HHS Public Access

Author manuscript

Proc Conf Assoc Comput Linguist Meet. Author manuscript; available in PMC 2023 September 06.

Published in final edited form as:

Proc Conf Assoc Comput Linguist Meet. 2023 July ; 2023: 15566–15589. doi:10.18653/v1/2023.acl-long.868.

Revisiting Relation Extraction in the era of Large Language Models

Somin Wadhwa,

Silvio Amir,

Byron C. Wallace

Northeastern University

Abstract

Relation extraction (RE) is the core NLP task of inferring semantic relationships between entities from text. Standard supervised RE techniques entail training modules to tag tokens comprising entity spans and then predict the relationship between them. Recent work has instead treated the problem as a *sequence-to-sequence* task, linearizing relations between entities as target strings to be generated conditioned on the input. Here we push the limits of this approach, using larger language models (GPT-3 and Flan-T5 large) than considered in prior work and evaluating their performance on standard RE tasks under varying levels of supervision. We address issues inherent to evaluating generative approaches to RE by doing human evaluations, in lieu of relying on exact matching. Under this refined evaluation, we find that: (1) *Few-shot* prompting with GPT-3 achieves near SOTA performance, i.e., roughly equivalent to existing *fully supervised* models; (2) Flan-T5 is not as capable in the few-shot setting, but supervising and fine-tuning it with Chain-of-Thought (CoT) style explanations (generated via GPT-3) yields SOTA results. We release this model as a new baseline for RE tasks[1].

1. Introduction

Relation extraction (RE) is the task of identifying entities and their semantic relationships from texts. Standard supervised approaches (Eberts and Ulges, 2019a) to RE learn to tag entity spans and then classify relationships (if any) between these. More recent work has shown that conditional language models can capably perform this task—achieving SOTA or near-SOTA results—when trained to output linearized strings encoding entity pairs and their relations (Paolini et al., 2021; Lu et al., 2022b; Huguet Cabot and Navigli, 2021). However, to date such work has considered only moderately sized pre-trained models for RE such as BART (Paolini et al., 2021; Huguet Cabot and Navigli, 2021).

In this work we investigate the use of very large language models—-including GPT-3 (Brown et al., 2020b)—for end-to-end relation extraction via generation. Our contributions are as follows.

[1] https://sominw.com/ACL23LLMs

wadhwa.s@northeastern.edu .

1. We show that few-shot learning with GPT-3 yields near SOTA performance on standard RE datasets, outperforming fully supervised models.
2. We find that Flan-T5 (large; Chung et al. 2022) is not as capable, even when fine-tuned. But we then propose an approach to training Flan-T5 with *Chain-of-Thought* (CoT) style "explanations" (generated automatically by GPT-3) that support relation inferences; this achieves SOTA results.
3. Evaluating the performance of *generative* models for RE is non-trivial because one cannot rely on exact matches to targets. We address this by collecting a small amount of annotations scoring generated outputs against targets. We use these annotations to quantify the problem, identify erroneous gold references and accurately evaluate our models.

Our results indicate that, in general, **LLMs should be the default approach to RE**, especially given that one can train Flan-T5—which is dramatically smaller than GPT-3, and publicly available—to achieve SOTA performance (Figure 1).

2. RE via Text Generation

We treat RE as a conditional text generation task. Concretely, for a dataset of size N, we model the probability of generating a *linearized* string y of a relation triplet (`entity_1`, `relation_type`, `entity_2`) conditioned on a context string $\mathscr{C}$. Specifically, $\mathscr{C}$ includes a chain of n linearized examples (x_i, y_i), with $n << N$. Formally:

$$p_{\mathrm{LM}}(y \mid \mathscr{C}, x) = \prod_{t=1}^{T} p(y_t \mid \mathscr{C}, x, y_{<t})$$

We provide examples of context strings in the Appendix. We conduct experiments over four standard RE datasets comprising varying numbers of entities and relation types, namely ADE (Gurulingappa et al., 2012), CoNLL (Roth and Yih, 2004), NYT (Riedel et al., 2010), and DocRED (Yao et al. 2019); details in Table 1 and Appendix A.

