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 can improve its utility and adoption across critical sectors such as healthcare. Unlike comparable tasks like Question Answering (QA) and Summarization, IE remains largely understudied in this context. In this work, we introduce a method that incorporates evidentiality in the form of textual snippets to substantiate the extracted IE outputs (i.e. concepts and relations). We propose a prompt-then-tune pipeline that sequentially extracts IE outputs and corresponding evidence passages from unstructured electronic health records (EHRs). This pipeline supports an ensemble of large language models (LLMs), self-verification, and fine-tuning for generating patient profiles from EHR notes. Beyond evidence-based enrichment, we advocate for semantic-alignment metrics over exact-match metrics, as the latter constrain LLM expressiveness. Our evaluation on three EHR-derived datasets shows that a small-LLM ensemble outperforms stronger standalone LLMs by up to 2.4% on average across IE tasks. Additionally, we find that iterative prompting and smaller batch sizes not only reduce the complexity of intermediate batch processing but also significantly improve multi-task performance. We further demonstrate that training on synthetic data helps mitigate data scarcity, narrowing, (and in some cases surpassing) the performance gap with larger models.

1 Introduction

Numerous efforts have aimed to enhance AI's trust and transparency in healthcare (Saraswat et al., 2022; Srinivasu et al., 2022; Amann et al., 2020), however most of them precede the advent of generative AI. Recent works have primarily focused on eliciting grounded explanations for answers in healthcare QA on social media crawled datasets that often contain biased or sentimental opinions (Yang et al., 2023; Chen et al., 2024; Zhu et al., 2024). Additionally, these studies predominantly evaluate and emphasize strengths of proprietary models with limited exploration of capabilities of open-source LLMs (Qin et al., 2024; Vatsal and Singh, 2024). Beyond clinical QA, clinical IE remains relatively understudied despite its critical role in accelerating access to key artifacts considered in clinical practice. We argue that, improving the understandability of IE outputs, particularly for open-source LLMs is essential moving forward.

To this end, we are motivated to further enhance the intuitiveness and utility of IE outputs. Focusing on Named Entity Recognition (NER) and Relation Extraction (RE), we first propose a strategy that enriches these outputs with contextually relevant evidence (explanations) for better interpretation in clinical settings. Combining these tasks, 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 unstructured patient EHR records. Unlike prior explanation generation works, EHR profiling constrains evidence generation to the input context text itself rather than relying on the LLM's pretrained knowledge, which ensures traceable context-aware justifications. 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 leverage prompt augmentation (Munnangi et al., 2024), iterative prompting (IP) and Instruction Tuning (IT) (Zhang et al., 2023) to develop Generative Joint Entity, Relation and Evidence Extraction - **GenJERE**, a pipeline that decouples EHR profiling tasks to maximize compatibility of task-specific outputs while reducing information loss and reasoning burden

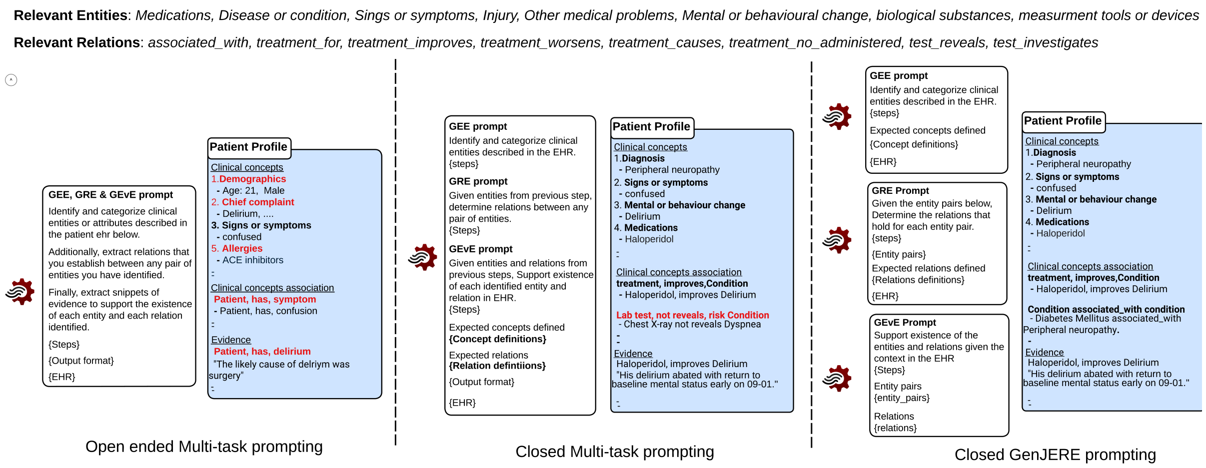

Figure 1: Prompt formulation for multi-task inference. Closed- differs from Open-ended by explicitly specifying predefined target concepts and relations and their definitions. GenJERE additionally employs iterative prompting querying an LLM with a single task on each turn. Red indicates false positive extracted profile elements.

associated with batch multi-task prompting (Sanh et al., 2021). Since EHR profiling is multi-tasking, we investigate both multi-task prompting (MTP), where multiple sub-tasks are simultaneously handled in a single inference call, and multi-stage prompting (MSP) where sub-tasks are executed across a series of inference calls (Figure 1).

