Towards understandable Generative Information Extraction

A case study on making LLMs more understandable EHR profilers.

Anonymous ACL submission

Abstract

Enhancing the understandability of Information Extraction (IE) outputs through Self-Verification (SV) can accelerate the adoption of large language models (LLMs) in critical domains such as healthcare. However, SV in Clinical IE has been largely limited to Named Entity Recognition (NER). Furthermore, recent work has shown that LLMs struggle to refine outputs without external feedback. In this work, we introduce External Verification (EV), where a separate LLM is used to directly verify and ground both NER and Relation Extraction (RE) outputs. We propose a prompt-then-tune pipeline that sequentially extracts structured outputs and supporting evidence from free-text electronic health records (EHRs). Our pipeline facilitates an LLM-ensemble, and fine-tuning for generating patient profiles. We further advocate for semantic alignment metrics over exact-match evaluation, which can overly constrain LLM expressiveness. Experiments across three EHR datasets show that our small-LLM ensemble can outperform larger standalone LLMs. We also find that smaller batch sizes improve multi-task performance, and that training on synthetic data mitigates data scarcity, narrowing, and in some cases surpassing, the performance gap with larger LLMs.

1 Introduction

Gero et al. (2023) recently demonstrated that SV, where an LLM verifies its outputs through multiple inference calls can both elicit supportive evidence for model-generated outputs and improve interpretability in Clinical IE. However, a common limitation shared with other efforts that elicit grounded-explanations is predominant evaluation and emphasis on the strengths of proprietary models, with limited exploration of capabilities of open-source LLMs (Qin et al., 2024; Vatsal and Singh, 2024). Furthermore, recent work has shown that LLMs struggle to refine outputs without external feedback (Kamoi et al., 2024). In this work, we argue that, improving understandability of Clinical IE outputs via open-source LLMs and cross-LLM verification is essential for moving forward.

While Gero et al. (2023) concentrated on interpretable grounding for NER, we propose a strategy that enriches both NER and RE outputs with contextually relevant evidence (explanations) for better interpretations. We introduce **EHR profiling**, a task that leverages LLMs to extract structured EHR profiles (characterized by entities, relations and their corresponding evidence) from real-world free-text patient EHR. Unlike prior explanation generation works, EHR profiling constrains evidence generation to the input context text itself hence ensuring traceable context-aware justifications rather than relying on the LLM's pretrained knowledge. Secondly, we argue that benchmarking IE using exact-match metrics is poorly suited for LLMs, because they generate expressive and contextually varied responses. To avoid discouraging their application in critical tasks like clinical IE, we propose evaluation metrics that assess semantic correlation or alignment to human annotations.

To tackle EHR profiling, we introduce EV, a training-free approach in which a separate LLM directly verifies and refines generated outputs (excluding intermediate feedback-generation step used in SV (Huang et al., 2023)) before a final answer is provided. We further incorporate definition augmentation (Munnangi et al., 2024), and Instruction Tuning (IT) (Zhang et al., 2023) to develop **GenJERE** (Generative Joint Entity, Relation and Evidence Extraction). **GenJERE** is a pipeline that decouples EHR profiling tasks to maximize compatibility of their outputs while reducing information loss and reasoning burden associated with batch multi-task prompting (Sanh et al., 2021). Verification is done at task-level to mitigate error propagation when reasoning across multiple tasks. The pipeline produces a collection of *instruction-ehr-profile* tuples which are subsequently used to

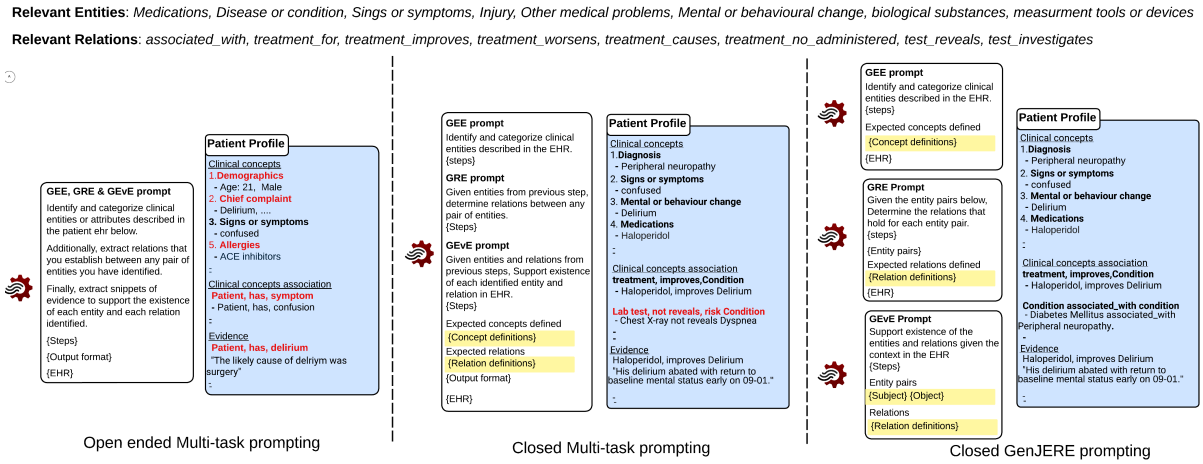

Figure 1: Prompting for multi-task IE. Closed- differs from Open-ended by explicitly specifying predefined ontologies i.e. concepts, relations and their definitions (yellow). GenJERE decomposes EHR profiling and sequentially tackles the tasks one by one. Extracted false-positives profile elements shown in red.

fine-tune individual LLMs for EHR Profiling.

Preempted by recent findings on LLM limitations in clinical IE such as sensitivity to instructions (Ceballos-Arroyo et al., 2024), **GenJERE** upgrades from an LLM to an LLM ensemble in order to leverage their collective strengths (Figure 2). Given the multi-task nature of EHR profiling, we evaluate both multi-task prompting (**MTP**), where all sub-tasks are handled in a single inference call, and multi-stage prompting (**MSP**) where sub-tasks are executed sequentially through task-specific inference calls (Figure 1).

Extensive evaluation on EHR data demonstrates **GenJERE**'s effectiveness in generating evidence-enriched outputs that exhibit stronger semantic alignment with human annotations compared to traditional multi-tasking for IE tasks. Our results show that a small-LLM ensemble augmented with EV can outperform powerful LLMs by up to 2.8% on average in IE tasks. We find that a smaller LLM fine-tuned on synthetic instruction-response pairs can narrow (and in some cases surpass) the performance gap with larger LLMs. We further demonstrate the effectiveness of semantic-alignment based metrics and hence their superiority over traditional exact-match metrics.

2 Prompting strategies

In previous works on IE, LLMs have been directly instructed to extract and classify entities and relations (or relation triplets) from text. While this approach, which we term **Open-ended MTP** due to the absence of predefined ontologies in the prompt, has been promising, it often fails to deliver satisfactory performance, particularly for relation extraction (**RE**) (Sainz et al., 2023; Han et al., 2023). Li et al. (2023b) demonstrated that Open-ended MTP is particularly ineffective for RE because it places a heavy reasoning burden on the model since it must both understand relation labels and accurately match the semantics of the relations between entities without explicit guidance.

To address this challenge when adapting LLMs to the clinical domain, we adopt **closed MTP**, where we provide predefined ontologies for both NER and RE, and instruct the LLM to ground its outputs in the input text. Our proposed GenJERE prompting is distinct from previous generative IE approaches that have used either Iterative Prompting (IP) (Li et al., 2023b; Munnangi et al., 2024) or SV (Ling et al., 2023; Gero et al., 2023), in two key ways: (i) we jointly evaluate both NER and RE, and (ii) we maximize the synergistic effects of combining multiple LLMs.

Formally, we categorize our EHR profiling IE as three separate tasks,

1. **G**enerative **E**ntity **E**xtraction (**GEE**): in which an LLM is prompted to detect and classify clinical entity mentions in an ehr.

2. **G**enerative **R**elation **E**xtraction (**GRE**): in which an LLM is prompted to classify the extracted entity pairs into relation types given the ehr document.

3. **G**enerative **Ev**idence **E**xtraction (**GEvE**): in which an LLM is prompted to extract an excerpt from an ehr as evidence for entities and relations, using either SV or EV.

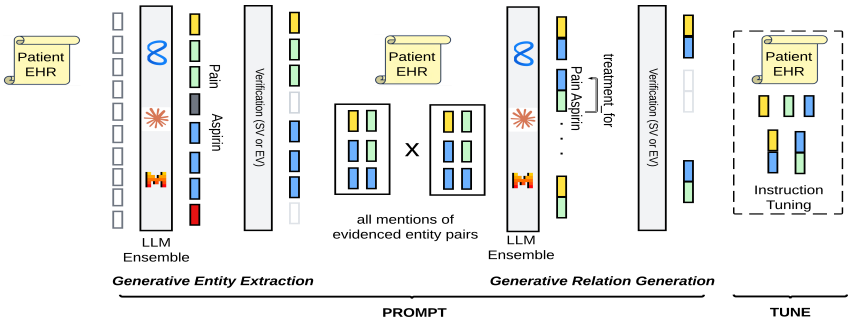

Figure 2: GenJERE prompt-then-tune pipeline.

