

# CARE: Extracting Experimental Findings From Clinical Literature

Aakanksha Naik<sup>1</sup> Bailey Kuehl<sup>1</sup> Erin Bransom<sup>1</sup>  
Doug Downey<sup>1</sup> Tom Hope<sup>1,2</sup>

<sup>1</sup>Allen Institute of AI, USA

<sup>2</sup>Hebrew University of Jerusalem, Israel

## Abstract

Extracting fine-grained experimental findings from literature can provide massive utility for scientific applications. Prior work has focused on developing annotation schemas and datasets for limited aspects of this problem, leading to simpler information extraction datasets which do not capture the real-world complexity and nuance required for this task. Focusing on biomedicine, this work presents CARE (Clinical Aggregation-oriented Result Extraction) — a new IE dataset for the task of extracting clinical findings. We develop a new annotation schema capturing fine-grained findings as n-ary relations between entities and attributes, which includes phenomena challenging for current IE systems such as discontinuous entity spans, nested relations, and variable arity n-ary relations. Using this schema, we collect extensive annotations for 700 abstracts from two sources: clinical trials and case reports. We also benchmark the performance of various state-of-the-art IE systems on our dataset, including extractive models and generative LLMs in fully supervised and limited data settings. Our results demonstrate the difficulty of our dataset — even SOTA models such as GPT4 struggle, particularly on relation extraction. We release our annotation schema and CARE to encourage further research on extracting and aggregating scientific findings from literature.

scientific studies, very few resources or tools capable of extracting such detailed information exist. However, once extracted and aggregated, scientific findings can power many critical applications such as producing literature reviews (DeYoung et al., 2021), proposing new hypotheses and ideas (Wang et al., 2023), and supporting evidence-based decision-making (Naik et al., 2022).

While there have been initial isolated efforts on building resources and tools to capture and represent aspects of experimental findings in various domains such as clinical trials (Lehman et al., 2019), computer science (Jain et al., 2020) and social and behavioral sciences (Magnusson and Friedman, 2021), one major obstacle to progress has been the complexity required in a representation scheme capable of capturing nuanced and salient information about findings. In this work, focusing on the goal of extracting clinical findings from biomedical literature, we make a first attempt at designing a suitable complex representation schema. Our proposed schema represents fine-grained information about experimental findings and conditions as n-ary relations between entities and attributes, and includes several structural complexities such as discontinuous span annotation, variable arity in relations and nestedness in relations. These aspects have been studied individually in previous datasets, but our schema is the first that unifies them to enable reliable representation of salient aspects of experimental findings. Though we design our schema with a focus on clinical findings, we conduct additional small-scale annotation studies demonstrating that our schema generalizes to two other scientific domains, computer science and materials science, with minor updates.

Using our new annotation schema, we collect extensive annotations for 700 abstracts from two types of biomedical literature: clinical trials and case reports. Our resulting dataset CARE is slightly larger in size than prior biomedical cor-

## 1 Introduction

“It is surely a great criticism of our profession that we have not organised a critical summary, by specialty or subspecialty, adapted periodically, of all relevant randomised controlled trials.” - Archie Cochrane, 1979

Though his original critique focused on clinical trials, this statement arguably applies to much of science today. Despite the potential utility of extracting, structuring and aggregating fine-grained information about experimental findings, and the conditions under which they were achieved, across

pora which also performed fine-grained annotation. Moreover, though our annotation task is more complex, we achieve good agreement scores (0.74-0.78 partial F1), comparable to prior work which uses simpler schemas (Luan et al., 2018; Nye et al., 2018), through multiple pilot rounds and consensus discussions. Our final dataset annotation is extremely rich; at 16.23 relations per abstract, our relation density is nearly 4x that of prior work on annotating findings from clinical trials (Lehman et al., 2019). Finally, a key differentiator for our dataset is that our annotations capture numeric findings in addition to their interpretation (e.g., significance, utility, etc.). Many prior datasets focus on the latter but not on the underlying findings, which are key to studying alternate hypotheses/interpretations or to conducting follow-up analyses.

We evaluate the performance of a wide range of information extraction baselines on our dataset, including both extractive systems and generative LLMs. Given the high annotation burden, we further test generative LLMs in both fully supervised as well as limited data settings. Our results demonstrate the difficulty of our dataset, with even SOTA models such as GPT4 struggling to accurately extract clinical findings.

Our work makes the following contributions:

- We propose a new annotation schema for fine-grained annotation of experimental findings, demonstrating its applicability to two types of biomedical literature as well as its potential generalizability to two other scientific domains (computer science and materials science)
- We release CARE, a dataset of 700 biomedical abstracts containing extensive annotations for the task of extracting clinical findings. Given the complexity of the task and annotation schema, we hope it presents a challenging new IE task with real-world utility.
- We present benchmarking results of several SOTA IE systems on our dataset. Our results show interesting trends such as the continuing superiority of encoder-only models on span extraction tasks and strong performance from GPT4 5-shot in-context learning on relation extraction (comparable to fully finetuned LLMs). However, there is still much room for improvement, which further established the complexity of CARE.

## 2 Related Work

### 2.1 Extracting Findings from Scientific Literature

Prior work on extracting findings or results from scientific literature has explored limited aspects of this problem. Gábor et al. (2018) and Luan et al. (2018) annotated *associative* relations between entities being compared or producing a result, as part of their broader goal of developing relation extraction datasets for computer science literature. However, this schema did not capture any nuanced information about the results (e.g., directionality, causality, etc.). Conversely, Magnusson and Friedman (2021) developed a schema solely focused on capturing associations between experimental variables and evidence. However, their schema targeted scientific claims, and is thus restricted to sentence-level annotation and limited in terms of how much additional nuance about the experimental setting can be captured.

