Proceedings of Machine Learning Research 298:1–36, 2025 Machine Learning for Healthcare

FactEHR: A Dataset for Evaluating Factuality in Clinical Notes Using LLMs

Monica Munnangi*[1,2] munnangi.m@northeastern.edu
Akshay Swaminathan*[3] akshaysw@stanford.edu
Jason Alan Fries*†[2] jfries@stanford.edu
Jenelle Jindal[4] jjindal@stanford.edu
Sanjana Narayanan[2] sanjana.books@gmail.com
Ivan Lopez[3] ivlopez@stanford.edu
Lucia Tu[2] ltu@wellesley.edu
Philip Chung[7] chungp@stanford.edu
Jesutofunmi A. Omiye[3,8] tomiye@stanford.edu
Mehr Kashyap[3] mkashyap@stanford.edu
Nigam Shah[5,6,9] nigam@stanford.edu

[1] *Khoury College of Computer Sciences, Northeastern University*
[2] *Center for Biomedical Informatics Research, Stanford University*
[3] *Department of Biomedical Data Science, Stanford School of Medicine*
[4] *Stanford Health Care*
[5] *Department of Medicine, Stanford School of Medicine*
[6] *Clinical Excellence Research Center, Stanford School of Medicine*
[7] *Department of Anesthesiology, Perioperative & Pain Medicine, Stanford School of Medicine*
[8] *Department of Dermatology, Stanford School of Medicine*
[9] *Technology and Digital Solutions, Stanford Health Care*

Abstract

Verifying and attributing factual claims is essential for the safe and effective use of large language models (LLMs) in healthcare. A core component of factuality evaluation is *fact decomposition*, the process of breaking down complex clinical statements into fine-grained atomic facts for verification. Recent work has proposed fact decomposition, which uses LLMs to rewrite source text into concise sentences conveying a single piece of information, to facilitate fine-grained fact verification. However, clinical documentation poses unique challenges for fact decomposition due to dense terminology and diverse note types and remains understudied. To address this gap and explore these challenges, we present FactEHR, an NLI dataset consisting of document fact decompositions for 2,168 clinical notes spanning four types from three hospital systems, resulting in 987,266 entailment pairs. We assess the generated facts on different axes, from entailment evaluation of LLMs to a

*. Equal Contribution
†. Corresponding author

qualitative analysis. Our evaluation, including review by the clinicians, reveals substantial variability in LLM performance for fact decomposition. For example, Gemini-1.5-Flash consistently generates relevant and accurate facts, while Llama-3 8B produces fewer and less consistent outputs. The results underscore the need for better LLM capabilities to support factual verification in clinical text.

1. Introduction

Verifying and attributing factual claims is essential for the safe and effective use of large language models (LLMs) in healthcare. Evaluation strategies have been proposed to assess summarization quality, ensure claims are properly grounded (i.e., attributed to specific text in the clinical note), and reduce hallucinations and related errors (Zheng et al., 2023). A core component of these strategies is *fact decomposition*, which rewrites complex text into concise, atomic statements, each conveying a single piece of information(Fabbri et al., 2022; Chen et al., 2023; Min et al., 2023).

Fact decomposition is used to verify facts against source documents using QA systems, textual entailment, or other model-based evaluation methods. Textual entailment, in particular, has proven effective for automating factuality assessment (Kamoi et al., 2023; Ru et al., 2024; Xie et al., 2024). Although these verification techniques are well studied in both general and scientific texts (Wadden et al., 2020; Wright et al., 2022), their application to clinical texts (Xie et al., 2024) remains underexplored. This represents a significant gap, as performance metrics are highly sensitive to the quality of fact decomposition (Wanner et al., 2024).

In healthcare, many documentation tasks require summarizing information from electronic health records (EHRs) (Fleming et al., 2023). EHRs encompass multiple facets of patient care, including various types of clinical documents, tabular data, and a variety of unstructured data types (e.g., medical imaging, waveforms). Extracting, rewriting, or verifying evidence from EHRs is necessary for tasks like summarization (Van Veen et al., 2024; Hegselmann et al., 2024), patient phenotyping (Yang et al., 2024), clinical trial recruitment (Wornow et al., 2024), medical text simplification (Devaraj et al., 2021), and knowledge graph construction (Arsenyan et al., 2023). Just as facts in radiology reports must be grounded to specific pixel data (Bannur et al., 2024), text generated for documentation tasks must be grounded in data present in the patient's EHR. Evaluating the ability of LLMs to perform this attribution is necessary for the successful use of LLMs in healthcare.

The decomposition and verification of the facts in clinical notes using LLMs presents significant challenges. Clinical notes employ medical terminology, special acronyms, and non-grammatical shorthand. Clinical observations are often compositional, as in: "*lobulation at the apex of the left hemithorax along the mediastinal border is residual of slowly resolving hematoma.*" For research purposes, clinical notes are frequently stripped of their original markup—like tables, lists, and section headers—thereby increasing ambiguity and parsing difficulties. The de-identification process, which masks patient names and dates, further introduces noise that can confound LLMs. Moreover, medical documentation serves diverse functions: nursing notes offer brief, time-stamped updates on a patient's physiological status, whereas discharge summaries integrate an entire hospital stay's details for billing and compliance purposes. Consequently, these documents differ substantially in presentation, length, fact density, and temporal scope.

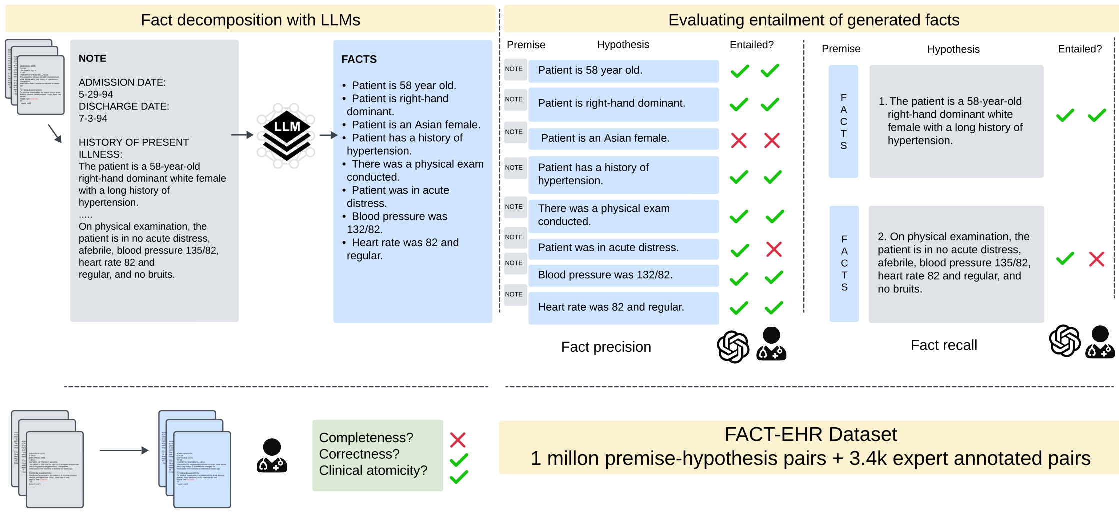

Figure 1: End-to-end pipeline of our work. We generate fact decompositions of clinical notes with LLMs. Evaluate the generated dataset in an entailment setting; For calculating Fact-Precision, we check if the **generated fact** (the hypothesis) logically follows from original **clinical note** (the premise) and vice versa for Fact-Recall. We validate the LLM-as-a-Judge with a subset of human evaluation. We also perform a modest amount of qualitative evaluation with experts.

Existing entailment datasets – which denote whether a piece of text (the *hypothesis*) logically follows from another piece of text (the *premise*) – do not capture the diversity of clinical documentation, either because they focus on a single note type (Miura et al., 2021), use artificial data (Romanov and Shivade, 2018a), or have a limited representation of hospital systems. Fact decomposition *across* different types of clinical documents has not been evaluated. In addition, it is frequently unclear how to evaluate the quality of a decomposition because determining entailment requires medical expertise. Finally, frontier model APIs are often not HIPAA-compliant, hindering evaluations using popular datasets such as MIMIC (Johnson et al., 2016).

To address this gap and to explore these challenges, we present FactEHR, a fact decomposition and textual entailment dataset derived from 2,168 clinical notes across four document types from three hospital systems. In Figure 1 provides an overview of the pipeline used to generate fact decompositions, evaluate them using LLM-as-a-Judge (a method that uses a large language model to assess whether a claim is entailed, contradicted, or unsupported by a source text), and validate the results against a subset of expert evaluations. We also include a small set of qualitative human evaluations. Using this dataset, we explore the following questions and contribute toward answering them:

RQ1: Fact Decomposition How do fact decompositions vary in length, quality, and similarity across LLMs? We compare outputs from four models and analyze key differences.

RQ2: Entailment How well can LLMs assess textual entailment in clinical documents? We evaluate performance on existing NLI benchmarks and a new set of expert-labeled entailment pairs derived from clinical notes.

RQ3: Fact Verification How accurate and complete are LLM-generated fact decompositions of clinical notes? We assess performance using formal metrics and detailed clinical expert review.