3 GenJERE Pipeline

3.1 Overview

Figure 2 illustrates the sequence of steps in the pipeline. Given a patient's ehr, the initial stage prompts an ensemble of LLMs $(l_i, \ldots, l_n)$ to generate EHR profiles, consisting of entities and relations. This stage also incorporates task-specific SV or EV powered by a separate LLM G. The output of this stage is a JSON-formatted EHR profile, enriched with contextually relevant evidence that grounds each extracted element. The EHR profiles then serve as multi-level labels (at entity, relation and evidence levels) for training an LLM to generalize across unseen EHR notes.

3.2 Stage 1: GEE, GRE and GEvE

(a) GEE: Using prompts augmented with concept definitions[1], each l_i is instructed to extract an entity list E_i resulting into n lists. These n entity lists are then merged and processed into a single deduplicated entity list $\hat{E}$. See App. A.2 for post-processing steps.

(b) GEvE (Entity-level): We instruct each l_i (via SV) and separately with G (via EV) to verify each entity in $\hat{E}$ at a time by predicting a binary verdict "Yes" if the entity exists (in which case l_i further provides an evidence excerpt to justify the presence of e') or "No" if it does not.

(c) GRE: In this step, the grounded entity list $\hat{E}$ is enumerated to create (*subject, object*) entity pairs and use prompts augmented with relation definitions[1] to ask each l_i to predict which (if any) relations from predefined ontology are expressed by the entity pair. The resulting n relation lists are merged into a single deduplicated list of relations $\hat{R}$. Post-processing details are in App. A.2.

(d) GEvE (Relation-level): Similar to the entity-level, this step instructs each l_i (via SV) as well as G (via EV) to verify the relations in $\hat{R}$ and return the relevant evidence excerpts for the relations.

3.3 Stage 2: IT for EHR Profiling

We learn a model $f(i; ehr) \rightarrow y$ that generates an EHR profile y given an input prompt text.

Prompt structure: We feed the LLM with an instance [I; def; c; ehr; profile], formed by concatenating the task instruction (I), a schema defining the EHR profile elements i.e. clinical concepts and relations (def), Chain-of-Thought (COT) steps (c), the EHR note (ehr), and the EHR profile (profile) obtained from stage 1 in previous section. We linearize the profile into JSON sequences, because most LLMs are likely to have encountered them more than BIO or YAML (Goel et al., 2023).

3.3.1 Long context Tuning

EHR notes can be extremely long, and when combined with the IT prompt context detailed above, some instances exceed the maximum sequence lengths (8K) of some ensemble LLMs as shown in Table 5. However, we are also aware that when instruction-tuned with long context (>8K), LLMs have demonstrated not only the ability to preserve their shorter context processing capabilities, but also rival larger LLMs (Li et al., 2023a). We therefore explore IT where the pretrained context length is extended using LongLoRA[2] (Chen et al., 2023).

4 Experiments

Data: We conduct experiments using n2c2 (National NLP Clinical Challenges) which contains de-identified EHR records[3]. We specifically use IE-suitable datasets detailed below, to facilitate evaluation and include the dataset statistics in Table 9.

[1]Definitions drawn from the 2010 concept, asserts and relation extraction Challenge (Uzuner et al., 2011)

[2]LongLoRA for Long context fine-tuning

[3]Data Collected from medical facilities in the US (n2c2)

- **i2b2 2009 Medication Extraction Challenge (MEC)**, annotated for the extraction of medication regimen (medications, dosages, modes, frequencies, durations) and reasons for starting medications (Uzuner et al., 2010).

- **i2b2 2010 Relations Challenge (RC)**, annotated for extraction of (1) medical problems, tests, treatments and assertions made on medical problems; and (2) relations across the aforementioned concepts (Uzuner et al., 2011). All relations listed in App. B.

- **2018 Adverse Drug Events & Medication Extraction (ADE)**, annotated for extraction of medications and their relations to adverse events (Henry et al., 2020).

Baselines and Ensemble LLMs: For our baselines, we finetune two state of the art transformer-based biomedical LMs, BioBERT (Lee et al., 2020) and PubMedBERT (Gu et al., 2021). We formulate GEE as a token classification problem maximizing the prediction of the gold B.I.O. tagged labels. We formulate GRE as a sequence classification where we concatenate the embeddings of the [CLS] token, and the entities and maximizing the likelihood that the model assigns the highest probability to the correct relation (Gu et al., 2021). We do not use these for GEvE because of absence of gold evidence annotations limiting supervised training.

For our ensemble, we consider two open-source LLMs and a chat model i.e. Llama-3.1-8B-Instruct (L3.1-8B-Ins), Mistral-7B-Instruct (M-7B-Ins) and claude-3-haiku (Cld-3) respectively. We additionally use Llama-3.3-70B-Instruct (L70) and GPT-4 (2024-08-06) as EV llms. We only consider zero-shot in our experiments because of (1) the context is already substantively long (as discussed in 3.3.1) and we set the maximum number of tokens to generate to 2048 for MTP and 1024 for GenJERE prompting and (2) we hypothesize that the embedded definitions contain guiding signals during inference.

Inference and Fine-tuning: We deploy two workers each with 4 $\times$ A100 GPUs cluster and use a learning rate of $2e-5$ and keep the rest of the training and evaluation setting to their defaults from the pretrained versions of the models. To optimize inference, we implement Langchain's prompt templating and use VLLM's efficient batching capability for inference on large datasets.

Metrics: Exact-match (**EM**) metrics like F1 may fail to fully capture the true capabilities of LLMs, when used for IE (Han et al., 2023). Despite producing open-ended and expressive responses, LLMs can still generate semantically accurate, understandable and task-relevant outputs (Es et al., 2024). For example an LLM may extract "the pain" instead of "pain", which both indicate the same meaning, but exact-match metrics will not account for this equivalence. In this work, we advocate for semantically accurate extractions and benchmark this strategy against traditional metrics like F1 and recently proposed soft-matching score (**soft**)[4] which relies on edit-distance to identify the most similar prediction to the gold annotations (Han et al., 2023).

For GEE, we propose a **Semantic Coverage** (**SC**) score, to measure how well the generated entities semantically align with the gold annotations. For each gold entity $\mathbf{e} \in \mathbf{E}$, we identify the most similar predicted entity in set $\mathbf{P}$ using cosine similarity (cos). If the similarity exceeds a predefined threshold $\phi = 0.95$[4], the entity is considered matched, otherwise it is not. SC is computed as the fraction of gold entities matched,

$$\text{SC} = \frac{1}{|\mathbf{E}|} \sum_{\mathbf{e} \in \mathbf{E}} \mathbb{1} \left(\max_{\mathbf{p} \in \mathbf{P}} (\text{sim}(\mathbf{e}, \mathbf{p})) \geq \phi \right) \quad (1)$$

where $\mathbb{1}$ is an indicator function, $\mathbb{1}(..) = \{1, 0\}$.

For GRE, we adopt multi-aspect evaluation framework **GenRES** (Jiang et al., 2024), which emphasizes semantic similarity. We compute five metrics: **Topical Similarity Score (*TS*)**, measuring the information abundance of extracted triples compared to the source text, **Granularity Score (*GS*)**, measuring the level of detail of extracted triples, **Factualness Score (*FS*)**, quantifying the alignment between extracted triples and source text information and **Uniqueness Score (*US*)**, evaluating the diversity of the extracted triples and **Completeness Score (*CS*)**, measuring how comprehensively the extracted triples cover the source text[5].

For GEvE, we use a prompt-based evaluator to evaluate whether the extracted evidence supports existence of the extracted entity (for GEE) or relation (for GRE)[6]. Given an evidence passage, prompt an LLM to either support or refute the claim

[4] See App. D.2 for estimating thresholds details

[5] App. D.1 has equations formally used for GRE metrics.

[6] GPT3.5-Turbo-Instruct is employed as a fact-checker (FS), Granularity-checker (GS) and evidence-checker (EP).