Several prior efforts have also explored this problem in the domain of biomedicine. The EBM-NLP (Nye et al., 2018) and Evidence Inference (Lehman et al., 2019) datasets contain annotations for experimental findings from clinical trials, following the well-established PICO (participant, intervention, comparator, outcome) framework (Richardson et al., 1995). Similarly, Sanchez-Graillet et al. (2022) develop a PICO-inspired schema-based annotation format to annotate diabetes and glaucoma clinical trials. Some work (Chen et al., 2022) focuses on aggregating findings from clinical trials, which are already manually organized in structured formats in databases such as AACT (Aggregate Analysis of ClinicalTrials.gov) (Tasneem et al., 2012). These efforts, however, are heavily tailored for clinical trials and do not translate easily to other domains. Finally, Luo et al. (2022a) conducted *novelty* annotation for relations, indicating whether they were presented as new observations; however they annotated a broad set of relations, not focused on experimental findings.

In our work, we develop a schema which attempts to balance both, ability to capture fine-grained information about experimental findings and potential to generalize across scientific domains.

### 2.2 Extracting Numeric Information

Another unique aspect of our schema is our focus on extracting and linking numeric information from

Type	EBM	CTKG	Example
Population	✓	✓	This study compared rizatriptan 5 mg and placebo in 1268 outpatients treating a single migraine attack
Subpopulation	✓	✓	We found low-certainty evidence of little or no difference in delirium (RR 1.06, 95% CI 0.55 to 2.06; 2 studies, 800 participants)
Intervention	✓	✓	Dialysate magnesium was 0.375 mM/L for the hemodialysis
Measurement	✓	✓	Headache relief rates after rizatriptan 10 mg were higher
Temporal	✗	✓	After a 48-hour run-in period, oral verapamil 480 mg/day and placebo were administered
Numeric Finding	✗	✓	The number of attacks during treatment periods were 31 and 23
Qualifier	✗	✗	Pindolol and metoprolol lowered blood pressure to the same extent

Table 1: Examples of entity types

Type	EBM	CTKG	Example
Age	✓	✗	for those age 60-67 years
Sex	✓	✗	210 females
Size	✓	✓	12 patients
Condition	✓	✓	patients getting hemodialysis
Demographic	✗	✗	A 40's Japanese man
Route	✗	✗	oral verapamil
Dosage	✗	✗	verapamil 480 mg/day
Strength	✗	✗	rizatriptan 5 mg
Duration	✗	✗	for 4 weeks

Table 2: Examples of attribute types

experimental findings and setup, which is often understudied.

Some prior work on open information extraction has explored extraction and linking of numeric spans (Madaan et al., 2016; Saha et al., 2017), including linking to implied entities (Elazar and Goldberg, 2019) (e.g., “it’s worth two million” can be linked to currency). However, these models broadly focused on sentence-level extraction and their applicability to the scientific domain was untested.

Within the scientific domain, some studies have focused on numeric information extraction from various biomedical/clinical text sources. Kang and Kayaalp (2013) and Claveau et al. (2017) extract numeric spans from FDA-released decision summaries and clinical trial eligibility criteria respectively. The EBM-NLP corpus (Nye et al., 2018) annotates some categories of numeric information

associated with cohorts participating in a clinical trial, but ignores information associated with trial outcomes and findings. Among non-medical scientific domains, numeric span extraction work has mainly focused on extraction from table cells (Hou et al., 2019). None of these studies focus extensively on linking numeric spans with entities that can help in interpreting and structuring this information, which is a key component of our work.

### 3 Annotation Schema

We develop a new annotation schema to represent fine-grained clinical findings present in biomedical abstracts. Our schema captures this knowledge via three main elements, commonly used in information extraction tasks:

**1. Entities** involved in a study, which are spans of text, either contiguous or non-contiguous, belonging to one of the seven types listed in Table 1.

**2. Attributes** associated with entities, which are also contiguous or non-contiguous spans of text, belonging to one of the nine types listed in Table 2. The first five attribute types are associated with population and subpopulation entities, while the remaining four types are associated with intervention entities. Other entity types do not have any associated attributes.

**3. N-ary Relations** linking together various entities and attributes, where N (i.e., relation arity) is variable and nesting is allowed. Thus, a relation is an n-tuple, where each element in the tuple can be an entity, attribute or another n-ary relation. Relations are categorized into four types listed in Table 3.

Type	Arity	EI	CTKG	Example
AttributeOf	N-ary	✗	✗	( <i>Subpopulation</i> : 144 had the U-type method, <i>Size</i> : 144)
SubpopulationOf	N-ary	✗	✗	( <i>Population</i> : 285 women, <i>Subpopulation</i> : 144 had the U-type method, <i>Subpopulation</i> : 141 had the H-type method)
InterventionOf	Binary	✗	✓	( <i>Subpopulation</i> : 144 had the U-type method, <i>Intervention</i> : U-type method)
Result	N-ary	✓	✓	( <i>Subpopulation</i> : 144 had the U-type method, <i>Measurement</i> : objective cure rates, <i>NumericFinding</i> : 87.5%)

Table 3: Examples of relation types. Note that while the EI and CTKG datasets contain 4-ary and binary result relations respectively, our n-ary schema allows fine-grained information to be captured more flexibly.

