drafts the BHC based on all of the EHR notes available for a patient until their discharge from the hospital, a BHC is considered grounded if the clinical details present are well-substantiated by the prior EHR notes.

Given a pair of BHCs drafted by two different LLMs for the same hospital admission, the goal of this annotation task is to determine the model that produces better grounded BHCs.

What You Will See

Each annotation example will include the following items:
- Two drafts of the BHC section for the same hospital admission
- Prior EHR notes available until discharge time

Consecutive annotation examples may come from the same hospital admission, in which case they share the same prior notes to assess groundedness against.

What To Do

For each pair of BHCs, provide the following information:

1. "Groundedness" Rating
- A > B: Clinical details in draft A are better grounded in the prior EHR notes.
- A < B: Clinical details in draft B are better grounded in the prior EHR notes.
- A ≈ B: The two drafts exhibit similar levels of groundedness. They are either similarly grounded or similarly ungrounded in the prior EHR notes, but one draft is not significantly worse than the other in terms of groundedness.

2. Free-Text Comment
Please briefly explain your reasoning.

Important Notes (Please Read)

- Compare the two drafts only based on their levels of groundedness, and not on whether one draft is e.g., more clinically useful, thorough, better organized than the other. If two drafts exhibit similar levels of groundedness, do not account for the clinical significance, severity, or magnitude of deviations from the prior EHR notes to break the tie. For instance, if both drafts each respectively have one atomic clinical detail that is unsubstantiated, then they should likely be labeled as "A ≈ B".
- Because all BHCs have undergone a de-identification process, the BHC drafts may contain redactions (e.g., "_____", "**First Name4**", "**Hospital1**", "NamePattern2") or imputed dates (e.g., "[**2122-6-20**]") that render them awkward or grammatically incorrect. Please ignore such de-identification artifacts when comparing the drafts.

F.1 Inter-Annotator Agreement

As discussed in Section 7.3, for every comparison of groundedness between a BHC generated by one of the SFT models and the BHC zero-shot-generated by the base model, we obtain three independent NP annotations. Across $10 \times 3 = 30$ total pairwise comparisons (10 admissions, 3 comparisons per admission), we observe unanimous agreement on 63% (19/30) of cases, majority agreement on 27% (8/30), and a three-way disagreement in the remaining 10% (3/30) of cases. Since a three-way disagreement indicates that all three possible ratings—A > B, A ≈ B, A < B—are given, we treat the consensus label in such cases as a tie (i.e., A ≈ B). In Table F1, we also provide a detailed breakdown of the comparison results.

Table F1: Nurse-practitioner (NP) groundedness preferences over 10 admissions, where the BHCs from each SFT setting (**Step 2**, *Grounded*, *Grounded + Global*) are compared against the zero-shot-generated BHCs from the base model. Labels are consensus verdicts over three independent reads; comparisons on which the three reads disagreed entirely are counted as ties.

SFT Target	Groundedness vs. Base Model		
	Better	Equal	Worse
Step 2	20% (2/10)	0% (0/10)	80% (8/10)
Filtered: *Grounded*	30% (3/10)	60% (6/10)	10% (1/10)
Filtered: *Grounded + Global*	50% (5/10)	10% (1/10)	40% (4/10)
All comparisons	33.3% (10/30)	23.3% (7/30)	43.3% (13/30)



Figure G1: SFT on the post-edit BHCs filtered using our edit tagger (Section 5) enables performance improvements (left) while preserving groundedness/factuality with respect to the input prior EHR notes (right). *Clinician-Written* denotes the setting when we fine-tune each model on the original clinician-written BHCs for each admission. *Post-Edit (After Step 1/2)* denotes the setting where we fine-tune on the BHCs resulting from **Step 1** and **Step 2** of our post-edit generation pipeline (Section 4). *Filtered Post-Edit: Grounded* denotes the setting where we fine-tune on the post-edit BHCs from **Step 2**, after filtering out the ungrounded edits. *Filtered Post-Edit: Grounded + Global* denotes the setting where we fine-tune on the post-edit BHCs from **Step 2**, after filtering out the ungrounded and personal edits. Error bars indicate the 95% bootstrap CIs, computed by resampling the test set with replacement 10,000 times.