	2010 RC							2018 ADE							2009 MEC	
	GEE	GRE					GEvE	GEE	GRE					GEvE	GEE	GEvE
	SC	*TS*	*GS*	*FS*	*US*	*CS*	*EP*	*SC*	*TS*	*GS*	*FS*	*US*	*CS*	*EP*	*SC*	*EP*
Ground Truth	100	43.0	94.3	91.5	99.6	100	_	100	41.5	94.7	92.0	99.2	100	_	100	_
BioBERT	83.4	52.2	76.4	76.1	87.8	42.6	_	84.5	54.7	67.2	73.2	88.0	49.5	_	88.9	_
PubMedBERT	82.6	52.0	74.1	75.4	89.0	41.5	_	82.9	53.8	65.1	72.8	90.6	50.0	_	87.5	_
Multi-task prompting																
Mistral 7b Instruct	84.7	39.3	48.7	58.1	83.5	28.4	48.9	84.5	29.7	44.5	61.3	88.4	36.5	51.9	82.4	64.6
Llama 3.18B Instruct	82.5	38.6	58.5	56.5	83.3	33.9	50.7	83.3	32.5	49.8	60.7	86.5	34.4	49.7	82.4	66.4
Claude 3 Haiku	83.4	39.7	64.5	47.3	88.4	37.2	59.9	89.8	41.5	57.8	64.7	89.4	39.2	53.2	86.9	69.8
Ensemble [1:2:3]	85.4	41.7	61.5	61.3	89.7	36.6	62.0	91.2	39.5	58.3	64.9	90.4	41.4	53.3	87.7	70.1
GenJERE Prompting																
Mistral 7b Instruct + SV	87.9	46.3	67.5	71.5	99.7	39.7	69.9	89.1	44.5	59.4	73.1	99.1	45.6	63.9	86.8	73.9
Llama 3.18B Instruct + SV	86.5	45.6	62.5	69.8	99.5	41.0	68.5	88.9	48.9	67.2	73.8	98.3	49.4	60.5	85.8	74.5
Claude 3 Haiku + SV	88.5	48.7	72.5	77.5	99.4	42.2	77.1	91.0	52.3	68.0	75.4	99.4	51.4	66.1	87.9	77.9
Mistral 7b Instruct + EV	88.4	49.1	70.5	74.2	99.7	44.0	72.8	91.1	50.2	61.3	77.2	99.5	49.1	66.1	88.7	74.5
Llama 3.18B Instruct + EV	88.1	47.4	69.1	72.2	99.7	45.7	71.4	89.5	52.3	67.9	74.3	99.0	50.8	62.7	87.0	75.9
Claude 3 Haiku + EV	88.9	51.3	74.8	78.1	99.4	45,7	78.6	91.9	54.8	68.8	76.1	99.9	51.4	67.3	89.0	78.5
Ensemble	89.1	55.2	79.1	80.3	99.8	44.9	81.4	93.4	56.7	69.4	79.6	99.8	51.9	69.4	91.4	79.4
Ensemble + EV (L70)	90.2	58.5	81.5	81.2	**99.9**	49.8	85.9	95.2	61.3	72.4	80.8	99.9	53.5	**71.0**	92.7	**80.7**
Ensemble + EV (GPT-4)	**93.4**	**60.4**	**85.2**	**84.2**	**99.9**	**50.1**	**88.7**	**96.3**	**62.8**	**73.3**	**82.1**	**99.9**	**56.1**	70.7	**94.9**	**80.7**

Table 1: Evaluation of closed MTP and GenJERE prompting for GEE, GRE, and GEvE. "+ SV" denotes using the same model for SV, whereas "+ EV" denotes using L70 for EV. Ensemble combines (union) outputs from all models (non-verified or without SV). Best scores across metric are shown in bold.

in evidence passage respectively returning "True" or "False". We compute an **Evidence Precision Score** (**EP**) as a fraction of the supported excerpts,

$$\text{EP} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} \mathbb{1}(\text{passage supports claim}_{\text{i}}) \quad (2)$$
$$\text{i} \in \text{entities,relations}$$

where N is the total number of extracted entities and relations (combined) and $\mathbb{1}(..) = \{1, 0\}$.

4.1 Results

Supervised Biomedical LMs vs Multi-task- vs GenJERE-prompting: Despite strong multi-instruction-following capabilities, Table 1 shows that LLMs struggle with MTP for IE, particularly for GRE. Except for GEE, GenJERE variants consistently outperform their MTP counterparts by an average of 12 percentage points across GRE and GEvE, with EV yielding the largest gains. We also observe that, although they underperform relative to GenJERE, supervised SOTA baselines, BioBERT and PubMedBERT achieve superior results compared to MTP. We attribute MTP's struggles to 1) complexity of intermediate batch processing across chained tasks, 2) EHR notes require domain-expert knowledge that is also challenging for SOTA LLMs (Liu et al., 2024), and 3) long input sequences that exceed context limits in some cases (Table 5). In contrast, GenJERE uses shorter prompts and benefits from intermediate processing, as well as task-specific SV or EV and post-processing steps such as output deduplication.

External Verification (EV) impact: As discussed in section 3.2 and section 3.2, verification during GEvE refines outputs at the entity and relation levels, thereby impacting GEE and GRE performance, respectively. We observe that replacing SV with EV yields better results, outperforming SV by 0.4–2.9 points for GEE, 0.0–6.6 points for GRE, and 0.6–2.2 points for GEvE across all experiments. This performance suggests that LLMs have the ability to enhance each others outputs, especially in IE such as EHR profiling.

LLM Ensemble superiority: We observed significant performance gains from the ensemble over individual models across all metrics, highlighting the synergistic benefits of combining multiple LLMs for IE. Incorporation of EV yields further gains, with GPT-4, attaining the best overall performance across datasets, demonstrating the effectiveness of LLM-based verification in mitigating error propagation within the GenJERE pipeline.

Ensemble vs Stand-alone Verifier: We separately investigate each of the EV llms for their

	GEE	GRE					GEvE
	SC	*TS*	*GS*	*FS*	*US*	*CS*	*EP*
Multi-task Prompting							
L70	89.3	55.9	79.4	74.5	**99.9**	44.3	79.4
GPT-4	91.2	56.8	85.1	**84.9**	**99.9**	46.9	86.0
GenJERE Prompting							
Ensemble	89.1	55.2	79.1	80.3	99.8	44.9	81.4
+EV (L70)	90.7	58.8	81.3	80.7	**99.9**	**51.9**	85.3
+EV (GPT-4)	**93.4**	**60.4**	**85.2**	84.2	**99.9**	50.1	**88.7**

Table 2: Comparing the ensemble to the superior LLMs (closed MTP) on the 2010 Relations Challenge dataset.

	2010 RC				2018 ADE				2009 MEC	
	BioBERT		Ens + EV(L70)		BioBERT		Ens + EV(L70)		BioBERT	Ens + EV(L70)
	GEE	GRE	GEE	GRE	GEE	GRE	GEE	GRE	GEE	GEE
P	72.5	44.6	71.2	51.2	79.0	47.7	80.4	51.8	84.3	89.1
R	71.2	43.1	69.4	49.4	78.8	46.5	78.7	50.0	85.9	88.4
F	71.8	43.8	70.3	50.3	78.9	47.1	79.5	50.9	85.1	88.7
soft	81.0		87.3		81.1		90.6		86.4	90.4
SC	**83.4**		**90.2**		**84.5**		**95.2**		**88.9**	**92.7**
TS		52.2		58.5		54.7		61.3		
GS		76.4		81.5		67.2		72.4		
FS		76.1		81.2		73.2		80.8		
US		**87.8**		**99.9**		**88.0**		**99.9**		
CS		42.6		49.8		49.5		53.5		

Table 3: Comparing traditional EM metrics (Precision (P), Recall (R), F1 (F1)), edit-distance based metrics (**soft**) and semantic-alignment based metrics (italics) of the best performing biomedical LM (BioBERT) and the Ensemble (Ens+EV(L70)). Empty cells denote task-metric incompatibility.

stand-alone performance and discover that, the ensemble on its own is still very competitive, and when augmented with GPT-4 performs best (Table 2). Standalone GPT-4 dominates ***FS*** and ***GS***, which we hypothesize is due to the use of a closely related base LLM (GPT-3.5-Turbo-Instruct) for evaluation, likely biasing the results toward GPT-4 generations.

4.1.1 Performance of Semantic alignment (SA) metrics against traditional EM benchmarks

In Table 3, we report evaluation results using **EM** metrics, Precision, Recall, and F1 alongside previously computed SA metrics. While SA achieves the highest overall scores, both SA and edit-distance–based (**soft**) metrics substantially outperform EM F1, with gains of up to 19.9 and 49.0 F1 points for GEE and GRE, respectively. The larger margins observed for GRE arise from the need to compare multiple phrases and preserve relation-triple structure as seen in Figure 4, which limits EM metrics' ability to account for partial or semantically equivalent matches that LLMs are particularly prone to produce.

5 Analysis and Discussion

We assess LLM consistency and interpretability via (1) semantic alignment with annotations and (2) the ratio of relevant to irrelevant extractions. Semantic alignment is measured using distance between embeddings[7] of the annotations and extractions. For GEE, document-level average embeddings are compared using L2 distance, while GRE computes relation triple embeddings via element-wise addition of subject, object, and predicate embeddings. To assess relevance, we perform an open-ended MTP experiment excluding target concepts and relations (Figure 1). Ens-L70 and Ens-GPT-4 refer to ensembles with corresponding verifiers (EV).

Semantic alignment to annotations As shown in Figure 3, GEE distances are generally shorter than GRE, suggesting a stronger semantic alignment for entities. Notably, the ensemble consistently yields the smallest distances, suggesting that the combined outputs are more semantically faithful to human annotations than individual models and the standalone verifiers. To further contextualize the idealness of semantic alignment evaluation, Figure 4 illustrates that despite not generating exact matching spans, LLM outputs are semantically relevant and interpretable. Even without exact span matches (scores=1.0), partial matches matches between ground truth and predictions receive meaningful scores. For example, "a desaturation" vs. "desaturation" scores 0.9 in GEE, and "PO predis-

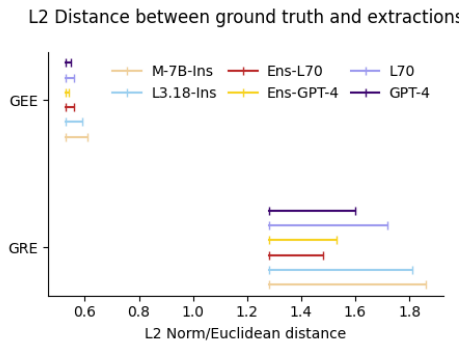

Figure 3: Distance between overall L2 norms of the embeddings of the annotations (left) and extractions (right). Larger distances depict lower semantic alignment.