Our findings reveal considerable variation in the number and atomicity of fact decompositions across clinical documents and across LLMs, with some LLMs generating 2.6x more facts per sentence than others. The quality of fact decompositions also vary highly between LLMs, especially closed models generating factually consistent and covering most information as opposed to smaller open source models, where they are incomplete or they get into infinite generation loops. Consequently, higher fact counts do not correspond to better coverage or consistency—smaller open models often omitted key clinical details or hallucinated, while closed models more reliably captured core information. This raises questions about the validity of metrics that rely on fact decompositions for evaluating LLMs in healthcare documentation tasks. To facilitate future research in this direction, we release code[1], and data[2].

Generalizable Insights about Machine Learning in the Context of Healthcare

Our study highlights that fact decomposition is a crucial yet overlooked factor in evaluating LLMs for healthcare applications, as variations in decomposition quality significantly impact factuality assessments. We demonstrate that textual entailment, a key method for fact verification, faces unique challenges in clinical settings due to domain-specific complexities like shorthand, compositional observations, and structural inconsistencies. Additionally, we find substantial variability in how different LLMs generate fact decompositions, affecting downstream evaluations and raising concerns about the reliability of existing benchmarks. These insights underscore the need for standardized fact decomposition methodologies and more robust, domain-specific evaluation strategies to ensure accurate and trustworthy use of LLMs in clinical NLP tasks.

We release a natural language inference (NLI) dataset generated from real clinical notes spanning multiple note types and sources. Fact decompositions are produced by LLMs, with a subset evaluated by medical experts and scaled evaluation provided by LLM-as-a-Judge. This dataset can be used as (a) a benchmark for evaluating models on the fact decomposition task and (b) training models using both LLM-generated decompositions and their corresponding evaluations.

2. Related Work

Large Language Models (LLMs) in Healthcare LLMs encode extensive clinical and biomedical knowledge, recent models specifically, GPT-4 (OpenAI, 2023), as well as Gemini

1. https://github.com/som-shahlab/factehr
2. https://som-shahlab.github.io/factehr-website/

(Team et al., 2024), have emerged as state-of-the-art LLMs, showcasing impressive capabilities in a wide range of domain-specific applications (Singhal et al., 2023a; Agrawal et al., 2022; Munnangi et al., 2024). Several evaluations have shown that LLMs can achieve performance comparable to fully supervised models on tasks such as entity extraction and relation extraction (Wadhwa et al., 2023), under both few-shot and zero-shot settings. While considerable progress has been made on atomic fact generation in general domains (Min et al., 2023; Gunjal and Durrett, 2024) and biomedical literature (Wadden et al., 2020; Wright et al., 2022), there is little research focusing on clinical notes. Existing frameworks such as MAIRA 2 (Bannur et al., 2024) address fact-checking in radiology reports but not the broader range of clinical documents.

Clinical NLI Datasets Several related NLI datasets have been proposed in the clinical domain. The NLI4CT dataset focuses on clinical trial reports (Jullien et al., 2023) and represents one of the earliest efforts to apply NLI to scientific text, though it is limited in size, containing only 2,400 instances. Another dataset targets social determinants of health extracted from clinical notes (Lelkes et al., 2023). Our work differs substantially from both in terms of scale and the diversity of EHR note sources.

LLM-as-a-Judge Evaluating language model outputs remains an open research challenge, particularly for generative tasks where multiple valid responses exist (Chang et al., 2023). In closed-world tasks with a small, discrete set of correct answers—such as classification or multiple-choice question answering—models typically use exact match or likelihood-based selection to produce outputs, which are then evaluated using standard metrics like accuracy. For short-form text generation, token-overlap metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are commonly used. However, these metrics often fail to reward correct but paraphrased outputs. Semantic similarity metrics like BERTScore (Zhang et al., 2020) address this limitation by comparing contextualized token embeddings to measure some aspects of meaning, rather than relying on exact lexical overlap.

For complex generative tasks, evaluation is challenging due to the large space of plausible outputs and the potential mixing of correct and incorrect information within a single response. Recent work (Min et al., 2023; Xie et al., 2024; Tian et al., 2023) addresses this by decomposing text into atomic facts to assess factuality and completeness; for example, (Xie et al., 2024) compute claim-level precision and recall using entailment models. However, current models often struggle to generate atomic facts that are both accurate and comprehensive, especially in the clinical domain. In our experiments, entailment-based evaluation using LLMs showed strong correlation with human judgments. Given the high cost and limited scalability of expert annotation, we adopt an LLM-as-a-Judge framework—validated against human evaluations—to assess our dataset.

3. Methods

3.1. Experimental Setup

We use LLMs to perform fact decomposition, the process of breaking down source text into concise, atomic statements that convey individual pieces of information. We then examine the resulting fact decompositions, comparing their characteristics such as number of decomposed facts, length of the generated facts and similarity across different LLMs.

Finally, we pair each decomposition with its corresponding source note to form entailment pairs, evaluating how accurately and completely LLMs capture the original information. In this section, we describe the models, prompts, datasets, types of clinical notes and evaluation, in depth, with more information in the appendix as appropriate.

Models We consider four LLMs to perform fact decomposition: GPT-4o (OpenAI, 2024), o1-mini (Singhal et al., 2023b), Gemini-1.5-Flash-002 (Team et al., 2024), and Llama3-8b-Instruct (Dubey et al., 2024). For brevity, we will refer to Llama3-8b-Instruct as Llama3 or Llama3-8B and Gemini-1.5-Flash-002 as Gemini-1.5 in the rest of the paper (unless specified otherwise). We also conducted preliminary experiments with domain specific models including Google's MedLM but found its performance subpar[3], and so did not pursue further. All models were run using HIPAA-compliant compute environments and APIs. Model generation hyperparameters are included in Appendix D.

Prompts We use two distinct prompts: one for fact decomposition and another for entailment evaluation. The fact decomposition prompt, adapted from (Min et al., 2023), includes two in-context examples of clinical note fact decompositions and instructs the LLM to output independent facts as a delimited string (Appendix L, Figure 11). The entailment evaluation prompt is tuned using 40 premise-hypothesis pairs sampled from our datasets, optimizing for F1 score. The final version, adapted from (Xie et al., 2024), instructs the LLM to produce a binary entailment judgment in JSON format (Appendix 12).

3.2. Data Sources & Preprocessing

We sample clinical notes from three de-identified research datasets: MIMIC (MIMIC-III (Johnson et al., 2016) and MIMIC-CXR (Johnson et al., 2019)), from Beth Israel Deaconess Medical Center in Boston, MA; CORAL, from the University of California, San Francisco (UCSF) (Sushil et al., 2024); and MedAlign, from Stanford Health Care (SHC), Palo Alto, CA (Fleming et al., 2023). We randomly sample up to 250 notes per type from each dataset, limiting note lengths to 64–3840 whitespace-delimited tokens, for a total of 2,168 notes as summarized in Table 5 and Table 6 for token length. Additional details are in Appendix A.1.

3.3. Note Types

We consider four clinical note types: (1) *Procedure Note*, which typically includes procedures, clinical indications, findings, and follow-up recommendations. (2) *Nursing Note*, which provides a systematic assessment of a patient's condition across body systems (e.g., cardiovascular, neurologic) at a specific time, with less emphasis on future care planning. (3) *Progress Note*, which summarizes a patient's medical status from the previous day and outline the care plan for the next. (4) *Discharge Summary*, which provides a concise overview of the patient's presentation, past medical history, key findings, future medical plans, and discharge medications for subsequent care providers. We provide more details about note types in Appendix B.

3. More information in Appendix C.

3.4. Entailment evaluation (Validating LLM-as-a-Judge)

This study benchmarks the LLM-as-a-judge approach for fact verification in clinical texts, a previously under-evaluated component. To ensure statistical validity of performance metrics like sensitivity and specificity, power calculations (Buderer, 1996) guided the scale of human annotations based on NLI results from the FactEHR dataset. We annotated 1000 unique entailment pairs (250 per note type) and included 200 duplicates (50 per note type) to assess inter-annotator agreement, totaling 1200 annotations in this tranche—sufficient for high-confidence estimates (99% overall, 80% per note type). Combined with the 2468 previously annotated pairs, our dataset now comprises 3500 human-annotated clinical entailment pairs, making it one of the largest of its kind in the domain.

To evaluate the accuracy and completeness of fact decompositions, we pair each decomposition with its corresponding source note to form premise–hypothesis pairs for textual entailment. We assess whether LLMs can reliably evaluate entailment in this setting (see Appendix I, Table 9 for details). As part of early model development, we also benchmarked prior models such as RoBERTa, finding that their performance was consistently lower than that of frontier LLMs like GPT-4o. Additionally, we explored entity-level metrics such as RadGraph F1 (Jain et al., 2021), but found they did not correlate well with expert judgments of factuality in our domain, highlighting the need for more nuanced and holistic evaluation approaches.

We first benchmark five LLMs (GPT-4o, GPT-4o-mini, Gemini-1.5, Llama3-8B, and Llama3-70B) using four publicly available NLI benchmarks: SciTail (Khot et al., 2018), MedNLI (Romanov and Shivade, 2018a), MultiNLI (Williams et al., 2018), and SNLI (Bowman et al., 2015). Additional information about these datasets is provided in Appendix A.2. We also evaluate performance on the FactEHR development set, which is described below. All models use the entailment prompt shown in Figure 12, applied to all 987,266 entailment pairs in the FactEHR dataset. Of these, 1,036 pairs were manually annotated by clinical experts, enabling direct assessment of model accuracy.

3.5. Fact Verification

To assess the accuracy and completeness of model-generated fact decompositions, we use *fact-precision* and *fact-recall*, entailment-based metrics proposed by DocLens (Xie et al., 2024).