G Additional Results

G.1 SFT on Clinician-Written BHCs and Step 1 BHCs

First, we consider fine-tuning on the original *clinician-written* reference BHCs. This setting presumes access to human-written notes, which does not align with the main setting that we focus on, where clinicians always edit LLM drafts to complete documentation tasks (Section 6). We nonetheless include it as a reference point, as it reflects the gains achievable when fully human-written notes are available for supervision targets. As expected, we see that fine-tuning on clinician-written BHCs



Figure G2: Differences in the average groundedness scores between the ground-truth BHCs used for SFT, when considering the clinician-written reference BHC (left), the simulated post-edit BHCs from **Step 1** and **Step 2** of the MIMIC-III-EXT-BHC-EDIT dataset construction process (middle), and the BHCs obtained from applying our *Grounded* and *Grounded + Global* edit filters (Section 5) (right). We use the MedScore, DocLens-Attribution, and PDSQI-9 ("Accurate") metrics considered in our experiments (Section 6), and show the scores on edit pairs generated for each base model. The reference BHCs from MIMIC-III generally exhibit low groundedness scores, as suggested by prior works (Chung et al., 2026). The simulated post-edit BHCs, which are aligned in content and style to the reference BHCs, also show low groundedness scores. Filtering out the ungrounded edits leads to noticeably higher groundedness scores across all models and our two filtering strategies.

produces meaningful gains on alignment-based metrics across all 9 (model, metric) pairs relative to the base model (Figure G1). Critically, we observe that these gains come at the cost of degraded groundedness across all 9 (metric, model) pairs. Such degradation may stem from a mechanism we note in Sections 1 and 3.2: when clinicians draft or edit notes, they often draw information from sources external to the model's inputs, as noted in e.g., Chung et al. (2026).

Second, we assess the effect of using the intermediate (**Step 1**) output of our edit generation pipeline as the supervision target. In Figure G1, the trends mirror those of SFT on clinician-written BHCs and SFT on post-edit (**Step 2**) outputs: meaningful gains on alignment-based metrics across all models, but at the cost of degradation on all groundedness-based metrics. Relative to **Step 2**, the alignment gains are smaller for all 9 (metric, model) pairs. For groundedness, we observe all metrics degrade more than SFT on **Step 2** for Gemma-3-4B-IT, while the other two models show the opposite trend.

G.2 GROUNDEDNESS OF POST-EDIT BHCS IN MIMIC-III-EXT-BHC-EDIT

In Figure G2, we analyze the differences in average groundedness scores between the ground-truth BHCs used for SFT, when considering the clinician-written BHC, the simulated post-edit BHCs from **Step 1** and **Step 2** of the MIMIC-III-EXT-BHC-EDIT dataset construction process, and the

BHCs obtained from applying our *Grounded* and *Grounded + Global* edit filters (Section 5). We use the MedScore, DocLens-Attribution, and PDSQI-9 ("Accurate") metrics considered in our experiments (Section 6), and show the scores on edit pairs generated for each base model. The clinician-written BHCs from MIMIC-III generally show low groundedness scores, as suggested by prior works (Chung et al., 2026). The simulated post-edit BHCs, which are aligned in content and style to the clinician-written BHCs, also exhibit low groundedness. Filtering out the ungrounded edits leads to noticeably higher groundedness scores across all models and our two filtering strategies.

G.3 Varying the Level of Ungroundedness in Training Targets

Our main experiments in Section 7 compare between two fine-tuning setups: (i) SFT on raw post-edit BHCs, which contain all ungrounded edits simulated via our dataset construction process (Section 4); and (ii) SFT on the filtered post-edit BHCs (*Grounded*, *Grounded + Global*), where all edits classified as ungrounded by the edit tagger have been reverted (Section 5). To characterize how the level of ungroundedness in the training targets affects SFT performance between these two extremes, we conduct additional experiments, where we revert a randomly selected fraction of the ungrounded edits identified by our edit tagger in each post-edit BHC, at fixed levels of 0%, 30%, 50%, 75%, and 100%, using the same filtering procedure as in Appendix D.1. The 0% setting corresponds to SFT on the raw post-edit BHCs (**Step 2**), and the 100% setting corresponds to applying the *Grounded* filter as in Section 5. We conduct this analysis with Qwen-2.5-7B-Instruct, using the same fine-tuning and evaluation setup as in Section 6.