[7] Using OpenAI's text-embedding-3-small.

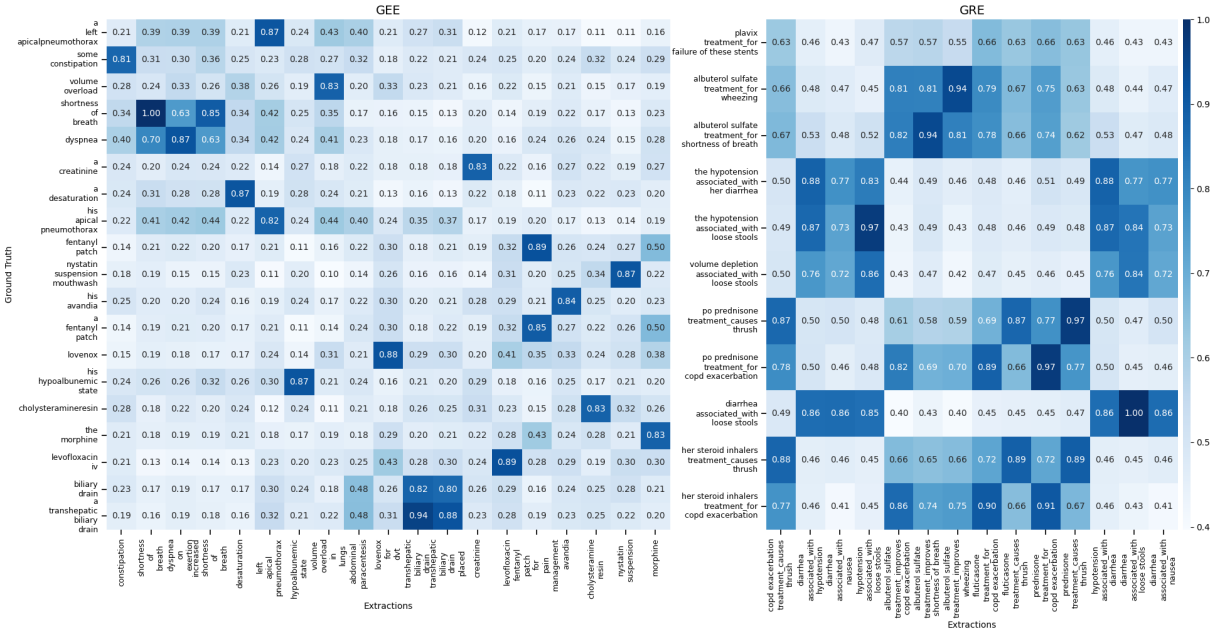

Figure 4: Analysing semantic similarity between ground truth and extractions for the best ensemble model using cos scores. Color intensity reflects strong semantic alignment between ground truth and predictions (App. E).

one treatment_causes thrush" vs "predisone treatment_causes thrush" scores 0.97 in GRE (as PO denotes oral administration). Under traditional P,R,F metrics, both cases would score 0.

Relevant vs Irrelevant Extractions: We analyze errors made by LLMs in EHR profiling, focusing on the generation of irrelevant concepts and relations. As shown in Figure 5, constraining LLMs with a predefined ontology i.e., specifying target concepts and relations in Closed MTP and Closed GenJERE substantially reduces the generation of irrelevant items and increases the relevance of extracted outputs. This observation aligns with Webson and Pavlick (2022), who highlight the importance of explicitly specifying target or expected terms in prompts to mitigate misleading prompt

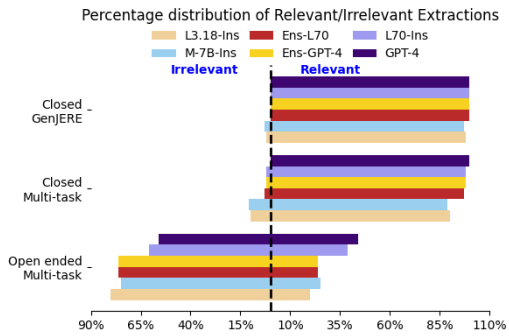

Figure 5: Percentage distribution of the relevant and irrelevant features (entity classifications combined with relation classifications). Full list of Irrelevant and relevant concepts and relations are included in the App. B.

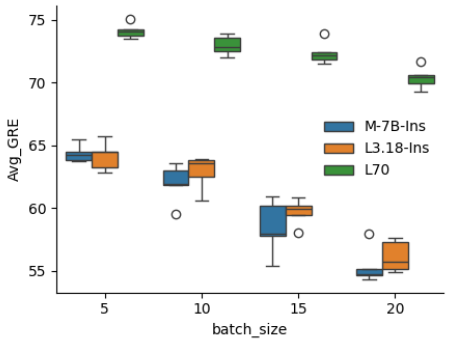

Figure 6: Investigating the average GRE scores over 5 runs for different batch sizes.

semantics. In contrast, omitting such constraints allows greater flexibility that enables LLMs to generalize across datasets and perform well in open IE settings (Jiang et al., 2024).

Varying In-context batch prompting: To assess the impact of intermediate batch prompting, we use entity pairs from closed MTP GEE outputs and query non-proprietary LLMs via closed MTP with batch sizes of 5–20 for GRE and GEvE. We report average performance across five GRE metrics. We observe performance declining as batch size increases (Figure 6), yet all models outperform their original MTP results (Table 1), supporting our hypothesis that LLMs struggle with intermediate batch processing in single-turn multi-task settings.

5.1 Ablation

We then probe the impact of the SV and EV by eliminating the GEvE task hence the verification

	GEE	GRE					GEvE
	SC	***TS***	***GS***	***FS***	***US***	***CS***	***EP***
L70	87.9	53.5	74.7	77.3	99.5	50.3	68.1
GPT-4	91.2	57.7	**80.4**	**80.9**	**99.9**	51.8	**72.6**
L3.1-8B-Ins^{+}	92.0	58.1	75.0	77.7	99.5	53.6	66.8
L3.1-8B-Ins^{++}	**93.4**	**60.0**	78.4	79.5	99.2	**57.0**	69.4

Table 4: Evaluation performance on the synthetically generated instruction-ehr-profile triples on the 2010_Relations_Challenge dataset. "$^{+}$" is standard LoRA tuning and "$^{++}$" is LongLoRA tuning.

from the GenJERE pipeline. As shown in Figure 7, we notice a drop in performance across all the individual models as well as the ensemble from original SV and EV. In particular, we notice a bigger drop in performance for EV, especially for GRE, suggesting that relation extraction would benefit most from SV and EV.

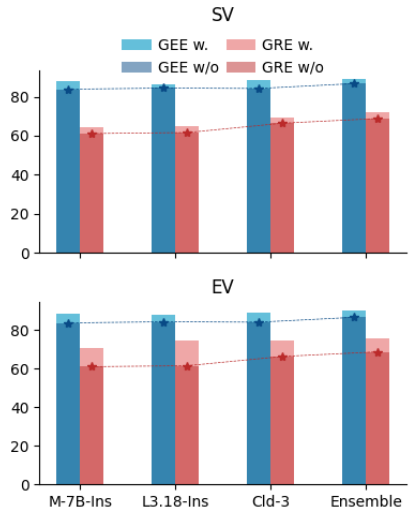

Figure 7: Impact of SV and EV. We plot SC (GEE), Average GRE (GRE), and EP (GEvE). "w" indicates the use of SV/EV, whereas "w/o" indicates their absence.

6 GenJERE Tuning

To complete the final stage of GenJERE, we fine-tune Llama-3.8B-Ins using the *instruction–EHR–profile* tuples generated in Stage 1 (3.3, App. 13), specifically using the Ens-L70. We use the provided train splits for generating the synthetic tuples and subsequently use them for fine-tuning and evaluate on the test split. We first adapt the model using LoRA and later using LongLoRA for long context tuning.

Results: Despite finetuning on a large number of inputs with a relatively larger context window (Table 5), Llama-3.8B-Ins performs competitively with standalone LLM-verifiers, surpassing them on GEE and on the ***CS*** metric for GRE. Applying LongLoRA further improves performance, highlighting the importance of long-context tuning for EHR-scale inputs. Although standalone verifiers achieve higher scores on some GRE metrics (***GS***, ***FS***, and ***US***), the LongLoRA-tuned 8B model attains the best overall performance, exceeding GPT-4 by 2.8 summed points across all experiments.