Following Huguet Cabot and Navigli (2021), we linearize our target relation triplets. However, we adopt a much simpler scheme than prior work: We linearize inputs with a single relation type (e.g. ADE) as a list of tuples:

```
[(drug, effect), ... ,(drug, effect)]
```

For inputs with multiple relation types (as in CoNLL04 and NYT), we form *triplets* comprising a `subject`, `relation`, and `object` (along with their corresponding types), in the order of appearance of the subject entity:

```
[(entity_1:entity_1_type, relation_type, entity_2:entity_2_type),..]
```

A training instance is then a pair of input text and a linearized target string:

```
Input Bill Nelson, NASA administrator announced the mars mission today.
Target [(Bill Nelson:Per, Work_For, NASA:Org)]
```

Challenges inherent to evaluating generative large language models for RE

The expressivity of language models coupled with the open-endedness of RE makes evaluation difficult. This has led to inconsistent approaches to evaluation (Taillé et al., 2020). Past work, especially that pre-dating LLMs for the task, has tended to perform "strict" evaluation, requiring exact matches between generated linearized relation tuples and references. This may be appropriate when is evaluating smaller conditional generation models (such as BART) for RE, which have been *fine-tuned* on large training sets, because after training such models consistently generate standardized outputs. By contrast, however, models like GPT-3 (or other large language models capable of zero- or few-shot application) can produce a wide variety of output formats which convey similar content.

For example, given an input from ADE and prompted to *list all drugs and associated adverse events*, a large language model might yield *Aspirin: stomach pain, chest pain*. Or it may instead output: *Side effects of aspirin include cramping and stomach pain, and pain in the chest*. There are countless possible variants which may all communicate the correct answer; we provide additional real examples in the Appendix D. The flexibility of language means that parsing out the structured result to compare it to a reference (to calculate standard metrics like precision, recall, and F-1) is a non-trivial problem. This is in stark contrast to traditional approaches to tasks like NER and RE where models effectively classify input tokens instead of generating new ones from a vast vocabulary.

Training models, either via traditional supervised learning or in-context few-shot learning, encourages models to comport with the structure of training instances. We therefore focus our analysis on such supervised settings in this work, starting with an evaluation of few-shot learning with GPT-3 for RE. Nonetheless, even when supervised, LLMs used for RE are prone to generating outputs which may be accurate but nonetheless differ from the target. To address this, we enlist human annotators to judge whether the model outputs convey the same information as the reference targets.

3 In-Context Few-Shot Learning with GPT-3 for RE

In this section we first describe our few-shot prompting strategy for GPT-3, and report the results realized by this approach across a set of RE corpora. We adopt forms of instructional in-context few-shot prompting to GPT-3.[2] Motivated by the preceding discussion regarding evaluation challenges, we collect human annotations judging the model's generations against the gold references. Finally, using these annotations we report results achieved using GPT-3 with few-shot prompting for RE (Table 2). All references to GPT-3 in this work refer to the "text-davinci-002" variant.

[2]We provide details on the costs incurred for each of these experiments in the Appendix B.1.

3.1 Prompts

We describe the prompts we use for each of the datasets considered in turn.

ADE—To construct prompts for ADE, we use the instructional prompt: *List all (drug: adverse effects) pairs in the following text*, followed by an input text. We then select 12 examples ("shots") at random from the training set, and for each we append the corresponding input followed by linearized target relations to the instructional prompt; this yields a prompt featuring 12 examples, comprising 755 tokens. To make a prediction for a new example we append one last *List all (drug: adverse effects) pairs in the following text* instruction followed by the corresponding text and then ask GPT-3 to generate text conditioned on this final prefix. Specifically, we perform this generation using default parameters save for sampling temperature, which we set to 0.5.[3] We impose a maximum output length of 256 tokens.

CoNLL—As an instructional prefix for CoNLL, we use: *List the entities of the types [LOCATION, ORGANIZATION, PERSON] and relations of types [Organization Based In, Work For, Located In, Live In, Kill] among the entities in the given text*. Since CoNLL is composed of four entity and five relation types, we constructed our prompt manually to contain at least one example of each entity and each relation type, for a total of 12 exemplars in the prompt. The total length of the CoNLL prompt was 960 tokens. To ensure fair comparison to prior work on generative RE over CoNLL, we use the same validation set as Eberts and Ulges (2019a).

NYT—The large number of relations (24 in total) in the NYT dataset precludes the possibility of providing detailed instructions enumerating all entity and relation types. We instead shorten the instructional prefix by removing specific relation-type descriptors and create a prompt with only 20 exemplars capturing all entity and relation types. The size of this prompt was 2095 tokens.