### 3.1 Comparison to Previous Clinical IE Schemas

Much prior work such as EBM-NLP (Nye et al., 2018), Evidence Inference (Lehman et al., 2019; DeYoung et al., 2020), and CTKG (Chen et al., 2022) has focused on developing schemas to represent clinical knowledge in a structured format. However, these schemas suffer from a few shortcomings: (i) most are designed for clinical trials; their applicability to other types of biomedical literature is untested, (ii) focus on a small set of broad entity types, which leaves out fine-grained details, (iii) follow strict relation formats, which makes it hard to capture additional nuance that might be useful for interpreting findings

Our schema makes several enhancements to tackle these issues. First, it is extensible to other categories of biomedical literature beyond clinical trials, and we demonstrate this by applying our schema to case reports. Second, our schema captures more fine-grained information about various entities than prior work via attributes (see Table 2). Third, allowing for variable arity and nesting in relation annotation provides the flexibility which makes our schema capable of representing both atomic findings (e.g., value of primary outcome observed for a given intervention) as well as composite findings (e.g., outcome improvement observed for intervention vs control groups).

Table 1 indicates whether these entity types are present in the EBM-NLP and CTKG data. Table 2 indicates whether these attributes are present in the EBM-NLP and CTKG data. Table 3 indicates whether these relations are present in the Evidence Inference (EI) and CTKG data.

### 3.2 Annotation Complexity

In addition to using an expanded set of entity, attribute and relation types, our annotation schema allows the following phenomena:

**Discontinuous spans:** Biomedical abstracts often present multiple entities as conjunctive phrases or lists of items, so we allow discontinuous span annotation to capture each entity separately. For example, given the phrase “maximal diameters and volumes”, our annotation scheme captures two measurement entities: “maximal diameters” and “maximal volumes”, with the latter being a discontinuous span.

**Nested/overlapping spans:** Attributes, as defined in our annotation scheme, are often present within an entity span or overlap with an entity span. This motivates our decision to allow nested and overlapping spans to be annotated.

**Variable arity in relations:** Owing to variation in clinical studies, findings are often described in a wide range of formats (e.g., outcome for a single population, outcome for a pair of populations, outcome for a single population at different time periods, etc.). This diversity motivated our choice of *variable arity* for relation annotation, similar to Tiktinsky et al. (2022).

**Nested relations:** In addition to outcomes for individual populations/groups, clinical studies often present comparative findings and analyses, such as improvement on an outcome given a pair of interventions. Our scheme allows for annotation of nested relations to link these higher-order observations with their associated atomic findings. Our complete annotation guidelines are available at <https://github.com/aakanksha19/clinical-findings-extraction>. Figure 1 presents partial entity, attribute and relation annotations for an example clinical trial abstract.



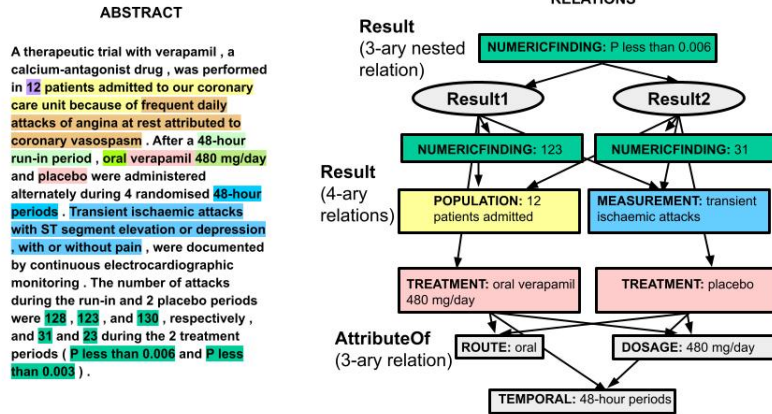


Figure 1: Partial entity, attribute and relation annotation using our schema for a clinical trial abstract.

## 4 Dataset Collection

**Annotation Tool:** For our data annotation process, we use TeamTat<sup>1</sup> (Islamaj et al., 2020), a web-based tool for team annotation projects. We choose this interface since it allows for n-ary and nested relation annotation, a core component of our annotation schema.

**Annotator Background:** To annotate CARE, we recruit two in-house annotators<sup>2</sup> with backgrounds in data analytics and data science, both having extensive experience in reading and annotating scientific papers. Additionally, one of our annotators has a background in biology. Both annotators went through several pilot rounds to gain familiarity with our annotation tasks and schema. Additionally, we used their feedback and insights from pilot annotation rounds to further solidify our schema design (see §4.1). We also solicited feedback from two medical students and an MD to validate our final schema.

**Data Sources:** CARE covers two categories of biomedical literature: (i) clinical trials, and (ii) case reports. Clinical trials are research studies that test a medical, surgical, or behavioral intervention in people to determine whether a new form of treatment or prevention or a new diagnostic device is effective. Case reports are detailed reports of the symptoms, signs, diagnosis, treatment, and follow-up of an individual patient, usually motivated by unusual or novel occurrences.

We sample clinical trials from the EBM-NLP (Nye et al., 2018) dataset, which consists of 4993

Category	Exact F1	Partial F1
Entity	0.5764	0.7578
Attribute	0.6174	0.7801
Relation	0.4209	0.7414

Table 4: Final inter-annotator agreement scores on a sample of 28 abstracts, measured during full-scale data annotation.

abstracts annotated with PICO spans, only retaining abstracts containing at least one number (4685 in total). Since EBM-NLP sampled randomized clinical trials from PubMed with an emphasis on cardiovascular diseases, cancer, and autism, the clinical trials portion of our dataset also heavily features these topics. To sample case reports, we extract all reports with at least one number in the abstract from PubMed (907,862 in total) and randomly sample abstracts from this pool. Comparing MeSH term distributions across all case reports (2M abstracts) and case reports containing numeric information, we see a massive reduction ( $> 30\%$ ) in terms associated with the following topics: surgery and post-surgery care, dentistry, ophthalmology, prostheses and rehab, patient care and nursing, some mental disorders and circulatory diseases/issues. Hence, we expect these topics to be relatively undersampled in our pool of case reports. We sample 350 abstracts each from EBM-NLP and our filtered case reports, resulting in our final dataset size of 700 abstracts, which is slightly larger than prior corpora that also perform fine-grained annotation (§ 4.3).