7 Related work

Efforts to enhance interpretability of LLM predictions in healthcare have mainly concentrated on QA and Text Summarization, inadvertently neglecting other essential tasks such as Clinical IE (Vatsal and Singh, 2024; Zhu et al., 2024; Qin et al., 2024). COT prompting has been predominantly adopted for eliciting grounded explanations for LLM outputs to enhance interpretability of critical healthcare aspects like mental health (Qin et al., 2024; Yang et al., 2023). Our work mostly aligns with Qin et al. (2024) as they also leverage IP and compute a weighted average of results from iterations and Chen et al. (2024), because they finetune an LLM for explanation generation. Besides redirecting attention to IE, our work distinguishes itself in three ways, 1) we consider enhancing understandability when handling multiple tasks and 2) we interrogate the capabilities of smaller LLMs that have largely been underexplored in this regard and 3) we evaluate on real world EHR data rather than social media data in order to focus on more objective opinions and evidence-based decisions.

8 Conclusion

We proposed GenJERE, a prompt-then-tune pipeline that leverages a small-LLM ensemble and EV to improve performance and understandability of Clinical IE. We empirically demonstrate that ensembles of smaller open-source LLMs can outperform larger standalone models in generating semantically faithful and relevant outputs. We compared single-turn MTP and MSP (GenJERE) prompting and observed that LLMs struggle with simultaneous multi-task inference for IE on EHR data, due to the complexity of intermediate processing and information transfers across sub-tasks. Our experiments show that smaller batch sizes can mitigate this challenge. Furthermore, we demonstrate that fine-tuning a small LLM on synthetic data improved performance and outperformed stronger LLMs in Clinical IE tasks.

Limitations

Using a separate LLM to verify LLM outputs (or EV) is intuitive and, as demonstrated in this work, can outperform SV. However, the task-specific EV strategy as deployed in this work can introduce computational overhead, as multiple inference calls (at task level) are required producing a final output. Even though the ensemble verified with Llama-3.3-70B-Instruct performs competitively, it's performance is still slightly inferior to the ensemble verified with GPT-4, suggesting that GPT-4-based verification could further improve performance albeit at substantially higher cost, particularly at scale. Having said this, we demonstrated that the open-source LLM ensemble verified by Llama-3.3-70B-Instruct outperforms stand-alone GPT-4 for the clinical IE tasks considered.

Clinical IE is language agnostic and applicable to clinical notes regardless of the language. However, our empirical evaluation is limited to clinical notes prepared in English, a morphologically limited language.

We do not assess the quality of the extracted evidence passages accompanying extracted concepts and relations, mainly because, there was no corresponding ground truth annotations.

Adopting LLM-based evaluators (like we do on some GRE and GEvE metrics) can minimize reliance on expertly curated data, however, it always raises concerns about potential biases favoring LLM generated text over human annotations. This concern has been well documented in the NLP community. Furthermore, although semantic alignment offers a more accurate reflection of LLM capabilities as we extensively demonstrate, there is a need for a well thought through trade-off between exact-match and semantic alignment in order to establish an adequate and robust evaluation system for generative IE. Our work will trigger future endeavors in searching for metrics to effectively assess LLM outputs (at any scale) for clinical utility.

Ethics Statement

The EHR-derived datasets we use in our evaluation experiments were de-identified by the owners and publishers, and can confirm that we used the publicly available versions of these datasets as reported in the paper. We also do not annotate or release any datasets as part of this work. Prior work has reported that LLMs encode various types of social biases such as gender and race related biases (Kaneko and Bollegala, 2022). We do not evaluate how such biases may influence or be amplified by the prompting methods proposed in this paper. We therefore strongly believe it is important to conduct such a social bias analysis before the methods proposed in this paper are deployed in systems that are interacted by human users.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered Mcinerney, Byron C Wallace, and Silvio Amir. 2024. Open (clinical) llms are sensitive to instruction phrasings. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 50–71.

Yanda Chen, Chandan Singh, Xiaodong Liu, Simiao Zuo, Bin Yu, He He, and Jianfeng Gao. 2024. Towards consistent natural-language explanations via explanation-consistency finetuning. *arXiv preprint arXiv:2401.13986*.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6894–6910.

Zelalem Gero, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, and Hoifung Poon. 2023. Self-verification improves few-shot clinical information extraction. *arXiv preprint arXiv:2306.00024*.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*, page 48.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Anna Huang et al. 2008. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Pengcheng Jiang, Jiacheng Lin, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2024. Genres: Rethinking evaluation for generative relation extraction in the era of large language models. *arXiv preprint arXiv:2402.10744*.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask–evaluating social biases in masked language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11954–11962.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023a. How long can context length of open-source llms truly promise? In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023b. Revisiting large language models as zero-shot relation extractors. *arXiv preprint arXiv:2310.05028*.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Yanchi Liu, Wei Cheng, Haoyu Wang, Zhengzhang Chen, Takao Osaki, Katsushi Matsuda, Haifeng Chen, et al. 2023. Improving open information extraction with large language models: A study on demonstration uncertainty. *arXiv preprint arXiv:2309.03433*.

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. 2024. Large language models are poor clinical decision-makers: A comprehensive benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13696–13710, Miami, Florida, USA. Association for Computational Linguistics.

Monica Munnangi, Sergey Feldman, Byron C Wallace, Silvio Amir, Tom Hope, and Aakanksha Naik. 2024. On-the-fly definition augmentation of llms for biomedical ner. *arXiv preprint arXiv:2404.00152*.

Jeremy Qin, Bang Liu, and Quoc Dinh Nguyen. 2024. Enhancing healthcare llm trust with atypical presentations recalibration. *arXiv preprint arXiv:2409.03225*.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction. *arXiv preprint arXiv:2310.03668*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Shubham Vatsal and Ayush Singh. 2024. Can gpt redefine medical understanding? evaluating gpt on biomedical machine reading comprehension. *arXiv preprint arXiv:2405.18682*.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Jingwei Zhu, Ancheng Xu, Minghuan Tan, and Min Yang. 2024. Xinhai@ clpsych 2024 shared task: Prompting healthcare-oriented llms for evidence highlighting in posts with suicide risk. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 238–246.

Appendices

A GenJERE pipeline

A.1 Prompts

For the GenJERE setup, we design task specific prompts for GEE, GRE and GEvE and for the MTP setup, we enclose the task specific instruction in a single prompt (Figure 11). As seen in each of task-specific prompts in Figure 8, Figure 9 and Figure 10, we embedd definitions of the target concepts and relations denoted by "{Entity definitions Schema}", and "{Relations definitions Schema}" respectively.

A.2 Post-processing

As earlier indicated in section 4, LLMs are typically elaborative and expressive, and they quite often return erroneous responses especially when forced to follow specific structures or formats. We focus on both the syntactic and semantic errors made during post-processing in order to parse their outputs into the desired format as they are being transferred from one sub-task to another as shown in 2. For the syntactic parsing, three steps are followed, (1) using langchains inbuilt output parsers "OutputFixingParser" (line 8) which can pass the misformatted output, along with the formatted instructions, to the model and ask it to fix it, (2) Use the PydanticParser (line 9) which follows the defined schema and extracts only specified objects (concepts or relations) and (3) searches and removes unwanted patterns (lines 12-21) in outputs and finally retain a unique list of elements, For the semantic parsing, we initialise a transformer model microsoft/deberta-xlarge-mnli via sentence transformers, use it to

GRE Prompt

"""
"You're an expert medical knowledge assistant capable of processing medical records. Given a schema with various relations that can occur between clinical entities"

{Relations Schema}

"Your task is to determine if there are any relations from the given schema that exist between the specified subject_entity and object_entity in the context of the given patient EHR record (EHR)."

{patient_ehr_note}

Subject: {subject_entity}
Object: {object_entity}

"Note that some entity pairs may have no relations existing between them."
"Output a list of the determined or recognised relations between the entity pair (e.g., ["relation1", "relation2"]); otherwise, an empty list [] if none is identified."
"Do not provide any extra details as part of your response, such as code snippets, the prompt, or the relation definitions included in the given schema."
"Respond using the JSON output format, wrapping the entire answer in ```json and ``` tags."
"""

Figure 9: GRE prompt

GEE Prompt

"""
"You're an expert medical knowledge assistant capable of processing medical records. Given a patient electronic health record (ehr) below,"

{patient_ehr_note}

"Your task is to identify and categorize clinical entities or attributes described in the patient ehr above. The clinical categories expected are described in the Schema given below:

{Entity defintions Schema}

"Return 'None' for any clinical category that has no identifiable information"
"Think step by step, keep your answers precise and concise"
"Do not repeat or regenerate your answers"
"Do not provide extra details such as descriptions of the identified entities, details to support your answers, functions or algorithms used in the generation or processing"
"Respond using the JSON output format, wrapping the entire answer in ```json and ``` tags"
"""

Figure 8: GEE Prompt

	>8k
Llama Tokenizer	51
Mistral Tokenizer	93

Table 5: Number of instances in 2010 Relations Challenge dataset whose sequence length is above the trained context window (8192 tokens).