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 and maximise diversity (Figure 2). To mitigate error propagation, GenJERE uses IP which lessens the burden of reasoning across multiple tasks in a single turn while enabling self-verification, and also incorporates a denoising LLM which extracts snippets of contextual evidence to enhance output interpretability. This pipeline results in a collection of *instruction-ehr-profile* tuples which are subsequently used to fine-tune individual LLMs for EHR Profiling.

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 can outperform powerful LLMs by up to 2.4% on average in IE tasks. Unlike traditional multi-tasking, GenJERE is able to consistently generate relevant and expected target profile elements. Moreover, it alleviates the complex reasoning burden typically associated with batch prompting (Cheng et al., 2023). Finally, we demonstrate that a smaller LLM fine-tuned on synthetic instruction-response pairs can narrow (and in some cases surpass) the performance gap with larger LLMs.

2 Preliminaries

We formulate EHR profiling 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 document into a predefined set of entity types.

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

3. **G**enerative **Ev**idence **E**xtraction (**GEvE**): in which an LLM is prompted to retrieve a passage from an ehr document to support existence of extracted entities and relations.

Method overview: We approach EHR profiling using a two-stage pipeline as illustrated in Figure 2. The initial stage leverages the an LLM ensemble to generate EHR profiles, which are linearized into a structured JSON format. This stage also incorporates an additional LLM (superior to any model in the ensemble) to refine the EHR profiles and enrich them with contextually relevant evidence textual snippets. The second stage treats the refined EHR profiles as labels at entity, relation and evidence levels for training an LLM to generalizE across unseen EHR notes. We apply IT inorder to enhance the multi-task learning of GEE, GRE and GEvE.

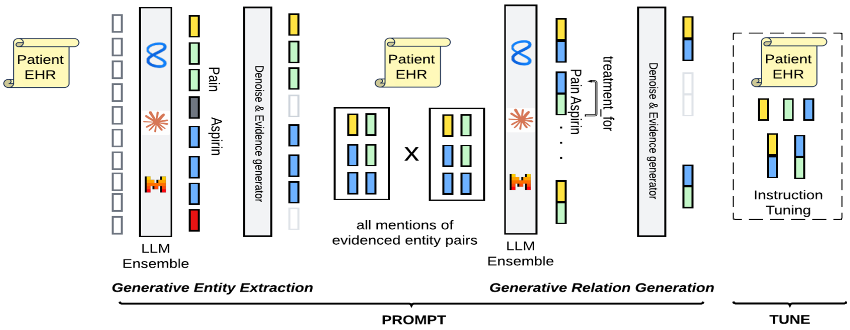

Figure 2: GenJERE prompt-then-tune pipeline.

3 GenJERE Pipeline

To implement GenJERE (Figure 2), we assume access to a dataset D of unstructured EHR notes, an ensemble of LLMs $l_{i\in\{1,n\}} \in \mathcal{L}$, where n = number of LLMs, a denoising and evidence generating LLM $\mathcal{M}$ superior to all $l_i \in \mathcal{L}$.

3.1 Stage 1: GEE, GRE and GEvE

(a) GEE: In the first component, each l_i is instructed to extract an entity list E_i (following GEE's definition) resulting into n lists. The 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 Denoiser): The next component iteratively prompts $\mathcal{M}$ with one entity $e \in \hat{E}$ at a time, instructing it to (1) predict a binary verdict (Yes or No) indicating whether the entity exists in the ehr and (2) extract an evidence snippet to justify e's presence otherwise returning "No mention of entity", The prompts with the heuristics for all sub-tasks is provided in the App. A.2.

(c) Entity Pairing: The third component enumerates the list of evidenced entities $\hat{E}$ to create entity pairs. Using a set of heuristics, this component infers the directionality in the entity pair based on the relation type in a schema of predefined relation types $\mathcal{R}$ i.e. given a relation $r \in \mathcal{R}$, it determines which entity constitutes a subject or an object. For example, for r = **treatment_for**, the subject entities will be of entity types *[Treatments, Drugs]* and the object entities are *[Disease or condition, Signs or symptoms, Injury, Other medical problems, Mental or behavioural disorder]*. See App. A.3.