Table G1: SFT results for Qwen-2.5-7B-Instruct when reverting increasing fractions of randomly subsampled ungrounded edits in the post-edit BHCs. 0% corresponds to SFT on the raw post-edit BHCs (**Step 2**); 100% corresponds to applying the *Grounded* filter.

	Alignment to Clinician-Written Reference			Groundedness		
Setting	DocLens-P	DocLens-R	MEDCON	MedScore	DocLens-Attr.	PDSQI-9
Base Model	0.271	0.174	0.253	0.948	0.900	0.599
SFT (0% removed; **Step 2**)	0.303	0.331	0.380	0.874	0.753	0.236
SFT (30% removed)	0.288	0.347	0.382	0.877	0.783	0.303
SFT (50% removed)	0.314	0.340	0.372	0.905	0.811	0.349
SFT (75% removed)	0.321	0.326	0.376	0.927	0.828	0.429
SFT (100% removed; *Grounded*)	0.320	0.325	0.371	0.953	0.880	0.580

Table G1 shows that all three groundedness metrics improve monotonically as more ungrounded edits are removed. At 100% removal, the MedScore and PDSQI-9 scores recover to near or above the level of those achieved by the base model. In contrast, alignment to clinician-written references remains largely stable across all levels (e.g., MEDCON score between 0.371 and 0.382), and remains substantially above the base model throughout all fractions. These results suggest that ungrounded edits account for most of the groundedness degradation observed after SFT on raw post-edit BHCs (Section 7.1), while contributing little to improvements in alignment to clinician-written references.

These results also indicate how errors made by the edit tagger, if present, can affect filtering. Failing to identify a subset of ungrounded edits can be considered equivalent to removing only a fraction of ungrounded edits, and the gradual improvement in groundedness suggests that filtering remains beneficial under moderate error rates in the edit tagging outputs, while improvements in the edit tagger's accuracy translate to additional gains.

G.4 Average Lengths of Model-Generated BHCs

In Table G2, we summarize the average lengths of BHCs generated by the base model and the different fine-tuned models, including the "oracle" setting where models are fine-tuned on the original clinician-written reference BHCs. Across all model families, we overall see that the BHCs generated from the fine-tuned models tend to be much longer in length than the BHCs zero-shot-generated from the base models, with the filtered settings resulting in slightly shorter BHCs compared to the raw post-edit and clinician-written reference SFT settings.

Table G2: Average BHC length (word counts) by base model and SFT setting, on all examples from the *grounded* VERIFACT-BHC test set and the MIMIC-III-EXT-BHC-EDIT test set.

Model	Setting	Mean / Std		Min / Max		Quartiles		
		Mean	Std	Min	Max	Q1	Q2	Q3
Gemma-3-4B-IT	Base Model	192.9	69.5	29	851	153	182	221
	SFT (Clinician-Written BHC)	329.2	399.0	38	3751	132	234	380
	SFT (Post-Edit)	350.7	413.0	40	3031	145	259	393
	SFT (Filtered Post-Edit: *Grounded*)	295.3	386.3	45	3270	126	213	331
	SFT (Filtered Post-Edit: *Grounded* + *Global*)	310.7	379.4	76	3566	156	214	346
Qwen-2.5-7B-Instruct	Base Model	109.3	41.0	40	389	81	100	127
	SFT (Clinician-Written BHC)	294.5	285.1	51	3013	135	245	397
	SFT (Post-Edit)	279.8	271.1	62	2611	138	223	347
	SFT (Filtered Post-Edit: *Grounded*)	219.9	139.9	50	800	103	178	320
	SFT (Filtered Post-Edit: *Grounded* + *Global*)	177.3	221.9	50	2887	90	123	195
Llama-3.1-8B-Instruct	Base Model	257.2	148.9	81	2098	169	232	319
	SFT (Clinician-Written BHC)	373.7	437.0	37	2969	151	287	426
	SFT (Post-Edit)	499.3	542.9	91	3483	242	374	509
	SFT (Filtered Post-Edit: *Grounded*)	349.4	346.4	66	3327	176	310	413
	SFT (Filtered Post-Edit: *Grounded* + *Global*)	287.9	255.2	66	2989	167	232	349

Table G3: Average BERTScore, ROUGE-L, and BLEU-2 by base model and SFT setting, on all examples from the *grounded* test set based on VERIFACT-BHC (Section 6).