We next aim to evaluate the performance of GPT-3 for RE when provided the above prompts. But doing so requires addressing the challenges inherent to evaluating LLMs for RE outlined above (and in prior work; Taillé et al. 2020).

3.2 Manually re-evaluating "errors"

We quantify the errors in evaluation that occur when one uses "strict" measures of performance while using few-shot prompted LLMs for RE across each dataset. We do this by acquiring human annotations (collected via Mechanical Turk; details in Appendix D) on model outputs, with respect to reference labels provided in the accompanying datasets. In particular, we show annotators ostensible "false positive" and "false negative" outputs produced by GPT-3 for these corpora—as would be computed using exact matching against references—and ask them to judge whether these are accurately categorized.

[3] In preliminary manual assessments, this seemed to yield qualitatively better outputs here than the default temperature.

On **ADE** we find that 51.67% of "false positives"—a slight majority—are more accurately viewed as *true* positives, and 32.61% of "false negatives" are deemed as, in fact, true negatives. On **CoNLL** outputs, annotators marked 50.27% of "false positives" as valid, and 36.6% of "false negatives" as being accurate.

As mentioned above, we were unable to design a prompt for **NYT** that yielded reasonable few-shot results with GPT-3. So we instead ask annotators to evaluate outputs from Flan-T5 fine-tuned on the NYT train set. In this case, they deemed 36.9% and 22.97% of "false positives" and "false negatives", respectively, to in fact be accurate. We present some illustrative cases in Figure 2 and additional examples in Appendix Tables 8 and 7.

These findings imply that strict (exact-matching) evaluation against references for RE will be inaccurate (and pessimistic). In the results we later report for LLMs, we therefore take into account these manual assessments.[4]

3.3 Results

Using the above prompts and manual annotation process just described, we find that in most cases **GPT-3 performs comparably to current *fully supervised* SOTA RE models without fine-tuning and given only 12–20 training examples**. This can be seen in Table 2 (2.a). We also find a substantial number of instances where the model correctly identifies relation pairs, which in fact are incorrectly marked in the references (detailed below in Section D). We observe additional issues with the NYT and CoNLL datasets which we discuss below.

CoNLL—We find a number of relation triplets where the output does not conform to the set of valid relation types (~% of relation triplets in the validation set). Examining these triplets, we often find the out-of-domain relation-types to be either closely related to a correct CoNLL relation-type (e.g., *shoot* →*kill*) or otherwise correct even if not related to a CoNLL relation-type. There were a total of 18 input validation instances in which at least one of the generated relation triplet did not conform to a valid CoNLL relation; we provide a full list of these instances and the generated relation triplets in the Appendix D.1.

NYT—We find the strategy of omitting the relation descriptions in the prompt to be detrimental to the model's performance. Contrary to our findings in ADE and CONLL, we observe a *sharp decline* in Micro-F1 scores in case of NYT (~30 point reduction) as compared to the fully supervised SOTA. Further, we observe a non-trivial number of invalid or empty output instances (~10.6% of all generated sequences). These results highlight a remaining limitation of in-context learning with large language models: for datasets with long texts or a large number of targets, it is not possible to fit detailed instructions in the prompt. In light of the issues we were unable to evaluate this approach on the DocRED dataset, which we leave for future work. In such cases, traditional fine-tuning is the practical option.

[4]One could also *train* a model on manual assessments of "false positives" and "false negatives" to semi-automate this evaluation (avoiding the need to collect such judgments on entire testing sets); we provide results showing the feasibility of doing so in the Appendix D.

Despite these limitations, the fact that GPT-3 is able to (marginally) outperform the current SOTA with in-context learning from tens of examples is encouraging. But GPT-3 is a massive opaque model available only via OpenAI's API (at cost). Further, fine-tuning GPT-3 would incur additional cost, and one would have access to the resultant model only via the OpenAI interface. For these reasons, smaller, open-source LLMs for RE would be preferable. Next we show that by enriching supervision with *Chain-of-Thought* (CoT) outputs elicited from GPT-3, we can achieve SOTA performance using Flan-T5 (Large).

4 SOTA RE Performance with Flan-T5

We use Flan-T5 (Large), an LLM trained on a large number of tasks with instructional prompts. We first evaluate this in a few-shot setting (Section 4.1), shortening prompts in light of T5's smaller size, compared to GPT-3. We then consider fine-tuned variants, including a novel approach in which we train Flan-T5 using *chain-of-thought* (CoT) style explanations for RE elicited from GPT-3. The latter strategy yields SOTA results across all datasets considered.