(d) GRE: The fourth component reintroduces the ensemble, where each l_i is iteratively prompted to process each entity pair $(e_1, e_2) \in \mathcal{E} \times \mathcal{E}$ one at a time, predicting which (if any) relations from $\mathcal{R}$ are expressed by the entity pair. Each l_i generates a list of relation triples R_i, and the n relation lists are merged into a single deduplicated list of predicted relations $\hat{R}$. Post-processing details in App. A.2.

(e) GEvE (Relation-level Denoiser): The penultimate component reintroduces $\mathcal{M}$ to denoise $\hat{R}$ and extract evidence snippets for $r \in \hat{R}$, similar to what was done at entity level (component (b)).

3.2 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 $[I; def; c; ehr; profile]$, where ; indicates concatenation, I denotes the main task instruction, def denotes a schema of definitions of target EHR profile elements i.e. clinical concepts and relations, c implies the Chain-of-Thought (COT) steps to follow, ehr denotes the EHR notes, and $profile$ implies profiles that would have been obtained from stage 1 in previous section.

EHR profile structure: All elements of the output y are best interpreted if structurally organised rather than in an amorphous manner. We opt to linearize responses into JSON sequences, as JSON is a common format that most LLMs are likely to have encountered more frequently (e.g. in code) than BIO or YAML formats (Goel et al., 2023).

3.2.1 Long context Tuning

EHR notes can be extremely long, and combined with the IT prompt context detailed earlier, some instances exceed the maximum sequence lengths (8K tokens) of some LLMs in the ensemble (Table 1). However, we are also aware that when instruction-tuned and exposed to long context (>8K), LLMs have demonstrated not only an ability to preserve their shorter context processing capabilities, but also rival larger LLMs (Li et al., 2023). We therefore explore IT where the pretrained context length is extended using LongLoRA[1] (Chen et al., 2023).

4 Experiments

Data: We conduct experiments using n2c2 (National NLP Clinical Challenges) which contains de-identified EHR records from health facilities in the US[2]. We specifically use the datasets detailed below, because they're annotated for IE tasks, hence making them suitable for evaluating our EHR Profiling tasks.

- **i2b2 2009 Medication Extraction Challenge**, which was 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**, which was 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)**, which was annotated for extraction of medications and their relations to adverse events (Henry et al., 2020).

Models: For our ensemble, we consider 2 instruction tuned open-source LLMs and a chat model i.e. Llama-3.1-8B-Instruct, Mistral-7B-Instruct and claude-3-haiku respectively. For the denoiser, we explore Llama-3.1-405B-Instruct (L405) and GPT-4 (2024-08-06). We only consider zero-shot setting in our experiments because of (1) the context is already substantively long (as discussed in 3.2.1) 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 good signals that would guide the LLM during inference.

	>8k
Llama Tokenizer	51
Mistral Tokenizer	93

Table 1: Number of instances in 2010 Relations Challenge dataset whose sequence length is above the trained context window (8192 tokens).

4.1 Implementation

Prompting: We design task-specific prompts with plain text instructions and COT. We further augment them with definitions of the entity types (for GEE) and relation types (for GRE). Definition augmentation (DA) has enhanced instruction-only prompting in the past (Munnangi et al., 2024).

Inference and Fine-tuning: We deploy 2 workers each with 4 $\times$ A100 GPUs cluster and use a learning rate of 2×10^{-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[3] and use VLLM's[4] efficient batching capability for inference on large datasets.

Metrics: Exact-match metrics may discourage adoption of LLMs for clinical IE due to their rigid requirement of exactly matching reference annotations. However, this fails to reflect the true capabilities of LLMs, which despite producing open-ended and expressive responses, can still generate outputs that are semantically accurate, understandable and task-relevant (Es et al., 2024; Liu et al., 2023). Motivated by this, we advocate for evaluation of the semantic alignment of the outputs in this work. For the GEE, we propose a **Semantic Coverage** (**SC**) score, to evaluate how comprehensively the generated entities cover the information in the source text. For each ground truth entity $\mathbf{e} \in \mathbf{E}$, we use cosine similarity (cos) to search for the most semantically similar entity from the predicted entity set $\mathbf{P}$. If the similarity between $\mathbf{e}$ and the best matched $\mathbf{p} \in \mathbf{P}$ exceeds a predefind threshold ϕ, we consider $\mathbf{e}$ to be successfully matched, otherwise not. We set $\phi = 0.95$ and compute SC as,

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

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

For GRE, we adopt Jiang et al. (2024)'s multi-aspect evaluation framework (GenRES), which emphasizes semantic similarity. We compute a) **Top-**