GEvE Prompt

"""
"You'll be given a patient's eelectronic health record (EHR), a clinical entity, its corresponding classification and a Schema with descriptions of various clinical entity classifications"
"Please follow these steps carefully:"

"1. Check if the clinical entity is mentioned in the EHR."
"2. Determine if the context in the EHR supports the given classification for the clinical entity based on the classification's description in the Schema provided."
"3. Respond as follows:
  '[[Yes]]' if the clinical entity is mentioned in the EHR and the classification is correct or,
  '[[No]' if the clinical entity is not mentioned or the classification is incorrect."
"4. Provide a relevant sentence from the EHR as evidence to support your answer."
"5. If your answer is '[[No]]', search through the Schema for an appropriate classification. If a sutiable classification is found, include it in your response, otherwise, leave the correct classification blank."
"6. Output your answer using JSON format, wrapping the entire answer in ```json and ``` tags such as below
 {
  {'verdict': '[[Yes]]',
  'evidence': 'Relevant sentence from the EHR supporting the verdict, Yes'},
  {'verdict': '[[No]]',
  'evidence': 'Relevant sentence from the EHR supporting the verdict, No',
  'correct classification': 'Appropriate classification if found'}
 }

"7. Do not provide extra details such as descriptions of the identified correct relations, functions or algorithms used in the generation or processing."

{patient_ehr_note}

{Schema}"
"""

Figure 10: GEvE, Entity denoiser

GEE Prompt

"""
"You're an expert medical knowledge assistant capable of processing medical records. Given a patient electronic health record (ehr) below,"

{patient_ehr_note}

"Your task is to identify and categorize clinical entities or attributes described in the patient ehr above. The clinical categories expected are described in the Schema given below"

{Entity_definitions_Schema}

"Return 'None' for any clinical category that has no identifiable information"

GRE Prompt

"Given entities from previous steps, determine if there are any relations from the given schema that exist between any any pair of entities in context of the EHR. Enumerate the entity list checking every possible pair of entities and determine if any of the relations in schema below is expressed between them.

{Relations_definitions_Schema}

"Note that some entity pairs may have no relations existing between them."

GEvE Prompt

Given entities and relations from previous steps, Extract a textual snippet or evidence passage from the EHR to justify or support their presence in context of the EHR.

"Output your answer using JSON format, wrapping the entire answer in ```json and ``` tags such as below

{Output_format}

"Do not provide extra details such as descriptions of the identified correct relations, functions or algorithms used in the generation or processing."
"""

Figure 11: Closed MTP prompt

	Multi-task Prompting		GenJERE Prompting	
	Avg. Len		Avg. Len	
	Instruction	Context	Instruction	Context
2010 Relations Challenge	365	2391.8	121.9	1377.5
2010 ADE	348	3612.7	124.2	1398.2
2009 Medical Challenge	119	2019.5	98.5	1289.9

Table 6: Average length (measured in number of words) of Instructions (task query + CoT steps) and Context (EHR + definitions) for Multi-task and GenJERE Prompting. Multi-task Instructions combine GEE-Ins, GRE-Ins, and GEv-Ins, whereas GenJERE Instructions are calculated as (GEE-Ins + GRE-Ins + GEv-Ins) / 3

compute pair-wise similarity across the entity list and then eliminate one of each pair when their similarity exceed a threshold of 0.99 (lines 25 - 34). After these two stages we retain deduplicated list of entities. A similar process is followed for relations.

B Relevant and Irrelevant Entity and Relations Types

We performed qualitative analysis of the LLM's extractions during the comparison between MTP and GenJERE PTP. We observed that, In Open-ended MTP, where the target concept classifications and relations are not specified and to a lesser extent, in closed MTP, the LLMs often classified the ex-

Algorithm 1 Semantic Coverage

Input: Entity-level gold annotations $\mathbf{E}$, a predicted entity $\mathbf{p}$, $similarity = 0$,
Output: 1 or 0.
for $\mathbf{e}$ in $\mathbf{E}$ **do**
  $s \leftarrow \cos(\mathbf{e}, \mathbf{p})$
  if $s > similarity$ **then**
    $similarity \leftarrow s$
  end if
end for
if $similarity > \phi$ **then**
  return 1
else
  return 0
end if

tracted entities and relations with arbitrary concept and relation classification/types as shown in Figure 5. The list of these irrelevant classification generated is provided in Table 7.

C Context Length

To further contextualize the length of the context in the prompts, and thereby assess the need for extending the pretrained context length during fine-tuning was necessary, we calculated the average prompt length measured in words (Table 6) and tokens (Table 8). As seen in Table 8, while many instances fell within the trained context window, a substantial portion exceeded this window i.e. 32% and 40% based on LLama and Mistral Tokenizers respectively.

Relevant	Irrelevant
GEE	
Medication Regimen [drug, dosage, frequency, duration, route], Disease or condition, Signs or symptoms, Injury, Other medical problems, Mental or behavioural disorder, Test or procedure, Measurement tool or devices biological substances	Admission date, Discharge date, Demographics, Chief complaint, Allergies, Physical examination, Lab results, Family history, Immunization, Social history, Imaging results, Care coordination, Medical service
GRE	
associated_with, treatment_for, treatment_improves, treatment_worsens, treatment_not_administered, test_reveals, test_investigates	has_Diagnosis, has_SideEffects, performed_on, ordered_for, associated_with, riskfactor_for, measured_on

Table 7: List of relevant or expected concepts (GEE) and relations (GRE) and the irrelevant or unexpected concepts and relations.

Algorithm 2 Deduplication and output parsing

Input: LLM Results, **Output**: Deduplicated results
Initialise: Langchain's **OutputFixingParser, PydanticOutputParser**
Initialise: sim_model
Initialise: null_set = ["$none$", "$notindicated$", "not", "$notprovided$", "$null$", "$unknown$"]
Initialise: un_wanted_patterns = ['“‘json|“‘']
Initialise: dedup_results = {}
for (client, patient_profile) in results **do**
  results = OutputFixingParser(results)
  results = PydanticParser(results)
  Initialise dedup_entity_list = {}
  for element in results **do**
    for pattern in un_wanted_patterns **do**
      element = remove(pattern, element)
    end for
    if element exists **then**
      if lowerCase(element) $\notin$ null_set **then**
        closeoffunclosedquots(element)
        remove_large_spaces(element)
        if element $\notin$ dedup_entity_list **then**
          dedup_entity_list.add(element)
        end if
      end if
    end if
  end for
  Compute **elem_embeddings** = sim_model-
  .encode(dedup_entity_list)
  Compute similarity_matrix = sim_model-
  .similarity(elem_embeddings)
  for (i, sim_row) in enumerate(similarity_matrix) **do**
    for (j, sim) in enumerate(sim_row) **do**
      if $sim > 0.99$ **then**
        Print similar concepts information
      else
        deduplicated_entity_list.remove(element @ j)
      end if
    end for
  end for
  deduplicated_results[client] = deduplicated_entity_list
end for

	<8k	>8k
Llama Tokenizer	179 (78%)	51 (32%)
Mistral Tokenizer	137 (60%)	93 (40%)

Table 8: Number of instances in 2010 Relations Challenge dataset whose total length is above the trained context window (8192 tokens) using different Tokenizers.

D GRE metrics

We adopt the GENRES framework (Jiang et al., 2024) a multi-dimensional assessment in terms of topic similarity, uniqueness, granularity, factualness, and completeness of GRE results.

D.1 GENRES Formulation

Definition 1 (Source document): A source document is denoted as $\mathcal{D}$ and can be a sentence, passage or a document.

Definition 2 (Extracted Triples): A triple $\mathcal{T} = \langle s|r|o \rangle$ is a structure formatting a piece of free text into a subject s, a relation r, and an object o. The list of triples extracted from the source document $\mathcal{D}$ is denotes as $\mathcal{T}_\mathcal{D} = [\mathcal{T}_1, \mathcal{T}_2], ...]$.

Topical similarity (*TS*): ***TS*** measures the information abundance of the extracted triples $\mathcal{T}_\mathcal{D}$ relative to the source text $\mathcal{D}$. The elements of each triple are concatenated so that $\mathcal{T}_\mathcal{D}^\Delta = [\mathcal{T}_1', \mathcal{T}_2', ...] = [s_1 \oplus r_1 \oplus o_1, s_2 \oplus r_2 \oplus o_2, ...]$, and ***TS*** is computed as the *KL-divergence* of two topical distributions.

$$t(\mathcal{D}, \mathcal{T}_\mathcal{D}^\Delta) = e^{-\sum_{i=1}^{K} LDA(\mathcal{D})_i \log\left(\frac{LDA(\mathcal{D})_i}{LDA(\mathcal{T}_\mathcal{D}^\Delta)_i}\right)} \quad (3)$$

where LDA is Latent Dirichlet Allocation (Blei et al., 2003). A higher ***TS*** indicates that the extracted triples are closely aligned with the topical content in $\mathcal{D}$.

Uniqueness Score *US*: ***US*** assesses the diversity of the extracted triples $\mathcal{T}_\mathcal{D}$, ensuring that they are distinct and varied. It is computed as,

$$u(\mathcal{T}_\mathcal{D}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq 1}^{n} (\mathrm{CosSim}(\mathrm{v}_i, \mathrm{v}_j) < \phi) \quad (4)$$

where $\mathrm{CosSim}(\mathrm{v}_i, \mathrm{v}_j)$ is the cosine similarity between the embeddings of encoded triples v_i and v_j, ϕ is the similarity threshold and the normalization factor $n(n-1)$ accounts for all pairing such that $i \neq j$. A higher ***US*** indicates greater diversity among the triples, while a lower US suggests more similarity and potential redundancy.