<sup>1</sup><https://www.teamtat.org>

<sup>2</sup>also included as co-authors on this paper

Metric	Train	Dev	Test
#Docs	500	100	100
#Tokens	135,363	27,120	25,219
#Entities	12022	2367	2286
#Attributes	3992	804	762
#Relations	8205	1594	1560

Table 5: Statistics for final collected dataset.

#### 4.1 Annotation Pilots

We conducted three pilot rounds with the following goals: (i) training annotators to apply our annotation schema, (ii) evaluating agreement between annotators, and (iii) assessing whether our schema was capable of representing the clinical knowledge of interest. During each pilot round, annotators were provided a fresh set of 5-10 biomedical abstracts to perform entity, attribute and relation annotation. After every round, inter-annotator agreement was computed and all disagreements were discussed. For entity and attribute annotation, agreement is computed as entity-level F1 between the two annotators, using both strict matching (entity boundaries should match exactly) and partial matching (entity boundaries overlap on at least one token) criteria. For relation annotation, we first align relation annotations from both annotators by linking together pairs of relations which share  $\geq 50\%$  of participating entities. Then agreement is computed as F1 score between the two annotators, using both strict (100% of entities should match) and partial matching criteria. During discussions between rounds, annotators also highlighted important spans/relations being missed by the schema, which led to the addition of the subpopulation entity, demographic attribute, and subpopulationof and treatmentof relations. Despite the introduction of some new elements, inter-annotator agreement increased steadily over the pilot rounds, as shown in Table 12 in the appendix. After achieving reasonable agreement levels by round 3 (partial F1 scores of 0.79, 0.68 and 0.79 for entity, attribute and relation annotation respectively), we started the full-scale data annotation process.

#### 4.2 Full-Scale Annotation

The full-scale data annotation process was conducted in six rounds. During rounds 1-3, annotators were provided batches of 25 abstracts to annotate. As their familiarity with the annotation schema and

Phenomenon	Train	Dev	Test
#Discontinuous Spans	8.9%	10.1%	9.3%
#Nested Spans	3.4%	4.3%	2.5%
#Overlapping Spans	1.6%	2.0%	0.7%
#Nested Relations	11.4%	11.2%	11.9%

Table 6: Prevalence of interesting annotation phenomena in final collected dataset.

ability to handle ambiguous cases improved, we provided larger batches of 100 abstracts each during rounds 4-6. To continue monitoring agreement, a small agreement set of 5 abstracts (not identified to the annotators) was included in every round. After each round, agreement was assessed and disagreement discussion meetings were conducted to discuss ambiguous cases, if needed. Table 12 in the appendix presents inter-annotator agreement during each annotation round, while Table 4 shows overall agreement scores. We can see that both overall and per-round agreement scores continue to remain in the same range as agreement scores from later pilot rounds, demonstrating consistency in annotation quality. Appendix Tables 13, 14 and 15 present agreement scores per entity type, attribute type and relation type respectively. From these tables, we can see that Subpopulation and Intervention entities are the trickiest to annotate, leading to lower agreement on SubpopulationOf and InterventionOf relation types due to error cascading (i.e., if entity annotations don’t match, relation annotations are unlikely to match either).

**Consensus Annotation:** For every abstract annotated by multiple annotators, either during pilot rounds or full-scale annotation (55 in total), we construct a “consensus” version post disagreement discussion. The final dataset contains these consensus annotations for these abstracts. Since this subset has been annotated by multiple annotators and discussed extensively, we expect the annotations to be higher-quality and include all these abstracts in the test split.

#### 4.3 Dataset Statistics

Table 5 gives an overview of statistics for our final collected dataset. Our dataset size is comparable to other prior biomedical corpora which performs exhaustive fine-grained annotation (though not always with a clinical knowledge focus) such as BioRED (Luo et al. (2022a); 600 abstracts) and Sanchez-Graillet et al. (2022) (211 abstracts). Table 6 presents the proportion of various interest-

Model	Entity F1			Attribute F1		
	No Disc	Strict	Relaxed	No Disc	Strict	Relaxed
<b>OneIE</b>	55.07	56.91	<b>76.18</b>	48.82	48.82	57.22
<b>PURE</b>	<b>55.94</b>	<b>57.48</b>	74.03	<b>61.04</b>	<b>61.04</b>	<b>69.83</b>
<b>LocLabel</b>	53.69	51.30	70.75	55.25	50.58	62.77
<b>W2NER</b>	–	51.84	69.53	–	57.98	67.57

Table 7: Performance of extractive IE systems on entity and attribute extraction. For models which cannot extract discontinuous spans, no disc and strict metrics refer to performance without span merging and with oracle span merging.

ing phenomena allowed by our schema in the final dataset. Interestingly, CARE contains 9% discontinuous spans, making it one of the rare datasets containing a large proportion of discontinuous mentions.<sup>3</sup> At 11%, the final data also contains a high proportion of nested relations.

## 5 Benchmarking IE Models

We benchmark the performance of two categories of models on CARE: (i) extractive models, and (ii) generative LLMs. Additionally, we test generative LLMs in two settings: (i) finetuning on the full training set, and (ii) zero-shot and in-context learning.

**Experimental Setup:** We first evaluate each model on the three sub-tasks—entity extraction, attribute extraction and relation extraction—in isolation. Model performance on entity and attribute extraction is evaluated using entity-level F1. Relation extraction performance is evaluated using a relaxed overlap F1 score metric inspired by Tiktinsky et al. (2022), which assigns partial credit to correctly identified subsets of entities in a relation, even if all identified entities do not match. As with agreement score calculation, predicted relations are first aligned with gold relations by choosing the gold relation with highest overlap per predicted relation. Then a partial match score is computed as  $\#shared\_entities/total\_entities$  and used in the F1 computation instead of binary 0/1 score. Finally, some models we evaluate are incapable of handling all complexity present in our dataset (e.g., many extractive models do not extract discontinuous spans). For such cases, we include additional metric variants, discussed in the corresponding sub-

section.