[1] LongLoRA for Long context fine-tuning

[2] This data is collected from Partners Healthcare, Beth Israel Deaconess Medical Center, and the University of Pittsburgh Medical Center. About n2c2

[3] Langchain prompt templates

[4] VLLM for fast inference and serving

	2010 Relations Challenge							2018 ADE							2009 Medical Challenge	
	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	_
Multi-task prompting																
1. 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
2. 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
3. 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																
4. Mistral 7b Instruct	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
5. Llama 3.18B Instruct	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
6. Claude 3 Haiku	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
Ensemble [4:5:6]	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
+Denoiser (L405)	90.7	58.8	81.3	80.7	99.9	**51.9**	85.3	95.6	59.9	71.1	80.4	99.9	54.2	**71.1**	93.3	80.9
+Denoiser (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 2: Evaluation of closed- Multi-task and GenJERE prompting for GEE, GRE and GEvE. Ensemble [x:y:z] indicates a score of aggregated (union) outputs of the models x,y and z. With 4-6, the respective LLMs are used for denoising whereas with +Denoiser(m), m handles the denoising as described in section 3. Best scores are in bold.

ical Similarity Score (TS)**: which measures the information abundance of extracted triples compared to the source text, b) **Granularity Score (GS)**: measures the level of detail (granularity) of extracted triples from source text, c) **Factualness Score (FS)**: quantifies the extent of alignment of extracted triples with source text information[5] and d) **Uniqueness Score (US)**: assess the diversity of the extracted triples, e) **Completeness Score (CS)**: How comprehensively the extracted triples cover the information present in the source text.

For GEvE, we consider using a prompt-based evaluator to evaluate whether the extracted evidence passage supports existence of the clinical entity (for GEE) or the extracted relation (for GRE).[5] Given an evidence passage 1) prompt an LLM to either support or refute the claim in evidence passage respectively returning "True" or "False". We compute an **Evidence Precision** (**EP**) score as,

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

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

4.2 Results

We evaluate both stages of the GenJERE described in section 3. For stage 1, we evaluated the individual models, the ensemble, and the ensemble augmented with the denoiser and evidence generator (denoted as "+Denoiser") on GEE, GRE, and GEvE. The ensemble aggregates model outputs by taking the union of predictions across all models, while "+Denoiser" further refines the ensemble outputs via an LLM-based denoising mechanism described in section 3.1. Since stage 1 is mainly In Context Learning (ICL), we combine the train and test sets provided for the datasets and for the second stage, we finetune Llama-3.8B-Ins on the *instruction-ehr-profile* tuples (from 3.2, shown in App 13) using the train set and evaluate on the test sets.

4.2.1 Multi-task Vs GenJERE prompting

Despite their strong capabilities in following multiple instructions simultaneously, our experiments revealed their struggles in MTP for IE tasks, especially for GRE. With the exception of GEE, models prompted via GenJERE consistently outperform their MTP variants by a range of 4-21% across GRE and GEvE. We attribute the struggles of MTP to (1) Complexity of intermediate batch processing across chained tasks i.e. as task-specific outputs are being transferred from one task to another, (2) the nature of EHR notes demands sophisticated domain understanding which even SOTA LLMs

	GEE	GRE					GEvE
	SC	*TS*	*GS*	*FS*	*US*	*CS*	*EP*
Multi-task Prompting							
L405	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
+Denoiser (L405)	90.7	58.8	81.3	80.7	**99.9**	**51.9**	85.3
+Denoiser (GPT-4)	**93.4**	**60.4**	**85.2**	84.2	**99.9**	50.1	**88.7**

Table 3: Comparing the ensemble to the superior LLMs (closed MTP) on the 2010 Relations Challenge dataset.

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

struggle with (Liu et al., 2024), (3) the long input sequences comprising instruction, COT, EHR, and definitions which exceeds context window on some data points Table 1. In contrast, with GenJERE, models process relatively shorter prompts and additionally benefit from IP (which allows self-verification) and intermediate post processing such as deduplication of extractions.

4.2.2 LLM Ensemble superiority

We observed significant performance gains made by the ensemble over all the individual models across all metrics, thereby demonstrating the synergistic effects of combining multiple LLMS for IE. We also saw further gains with the incorporation of denoisers particularly GPT-4, which achieved most of the overall best scores across all datasets. This demonstrates the benefit of LLM-based denoising/error correction mechanism in mitigating error propagation as outputs are transferred from one sub-task to another through the GenJERE pipeline.

Ensemble vs Stand alone Denoiser: We separately investigate each of the denoisers for their stand-alone performance on the EHR profiling tasks and discover that, the ensemble on its own is still very competitive, and when augmented with a denoiser GPT-4 performs best (Table 3). Standalone GPT-4 is dominant in FS and GS which we hypothesize arises from employing a similar fabric LLM (GPT-3.5-Turbo-Instruct) in evaluation, thus likely to skew towards GPT-4 generations.