4.1 Few-Shot RE with Flan-T5

For few-shot learning with Flan-T5, we use the same instructional prefixes (with examples) as we did for GPT-3 above, but we reduce the number of exemplars in the prompts to make them more concise. We summarize our findings from these experiments on ADE and CoNLL below, and provide a full set of results in Appendix B.

ADE—We include 7 (instead of the 12 used for GPT-3) randomly selected in-context examples for ADE. We observe a significant increase in non-conforming relation pairs in outputs (13.9% of generations). These often include outputs where the model generates the same token (or a set of tokens) repeatedly, or where relation tuples contain greater or fewer than 2 entities. Unsurprisingly given these qualitative impressions, the model fares poorly under strict evaluation on the validation set, resulting in a ~ 20 drop in F1 score compared to GPT-3.

CoNLL—The prompt for CONLL consisted of 7 (in place of the 12 for GPT-3) exemplars inserted into the instructional prefix described above. Again we found that Flan-T5 generated many non-conforming outputs (12.5%). Additionally, we find that Flan-T5 generates a large number of out-of-domain relations between entities (over 120 unique relations), most of which are unrelated to CoNLL, making it impossible to meaningfully evaluate outputs (details in Appendix D).

NYT—We exclude this dataset given the large set of relation and entity types, which—as discussed above—makes designing a prompt with sufficient instructions that also fits within the in-context window impossible. (We address this below via fine-tuning, which sidesteps the issue.)

These results indicate that few-shot learning with Flan-T5 is not competitive with GPT-3, and so is not comparable to SOTA RE models. However, we next show that fine-tuning

Flan-T5 can yield substantially better results, especially if one includes *reasoning* about RE in the supervision.

4.2 Fine-tuning Flan-T5 for RE

We first perform standard fine-tuning for Flan-T5 (Large) using available training datasets. We report results from the test set in Table 2 (1.e.). This yields performance equivalent to, but not better than, existing fully supervised models such as REBEL.

As a potential mechanism to improve the performance of Flan-T5 for RE, we propose enriching the supervision used to fine-tune the model with *chain-of-thought* (CoT; Wei et al. 2022b) explanations, which we elicit automatically from GPT-3 over the training instances. Specifically, we craft a handful of such reasoning chains describing how target relations can be derived from the input texts. We provide the following three illustrative examples below.

Example Input (ADE) To describe a case of severe skin necrosis resulting from peripheral intravenous administration of low-dose vasopressin in a patient with catecholamine-resistant septic shock.
Target [(vasopressin, skin necrosis)]
Explanation A case of skin necrosis was described after administration of low-dose vasopressin.

Example Input (CONLL) In Colorado , 13 inches of snow in Denver Wednesday prompted officials to close Interstate 270 temporarily.
Target [(Denver, ‘Located In’, Colorado)]
Explanation - Denver officials closed Interstate 270 in Colorado, consequently we can see that Denver is located in Colorado.

Example Input (NYT) It will be the final movie credited to Debra Hill, a film producer and native of Haddonfield, who produced “Halloween” and was considered a pioneering woman in film.
Target [[Debra Hill:Per,’place-of-birth’, Haddonfield:Loc]]
Explanation - Debra Hill was a film producer born (native of) in Haddonfield.

Next we evaluate the impact of CoT explanations in two settings: As additional context for prompting GPT-3, and then as additional supervision signal with which to train Flan-T5.

4.2.1 Eliciting CoT reasoning for RE—We use the same prompts from the few-shot experiments above but augment them with CoT-style explanations (one per shot) written by one of the authors. This yields moderate gains in the overall performance for GPT-3 (~3 and ~2.2 micro-F1 points for ADE and CONLL, respectively; Table 2 2.b), and also reduces the number of non-conforming relations generated (from 13.9% to 0.8% on ADE, and from 12.5% to 1.1% on CONLL). Further, using CoT results in only one instance of an out-of-domain relation-type generated on CoNLL, compared to over 120 relations generated

without CoT explanations. In sum: using CoT in few-shot learning for RE with GPT-3 yields more standardized outputs, but does not much improve performance. Next we propose to capitalize on CoTs automatically generated over training sets to enrich the supervision with which we train Flan-T5.