Let S denote the set of sentences obtained by tokenizing the clinical note d, and let C represent the set of factual claims produced by decomposing d. We define $[d \models c] = 1$ if a hypothesis c is completely entailed by a premise d, and 0 otherwise. For example, the hypothesis, *"There is a pleural effusion on the left side,"* is completely entailed by the premise *"Left pleural effusion increased from prior scan."*

Fact Precision Fact precision measures the accuracy of decomposed claims using entailment, treating each fact as a hypothesis and the entire clinical note as the premise. It is defined as the proportion of facts in C that are entailed by clinical note d: $P^* = \frac{1}{|C|} \sum_c [d \models c]$

Fact Recall Fact recall measures the completeness of a fact decomposition using entailment, treating each sentence of the clinical note as a hypothesis and the entire set of facts as the premise. It is defined as the proportion of sentences in S that are entailed by the fact decomposition C:

$$R^* = \frac{1}{|S|} \sum_{s \in S} [C \models s],$$

where S is the set of sentences in the clinical note, and $[C \models s]$ is an indicator function that evaluates to 1 if s is entailed by C, and 0 otherwise.

We use a weighted fact recall metric to account for multiple facts within a single sentence and parsing issues. Instead of counting each sentence as a single fact, we estimate the number of facts per sentence. Sentences that contain incomplete or invalid information (e.g., just units, numbers, or parsing errors) receive a weight of zero. All other sentences are weighted by their estimated fact count, determined by the sentence atomicity measure described earlier. The weighted fact recall is then defined as:

$$R^*_w = \frac{\sum_{s \in S} w_s [C \models s]}{\sum_{s \in S} w_s},$$

where w_s is the weight assigned to sentence s, representing the estimated number of facts it contains.

4. FactEHR

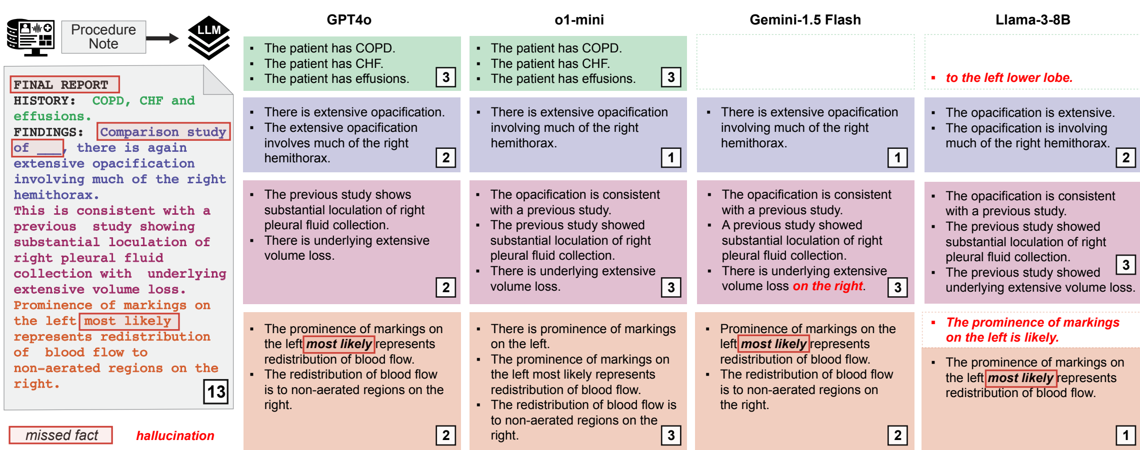

Figure 2: Example fact decompositions across LLMs. Colors indicate mappings from note (left) to corresponding facts, numbered boxes indicate fact counts, and red text indicates hallucination, red boxes indicated missed facts. All LLMs missed the the comparison study (a prior radiograph) and the note's final report status.

Complete details on data sources and preprocessing are provided in Subsection 3.2. Figure 2 shows example fact decompositions from each LLM evaluated in this study. Table 1 enumerates key dataset summary statistics.

4.1. Data Generation

We introduce FactEHR (*pronounced "factor"*), a dataset for fact decomposition and textual entailment created from de-identified clinical notes originating from three hospital systems. FactEHR includes:

- **Fact Decompositions**: 8,665 fact decompositions produced by four LLMs from 2,168 clinical notes spanning four document types.
- **Entailment Pairs**: 987,266 entailment pairs generated by pairing fact decompositions with their source text. Each pair has a binary entailment label from GPT-4o, and a subset of 1,036 pairs also includes labels from clinical experts.
- **Development Set**: A separate set of 2,468 entailment pairs annotated by clinical experts, used for entailment model development and tuning.

Name	Creation Rule	Size	Description
`FactNotes`	`note` $\in$ `D`	2,168	A collection of clinical notes from the dataset D.
`FactDecomp`	`note` $\rightarrow$ `fact-list`	8,665	Decomposition of notes into a list of structured, atomic facts.
`FactEntail_p`	`I[note` $\Rightarrow$ `fact]`	491,663	Entailment pairs for evaluating whether a note implies a given fact.
`FactEntail_r`	`I[fact-list` $\Rightarrow$ `sentence]`	495,603	Entailment pairs for evaluating whether a fact list implies a full sentence.
`FactEntail`	`FactEntail_p` $\cup$ `FactEntail_r`	987,266	Union of both entailment settings to support broad NLI evaluation.
`FactEntail_ann`	`x` $\sim$ `U(FactEntail)`	1,036	Human-annotated entailment data sampled from `FactEntail_p` and `FactEntail_r`.

Table 1: FactEHR dataset overview. We generate fact decompositions from 2,618 clinical notes and derive multiple NLI subsets.

Sampling FactEHR includes 8,665 fact decompositions generated by four LLMs from 2,168 clinical documents (4 x 2,168 = 8,172; five decompositions were not produced due to content moderation policies). To evaluate fact precision and recall, we formed 987,266 pairs from these decompositions. Each pair is classified as either a fact-precision pair, where the premise is the clinical note and the hypothesis is a fact from the decomposition, or a fact-recall pair, where the premise is the fact decomposition and the hypothesis is a sentence from the corresponding clinical note.

Entailment annotation From these 987,266, we randomly selected 1,036 pairs for manual entailment annotation by clinical experts. Clinical experts included two board certified physicians, two residents, two medical students, and a clinical researcher. Annotators were

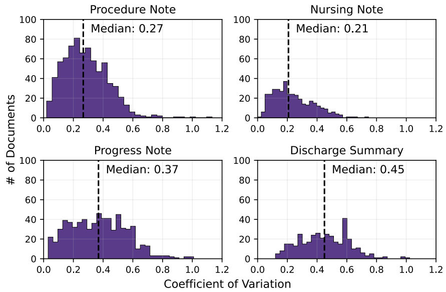

Figure 3: Distribution of coefficient of variation across documents. Higher values indicate higher variance in number of facts across LLMs.

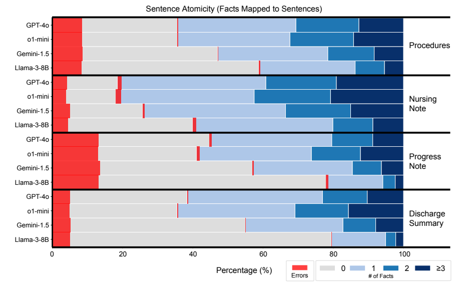

Figure 4: Sentence atomicity by note type and LLM, shown as the normalized distribution of the estimated number of facts-per-sentence

provided instructions for labeling entailment (Appendix E), and a subset of 100 entailment pairs were labeled in duplicate to calculate inter-rater agreement.

5. Results

5.1. RQ1 — Fact Decomposition

We examine the length and similarity of fact decompositions across LLMs, as well as the atomicity of source sentences.

Fact Counts We report the average number of facts generated per source note sentence across models and note types. To assess the variation in the number of facts generated across LLMs, we compute the coefficient of variation (standard deviation divided by the mean) of the number of facts generated by each LLM for each document (Everitt, 2006) of these counts. This metric, calculated across all documents within each note type, quantifies the degree of disagreement among LLMs on the total number of facts per document.

For all note types, GPT-4o and o1-mini produce more facts per sentence than Gemini-1.5 and Llama3-8B (Table 2). For discharge summaries, o1-mini generates 1.55 facts per sentence compared to 0.98 for Gemini-1.5 and 0.60 for Llama3-8B.

Figure 3 illustrates the distribution of the coefficient of variation (CV) in generated facts across clinical note types, showing the variability in the number of facts generated per document by LLMs. Discharge summaries exhibited the highest median CV (0.45), followed by progress notes (0.37), procedure notes (0.27), and nursing notes (0.21). The lower variability observed in procedure and nursing notes is expected given shorter token lengths.