5 Analysis and Discussion

To assess the quality of the LLM's extractions in terms of their consistency and interpretability, we investigate two key aspects, (1) the semantic alignment between their outputs and the annotations, and (2) the ratio of relevant to irrelevant extractions. For (1), we compute the overall distance between the embeddings of annotations and the extractions[6]. For GEE, we compute an average embedding per document for both annotations and extractions, then compute their L2 norms across the dataset. The document-level L2 norms are then averaged across the dataset for both annotations and extractions. GRE follows a similar process, except that relation triple embedding are obtained via element-wise addition of the subject, object and predicte embeddings. For (2), we set up an additional experiment, Open ended MTP, which excludes the target concepts, relations and their definitions from the prompt (Figure 1). From this point, Ens-L405 and Ens-GPT-4 refer to the ensemble combined with denoisers, respectively.

[6] Using Openai's text-embedding-3-small to obtain their respective embeddings

5.1 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 aggregated outputs are more semantically faithful to human annotations than individual models and the standalone denoisers. To further contextualize the idealness of semantic alignment evaluation, Figure 4 illustrates that despite not generating exact matching spans, LLM extractions are semantically relevant and would be understood if manually verified.

5.2 Relevant Vs Irrelevant Extractions

Open ended MTP performs poorly, often generating significantly more irrelevant profile items than the compared methods. We attribute this to the lack of guiding context in prompt (i.e. target words). This supports findings by Webson and Pavlick (2022), which highlight the importance of specifying target/expected words in the prompt that can substantively override the misleading prompt semantics. Even with context (target concepts, relations and their corresponding definitions), some models generate slightly more irrelevant items in closed MTP compared to Closed GenJERE prompting. As earlier noted, we attribute this to complexity of intermediate batch processing when handling multiple tasks simultaneously.

Varying In-context batch prompting: To analyse the impact of intermediate batch prompting or

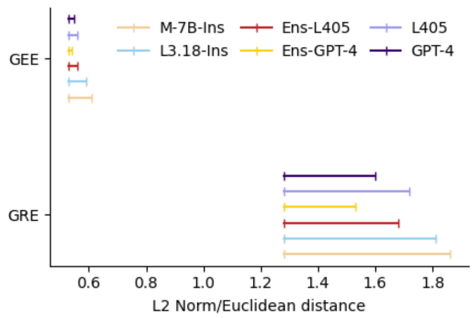

Figure 3: Distance between overall L2 norms of the embeddings of the annotations (left) and the extractions (left). Larger distances depict lower semantic alignment.