### 5.1 Extractive IE Baselines:

We evaluate the performance of the following state-of-the-art extractive systems:

- **OneIE** (Lin et al., 2020): A sentence-level joint entity, relation and event extraction system, which aims to extract an information network representation where entities and events are represented as nodes, and relations are represented as edges. Beam search is used to explore the space of possible networks and output the highest-scoring one.
- **PURE** (Zhong and Chen, 2021): A sentence-level pipelined extraction system, which learns separate contextual representations for entity and relation extraction, and uses entity representations to further refine relation extraction.
- **LocLabel** (Shen et al., 2021): A sentence-level two-stage named entity recognition (NER) system capable of handling nested span extraction. Inspired by object detection research, this system first produces boundary proposals for candidate entities, followed by labeling them with correct entity types.
- **W2NER** (Li et al., 2022): A sentence-level unified NER model, capable of handling both nested and discontinuous span extraction. This is achieved by recasting NER as word-word relation classification on a 2-D grid of word pairs, followed by decoding word pair relationships into final span extractions.

For comparability and better adaptation to our dataset, we substitute BERT-based encoders in all the previous systems with PubmedBERT (Gu et al., 2021), and follow the best-reported hyperparameters per system. Table 7 presents the performance of these systems on entity and attribute extraction. Unfortunately, applying these systems to our relation extraction task is infeasible, since none of

<sup>3</sup>Dai et al. (2020) considers 10% discontinuous spans to be a high proportion, identifying only three biomedical datasets that satisfy this criterion: CADEC (Karimi et al., 2015), ShArE 13 (Pradhan et al., 2013) and ShArE 14 (Mowery et al., 2014).

Model	Ent F1	Attr F1	Rel F1
<b>FLAN-T5</b>	<b>45.08</b>	23.27	<b>33.26</b>
<b>BioGPT</b>	14.43	<b>29.84</b>	32.95
<b>BioMedLM</b>	1.50	10.62	13.53

Table 8: Performance of finetuned LLMs on entity, attribute and relation extraction tasks.

Model	Setting	Ent F1	Attr F1
<b>GPT-3.5</b>	0-shot	11.14	5.06
	1-shot	21.40	8.61
	3-shot	23.40	8.85
	5-shot	8.92	9.92
<b>GPT-4</b>	0-shot	26.89	9.02
	1-shot	<b>31.07</b>	11.82
	3-shot	16.68	13.16
	5-shot	5.04	<b>13.90</b>

Table 9: Performance of GPT-3.5 and GPT-4 on entity and attribute extraction tasks, in both zero-shot and few-shot settings.

them are designed for document-level relation extraction or n-ary relations. Tiktinsky et al. (2022) modify the PURE system to perform n-ary relation extraction with variable arity. However, given a set of candidate entities, their system takes as input all possible n-ary combinations of these entities and predicts relationships between them. This is tractable in their setting because they focus on extraction from a single sentence and one entity type (drugs), but not tractable for our document-level multi-type n-ary relation extraction task.<sup>4</sup> Therefore, we do not report the performance of any extractive models on relation extraction.

Another caveat with extractive models is that all of them, with the exception of W2NER, are not capable of identifying discontinuous spans which form a large proportion of our dataset. We account for this during evaluation by reporting two variants of entity-level F1: (i) No Disc, which simply scores entity predictions as-is, and (ii) Strict, which merges together entity predictions if they’re linked together in gold annotation (i.e., we assume oracle span merging). Finally, we also report a relaxed entity overlap F1 score.

<sup>4</sup>On limiting combination size to 10, every abstract produces on the order of 500,000 candidate combinations

Model	Setting	Rel F1	
		Typed	Untyped
<b>GPT-3.5</b>	0-shot	8.64	14.35
	1-shot	20.90	31.58
	3-shot	22.21	31.58
	5-shot	24.56	32.20
<b>GPT-4</b>	0-shot	21.53	32.04
	1-shot	34.36	42.81
	3-shot	44.01	53.69
	5-shot	<b>45.11</b>	<b>55.04</b>

Table 10: Performance of GPT-3.5 and GPT-4 on relation extraction in both zero-shot and few-shot settings.

## 5.2 Generative IE Baselines:

Recent work has demonstrated that large language models are strong relation extractors, and can achieve comparable results to fully supervised generative IE systems (Wadhwa et al., 2023). This motivates us to assess the ability of LLMs on our tasks, in both finetuning and zero-shot/in-context learning settings.

We evaluate the following finetuned LLMs:

- **FLAN-T5** (Chung et al., 2022): An enhanced version of the T5 encoder-decoder model (Raffel et al., 2020) which has been finetuned on a large mixture of tasks. We use FLAN-T5-XL, which has 3B parameters. This model is not specifically pretrained for the biomedical domain.
- **BioGPT** (Luo et al., 2022b): A 1.6B autoregressive model, pretrained from scratch on 15M abstracts and titles from PubMed. This model uses a GPT-2 style architecture with a custom tokenizer trained on Pubmed abstracts.
- **BioMedLM**<sup>5</sup>: A 2.7B autoregressive model, pretrained from scratch on all PubMed abstracts and full-texts from the Pile (Gao et al., 2020). This model also uses a GPT2-style architecture with a custom PubMed-trained tokenizer.