Fact Similarities To quantify the similarity of fact decompositions across LLMs, we use Earth Mover's Distance (EMD) (Rubner et al., 1998; Kusner et al., 2015), which calculates the minimum cost to transform one probability distribution into another. Specifically, we embed each fact using ClinicalBERT (Alsentzer et al., 2019) and represent a single decomposition as a matrix of these fact embeddings. When comparing two decompositions, one serves as the source distribution of fact embeddings and the other as the target distribution. We then compute EMD using a cosine cost function and uniform weighting to determine how

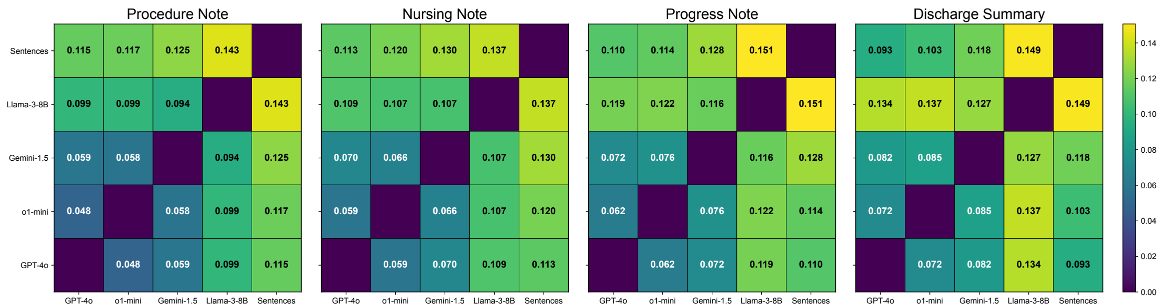

Figure 5: Mean EMD (cosine cost) across LLM fact decompositions and source sentences. This measures how embedded fact decompositions vary across LLMs and in comparison to the original document.

closely the source's set of facts aligns with the target's set, yielding a measure of similarity between the two decompositions.

Using (EMD) with a cosine cost function, we found that GPT-4o and o1-mini produced the most similar fact decompositions, closely followed by Gemini-1.5, whose pairwise EMD values against these models were consistently within 0.02 units of each other (Figure 5). In contrast, Llama3's decompositions were notably more divergent, with pairwise EMD scores roughly 0.03–0.05 units higher. Compared to the original source sentences, GPT-4o decompositions had the lowest mean EMD, while Llama3's were highest. Differences were also influenced by note type: procedure notes showed minimal variation due to their structured format, nursing notes had slightly higher EMDs reflecting more dynamic content, and discharge summaries—being lengthy and complex—exhibited the highest EMD values.

Sentence Atomicity To quantify the amount of information in each sentence of a source document, we estimate the number of facts contained within a single sentence, which we refer to as sentence atomicity. We assign each fact to its most similar sentence using a greedy matching approach based on cosine similarity of ClinicalBERT embeddings.

Overall, o1-mini produced the most facts per sentence, followed closely by GPT-4o, both maintaining similar rates of zero-fact sentences (Figure 4). Gemini-1.5 had more zero-fact sentences than GPT-4o and o1-mini, but fewer than Llama3.

Model Name	**Procedure Note**	**Nursing Note**	**Progress Note**	**Discharge Summary**
GPT-4o	1.45 ± 0.03	2.37 ± 0.14	1.28 ± 0.03	1.30 ± 0.02
o1-mini	1.45 ± 0.03	2.42 ± 0.13	1.46 ± 0.03	1.55 ± 0.03
Gemini-1.5	1.12 ± 0.03	2.05 ± 0.12	1.02 ± 0.02	0.98 ± 0.02
Llama3 8B	0.89 ± 0.03	1.58 ± 0.12	0.73 ± 0.04	0.60 ± 0.04

Table 2: Average facts per sentence by model and note type. Red indicates fewer facts than sentences, suggesting the LLM may fail to capture all facts.

Llama3 produced facts for only about 20% of sentences in progress notes and discharge summaries, and, except for nursing notes, generated fewer facts than sentences (Table 2).

	Procedure Note					Nursing Note				
	$P*$	$R*$	$R*_w$	$F1*$	$F1*_w$	$P*$	$R*$	$R*_w$	$F1*$	$F1*_w$
GPT-4o	**98.5**	**78.7**	92.4	**86.5**	95.0	**97.5**	**88.9**	92.2	**92.3**	94.2
o1-mini	97.8	78.4	**93.0**	86.2	**95.4**	96.6	88.5	**93.3**	91.6	**94.4**
Gemini-1.5	95.9	64.2	85.7	77.0	91.7	93.2	77.3	83.7	84.4	88.3
Llama3-8B	84.2	49.4	73.7	62.0	82.3	84.1	56.6	69.9	65.1	75.6
	Progress Note					Discharge Summary				
	$P*$	$R*$	$R*_w$	$F1*$	$F1*_w$	$P*$	$R*$	$R*_w$	$F1*$	$F1*_w$
GPT-4o	**97.4**	78.1	85.2	85.8	**89.9**	**97.0**	79.0	86.9	**86.8**	**91.1**
o1-mini	95.2	**79.7**	**85.7**	**86.6**	89.7	94.8	**81.1**	**88.2**	86.7	90.0
Gemini-1.5	96.0	66.2	75.2	77.9	83.9	96.5	65.4	76.3	76.9	83.8
Llama3-8B	83.4	41.9	57.4	53.8	67.0	84.6	38.9	57.3	50.4	67.2

Table 3: Comparison of average *fact precision* (P^*), unweighted (R^*) and weighted (R^*_w) *fact recall* and unweighted ($F1^*$) and weighted ($F1^*_w$) *fact F1* across documents for each note type and fact decomposition model. **Bolded** are the highest numbers (F1 and accuracy) of a model for each dataset. GPT-4o and o1-mini emerge as top performers across all note types.

Nursing notes have the highest proportion of sentences with three or more facts, reflecting their high information density. Progress notes, which tend to be longer and more narrative, have the lowest proportion of such sentences.

In summary, quantifying this variability is relevant for two reasons: (1) Fact Verification Approaches: Current workflows use LLM-generated decompositions to break claims into sub-claims for verification. However, our analysis reveals inconsistencies across LLMs. Ideally, facts should be atomic, addressing one property or action at a time. When decompositions fail to achieve this, they introduce uncertainty and lead to varying conclusions, and (2) Clinical Document Types: Clinical documents vary in length, structure, and complexity, challenging LLM-based decomposition. Radiology reports are concise, while discharge summaries are lengthy and complex. Procedure notes require detailed step-by-step breakdowns. LLMs struggle with longer, intricate documents, often producing incomplete decompositions. Addressing these challenges is key to improving LLM performance in clinical documentation.

5.2. RQ2 — Entailment

GPT-4o achieve the best F1 scores on four of the five benchmark datasets, including the FactEHR development set in table 10 (Appendix I), and was therefore selected as the entailment judge for all 987,266 entailment pairs in the FactEHR dataset. From these, we randomly sampled 1,036 pairs for manual annotation by seven clinical experts. On doubly annotated entailment pairs, Fleiss' Kappa was 0.67, indicating substantial inter-rater agreement. Under this evaluation, GPT-4o achieves a recall of 0.96 and precision of 0.86 (Appendix I, Table 11), closely mirroring its performance on the FactEHR development set. Overall, these findings suggest that GPT-4o serves as a reasonable proxy for human judgment in entailment assessments.

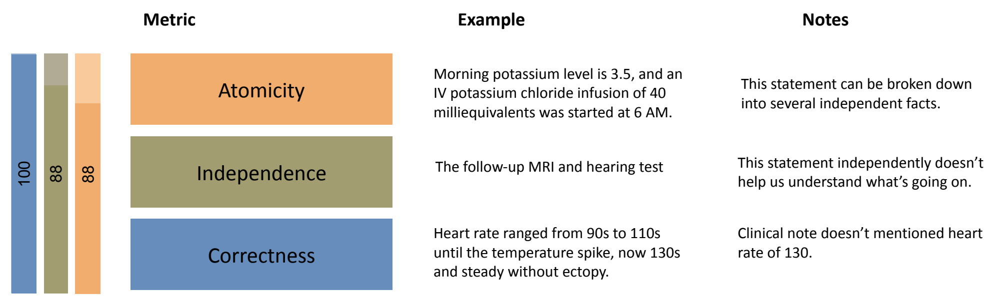

Figure 6: Overview of in-depth qualitative review with **GPT-4o** on the fact decomposition on twenty randomly-selected examples from FactEHR. We report **percentage** of correct, independent and atomic facts as annotated by medical expert.

5.3. RQ3 — Fact Verification

Using GPT-4o as the LLM entailment judge, we compute fact precision, fact recall (weighted and unweighted), and fact F1 (weighted and unweighted) for each fact decomposition across all note types (Table 3). GPT-4o and o1-mini emerge as top performers across all note types. For example, GPT-4o achieves a slightly higher fact F1 than o1-mini on procedure notes (86.5 vs. 86.2) and nursing notes (92.3 vs. 91.6), while o1-mini outperforms GPT-4o on progress notes (86.6 vs. 85.8). On discharge summaries, the two models are nearly tied (86.8 for GPT-4o vs. 86.7 for o1-mini). When considering the weighted fact F1 metric, o1-mini surpasses GPT-4o on procedure notes and nearly closes the gap on nursing notes. This suggests that the weighting scheme, which accounts for differences in fact counts per note, can influence which model is deemed superior.

Across all note types, fact precision remained consistently higher and less variable than fact recall—indicating that while the generated facts were generally accurate, many relevant facts were omitted, resulting in greater variability in recall and aligning with our qualitative findings (Appendix I, Figure 8).

Qualitative Expert Evaluation Expert reviewers assessed overall completeness of fact decompositions and the correctness, independence, and atomicity of individual facts. Completeness, defined as capturing all information in the source, is low: only 5% (1/20) of decompositions meet this criterion. For individual facts, reviewers judge whether each fact accurately reflected the source (correctness), stood on its own (independence), and represented a minimal unit of information (atomicity), we present detailed annotation guidelines in Appendix F.