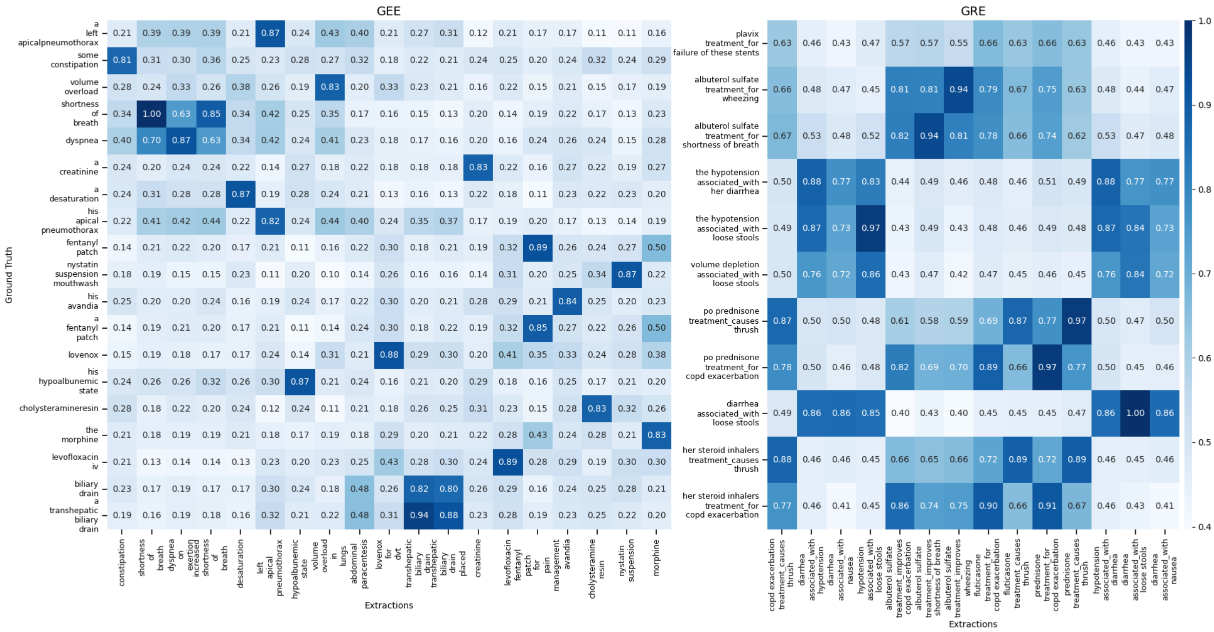

Figure 4: Analysing semantic similarity between ground truth and extractions for the best ensemble model using cos scores. Even when exact span matches (scores=1.0) are not achieved, partial matches still yield meaningful and understandable scores, e.g. for annotated span, "a desaturation", the closest match "desaturation" achieves score 0.9, which would have otherwise been Precision=0, Recall=0 and a micro F1=0 for traditional metrrics.

processing, we utilize GenJERE's entity pairing component to create entity pairs exclusively from GEE outputs of the closed MTP (ignoring GEE and GEvE outputs). We then query the LLMs (via closed MTP) with varying batch sizes [20, 15, 10, 5] of entity pairs for GRE and GEvE tasks. We only investigate the non-proprietary models and report the average score across all five GRE scores. We observe the performance progressively declining as batch size increases (Figure 6). Notably, there is a consistent performance improvement over the original MTP results for the respective models as reported in Table 2. These findings substantiate our hypothesis that LLMs struggle with intermediate batch processing during multi-tasking operations on a single turn.

5.3 Ablation

We then probe the relevance of the DA and IP. We set up two sets of experiments 1) Without definitions, in which the target concepts and relations are eliminated from the prompts and 2) Without IP, in which, we batch prompt the LLM with all extracted entities and their classifications and all relation triples (during denoising for GEE and GRE) in a single turn.

Multi-tasking without definitions: As shown in Figure 7, we observe a significant performance decline when definitions are eliminated from the prompts for all tested models. We also notice that the ensemble variants suffer more than the stand alone denoisers, which we attribute to the robust-

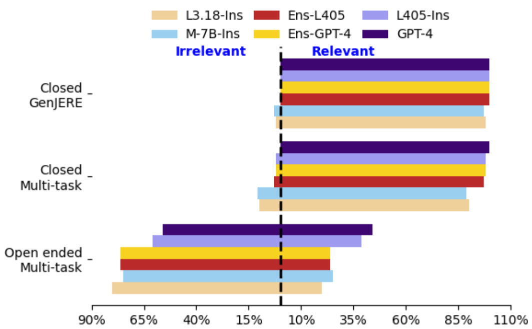

Figure 5: Percentage distribution of the relevant and irrelevant features (entity classifications combined with relation classifications). Full list of Irrelevant and relevant concepts and relations are included in the App. B.

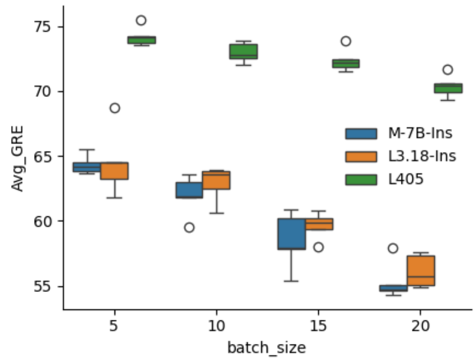

Figure 6: Investigating the average GRE scores over 5 runs for different batch sizes.

	GEE	GRE					GEvE
	SC	***TS***	***GS***	***FS***	***US***	***CS***	***EP***
L405	88.7	53.5	77.7	75.5	99.1	48.3	79.2
GPT-4	91.4	56.8	**80.1**	**81.9**	**99.9**	52.1	81.8
L3.1-8B-Ins^{+}	92.0	**58.1**	76.9	71.7	99.5	49.9	59.5
L3.1-8B-Ins^{++}	**92.8**	**58.1**	78.1	71.7	99.5	**53.2**	60.1

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.

ness of the denoisers, and their supreme ability to contextualize the EHR with respect to instructions.

Mult-tasking without IP: We further noticed in Figure 7, an even worse performance decline when IP was eliminated. This is especially seen with GRE, where there are several multi-labeled relation triples, which create ambiguity that obscures subtle differences between the different relations for the same entity pair. In summary, these results highlight 1) LLM batch prompting limitations on a single turn and 2) the critical contribution of both DA and IP for the overall LLM performance in multi-tasking such as EHR profiling.

6 GenJERE Tuning

We use LoRA to adapt Llama-3.18B-Instruction to the *instruction-ehr-profile* tuples generated in Stage 1 of the GenJERE pipeline. Preliminary analysis revealed that many input sequences (instruction + EHR record + profile) exceed Llama-3.18B's context window (8192 tokens), see Table 4.