Model	**Setting**	**BERTScore**	**ROUGE-L**	**BLEU-2**
Gemma-3-4B-IT	Base Model	0.827	0.152	0.138
	SFT (Clinician-Written BHC)	0.855	0.279	0.257
	SFT (Post-Edit: **Step 1**)	0.842	0.218	0.202
	SFT (Post-Edit: **Step 2**)	0.850	0.248	0.222
	SFT (Filtered Post-Edit: *Grounded*)	0.851	0.257	0.256
	SFT (Filtered Post-Edit: *Grounded* + *Global*)	0.843	0.217	0.233
Qwen-2.5-7B-Instruct	Base Model	0.836	0.173	0.094
	SFT (Clinician-Written BHC)	0.857	0.272	0.259
	SFT (Post-Edit: **Step 1**)	0.845	0.234	0.201
	SFT (Post-Edit: **Step 2**)	0.856	0.274	0.247
	SFT (Filtered Post-Edit: *Grounded*)	0.855	0.270	0.304
	SFT (Filtered Post-Edit: *Grounded* + *Global*)	0.849	0.238	0.218
Llama-3.1-8B-Instruct	Base Model	0.837	0.191	0.186
	SFT (Clinician-Written BHC)	0.858	0.280	0.271
	SFT (Post-Edit: **Step 1**)	0.843	0.217	0.207
	SFT (Post-Edit: **Step 2**)	0.849	0.233	0.167
	SFT (Filtered Post-Edit: *Grounded*)	0.855	0.270	0.248
	SFT (Filtered Post-Edit: *Grounded* + *Global*)	0.843	0.218	0.234

G.5 REFERENCE-BASED NLG METRIC EVALUATIONS

In Table G3, we report the performance using standard NLG metrics—BERTScore, ROUGE-L, BLEU-2—for all models and data configurations on all examples from the *grounded* test set based on VERIFACT-BHC. For all metrics, we use the implementations part of the `evaluate` package from HuggingFace. SFT on clinician-written references yields the highest BERTScore for all three models and the highest or near-highest ROUGE-L. Among edit-based settings, the *Grounded*-filter setting performs best or comparably, and achieves the highest BLEU-2 overall for Qwen-2.5-7B-Instruct. Filtering ungrounded edits consistently lifts BLEU-2 and ROUGE-L over the base model with comparable performance in terms of BERTScore.

G.6 LLM Ablations for Edit Tagging and Filtering

While we use a single LLM (GPT-5.5) for edit simulation (Section 4), tagging, and filtering (Section 5) in our experiments, we conduct additional experiments to demonstrate that our findings in Section 7 generally hold in settings where we use an alternative LLM for edit tagging and filtering. To do so, we replace GPT-5.5 with Claude Opus 4.8 (Anthropic, 2026) for both edit tagging and filtering, while still starting from the same GPT-5.5-simulated edit pairs in MIMIC-III-Ext-BHC-Edit and using the same fine-tuning and evaluation setup as in Section 6. We then measure their agreement at the edit-tagging level and in terms of the SFT performance trends (i.e., filtered vs. raw post-edit training) resulting from using each model for edit tagging and filtering.

Table G4: Agreement between edit taggers based on GPT-5.5 and Opus 4.8. Span similarity refers to the mean similarity (Jaccard score) of matched atomic edit spans; filter agreement is the fraction of matched edit pairs on which both taggers make the same keep/revert decision under each filter.

Base Model	Mean Span Similarity	*Grounded*	*Grounded* + *Global*
Gemma-3-4B-IT	0.909	0.820	0.702
Qwen-2.5-7B-Instruct	0.910	0.820	0.701
Llama-3.1-8B-Instruct	0.904	0.832	0.711

Table G5: SFT results with GPT-5.5 vs. Opus 4.8 as the edit tagger + filter, for Gemma-3-4B-IT and Qwen-2.5-7B-Instruct. Clinical claim/term-matching metrics (quantifying alignment to clinician-written references) are computed on the *grounded* test set, while groundedness metrics are computed on examples from both test sets (Section 6). PDSQI-9 denotes the rescaled "accurate" score.