4.2.2 Fine-tuning Flan-T5 with CoT explanations—We augment target relations used to train Flan-T5 with CoT strings automatically generated by GPT-3 over the training dataset. Specifically, we modify the prompt used in Section 3 to generate *CoT-style explanations* conditioned on the input *and* relation reference labels. The following is an example of the prompt we provide GPT-3 to elicit a *CoT-explanation*:

```
Text: This April 14 is the 125th anniversary of the night when Lincoln, the
16th president, was assassinated by John Wilkes Booth in the presidential
box at Ford’s Theatre.
Target [(John Wilkes Booth, ‘Kill’, Lincoln)]
Explanation - John Wilkes Booth assassinated Lincoln at the ford theatre.<s>

Text: Ray is being held in Tennessee ‘s Brushy Mountain State Prison on a
99-year sentence for the April 4, 1968, slaying of King.
Target [[Ray, ‘Kill’, King]]
Explanation -
```

We then use these explanations along with reference relation labels as targets to fine-tune Flan-T5 (Large), as depicted in Figure 3. Overall, we found this strategy to be effective obtaining state-of-the-art results across datasets, while being much faster to train compared with existing fully supervised models. We summarize our findings below, and report results in Table 1 (1.f.).

ADE: We obtain explanations for the entire training set and fine-tune Flan-T5 Large with an instructional prefix with a batch size of 8, learning rate 3e-5 for 6 epochs. The dataset defines 10 folds of train/test splits, and we evaluate using the best checkpoint for each fold in the dataset. Our model yields a 9.97 point gain in micro F-1 score (averaged over the folds) over the existing fully supervised generative SOTA (REBEL; Huguet Cabot and Navigli (2021)).

CONLL: For CONLL, we again obtain *CoT-style* explanations for the entire dataset via GPT-3. We then fine-tune with a batch size of 4 and learning rate 3e-5 for 10 epochs and evaluate using the best-performing checkpoint on the validation set. We see a 5.42 absolute point gain on the micro-F1 score over the existing fully-supervised generative SOTA.

NYT: comprises 56k training examples. In this case we generate CoT explanations via GPT-3 for only a subset of 25k examples (about half of the train set), due to its large size and the associated cost. We fine-tune the model with a batch size of 4, learning rate 2e-5 for

4 epochs and then evaluate using the best performing checkpoint on the validation set. We obtain a 3.37 point gain on the micro-F1 score over the existing fully-supervised SOTA.

In sum, **fine-tuning Flan-T5 (large) with both train labels and CoT explanations produced by GPT-3 yields SOTA performance across RE datasets by a considerable (5–10 points micro-F1) margin** (Figure 1).

4.2.3 "Fully Supervising" Flan with GPT-3—Above we showed that Flan-T5 (large) outperforms existing RE methods by substantial margins when trained using CoTs from GPT-3. Now we ask whether we can take this approach of distillation from GPT-3 even further by eliciting *both labels and CoT explanations* from GPT-3 in a few-shot setting, and then using these to train Flan-T5. That is, above we used the reference labels for training, whereas here we use "labels" produced by GPT-3 given just a handful (10s) of training instances as shots. We run this experiment only on CoNLL due to the cost of processing datasets in this way (which requires running few shot inference in GPT-3 over entire *training* sets).

To generate the targets in this case, we start with an instructional prefix and 12 training instances from CoNLL and their corresponding human-written explanations; this is the same setup as the in-context GPT-3 model (Table 1 2.b.), though here we apply this to the training instances. We then prompt GPT-3 on all training instances except for the 12 shots to produce pseudo labels (relations) and associated CoT explanations.

Using this new *GPT-generated training data*, we again fine-tune Flan-T5 (Large) as described above (Section 4.2.2), and evaluate it on the validation set. This approach marginally outperforms the existing fully-supervised SOTA (Huguet Cabot and Navigli, 2021), but underperforms fine-tuning Flan with references references and GPT-generated explanations (Table 2, 2.c.).

5 Related work

Standard NLP methods for identifying relations in free text have included Conditional Random Fields (Lafferty et al., 2001), structured SVMs (Tsochantaridis et al., 2004), and more recently, training large deep learning models with a joint objective (Eberts and Ulges, 2021, 2019a; Wang and Lu, 2020) to identify entities and relations simultaneously. More recently, the rise of massive language models (Radford and Narasimhan, 2018; Radford et al., 2019; Brown et al., 2020a) has also motivated research into prompt-based learning methods for structured prediction (Wang et al., 2022).