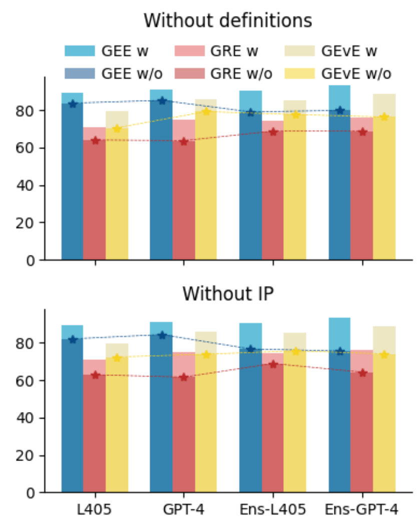

Figure 7: Probing relevance of DA and IP. Metrics plotted are **SC** for GEE, Average GRE for GEE and EP for GEvE, where "w". - with, and "w/o" - without.

We observed that the fine-tuned model performs competitively with the standalone denoisers, even outperforming them on GEE and the CS score for GRE. However, the denoisers still significantly outperform the fine-tuned model on GEvE. We also notice that LongLoRA enhances the performance of LoRA highlighting the benefit of long context tuning for long sequences such as EHRs.

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, IP and DA to improve generative IE performance while enhancing understandability. Across EHR datasets, the small-LLM ensemble outperformed larger models in generating semantically faithful and relevant outputs. We compared single-turn MTP and multi-turn (GenJERE) prompting, discovering that, LLMs struggle with simultaneous multi-task inference in IE on complex EHR data, largely due to intermediate batch processing complexity as outputs are transferred across sub-tasks. Our experiments show that IP and smaller batch sizes can mitigate this challenge, enhancing the MTP performance. Furthermore, fine-tuning a small LLM on synthetic data improved performance and outperformed stronger LLMs in some IE tasks.

Limitations

We introduce GenJERE, a pipeline that employs a small-LLM ensemble and MSP to execute multiple tasks sequentially, thereby alleviating the batch processing complexities during MTP. Despite GenJERE strong performance, some of its limitations should be noted as discussed below.

Implementing GenJERE end-to-end can be computationally expensive because 1) it processes one extraction at a time during self- or external-verification in its denoising mechanism and 2) it uses a brute-force approach by enumerating all possible entity pairs when inferring relations and 3) optionally facilitates for denoising using proprietary models which comes at a cost. Re-purposing an open-source LLM ensemble can be a potential alternative for denoising in future research endeavors especially when handling IE at scale.

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, so we can therefore only theorize its potential on corpora in other languages, especially those whose linguistic patterns deviate from english enormously. Moreover, the writing styles for clinicians can vary across countries, which can also affect model perplexity, and subsequently model performance on downstream tasks such as IE.

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.

References

Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20:1–9.

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*.

Zhoujun Cheng, Jungo Kasai, and Tao Yu. 2023. Batch prompting: Efficient inference with large language model apis. *arXiv preprint arXiv:2301.08721*.

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.

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.

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.

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*.

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

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

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*.

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*.

Deepti Saraswat, Pronaya Bhattacharya, Ashwin Verma, Vivek Kumar Prasad, Sudeep Tanwar, Gulshan Sharma, Pitshou N Bokoro, and Ravi Sharma. 2022. Explainable ai for healthcare 5.0: opportunities and challenges. *IEEE Access*, 10:84486–84517.

Parvathaneni Naga Srinivasu, N Sandhya, Rutvij H Jhaveri, and Roshani Raut. 2022. From blackbox to explainable ai in healthcare: existing tools and case studies. *Mobile Information Systems*, 2022(1):8167821.

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 4.1, 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 1. 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 compute pair-wise similarity

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

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

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

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.

A.3 Entity pairing

Following the 2010 Relations Challenge Dataset annotation guidelines, we retain only entity pairs eligible for the relation prediction task preserving directionality. For example, the "Treatment improves medical problem" relation annotated as "TrIP" id

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

defined to include mentions where a treatment improves or cures the problem. This suggests that concept classifications that constitute treatments would be subjects e.g. drugs, and those that constitute medical problems would be objects e.g. Disease or conditions. Table 5 summarises the relations, subjects and corresponding objects based on the annotation heuristics.