		Alignment			Groundedness		
Model	Setting	DocLens-P	DocLens-R	MEDCON	MedScore	DocLens-Attr.	PDSQI-9
Gemma-3-4B-IT	Base Model	0.200	0.173	0.230	0.929	0.864	0.330
	SFT (**Step 2**)	0.261	0.326	0.359	0.892	0.799	0.286
	SFT (*Grounded*, GPT-5.5)	0.299	0.303	0.355	0.934	0.876	0.541
	SFT (*Grounded*, Opus 4.8)	0.274	0.329	0.357	0.918	0.851	0.420
Qwen-2.5-7B-Instruct	Base Model	0.271	0.174	0.253	0.948	0.900	0.599
	SFT (**Step 2**)	0.303	0.331	0.380	0.874	0.753	0.236
	SFT (*Grounded*, GPT-5.5)	0.320	0.325	0.371	0.953	0.880	0.580
	SFT (*Grounded*, Opus 4.8)	0.322	0.299	0.373	0.932	0.835	0.463

Table G4 shows that the two models largely agree on edit spans and on which edits each filter reverts. The atomic edit outputs from each model were matched by computing the Jaccard similarity score between text spans for every possible pair of atomic edits with the same edit operation (e.g., addition, deletion, substitution) and identifying the highest-scoring pairs. The "Mean Span Similarity" in Table G4 refers to the average of these Jaccard similarity scores across all matched pairs. Their agreement in edit tagger outputs is generally lower with the *Grounded* + *Global* filter, consistent with the lower expert agreement on the *Global vs. Personal* axis (Section 5). With either model, SFT on *Grounded*-filtered BHCs substantially reduces the loss in groundedness caused by SFT on raw post-edit BHCs, while retaining most of the alignment gains (Table G5). However, the Opus 4.8-based filter recovers groundedness less fully: its scores remain slightly below the base model on MedScore and DocLens-Attribution for both models, and on PDSQI-9 for Qwen-2.5-7B-Instruct. Nonetheless, our overall conclusions in Section 7 remain unchanged: *with both GPT-5.5 and Opus 4.8-based edit taggers, filtering out the ungrounded edits mitigates degradation in groundedness after SFT, while enabling improvements over the base model*.

H Limitations

Semi-synthetic edits. While we obtain expert annotations to validate the quality of our simulated edits in MIMIC-III-Ext-BHC-Edit (Section 4.1), our simulated edits may not reproduce every aspect of how clinicians edit LLM drafts under real-world time and workflow constraints, and our dataset is not a substitute for evaluation on real post-deployment edit data. Rather, since every added or corrected detail in the post-edit BHCs comes from a real clinician-written BHC, the resulting summaries reflect information that clinicians actually document rather than arbitrary insertions. This

yields a controlled, reproducible testbed in which e.g., grounded, ungrounded, global, and personal edits can be isolated, which is difficult to establish at scale with *proprietary* post-deployment data. Our expert validation supports the quality of the simulated edits, but is limited in scale. Given the high cost of expert annotations, we leave a larger-scale expert validation study to future work.

Reliance on a single LLM. Our pipeline uses GPT-5.5 for edit simulation, tagging, and filtering. Our main findings largely hold with Claude Opus 4.8 (Anthropic, 2026) used for edit tagging and filtering (Appendix G.6), but our edit simulation process still relies on a single LLM.

Automated metrics for clinical documentation quality. In our experiments, we primarily assess improvements in BHC generation quality with reference-based clinical claim/term-matching metrics, which are imperfect proxies for measuring the clinical utility of model-generated BHCs. We leave additional expert evaluation to future work.

Personal edits. We mostly treat personal edits as noisy supervision signal, and our tagger is less reliable at distinguishing global from personal edits than along the other axes (Section 5). An interesting future direction is to leverage personal edits as a deliberate axis of adaptation to local settings.

Study scope. We focus on a documentation type in the clinical domain; natural extensions include constructing analogous benchmarks in non-clinical expert-edit settings (law, science).

Limited note coverage in MIMIC-III. As MIMIC-III (Johnson et al., 2016) does not capture all information available to clinicians, some clinical content in the reference BHCs may appear ungrounded only because the supporting prior notes are missing from the database, so the ungroundedness levels in our dataset may exceed those in real deployments.