Across models, Llama3 had the highest rate of incorrect facts. Discharge summaries yield the least number of atomic facts, reflecting their complexity, while progress and procedure notes contain more granular and independent facts. Detailed results for GPT-4o are shown in Figure 6; additional results for other models and a summary of key metrics are reported in Appendix J and Table 4, respectively.

Overall, fact decompositions by Llama3 on long notes are subpar, specifically they are either incomplete or they get into infinite generation loops. Fact decompositions by GPT-4o are better, but also not complete when there are long notes. Fact decomposition appears to

	Correct	**Independent**	**Atomic**
GPT-4o	100	88	88
Gemini-1.5	96	94	81
Llama3-8B	45	95	74

Table 4: Overview of qualitative (fact-level) review on the fact decomposition for 20 notes. We report **percentage** of correct, independent and atomic facts as annotated by medical expert.

be easier for shorter notes than super long/dense notes. Even though some long notes get decomposed to 200+ facts, they are still missing content that was not decomposed. Llama3 and GPT-4o seem to have a greater tendency to produce incomplete statements which is measured as a non-independent statement as compared to Gemini and o1-mini.

6. Discussion

In this study, we evaluate four LLMs' ability to decompose over 2,000 real-world, EHR-derived clinical notes of four types into independent facts. We evaluate the fact decompositions with an LLM-as-a-Judge (validated by expert evaluation), as an entailment task. For calculating Fact-Precision, we check if the generated fact (the hypothesis) logically follows from original clinical note (the premise) and vice versa for Fact-Recall[4]. To enable reproducibility and further research in this direction, we will release FACTEHR, a dataset encompassing 8,665 fact decompositions and 987,266 entailment pairs from four note types and three institutions, including 1,036 pairs annotated by seven clinical experts. The dataset was specifically designed to address a gap in the literature regarding the accuracy of fact decomposition in clinical text. While fact decomposition is a critical first step for many fact-checking and verification methods, such as FactScore (Min et al., 2023), there is limited research on its effectiveness, particularly when applied to both human- and LLM-generated clinical text. Our work aims to fill this gap by evaluating LLM performance on this foundational task, which is often overlooked in existing studies. Additionally, we examine the length and similarity of the decompositions and find that there was up to a 2.6-fold difference in the number of facts generated by different LLMs. Clinician review confirmed that while the generated facts were generally accurate, although, important details were often omitted.

While choosing the LLMs for fact decomposition task, we focused on evaluating a representative set of high-performing general-domain models, as these have demonstrated broad utility and scalability for real-world applications. That said, we acknowledge the importance of benchmarking against additional domain-specific models and discuss the need to evaluate them in future work. We also emphasize that the methodology proposed in this study is model-agnostic and can be applied to evaluate the performance of other models, including ones that are medically pretrained. The release of the FactEHR dataset directly supports future evaluation of LLMs by the community.

4. Detailed explanation in Section 3.5

In the future, we hypothesize that the dataset could be used to fine-tuning a smaller language model for fact-decomposition tasks, which could leverage the training and human-evaluated data splits from FactEHR to achieve higher quality decompositions tailored to specific needs. The results highlight the need to improve LLM-based fact decomposition methods for clinical documents to support the use of LLMs for healthcare use.

Limitations There are several limitations to this work. First, our study focuses on four common note types, but many other specialized note types (e.g., ophthalmology) may yield even greater variability in fact decomposition, suggesting our error estimates are conservative. Second, we rely on a combination of open-source and proprietary models within a HIPAA-compliant environment, ensuring privacy but limiting reproducibility; we mitigate this by releasing our dataset, prompts and code. Third, lack of access to ground truth fact decompositions for clinical notes. This constraint forces us to rely on estimates and proxy measures, such as Fact-Recall, to assess the completeness of model-generated fact decompositions. These metrics are inherently limited in their ability to capture the true completeness and accuracy of the decompositions. Lastly, since we examine only English documents, the generalizability of these findings to other languages remains uncertain. Additionally, there could be potential risks with the data. There could be data privacy concerns due to the de-identified but still potentially identifiable patient information, potential biases in the data. We acknowledge the potential for bias when using LLMs as judges. To mitigate this risk, we incorporated several safeguards; (a) LLM predictions are bench-marked against the human-annotated subset to ensure alignment and identify areas where LLMs might deviate (b) statistical power calculations guided the size of the human-labeled subset, ensuring it was large enough to rigorously validate LLM performance (c) our study explicitly discusses the limitations of LLM-based evaluation and highlights the importance of human-annotated benchmarks for future work.

In conclusion, we believe there is significant (and growing) utility to releasing large-scale synthetic and partially synthetic datasets in conjunction with LLM-based evaluators, reflected here by our FactEHRdataset and NLI clinician-curated dataset. This is especially important for academic labs that lack access to frontier model APIs that are HIPAA-compliant, creating barriers to exploring clinical text problems. Curating a sample of important clinical document types across multiple institutions and providing frontier LLM fact decompositions directly enables the research community to explore key problems in synthetic data quality evaluation, preference alignment, model distillation and more. By combining human-verified annotations with scalable LLM-based evaluations, our approach achieves a balance between scalability and the rigor needed for clinical applications, also keeping LLM bias in check (Li et al., 2024). This also ensures reliability as we rigorously validated the performance of LLMs by benchmarking them against a carefully selected subset of manually labeled examples. We also explicitly discusses the limitations of LLM-based evaluation earlier in this section and highlights the importance of human-annotated benchmarks for future work.

Acknowledgments This work is generously supported by the Mark and Debra Leslie endowment for AI in Healthcare. The authors also thank the Clinical Excellence Research Center (CERC) at Stanford for their support. AS thanks the Paul and Daisy Soros Foundation and Knight Hennessy Scholars for their support.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors, 2022.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909.

Vahan Arsenyan, Spartak Bughdaryan, Fadi Shaya, Kent Small, and Davit Shahnazaryan. Large language models for biomedical knowledge graph construction: Information extraction from emr notes, 2023. URL https://arxiv.org/abs/2301.12473. v2.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. Maira-2: Grounded radiology report generation, 2024. URL https://arxiv.org/abs/2406.04449.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

Nancy M. Fenn Buderer. Statistical methodology: I. incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. *Academic Emergency Medicine*, 3(9):895–900, 1996. doi: https://doi.org/10.1111/j.1553-2712.1996.tb03538.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1553-2712.1996.tb03538.x.

Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered McInerney, Byron C. Wallace, and Silvio Amir. Open (clinical) LLMs are sensitive to instruction phrasings. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii, editors, *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 50–71, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.5. URL https://aclanthology.org/2024.bionlp-1.5/.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models, 2023. URL https://arxiv.org/abs/2307.03109.

Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, Dan Roth, and Tal Schuster. PropSegmEnt: A large-scale corpus for proposition-level segmentation and entailment recognition. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8874–8893, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.565. URL https://aclanthology.org/2023.findings-acl.565.

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. Paragraph-level simplification of medical texts. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.395. URL https://aclanthology.org/2021.naacl-main.395.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and et al. Angela Fan. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Brian S Everitt. *The Cambridge dictionary of statistics*. Cambridge University Press 1998, 2002, 2006, 2006.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.187. URL https://aclanthology.org/2022.naacl-main.187.

Scott L. Fleming, Alejandro Lozano, William J. Haberkorn, Jenelle A. Jindal, Eduardo P. Reis, Rahul Thapa, Louis Blankemeier, Julian Z. Genkins, Ethan Steinberg, Ashwin Nayak, Birju S. Patel, Chia-Chun Chiang, Alison Callahan, Zepeng Huo, Sergios Gatidis, Scott J. Adams, Oluseyi Fayanju, Shreya J. Shah, Thomas Savage, Ethan Goh, Akshay S. Chaudhari, Nima Aghaeepour, Christopher Sharp, Michael A. Pfeffer, Percy Liang, Jonathan H. Chen, Keith E. Morse, Emma P. Brunskill, Jason A. Fries, and Nigam H. Shah. Medalign: A clinician-generated dataset for instruction following with electronic medical records, 2023.

Anisha Gunjal and Greg Durrett. Molecular facts: Desiderata for decontextualization in llm fact verification, 2024. URL https://arxiv.org/abs/2406.20079.

Stefan Hegselmann, Shannon Zejiang Shen, Florian Gierse, Monica Agrawal, David Sontag, and Xiaoyi Jiang. A data-centric approach to generate faithful and high quality patient summaries with large language models, 2024. URL https://arxiv.org/abs/2402.15422.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P.

Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports, 2021. URL https://arxiv.org/abs/2106.14463.

Daniel P. Jeong, Saurabh Garg, Zachary C. Lipton, and Michael Oberst. Medical adaptation of large language and vision-language models: Are we making progress?, 2024. URL https://arxiv.org/abs/2411.04118.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3, 2016. URL https://api.semanticscholar.org/CorpusID:33285731.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6 (1):317, 2019.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. NLI4CT: Multi-evidence natural language inference for clinical trial reports. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.1041. URL https://aclanthology.org/2023.emnlp-main.1041/.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-world entailment for claims in Wikipedia. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.470. URL https://aclanthology.org/2023.emnlp-main.470.

Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5189–5197. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.12022. URL https://doi.org/10.1609/aaai.v32i1.12022.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.