5.1 Relation extraction with pre-trained LMs

Several recently proposed RE approaches (which we have built upon here) have proposed addressing the task using conditional generative models to output string encodings—i.e., linearized forms—of target relations (Zeng et al., 2018, 2020; Nayak and Ng, 2020; Huguet Cabot and Navigli, 2021). Paolini et al. (2021) proposed a framework that formulated many structured prediction tasks, including relation extraction, as a seq2seq problem where they decode outputs into structured information. Huguet Cabot and Navigli (2021) extended this

line of work by training a SOTA BART-style (Lewis et al., 2020) model specifically for relation extraction using a unique triplet linearization strategy. Beyond these task-specific models, Wang et al. (2022) proposed a task-agnostic structured pre-training scheme which enables zero-shot transfer to several structured prediction tasks.

These past efforts focussed on *solely* fine-tuning seq2seq models, adopting standard supervised approaches to learning to generate the relations expressed in a given input. (REBEL incorporated a pretraining scheme designed for RE (Huguet Cabot and Navigli, 2021), but this was in addition to a fine-tuning step.) In this work we also evaluate the ability of large language models to perform *few-shot* relation extraction via in-context learning; to our knowledge this is the first such evaluation for RE specifically, although few-shot learning more generally is an active sub-area of research.

5.2 Few Shot In-Context Learning

Few shot in-context learning entails incorporating a few training examples into model prompts, effectively "learning" via the activations induced by passing these examples through the network at inference time. This has the advantage of completely forgoing model weight updates, which can be costly for LLMs (Wang et al., 2021). An active area of research concerns such cross-task generalization capabilities (Ye et al., 2021; Wei et al., 2022a; Min et al., 2022; Xu et al., 2022) of LLMs where a model learns a new, previously-unseen task efficiently with just a few examples. Chen et al. (2022) also proposed a self-supervised objective as an intermediate stage between pre-training and downstream few-shot learning. Recent work on few shot in-context learning has largely focused on the selection (Liu et al., 2022) and ordering (Lu et al., 2022a) of exemplars included in the prompt provided to the model.

6 Conclusions and Future Directions

We have evaluated the capabilities of modern large language models (LLMs)—specifically GPT-3 and Flan T5 (Large)—on the task of Relation Extraction (RE). We found that, when evaluated carefully, GPT-3 performs comparably to fully supervised state-of-the-art (SOTA) models, given only 10s of examples. We then proposed a distillation technique in which we augmented target RE labels with *Chain of Thought* (CoT) style explanations elicited from GPT-3 and used this to fine-tune Flan-T5; this yielded SOTA performance across all datasets considered, often by wide margins (5–10 points in F1). Our results suggest that where feasible, LLMs should be a standard baseline for RE.

Future directions

We have left several avenues open for further exploration. For example, evaluating LLMs like GPT-3 for RE required collecting manual annotations to identify ostensible "false positive" and "false negative" model outputs which were in fact accurate. Designing models to automate this evaluation might provide similar reliability without the accompanying costs; we provide preliminary work in this direction through the use of simple BERT-style classifiers in Appendix D.

Limitations

We have demonstrated that across three standard RE datasets, LLMs achieve SOTA results. In particular, GPT-3 yields such performance even given only 10s of training sample for in-context learning. We then showed that we can similarly achieve SOTA performance with the much smaller (and open-source) Flan T5 (Large) model, when trained using CoT generations produced by GPT-3. We also highlighted key challenges for evaluation in this setting.

But there are important limitations to these contributions. First, here we considered three standard RE datasets with binary relations but—as we discussed—we excluded more complex RE datasets. For example, we did not consider corpora containing *n*-ary relations between entities (Taboureau et al., 2010). We were also unable to run experiments on datasets with lengthy texts and a large number of relations, such as DocRED (Yao et al., 2021), due to the necessary prompt lengths for such inputs.

Second, while we found that CoT-style explanations generated by GPT-3 can be fruitfully used as additional supervision to fine-tune smaller language models, we made no attempt to evaluate the *quality* of these generated explanations which may have an impact on the model performance.

Third, we did not fine-tune GPT-3 on the RE datasets, mainly due to the cost of doing so. It is likely that a fine-tuned GPT-3 would yield performance superior to the results we achieved with Flan T5 (which constitute current SOTA). But, in addition to the costs necessary for fine-tuning this model, the resultant weights would not be accessible to run locally in any case; one would have access to it only via the OpenAI interface, which motivated our decision to fine-tune the smaller and open-source Flan T5 instead.