When training and testing on attribute and relation extraction tasks, these models are provided gold entities and attributes by surrounding them with entity markers (`< ent >` `< /ent >`) in the input text. All models are trained for 10 epochs with a learning rate of  $1e-5$ , input context length of 1024, output length of 128, and a batch size of 2. Table 8 presents the performance of these models on entity, attribute and relation extraction.

<sup>5</sup><https://crfm.stanford.edu/2022/12/15/biomedlm.html>



Original Type	Generalized Type	Description
Population	Research Problem Context	Setting/scenario in which the authors are testing their hypothesis (e.g., task or dataset being studied in ML/NLP).
Subpopulation	Problem Stages/Sub-parts	Subgroups or subsamples of overall setting (e.g., dataset splits in ML/NLP).
Treatment	Technique/Method	Key technique being proposed or investigated and other techniques being compared (e.g., model or metric in ML/NLP).
SubpopulationOf	Sub-PartOf	Links together problem context entities to stage/sub-part entities (e.g., for ML/NLP, this relation would link the overall task to low-data and fully supervised settings).
TreatmentOf	AppliedTo	Links together a technique to all the problem contexts/sub-parts it is being tested in.

Table 11: Changes required to construct a generalized version of our original schema developed for clinical finding extraction, which we use to test whether it applies to other domains such as computer science and materials science

Finally, we evaluate both GPT3.5 and GPT4 in zero-shot and in-context learning settings. We test the 16k and 8k context length versions of GPT3.5 and GPT4 respectively since our extraction tasks are abstract-level and require longer input contexts. We use the June 2023 release versions due to their *function calling* capabilities, which allow us to describe our desired IE schema in JSON format and produce the extractions in a clean JSON format that adheres to the specified schema.

**In-context learning setup:** For our in-context learning experiments, we follow (Liu et al., 2021) and choose the  $k$  most similar examples from the training set for every test instance. We measure similarity using the SPECTER v2.0 (Singh et al., 2022) PRX model, an embedding model trained on scientific titles and abstracts. Top examples for every test instance are appended to the prompt in decreasing order of similarity, with later examples dropped if they don’t fit within the input context. We run experiments for the  $k = 1, 3, 5$  most similar examples. All experiments are run with a temperature of 0 and max output length of 512 tokens.

Table 9 shows the performance of both models on entity and attribute extraction, and Table 10 shows the performance on relation extraction. A caveat with relation extraction evaluation is that model outputs sometimes contain correct entity/attribute spans assigned to the wrong type (e.g., a subpopulation misclassified as a population entity in a result relation). Since we are evaluating the performance of relation extraction in isolation, such mistyping should not be considered an error. Hence, we report relation extraction performance

in both *typed* (considering mistypes as mistakes) and *untyped* (evaluation agnostic of type) settings.

### 5.3 End-to-End Evaluation:

In addition to evaluating SOTA systems on each sub-task in isolation, we assess the feasibility of building an end-to-end system for the entire task. From Tables 7, 8, and 9, we can see that PURE is the best-performing system on entity and attribute extraction across all model categories, according to strict entity overlap. On the other hand, Tables 8 and 10 show that GPT-4 5-shot and FLAN-T5 are the best-performing systems on relation extraction (though GPT3.5 5-shot and BioGPT are close as well). Therefore, we test out a hybrid end-to-end extraction system in which entities and attributes are first extracted using PURE, then input text marked up with these extractions is provided to FLAN-T5 for relation extraction (untyped). This hybrid end-to-end system achieves an F1 score of 33.58, which is very similar to RE performance with gold entity/attribute markup. We hypothesize that this might be an indication that finetuned LLMs largely ignore entity/attribute markup while performing relation extraction

## 6 Discussion

### 6.1 How much does strict evaluation underestimate LLM performance?

Tables 8 and 9 show that generative models severely lag behind much smaller extractive baselines on entity and attribute extraction tasks, even in fully supervised settings. However, prior work (Wadhwa et al., 2023) has observed that strict evaluation met-

rics often underestimate the performance of LLMs since their outputs might contain minor variations from gold annotations, which can still be considered correct. They recommend performing human evaluation to obtain a more accurate estimate of LLM performance.

Motivated by this, we conduct a human evaluation of FLAN-T5 and GPT4 5-shot predictions on both entity and attribute extraction sub-tasks. For every model-task pair, we collect all abstracts containing at least one wrong prediction and randomly sample 10 abstracts to evaluate. Our evaluation of FLAN-T5 predictions shows that 4 out of 116 incorrect entity predictions and 13 out of 53 incorrect attribute predictions were labeled correct. From GPT-4 evaluation, we observe that only 5 out of 126 incorrect entity predictions and 13 out of 109 incorrect attribute predictions were labeled correct. This indicates that these models indeed struggle with our span extraction sub-tasks, and their low performance is not simply a consequence of using a strict evaluation metric.

## 6.2 How easily can we extend our schema to other domains?

Though we focus on the goal of extracting clinical findings from biomedical literature while designing our schema, we try to incorporate enough flexibility to allow our schema to be easily adapted to other scientific domains. To demonstrate this flexibility, we conduct small-scale annotation studies in two additional scientific domains: (i) Computer Science, and (ii) Materials Science.

We first develop a *generalized* version of our proposed schema to use for these studies. Of the three elements in our schema, entities and relations are largely transferable and only require minor renaming to be applicable to other domains. Table 11 provides an overview of the changes made to entity/relation nomenclature. Attributes on the other hand, were tailored more closely to our goal of extracting clinical findings. Therefore, we drop all attributes and ask our annotators to propose candidate attributes as they go through the annotation process. We use the same annotators who participated in dataset create, to leverage their existing familiarity with our schema, assigning one annotator to each domain. Their task is to annotate ten abstracts each while documenting: (i) potential attributes that can be added to the schema, and (ii) important experimental information missed by the

generalized schema.