Adam Lelkes, Eric Loreaux, Tal Schuster, Ming-Jun Chen, and Alvin Rajkomar. SDOH-NLI: a dataset for inferring social determinants of health from clinical notes. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4789–4798, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.317. URL https://aclanthology.org/2023.findings-emnlp.317/.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv: 2411.16594*, 2024.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, 2023.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.416. URL https://aclanthology.org/2021.naacl-main.416.

Monica Munnangi, Sergey Feldman, Byron Wallace, Silvio Amir, Tom Hope, and Aakanksha Naik. On-the-fly definition augmentation of LLMs for biomedical NER. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3833–3854, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.212. URL https://aclanthology.org/2024.naacl-long.212/.

OpenAI. Gpt-4 technical report, 2023.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Lakshmi Radhakrishnan, Gundolf Schenk, Kathleen Ferar, Boris Oskotsky, Habibeh Ashouri Choshali, Thomas Plunkett, Sharat Israni, and Atul Butte. A certified de-identification system for all clinical text documents for information extraction at scale. *JAMIA Open*, 6, 07 2023. doi: 10.1093/jamiaopen/ooad045.

Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium, October-November 2018a. Association for Computational Linguistics. doi: 10.18653/v1/D18-1187. URL https://aclanthology.org/D18-1187.

Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain, 2018b. URL https://arxiv.org/abs/1808.06752.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation, 2024. URL https://arxiv.org/abs/2408.08067.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 59–66. IEEE, 1998.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023a.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023b. URL https://arxiv.org/abs/2305.09617.

Madhumita Sushil, Vanessa E. Kennedy, Divneet Mandair, Brenda Y. Miao, Travis Zack, and Atul J. Butte. Coral: Expert-curated oncology reports to advance language model inference. *NEJM AI*, 0(0):AIdbp2300110, 2024. doi: 10.1056/AIdbp2300110. URL https://ai.nejm.org/doi/abs/10.1056/AIdbp2300110.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry

Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi

Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. Fine-tuning language models for factuality, 2023.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142, February 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02855-5. URL http://dx.doi.org/10.1038/s41591-024-02855-5.

David Wadden, Kyle Lo, Lucy Lu Wang, Shanchuan Lin, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Conference on Empirical Methods in Natural Language Processing*, 2020. URL https://api.semanticscholar.org/CorpusID:216867133.

Somin Wadhwa, Jay DeYoung, Benjamin Nye, Silvio Amir, and Byron C Wallace. Jointly extracting interventions, outcomes, and findings from rct reports with llms. In *Machine Learning for Healthcare Conference*, pages 754–771. PMLR, 2023.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. A closer look at claim decomposition. In Danushka Bollegala and Vered Shwartz, editors, *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.starsem-1.13. URL https://aclanthology.org/2024.starsem-1.13.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume*

1 (Long Papers), pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

Michael Wornow, Alejandro Lozano, Dev Dash, Jenelle Jindal, Kenneth W. Mahaffey, and Nigam H. Shah. Zero-shot clinical trial patient matching with llms, 2024.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. Generating scientific claims for zero-shot scientific fact checking. *ArXiv*, abs/2203.12990, 2022. URL https://api.semanticscholar.org/CorpusID:247627737.

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. DocLens: Multi-aspect fine-grained medical text evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 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.

Jingye Yang, Cong Liu, Wendy Deng, Da Wu, Chunhua Weng, Yunyun Zhou, and Kai Wang. Enhancing phenotype recognition in clinical notes using large language models: Phenobcbert and phenogpt. *Patterns*, 5(1), 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023.

Appendix A. Dataset Details

A.1. Data Sources for Clinical Notes

For FactEHR, we sampled 2,168 clinical notes from three de-identified research datasets. We describe each of these datasets in detail and summarize the number of notes from each of them in table 5 and token statistics in table 6. We describe each source in detail below:

MIMIC MIMIC-III (Johnson et al., 2016) is a database of de-identified EHR comprising over 40k patients admitted to the intensive care unit of the Beth Israel Deaconess Medical Center between 2001 and 2012. MIMIC-III includes several note types, we sample **progress notes, discharge summaries and nursing notes**.

MIMIC-CXR The MIMIC Chest X-ray Database v2.0.0 (Johnson et al., 2019) is a large publicly available dataset of chest radiographs in DICOM format with free-text radiology reports. The dataset contains 377,110 images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, MA. We only use the free-text **radiology reports** for fact generation. We mapped the test splits from MIMIC-CXR with MIMIC-III and used a sub-sample of the split for our experiments.

CORAL is a collection of a diverse set of 20 breast cancer and 20 pancreatic cancer patients from the University of California, San Francisco (UCSF) Information Commons, which contained patient data between 2012–2022, de-identified with Philter (Radhakrishnan et al., 2023). We use all the 172 progress notes for our experiments.

MedAlign is collection of de-identified EHR data from Stanford Hospital and Lucile Packard Children's Hospital. We sub-sample **progress notes, nursing notes and discharge summaries** from MedAlign (Fleming et al., 2023) which consists of 276 longitudinal EHRs, out of which we use 750 notes.

Note Type	**# of Notes**		
	MIMIC	MedAlign	CORAL
Progress Note	250	250	172
Nursing Note	250	129	-
Discharge Summary	250	117	-
Procedure Note	-	250	-
– *Radiology Note*	500	-	-
Total	**1250**	**750**	**172**

Table 5: Counts of notes by type and data source. Radiology notes are a subtype of procedure notes, and were sampled from both MIMIC-III and MIMIC-CXR.

A.2. Entailment Data Sources

For entailment evaluation, we use existing NLI datasets. We describe them in detail, in the following paragraphs:

Note Type	# Tokens	
	Mean (SD)	Min/Max
Progress Note	1510 (1199)	[65, 6758]
Nursing Note	247 (168)	[60, 977]
Discharge Summary	2253 (1041)	[155, 7264]
Procedure Notes	339 (314)	[62, 3047]

Table 6: Note token length summary statistics.

MultiNLI (Williams et al., 2018) The Multi-Genre Natural Language Inference corpus is a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information. The labels include *entailment*, *neutral* and *contradiction*.

MedNLI (Romanov and Shivade, 2018b) is a dataset annotated by healthcare professionals for the task of natural language inference (NLI) based on patient medical histories. The premise sentences in this dataset are sourced from the MIMIC-III (Johnson et al., 2016) database. They use *Past Medical History section* where annotators write alternative sentence which could be one of these *entailment*, *neutral* or a *contradiction*. The *test* split consists of 1422 such pairs along with the label.

Scitail (Khot et al., 2018) The SCITAIL dataset is created for answering school-level science questions by converting each question and its correct answer into an assertive hypothesis (H). Relevant sentences are extracted from a large text corpus to serve as premises (P). Each premise-hypothesis pair is then annotated (via crowdsourcing) as entails or neutral. There are 2,126 pairs in the test split.

SNLI (Bowman et al., 2015) The Stanford Natural Language Inference corpusis a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels entailment, contradiction, and neutral. Their test split is 10,000 sentence pairs.

Appendix B. Note Types

FactEHR includes samples from four clinical note types. We outline the types and their clinical purpose below.

Procedure Note Document medical procedures, including diagnostic and therapeutic interventions. Radiology reports, as a key example, detail findings from imaging studies like X-rays, CT scans, and MRIs, often comparing them to prior exams. Other examples include endoscopies, biopsies, and surgical procedures. These notes typically include the procedure performed, clinical indications, findings, and follow-up recommendations.

Nursing Note Written by nurses to provide a systematic assessment of a patient's condition across individual body systems (e.g., cardiovascular, neurologic) during a specific time frame. The focus of the note is primarily on the patient's current status, with less emphasis on detailed planning for future care.

Progress Note Written by physicians to summarize a patient's medical status over the preceding day and outline the care plan for the following day. They include a review of significant events and relevant diagnostic tests conducted during that period. They include key events, diagnostic tests, a current patient exam, active medical issues.

Discharge Summary Written by physicians to synthesize key medical information from a patient's hospitalization, including clinical notes, diagnostic reports, and treatment plans. They provide a concise overview of the patient's presentation, past medical history, key findings, future medical plans, and discharge medications for subsequent healthcare providers.

Appendix C. Choice of Models

While we acknowledge the potential value of domain-specific models, numerous studies have shown that general-domain models often outperform medically pretrained(Jeong et al., 2024; Van Veen et al., 2024; Ceballos-Arroyo et al., 2024), or fine-tuned models across a range of tasks such classification, information extraction, and summarization, including our initial experiments.

Note Type	**MedLM**	**Gemini-1.5**
Discharge summary	58.5	64
Nursing note	77.4	79.8
Progress note	74.8	77.5

Table 7: Comparison of scores of MedLM and Gemini. We report Fact-F1 scores. **Gemini-1.5 consistently outperformed MedLM** in our evaluations.

Appendix D. Hyperparameters

For the fact decomposition task, we evaluated four models. GPT-4o, Gemini Flash 1.5, and LLaMA 3 8B were configured with a maximum token generation length of 4000 tokens, temperature of 0.01 and a top-p value of 0.9. Additionally, we included the o1-mini model, which used a temperature of 1, a top-p value of 1, and a maximum of 16,000 new tokens to allow for extra tokens generated in its chain of thought reasoning. At the time of our experiments o1-mini did not support other temperature or top-p settings. For the entailment task, we evaluated five models: GPT-4o, Gemini Flash 1.5, LLaMA 3 8B, LLaMA 3 70B, and GPT-4o-mini. These models were configured with a maximum token generation length of 25 tokens, with the same generation parameters as the fact decomposition task: a temperature of 0.01 and a top-p value of 0.9.