Factualness Score *FS* ***FS*** quantifies the extent to which extracted triples, denoted as $\mathcal{T}_\mathcal{D}$, align with the information in the source text $\mathcal{D}$. It is computed as,

$$f(\mathcal{D}, \mathcal{T}_\mathcal{D}) = \frac{1}{|\mathcal{T}_\mathcal{D}|} \sum_{\mathcal{T} \in \mathcal{T}_\mathcal{D}} [\![\mathcal{T} \text{ is supported by } \mathcal{D}]\!] \quad (5)$$

where $[\![\mathcal{T} \text{ is supported by } \mathcal{D}]\!]$ is an indicator function and specifically an LLM[8] is prompted to answer the query, for which the function returns 1 if the triple is factual and 0 if it is not. A high ***FS*** signifies that a substantial portion of the extracted triples are factually consistent $\mathcal{D}$.

Granularity Score (*GS*) The Granularity Score (***GS***) evaluates the level of detail of the extracted triples $\mathcal{T}_\mathcal{D}$ from the source text $\mathcal{D}$.

$$g(\mathcal{T}_\mathcal{D}) = \frac{1}{|\mathcal{T}_\mathcal{D}|} \sum_{\mathcal{T} \in \mathcal{T}_\mathcal{D}} e^{-n_\mathcal{T}} \quad (6)$$

GS assesses the triple's potential to be split into more granular sub-triples. This can be performed by prompting an LLM[8] to evaluate if a given triple can be divided into additional, more specific triples. The number of possible splits is represented by $n_\mathcal{T}$ for each triple $\mathcal{T}$.

$e^{-n_\mathcal{T}}$ is the exponential decay function based on the number of splits $n_\mathcal{T}$, which assigns a lower score to triples that can be split into more subtriples (indicating they are too broad or general). Therefore, a higher ***GS*** indicates a higher score suggests that the triples are at an appropriate level of specificity.

	Train	**Test**
2010 Relation Challenge	341	85
2009 Medical Challenge	1125	124
2018 ADE	404	101

Table 9: Statistics of the dataset used for evaluation experiments. The Train split is used to evaluate stage 1, whose main results are reported in Table 1. Stage 1's resulting instruction-response pairs, are then used for fine-tuning and evaluation is performed on the Test set.

[8]GPT-3.5-Turbo-Instruct is employed as the fact checker

Completeness Score *CS*: ***CS*** evaluates how comprehensively the extracted triples $\mathcal{T}_\mathcal{D}$ cover information present in $\mathcal{D}$.

$$c(\mathcal{T}'_\mathcal{D}, \mathcal{T}_\mathcal{D}) = \frac{|\{\mathcal{T} \in \mathcal{T}'_\mathcal{D}\} | \exists_\mathcal{T} \in \mathcal{T}_\mathcal{D}, \text{sim}(\mathcal{T}, \mathcal{T}') \geq \phi\}|}{|\mathcal{T}'_\mathcal{D}|} \tag{7}$$

where $\text{sim}(\mathcal{T}, \mathcal{T}') = \text{CosSim}(emb(\mathcal{T}), emb(\mathcal{T}'))$, calculates the cosine similarity between the embeddings of the extracted triple and the gold standard triple. The threshold ϕ is pre-defined to determine the acceptable level of similarity for a match.

D.2 Estimating Semantic Coverage SC Thresholds

Semantic Coverage (SC) in our work ranges from 0 to 1, where 0 indicates complete semantic dissimilarity and 1 indicates complete semantic similarity. SC is based on cosine similarity. Prior work has empirically shown that semantic equivalence between sentence or document pairs is typically achievable at a cosine similarity of ≥ 80 (Gao et al., 2021; Huang et al., 2008). More recently, (Jiang et al., 2024) tuned a range of similarity thresholds (for their proposed cosine-distance based metrics) for matching LLM-generated triples to ground truth, identifying 0.95 as optimal. We therefore adopt a threshold of 0.95 for SC (ϕ) in our evaluation.

For the soft (Han et al., 2023) evaluation metric, we follow original formulation using the threshold of 0.5 to determine match.

E Semantic alignment to annotations

Figure 12 shows more examples how semantic alignment enhances the interpretation of LLM IE outputs, while also minimising the effect of errors in ground truth or annotations. In some cases where, no exact matching span was identified within the extractions, our proposed metric assigned semantically similar matches a similarity score (cos) > 0, which would have otherwise been a precision=0 and recall=0 in traditional metrics. Examples of such cases include the following (formatted as (*ground truth, llm extraction*) for GEE, ("the hypotension", "hypotension"), ("your aspirin", "aspirin") ("low dose spironolactone", "spironolactone") and for GRE ("cardiac catheterization, test_investigates, her aortic stenosis"),("cardiac catheterization, test_investigates, aortic stenosis").

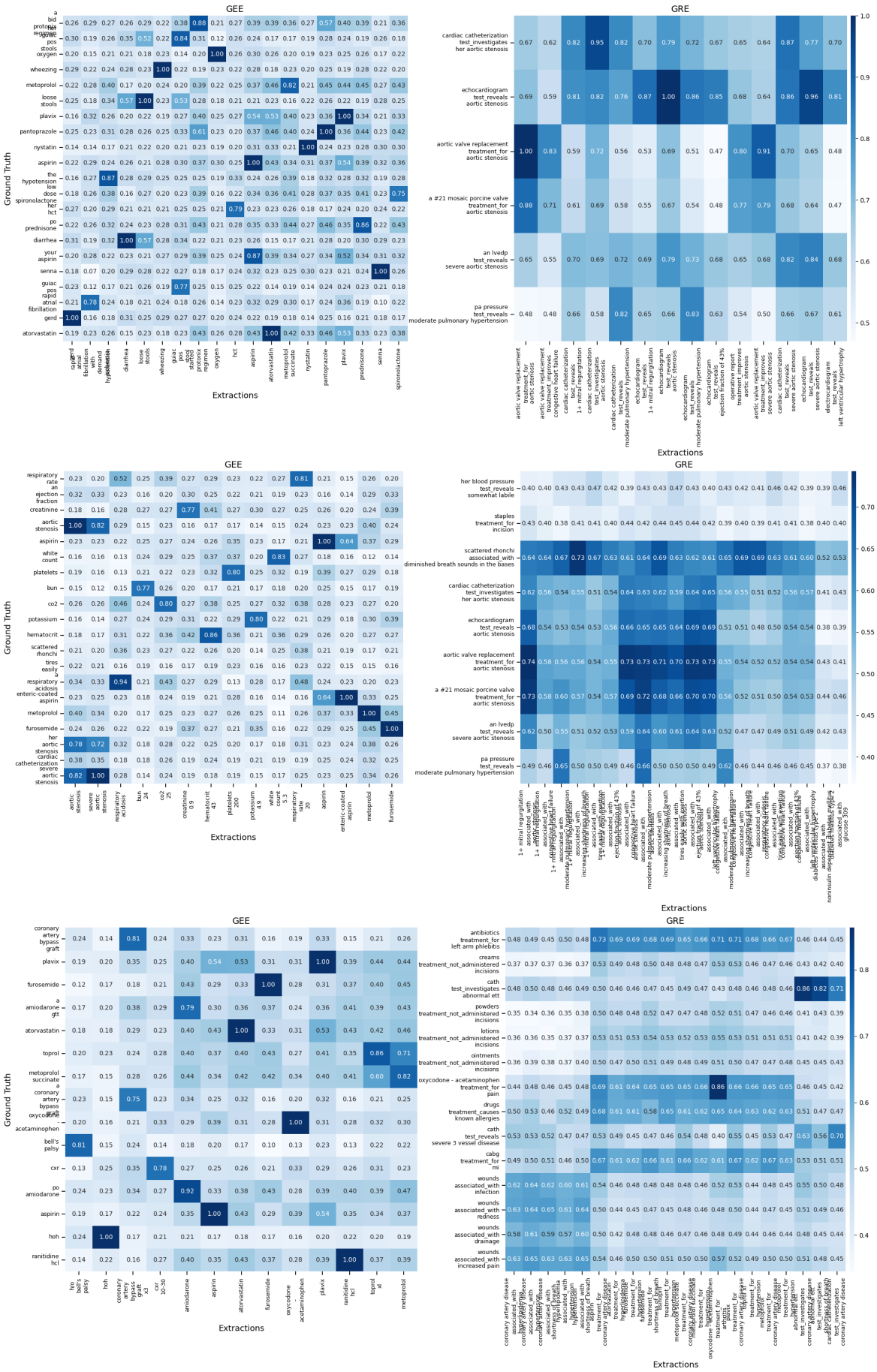

Figure 12: Analysing semantic similarity between ground truth and extractions for the best ensemble model using cos scores.