After completing the task, annotators reported that it was largely feasible to apply our proposed schemas to these two diverse scientific domains. Computer science posed some difficulty due to the presence of lots of relative results and references in the abstract, which made entity annotation ambiguous. However, there were no important aspects of experimental information, aside from potential attribute proposals, that our current schema could not account for.

## 7 Conclusion

In this work, we presented CARE, a new IE dataset for the task of extracting clinical findings from biomedical literature. To collect this dataset, we first developed a new annotation schema capable of capturing fine-grained information about experimental findings, which unified several challenging IE phenomena such as discontinuous spans, nested relations and variable arity n-ary relations. Using this annotation scheme, we collected an extensively annotated dataset of 700 abstracts from clinical trials and case reports. Our benchmarking experiments showed that state-of-the-art extractive and generative LLMs including GPT4 still struggle on this task, particularly on relation extraction. We release both our annotation schema and CARE as a challenging new resource for the IE community and to encourage further research on extraction and representation of findings from scientific literature.

## References

- Ziqi Chen, Bo Peng, Vassilis N Ioannidis, Mufei Li, George Karypis, and Xia Ning. 2022. A knowledge graph of clinical trials (ctkg). *Scientific reports*, 12(1):4724.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Vincent Claveau, Lucas Emanuel Silva Oliveira, Guillaume Bouzillé, Marc Cuggia, Claudia Maria Cabral Moro, and Natalia Grabar. 2017. Numerical eligibility criteria in clinical protocols: annotation, automatic detection and interpretation. In *Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine, AIME 2017, Vienna, Austria, June 21-24, 2017, Proceedings 16*, pages 203–208. Springer.

- Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. [An effective transition-based model for discontinuous NER](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5860–5870, Online. Association for Computational Linguistics.
- Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. [MS<sup>2</sup>: Multi-document summarization of medical studies](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132.
- Yanai Elazar and Yoav Goldberg. 2019. Where’s my head? definition, data set, and models for numeric fused-head identification and resolution. *Transactions of the Association for Computational Linguistics*, 7:519–535.
- Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haifa Zargayouna, and Thierry Charnois. 2018. [SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers](#). In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- 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.
- Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. [Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5203–5213, Florence, Italy. Association for Computational Linguistics.
- Rezarta Islamaj, Dongseop Kwon, Sun Kim, and Zhiyong Lu. 2020. Teamtat: a collaborative text annotation tool. *Nucleic acids research*, 48(W1):W5–W11.
- Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. [SciREX: A challenge dataset for document-level information extraction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.
- Yanna Shen Kang and Mehmet Kayaalp. 2013. Extracting laboratory test information from biomedical text. *Journal of pathology informatics*, 4(1):23.
- Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.
- Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717.
- Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973.
- Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. [A joint neural model for information extraction with global features](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.
- Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. [Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.
- Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022a. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022b. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.
- Aman Madaan, Ashish Mittal, Ganesh Ramakrishnan, Sunita Sarawagi, et al. 2016. Numerical relation extraction with minimal supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

- Ian H Magnusson and Scott E Friedman. 2021. Extracting fine-grained knowledge graphs of scientific claims: Dataset and transformer-based results. *arXiv preprint arXiv:2109.10453*.
- Danielle L Mowery, Sumithra Velupillai, Brett R South, Lee Christensen, David Martinez, Liadh Kelly, Lorraine Goeuriot, Noemie Elhadad, Sameer Pradhan, Guergana Savova, et al. 2014. Task 2: Share/clef ehealth evaluation lab 2014. In *Proceedings of CLEF 2014*.
- Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. 2022. [Literature-augmented clinical outcome prediction](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453, Seattle, United States. Association for Computational Linguistics.
- Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron C Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207.
- Sameer Pradhan, Noemie Elhadad, Brett R South, David Martinez, Lee M Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana K Savova. 2013. Task 1: Share/clef ehealth evaluation lab 2013. *CLEF (working notes)*, 1179.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- W Scott Richardson, Mark C Wilson, Jim Nishikawa, Robert S Hayward, et al. 1995. The well-built clinical question: a key to evidence-based decisions. *Acp j club*, 123(3):A12–A13.
- Swarnadeep Saha, Harinder Pal, and Mausam. 2017. [Bootstrapping for numerical open IE](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 317–323, Vancouver, Canada. Association for Computational Linguistics.
- Olivia Sanchez-Graillet, Christian Witte, Frank Grimm, and Philipp Cimiano. 2022. An annotated corpus of clinical trial publications supporting schema-based relational information extraction. *Journal of Biomedical Semantics*, 13(1):1–18.
- Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. [Locate and label: A two-stage identifier for nested named entity recognition](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2782–2794, Online. Association for Computational Linguistics.
- Amanpreet Singh, Mike D’Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*.
- Asba Tasneem, Laura Aberle, Hari Ananth, Swati Chakraborty, Karen Chiswell, Brian J McCourt, and Ricardo Pietrobon. 2012. The database for aggregate analysis of clinicaltrials. gov (aact) and subsequent regrouping by clinical specialty. *PloS one*, 7(3):e33677.
- Aryeh Tiktinsky, Vijay Viswanathan, Danna Niezni, Dana Meron Azagury, Yosi Shamay, Hillel Taub-Tabib, Tom Hope, and Yoav Goldberg. 2022. A dataset for n-ary relation extraction of drug combinations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3190–3203.
- Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. [Revisiting relation extraction in the era of large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.
- Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2023. Learning to generate novel scientific directions with contextualized literature-based discovery. *arXiv preprint arXiv:2305.14259*.
- Zexuan Zhong and Danqi Chen. 2021. [A frustratingly easy approach for entity and relation extraction](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.