Appendix E. Entailment Annotation Guidelines

The following instructions were provided to clinical annotators for evaluating entailment of premise-hypothesis pairs.

Annotation Instructions

1. You have two sheets: one containing the **reference texts** (labeled as 1 and 2) and one for **annotating claims**.

2. Start in the **annotation** sheet.

3. For each row, review the text in the `claim` column and determine if all the information in it can be **completely** inferred from the reference text specified in the `reference_ID` column.

4. Annotate as follows:

   - **1**: If all of the information in the `claim` column can be **completely** inferred from the reference text.
   - **0**: If any part of the `claim` contains information that is not present in the reference text.

5. If necessary, open the reference text sheet and annotation sheet in two separate windows for ease of comparison.

Examples

Example of a claim that can be completely inferred (annotation = 1): Reference text: Left pleural effusion increased from prior scan. **Claim:** There is a pleural effusion on the left side. **Explanation:** The claim is completely inferable from the reference text.

Example of a claim that can be completely inferred (annotation = 1): Reference text: FINDINGS: Nearly complete opacification of the left hemithorax is of increasing density since the recent prior CT. **Claim:** The results showed nearly complete opacification of the left hemithorax, which has increased in density since the last CT scan. **Explanation:** The claim is completely inferable from the reference text.

Example of a claim that cannot be completely inferred (annotation = 0): Reference text: Left pleural effusion increased from prior scan. **Claim:** The CT scan shows that the pleural effusion is increased from the prior scan. **Explanation:** There is no mention of a CT scan in the reference text.

Example of a claim that cannot be completely inferred (annotation = 0): Reference text: The patient has been evaluated for potential consolidation or pneumothorax. Multiple prior chest radiographs and a recent Chest CT were referred for comparison. **Claim:** Evaluate for evidence of consolidation or pneumothorax. COMPARISON: Multiple prior chest radiographs most recent on 1/15/22. **Explanation:** The reference text does not mention that the most recent scan was on 1/15/22.

Example of a claim that cannot be completely inferred (annotation = 0): Reference text: 1. The patient is intubated and has hypoxic respiratory failure. 2. A comparison is made with a prior study. 3. The cardiac size is normal. 4. The lines and tubes are in the standard position. 5. There are large right and moderate left pleural effusions that are grossly unchanged. 6. The right upper lobe opacity has improved consistent with improving atelectasis. 7. The pleural effusions are associated with atelectasis and are larger on the right side. 8. There is mild vascular congestion. Multiple prior chest radiographs and a recent Chest CT were referred for comparison. **Claim:** FINAL REPORT SINGLE FRONTAL VIEW OF THE CHEST REASON FOR EXAM: Intubated patient, hypoxic respiratory failure. **Explanation:** The claim contains several pieces of information that are not inferable from the reference text: 1) The fact that it is a final report, 2) It is a single frontal view, 3) The reason for the exam is hypoxic respiratory failure.

Appendix F. Qualitative Annotation Guidelines

Expert reviewers assessed overall completeness of fact decompositions and the correctness, independence, and atomicity of individual facts.

Completeness: The set of fact decompositions are defined complete if the facts capture all information present in the source document.

1. : **1**: If all of the information in the `note` is covered in the **decomposed fact list**

2. : **0** if any part of the `note` is missed by decomposed facts.

3. Example: `Note: "No evidence of pneumothorax.`
`Left pleural effusion increased from prior scan."`
`Facts: "No pneumothorax. There is a pleural effusion."`
Missing: `"Pleural effusion is increased from last scan."`
Score: 0

For individual facts, reviewers judge on three criterion:

Atomicity: A fact is atomic if it represented a minimal unit of information which cannot be decomposed further.

1. **1**: The sentence presents exactly one fact, statement, or concept without unnecessary complexity.

2. 0: The statement includes multiple facts, making it difficult to evaluate or extract key information.

3. Example: `Note: "No evidence of pneumothorax.`
`Left pleural effusion increased from prior scan."`
`Facts: "Pleural effusion is left sided and increased from last scan."`
Score: 0

Correctness: A fact is correct if it accurately reflects the contents of the source.

1. **1**: The fact presented accurately reflects the information in the note.

2. **0**: The statement includes multiple facts, making it difficult to evaluate or extract key information.

3. Example: `Note: "No evidence of pneumothorax.`
`Left pleural effusion increased from prior scan."`
`Facts: "There is a large pneumothorax."`
Score: 0

Independence: A fact is independent if it can be understood on its own.

1. **1**: The fact presented accurately reflects the information in the note.
2. **0**: The statement includes multiple facts, making it difficult to evaluate or extract key information.
3. Example: `Note: "No evidence of pneumothorax. Left pleural effusion increased from prior scan."`
   `Facts: "It is left sided."`
   Score: 0

Appendix G. Compute Environment

Experiments are performed in a HIPAA compliant APIs and local on-prem university compute environment using 4 Nvidia H100 GPUs. All compute environments supported HIPAA-compliant data protocols.

Appendix H. Sentence Tokenization

Sentence tokenization errors resulting from formatting characters and tabular data are show in Table 8.

Note Type	Junk	Partial	Examples
Discharge Summary	1.2%	3.6%	`.` , `Rel.` , `05:55` , `Trimethoprim/Sulfamethoxazole.` , `SIRS/Sepsis` , `2+` , `05/21/2018` , `5`
Progress Note	0.7%	10.4%	`In` , `Unknown;` , `jpg]` , `6.` , `101 mEq/L` , `fevers.` , `24 hours` , `28.6` , `11/23/2015`
Nursing Note	0.2%	3.7%	`Learning Preference` , `MAE's.` , `Pink/ruddy.` , `01/09/2016` , `PT.` , `NPN` , `AFSF.` , `bs:ronchi.` , `50`
Procedures	9.0%	6.0%	`___` , `_________` , `I.` , `*` , `(Over)` , `|` , `_______`

Table 8: Example sentence tokenization errors with separate *junk* and *partial* error rates. Junk sentences are formatting characters (e.g. lines or tokens denoting split sections). Partial indicates an incomplete sentence that was incorrectly split during sentence parsing.

Appendix I. Supplementary Results

I.1. Validating LLM-as-a-Judge

A key objective of this work was to evaluate the "LLM-as-a-judge" component within a fact decomposition verification framework. Despite the critical role of this initial step in fact verification pipelines, it has not been rigorously benchmarked in clinical text or across diverse clinical document types. The scale of human-annotated labels used in this study was driven by the need to validate our proposed NLI-based LLM-as-a-judge. To determine an appropriate sample size for these human-provided labels, we performed power calculations. These calculations ensured that our manually labeled dataset was sufficiently powered to accurately estimate sensitivity, specificity, and other performance metrics, using parameters derived from NLI benchmarking results on the FactEHR development set.

With these parameters, annotating 250 entailment pairs per note type was sufficient to provide high-confidence estimates (99% confidence level overall and 80% confidence level per note type). This resulted in a total of 1000 uniquely annotated entailment pairs across the four note types. Additionally, we included 200 duplicate annotations (50 per note type) to assess inter-annotator agreement, ensuring we could reliably estimate kappa with high confidence. This brought the total for this annotation tranche to 1200 entailment pairs, including both unique and duplicate annotations. Combined with the 2468 entailment pairs already annotated, our dataset now comprises approximately 3500 human-annotated entailment pairs, representing one of the largest datasets of its kind in the clinical domain.

Metric	**Performance**
Expected sensitivity	95%
Expected specificity	87%
Prevalence of entailed claims	70%

Table 9: Performance of LLM-as-a-Judge model on the 250 entailment pairs per note type, they provide high-confidence estimates.

I.2. Fact Atomicity

We plot the total number of generated facts in figure 7, it varies significantly across LLMs and note types. The red line indicates the total sentence count per note type.

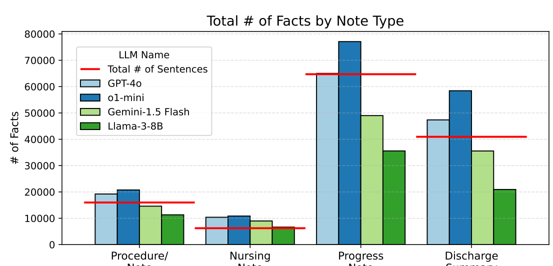

Figure 7: The total number of generated facts varies significantly across LLMs and note types. The red line indicates the total sentence count per note type.