Algorithm 1 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) ∉ null_set then
        closeoffunclosedquots(element)
        remove_large_spaces(element)
        if element ∉ 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(ele-

        ment @ j)
      end if
    end for
  end for
  deduplicated_results[client] = deduplicated_entity_list
end for
```

Relations	Subject	Object
associated_with	Disease or condition, Signs or Symptoms, Injury, Other medical problems Mental or behavioral, disorder	Disease or condition, Signs or Symptoms, Injury, Other medical problems, Mental or behavioral disorder
treatment_for, treatment_causes treatment_improves, treatment_worsens, treatment_not_- administered	Drug, Biological substances Other treatments	Disease or condition, Signs or Symptoms, Injury, Other medical problems, Mental or behavioral disorder
test_investigates, test_reveals	Test or procedure	Disease or condition, Signs or symptoms, Injury, Other medical problems, Mental or behavioural disorder

Table 5: Heuristics for entity pairing

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 extracted 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 6.

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

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 7) 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.

	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 7: 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

	<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 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").

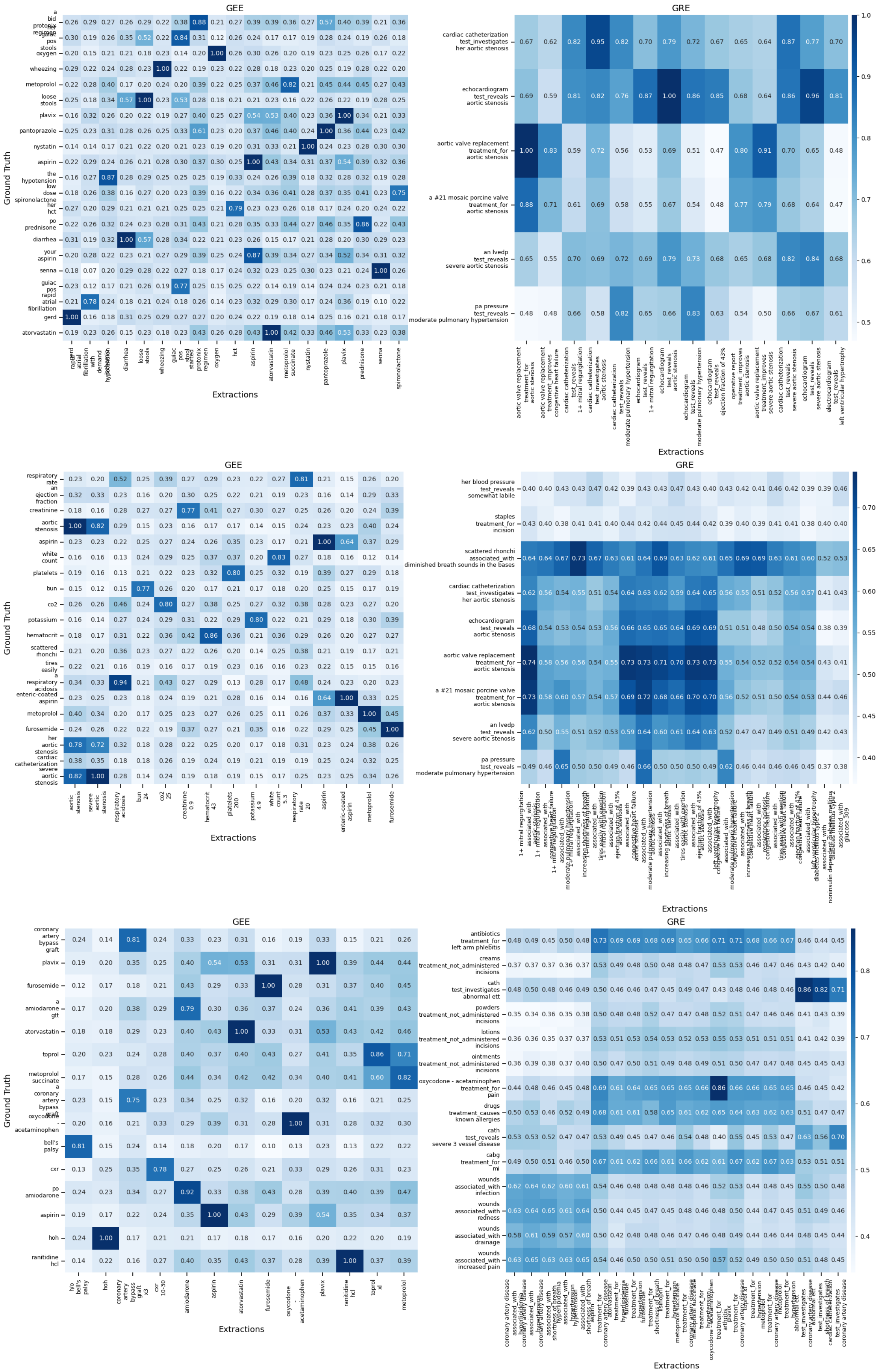

Figure 12: Analysing semantic similarity between ground truth and extractions for the best ensemble model using cos scores.

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