## A Schema Definitions

### A.1 Entity Types

Entities can belong to one of the following seven types:

1. **Population:** Patient groups/cohorts studied in an article.
2. **Subpopulation:** Slices/sub-groups of a population entity sharing some underlying characteristic.
3. **Intervention:** Treatment regimens, procedures, therapies etc. prescribed and/or tested to alleviate a population's conditions/symptoms.
4. **Measurement:** Tests used to assess population status and outcomes of the tested intervention.
5. **Temporal:** Temporal information such as time points at which outcomes are measured.
6. **NumericFinding:** All numeric information associated with study findings (e.g., p-values, hazard ratios, etc.).
7. **Qualifier:** Non-numeric information associated with study findings that provides important perspective for interpreting them (e.g., phrases indicating evidence directionality).

### A.2 Attribute Types

Attributes can belong to one of the following nine types:

1. **Age:** Numeric or non-numeric information about the age of the population under study.
2. **Sex:** Reported sex of the population under study.
3. **Size:** Size of the population sample under study.
4. **Condition:** Medical conditions prevalent in the study population, including diseases, symptoms, prior medical history and procedures, etc.
5. **Demographic:** Additional demographic information reported about the population such as location, race, etc.
6. **Route:** Description of the way an intervention is administered (e.g., a chemical may be administered orally, topically, intravenously, etc.).
7. **Dosage:** Quantity of administration for the intervention being studied. This is not necessarily limited to chemical/drug interventions (e.g., for an intervention like educational sessions, number of sessions is considered "dosage").
8. **Strength:** Strength of chemical/drug interventions administered.
9. **Duration:** Interval of time over which an intervention was administered.

### A.3 Relation Types

Our schema allows for both binary and n-ary relations (with variable n), to capture four types of structure:

1. **AttributeOf:** N-ary relations linking population and intervention entities with their associated attributes.
2. **Subpopulation:** N-ary relations capturing parent-child relationships between population and subpopulation entities.
3. **InterventionOf:** Binary relations linking population and subpopulations entities with the intervention(s) tested on them.
4. **Result:** N-ary relations capturing all numeric or non-numeric outcome results and comparisons reported by linking together the population, subpopulation, intervention, measurement, numericfinding and/or qualifier and temporal entities involved in each result/comparison.

All n-ary relations can contain multiple entities of a single type. For example, a result relation can involve multiple interventions or populations. The only cardinality constraints imposed are that every result relation should focus on a *single* measurement entity and always contain *at least one* population/intervention entity.

## B Additional Annotation Rules

While using this annotation schema to annotate clinical knowledge, we also keep in mind the following rules:

- For every entity/attribute span, only annotate its first occurrence in the text, unless there is a more descriptive span later. We follow this rule to avoid conducting an additional coreference annotation step to link all spans referring to the same entity.
- Ignore misspellings and include all associated modifiers and abbreviations while annotating spans
- Do not annotate generic or high-level spans (e.g., genetic disorder), or generic terms (e.g., complications, deficiency, disease, syndrome, gene, drug, protein, nucleotide, etc.).
- Do not annotate background occurrences of entities. For example, if a treatment Y is mentioned as "X is usually treated using Y,...", do not annotate Y unless Y was one of the treatments actually given to a population in the current study.

Round	Entity F1		Attribute F1		Relation F1	
	Exact	Partial	Exact	Partial	Exact	Partial
Pilot 1	0.6240	0.7579	0.7215	0.8163	0.2193	0.6379
Pilot 2	0.7206	0.8818	0.6923	0.7385	0.4997	0.7878
Pilot 3	0.6449	0.7900	0.5370	0.6852	0.4449	0.7960
Batch 1	0.5130	0.7318	0.7611	0.8496	0.3899	0.6979
Batch 2	0.6094	0.7900	0.6216	0.8508	0.6397	0.9137
Batch 3	0.5312	0.7797	0.6364	0.8182	0.3121	0.7595
Batch 4	0.5714	0.7817	0.7347	0.7755	0.5399	0.7343
Batch 5	0.5643	0.6929	0.4717	0.6762	0.3382	0.6766
Batch 6	0.6358	0.7930	0.5417	0.7582	0.3122	0.6890
Overall	<b>0.5764</b>	<b>0.7578</b>	<b>0.6174</b>	<b>0.7801</b>	<b>0.4209</b>	<b>0.7414</b>

Table 12: Evolution of inter-annotator agreement during pilots and full-scale annotation rounds

Type	Exact F1	Partial F1
Population	0.4333	0.8665
Subpopulation	0.4299	0.6168
Intervention	0.4333	0.5781
Measurement	0.5230	0.7554
Temporal	0.6230	0.6885
NumericFinding	0.7063	0.8812
Qualifier	0.6911	0.7749

Table 13: Inter-annotator agreement per entity type

Type	Exact F1	Partial F1
Age	0.8500	0.9756
Sex	0.9231	0.9231
Size	0.6462	0.7385
Condition	0.5091	0.7429
Demographic	0.6667	0.8000
Route	0.8000	0.8000
Dosage	0.6923	0.9630
Strength	-	-
Duration	0.0800	0.4800

Table 14: Inter-annotator agreement per attribute type. Note that the agreement sample did not include any strength entities.

Type	Exact F1	Partial F1
AttributeOf	0.7654	0.7654
InterventionOf	0.3797	0.3797
SubpopulationOf	0.1633	0.5185
Result	0.2561	0.7994

Table 15: Inter-annotator agreement per relation type

## C Inter-Annotator Agreement

Table 12 shows the evolution in inter-annotator agreement over our initial pilot rounds, as well as the level of inter-annotator agreement maintained during each round of the full-scale annotation process. We see a large increase in relation agreement from pilot 1 to pilot 2, and consistent agreement scores across all tasks in all rounds thereafter. Tables 13, 14 and 15 present inter-annotator agreement breakdown according to entity, attribute and relation types in our schema.