Instructions

Given a patient electronic health record (ehr) below,"
Identify and categorize clinical entities or attributes described in the patient ehr above.
In addition, extract relation triples if you establish an existing relation between an entity pair in context of the EHR.
Given entities and relations from previous steps, Extract a textual snippet or evidence passage from the EHR to justify or support their presence in context of the EHR.
Output your answer using JSON format, wrapping the entire answer in ```json and ``` tags such as below
Do not provide extra details such as descriptions of the identified correct relations, functions or algorithms used in the generation or processing."
Below are two schemas the first of target concepts and the second containing target relations

Entity_definitions_Schema
disease_or_condition: "This refers to a specific illness, disorder, or pathological state diagnosed in a patient. It is often documented alongside the diagnosis and includes information on the nature and duration of the disease. For example, Chronic Obstructive Pulmonary Disease (COPD) or Hypertension."
signs_or_symptoms: "Symptoms are subjective indications of a disease or condition experienced by the patient. They are described from the patient's perspective and can vary widely. Clinicians document symptoms as reported by the patient, such as persistent cough, fatigue, or nausea."
injury: "Refers to physical harm or damage to the body resulting from an external force, accident, trauma, or other external event. Injuries can range from minor to severe and may involve various body parts, including skin, muscles, bones, organs, and nerves. For example broken arm, thermal burn, sprain, muscle tear, Concussions, Corneal abrasions etc.)
:
Relations_definitions_Schema
associated_with: "This relation occurs when one medical problem influences or indicates another co-existing medical problem. This relation could involve a medical problems revealing aspects or even causing another medical problem. \nFor example, the sentence 'diabetes can complicate cardiovascular disease', depicts an associated_with relation between 'diabetes' and 'cardiovascular disease', 'Obesity increases load on joints accelerating wear and tear worsening osteoarthritis' depicts an associated_with relation between 'Obesity' and 'osteoarthritis'. Medical problems that can relate in this way include diseases, conditions injuries, signs and symptoms, mental disorders, behavioral disorder, abnormalities etc."
treatment_for: "This relation occurs when a treatment (e.g. drug) is administered for or given because of a medical problem or condition. For example, the sentence 'Lasix was given periodically to prevent congestive heart failure' depicts a treatment_for relation between 'Lasix' and 'heart failure', 'Insulin is administered for Type 1 Diabetes' depicts a treatment_for relation between 'Insulin' and 'Type 1 Diabetes', 'Ibuprofen is administered for fever', depicts a treatment_for relation between 'Ibuprofen' and 'fever'."
:

EHR

..... Soft, NT/ND +BS Ext : Warm, well-perfused - edema, - varicosities Neuro : A&Ox3, MAE, non-focal Pertinent Results : Echo 10-25 : PRE-BYPASS : Left ventricular wall thicknesses are normal . The left ventricular cavity size is normal . There is mild inferior wall hypokinesis . There is akinesis / dyskinesis and thinning of the mid to distal inferior septum and the apex . Overall left ventricular systolic function is mildly depressed . Right ventricular chamber size and free wall motion are normal . The descending thoracic aorta is mildly dilated . There are simple atheroma in the descending thoracic aorta . There are three aortic valve leaflets . There is no aortic valve stenosis . No aortic regurgitation is seen . The mitral valve leaflets are mildly thickened . Trivial mitral regurgitation is seen . POST-BYPASS : LV and RV function is unchanged . Aorta is unchanged . Other findings are unchanged . CXR 10-30 : Left lower lobe atelectasis has partially cleared . Upper lungs are clear . Mild postoperative widening of the cardiomediastinal silhouette is stable . No pneumothorax . 2018-10-25 11:15 AM BLOOD WBC - 18.3 # RBC - 3.42 Hgb - 10.9 Hct - 31.6 MCV - 92 MCH - 31.7 MCHC - 34.4 RDW - 13.3 Plt Ct - 134 2018-10-31 06:25 AM BLOOD WBC - 13.6 RBC - 2.72 Hgb - 8.6 Hct - 24.6 MCV - 91 MCH - 31.7 MCHC - 35.0 RDW - 14.0 Plt Ct - 314 2018-10-25 11:15 AM BLOOD PT - 13.3 PTT - 30.0 INR(PT) - 1.2 2018-10-29 06:50 AM BLOOD PT - 11.9 INR(PT) - 1.0 2018-10-25 12:36 PM BLOOD UreaN - 17 Creat - 0.7 Cl - 111 HCO3 - 23 2018-10-31 06:25 AM BLOOD Glucose - 91 UreaN - 19 Creat - 0.8 Na - 134 K - 4.0 Cl - 98 HCO3 - 26 AnGap - 14 Brief Hospital Course : Mr. Kammerer was a same day admit and on 10-25 was brought to the operating room where he underwent a coronary artery bypass graft x 3 . Please see operative report for surgical details . He tolerated the procedure well and was transferred to the CSRU for invasive monitoring in stable condition . Later on op day he was weaned from sedation, awoke neurologically intact, and extubated . Beta blockers and diuretics were initiated on post-op day one . He was diuresed towards his pre-op weight . He appeared to be doing well and was transferred to the SDU on this day . He did have burst of atrial fibrillation and was started on a Amiodarone gtt . His beta blockers were also titrated for maximal BP and HR control . Chest tubes were removed on post-op day two and epicardial paci One (1) Tablet PO DAILY (Daily). Disp : 30 Tablet (s) Refills : 0 6. Amiodarone 200 mg Tablet Sig : Two (2) Tablet PO BID (2 times a day): please take 400 mg twice a day until 11-02 then decrease to 400 mg once a day for 1 week and then decrease to 200 mg once a day . Disp : 40 Tablet (s) Refills : 0 7. Metoprolol Succinate 50 mg Tablet Sustained Release 24 HR Sig : One (1) Tablet Sustained Release 24 HR PO DAILY (Daily). Disp : 30 Tablet Sustained Release 24 HR (s) Refills : 0 8. Ferrous Gluconate 300 mg Tablet Sig : One (1) Tablet PO DAILY (Daily). Disp : 30 Tablet (s) Refills : 0 9. Ascorbic Acid 500 mg Tablet Sig : One (1) Tablet PO BID (2 times a day). Disp : 60 Tablet (s) Refills : 0 10. .....

Profile

"profile": {
"Entities":{
"disease_or_condition": ["Coronary artery disease", "Arthritis", "Carpal tunnel syndrome", "Hyperlipidemia", "Hypertension", "h/o Bell's Palsy", "Bell's Palsy"],
"signs_or_symptoms": ["Shortness of Breath"],
"other_medical_problems": ["HOH", "Hearing loss", "s/p Tonsillectomy"],
"treatments": ["Coronary Artery Bypass Graft x3"],
"test_or_procedures": ["Abnormal ETT", "Blood Tests", "CXR", "CXR 10-30", "Cardiac catheterization", "Chest X-ray", "Echo", "Echo 10-25", "Echocardiogram"],
"medication_regimen": [
{"drug_name": "Amiodarone", "dosage": "200 mg", "frequency": "BID", "duration": "11-02 then decrease to 400 mg once a day for 1 week and then decrease to 200 mg once a day", "route": "PO"},
{"drug_name": "Ascorbic Acid", "dosage": "500 mg", "frequency": "BID", "duration": "NA", "route": "PO"},
{"drug_name": "Aspirin", "dosage": "81 mg", "frequency": "DAILY", "duration": "NA", "route": "PO"},
.
.
.
"Relations": [
{"subject": "Coronary artery disease", "object": "Hyperlipidemia", "relation": "associated_with", "evidence": "Past Medical History: Carpal tunnel syndrome, Hypertension, Hyperlipidemia, Arthritis, h/o Bell's Palsy, HOH, s/p Tonsillectomy"},
{"subject": "Coronary artery disease", "object": "Hypertension", "relation": "associated_with", "evidence": "Past Medical History: Carpal tunnel syndrome, Hypertension, Hyperlipidemia, Arthritis, h/o Bell's Palsy, HOH, s/p Tonsillectomy"},
{"subject": "Coronary artery disease", "object": "Shortness of Breath", "relation": "associated_with", "evidence": "Chief Complaint: Shortness of Breath"},
{"subject": "Hyperlipidemia", "object": "Hypertension", "relation": "associated_with", "evidence": "Past Medical History: Carpal tunnel syndrome, Hypertension, Hyperlipidemia, Arthritis, h/o Bell's Palsy, HOH, s/p Tonsillectomy"},
{"subject": "Hypertension", "object": "Shortness of Breath", "relation": "associated_with", "evidence": "67 y/o male with worsening shortness of breath. Had abnormal ETT and referred for cath. Cath revealed severe 3 vessel disease."},
{"subject": "Coronary Artery Bypass Graft x3", "object": "Coronary artery disease", "relation": "treatment_for", "evidence": "The EHR states that the patient underwent a coronary artery bypass graft x3 for severe 3 vessel disease, which is a form of coronary artery disease."},
;
"Evidence":{
"[
{"entity": "Coronary artery disease", "evidence": "Discharge Diagnosis: Coronary artery disease s/p Coronary Artery Bypass Graft x3", "exact_match": true},
{"entity": "Amiodarone", "evidence": "He did have burst of atrial fibrillation and was started on a Amiodarone gtt...", "exact_match": true},
{"entity": "Ascorbic Acid", "evidence": "Ascorbic Acid 500 mg Tablet Sig: One (1) Tablet PO BID...", "exact_match": true},
{"entity": "Arthritis", "disease_or_condition", "evidence": "Past Medical History: ...Arthritis...", "corrected_from": "other_medical_problems"},
// ...
]"

Figure 13: Sample Instructions, EHR and profile