I.3. Fact Precision and Fact Recall

In Figure 8, we compare the distributions of fact precision and unweighted fact recall across documents for each note type and fact decomposition model. Across all note types and models, fact recall exhibits a much wider distribution than fact precision, indicating a higher

Dataset	**Model**	**N**	**Unparsable**	**P**	**R***	**F1**	**Acc.**
FactEHR	GPT-4o	2468	0	90.8	95.7	**93.2**	**90.0**
	Gemini-1.5	2369	0	89.1	96.5	92.6	89.0
	Llama3-8	2468	0.36	86.2	44.7	58.8	55.2
	Llama3-70	2468	0.166	88.18	90.12	89.14	84.24
	4o-mini	2468	0	92.16	90.97	91.56	87.97
MedNLI	GPT-4o	1422	0	94.7	76.0	**84.3**	**90.6**
	Gemini-1.5	1422	0	93.8	67.5	78.5	87.7
	Llama3-8	1422	0.001	65.2	86.7	74.5	80.2
	Llama3-70	1422	0.004	80.93	87.76	84.21	89.03
	4o-mini	1422	0	91.71	74.68	82.33	89.31
SciTail	GPT-4o	2126	0	89.7	83.6	**86.5**	**89.7**
	Gemini-1.5	2120	0	79.5	88.2	83.6	86.3
	Llama3-8	2126	0.002	47.9	98.8	64.6	57.0
	Llama3-70	2126	0.0117	70.49	93.35	80.33	81.89
	4o-mini	2126	0	80	90.74	85.03	87.35
MultiNLI	GPT-4o	9832	0	94.1	82.0	87.7	91.9
	Gemini-1.5	9832	0	91.4	86.6	**88.9**	**92.4**
	Llama3-8	9832	0.008	55.5	97.3	70.7	71.6
	Llama3-70	9832	0.007	79.93	92.23	85.64	89.11
	4o-mini	9832	0	90.87	86.49	88.62	92.18
SNLI	GPT-4o	10000	0	93.1	87.0	**90.0**	**93.4**
	Gemini-1.5	10000	0	93.0	86.1	89.4	93.1
	Llama3-8	10000	0.0036	56.5	97.6	71.6	73.9
	Llama3-70	10000	0.006	74.08	94.27	82.96	86.96
	4o-mini	10000	0	91.77	86.1	88.85	92.72

Table 10: Entailment evaluation benchmarking results. Note that Gemini-1.5 flash failed to process some examples due to content moderation ("N" column). Responses whose JSON output were not parsable were coerced to "0". **Bolded** are the highest numbers (F1 and accuracy) of a model for each dataset. Our results show that GPT-4o performs well on four out of five benchmarking datasets, including the FactEHR validation set. Therefore, we selected GPT-4o as the final LLM entailment judge.

variability in the model's ability to generate comprehensive fact decompositions. This variability in recall suggests that while some documents achieve relatively high recall, many documents leave significant portions of the original information from the source notes unaccounted for missing facts in the fact decompositions.

In contrast, fact precision is generally higher and shows tighter distributions across all models and note types. This finding suggests that, for the facts that are generated, they are largely accurate and correctly entailed from the source document.

I.4. Entailment Model Validation

In table 11, we present our results of GPT-4o (our evaluator model) on the human evaluation entailment dataset.

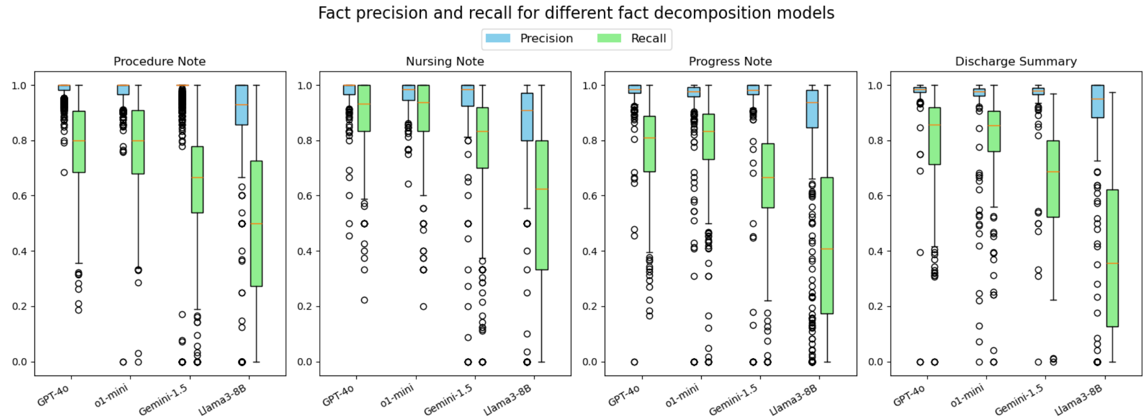

Figure 8: Comparison of fact precision and unweighted fact recall distributions across documents for each note type and fact decomposition model.

Sensitivity	Specificity	PPV	NPV	F1	Accuracy	N
96.1	64.7	86.3	87.8	90.9	86.7	1036

Table 11: Performance of GPT-4o on the 1,036 entailment pairs labeled by clinical experts.

Appendix J. Qualitative Evaluation

Expert reviewers assessed overall completeness of fact decompositions and the correctness, independence, and atomicity of individual facts. Detailed results for Llama3-8B and Gemini are shown in Figure 9 and Figure 10 respectively [5].

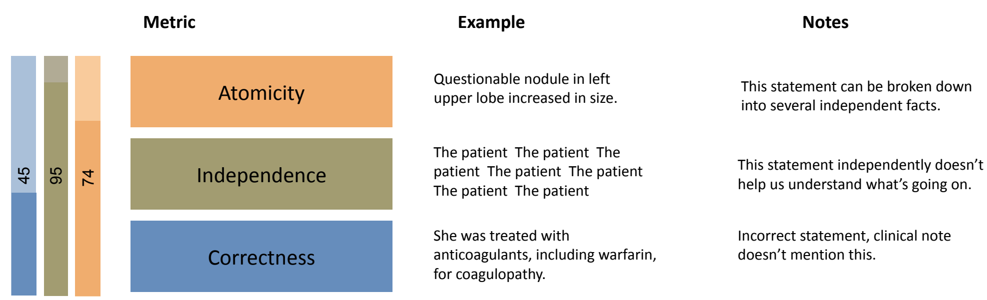

Figure 9: Overview of in-depth qualitative review with **Llama3-8B** on the fact decomposition on twenty randomly-selected examples from FactEHR. We report **percentage** of correct, independent and atomic facts as annotated by medical expert.

Appendix K. Cost

We report API costs for generating fact decomposition of our 2,168 notes and generating entailment labels for 987,266 entailment pairs in Appendix Table 12.

5. We did not have a comprehensive qualitative evaluation done for o1-mini due to time constraints.

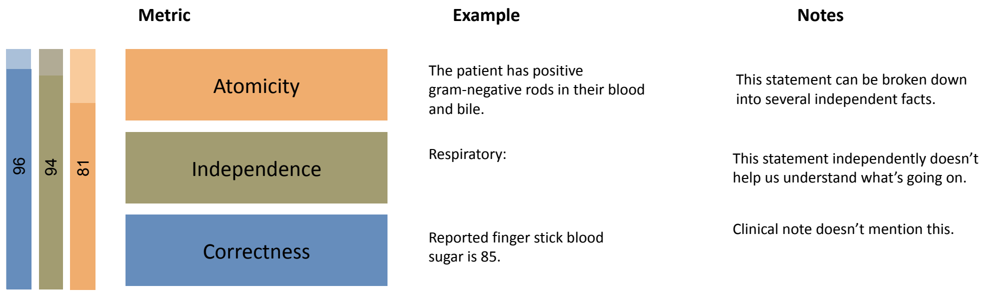

Figure 10: Overview of in-depth qualitative review with **Gemini** on the fact decomposition on twenty randomly-selected examples from FactEHR. We report **percentage** of correct, independent and atomic facts as annotated by medical expert.

Task	**Model**	**API cost (USD)**
Entailment	GPT-4o	10,162.72
Fact decomposition	GPT-4o	264.26
Fact decomposition	O1-mini	317.11
Fact decomposition	Gemini 1.5 flash	5.20

Table 12: API costs for generating fact decomposition of our 2,168 notes and generating entailment labels for 987,266 entailment pairs.

Appendix L. Prompts

We use two different prompts decribed as follows:

Fact Decomposition: We adopt the prompt from (Min et al., 2023), and include two in-context examples of fact decompositions of clinical notes written by a medical expert, illustrated in figure 11. The LLM is instructed to output independent facts as a delimited string. This approach resulted in fewer parsing errors than other methods.

Entailment Evaluation We adopt the prompt from (Xie et al., 2024), instructs the LLM to output a binary 0/1 indicator of entailment in JSON format, as illustrated in figure 12. This prompt was preferred over those requesting rationales or chain-of-thought reasoning due to its efficiency in minimizing output tokens, reducing inference times, and lowering computational costs, especially given the large number of premise-hypothesis pairs in the FactEHR dataset.

Please breakdown the following text into independent facts as a string delimited by "//" to separate the facts

Example 1 :
Note: "There is a dense consolidation in the left lower lobe."

Atomic facts:
There is a consolidation. // The consolidation is dense. // The consolidation is on the left. // The consolidation is in a lobe. // The consolidation is in the lower portion of the left lobe.

Example 2:

Note: "The patient has been having intermittent shortness of breath for the last two years."

Atomic facts:
The patient has been having shortness of breath. // The shortness of breath is intermittent. // The shortness of breath has been present for the last two years.

Do not include any other text, or say "Here is the list..."

Note: <note>

Figure 11: Prompt for fact decomposition.

You are an expert on natural language entailment.
Your task is to deduce whether premise statements entail hypotheses.
Return only '1' if the hypothesis can be fully entailed by the premise.
Return only '0' if the hypothesis contains information that cannot be entailed by the premise.
Generate the answer in JSON format with the following keys:
'entailment_prediction': 1 or 0, whether the claim can be entailed.
Only return the JSON-formatted answer and nothing else.

Premise: {premise}
Hypothesis: {hypothesis}

Here is the JSON-formatted answer:

Output:

Figure 12: Prompt for entailment evaluation with GPT-4o.