Finally, we *only* experiment with datasets curated in the English language and therefore, we do not know that the issues we have highlighted could replicate in the same way in other languages.

Ethics Statement

Our work required an extensive manual annotation and evaluation process which involved using Amazon Mechanical Turk. Turk requires we pay workers *per annotation*, so we have to estimate the time required for each task. To do so, we (the authors) carried out a small number of these annotations ourselves to determine fair approximate hourly compensation. We then set the price per annotation such that it averages out to $15/hour (we pay this rate irrespective of geographic location of the workers). We also provided our recruited AMT workers 20% additional time per annotation.

Acknowledgements

This work was supported in part by the National Institutes of Health (NIH) under the National Library of Medicine (NLM) grant R01LM012086 and by the National Science Foundation (NSF) grant III-1750978.

A: Datasets

We considered and conducted the evaluation of our methods on the following datasets. Basic data statistics are also reported in Table 1.

ADE

Adverse Drug Events (Gurulingappa et al., 2012) contains binary relations of (drug, adverse event) pairs. Drugs and adverse events are the only two entity types. This dataset provides a 10-fold split.

CONLL04

The CoNLL04 consists of sentences from news articles that were annotated for the mentioned entities and relations between entities (Roth and Yih, 2004). It includes four entity types (*PER, ORG, LOC, OTH*) and five possible relations (*KILL, WORK_FOR, LIVE_IN, LOCATED_IN, ORG_BASED_IN*).

NYT

The NYT comprises sentences sampled from New York Times news articles published between 1987 and 2007 (Riedel et al., 2010). The data was distantly annotated with relations triplets from FreeBase. We use a processed version of NYT (Zeng et al., 2018) containing three overlapping entity types (*LOC, PER, ORG*) and 24 relation types.

DocRED

Originally designed as a relation classification task, DocRED (Yao et al., 2019) differs considerably from the other datasets considered in this work in two important ways: (1) It comprises long texts which feature relations between entities at a *document-level*; (2) It contains annotations for 6 entity types and 96 relation types, with an average of 19.9 entities and 19.5 relation instances per document.

B: Models and Reproducibility

We provide average micro metrics over 5 seeds across each dataset in Table 3. On Flan-T5-Large, where we do fine-tuning, some hyperparameters were manually tuned but most left at their default values. The final values for the ones that were manually tuned are provided in Table 4.

We perform all experiments with a single NVIDIA Quadro RTX 8000 with 64GB of RAM on an Intel Xeon E502680v4 (2.4GHz).

B.1 Costs ($$$)

We provide details on the costs we incurred while running experiments on GPT-3 in Table 5.

C: Prompts

We use the following prompt elements as few-shot exemplars corresponding to each dataset in our evaluation. Inputs and target references are directly extracted from the original training sets while the explanations are human-written and were added when necessary for the experiments described in section 3 and 4.

ADE

Example Instructional Prefix: List all [drug, adverse effects] pairs in the TEXT provided below.
TEXT: We report on three observations of parkinsonian patients with levo-dopa-induced diphasic dyskinesias, who received subcutaneous apomorphine to reduce the duration of abnormal movements.
Relations: [[‘levo-dopa’, ‘diphasic dyskinesias’]]
Explanation: levo-dopa induced diphasic dyskinesias in parkinsonian patients.<s>

TEXT: A girl with cystic fibrosis and cyclic neutropenia developed an erythematous papular eruption without fever or neutrophilia 7 months after commencing therapy with G-CSF.
Relations: [[‘G-CSF’, ‘erythematous papular eruption’]]
Explanation: G-CSF therapy caused erythematous papular eruption in a girl with cystic fibrosis.<s>

TEXT: Hypersensitivity to carboplatin is a rare but real complication of therapy and should be considered in patients presenting with hyperacute changes on ECG whilst receiving carboplatin therapy.
Relations: [[‘carboplatin’, ‘hyperacute changes on ECG’], [‘carboplatin’, ‘Hypersensitivity’]]
Explanation: Patients who undergo carboplatin therapy are prone to hypersensitivity and hyperacute changes on their ECG.<s>

TEXT: The diagnosis of hypothermia was delayed until it was apparent for several days but resolved with the discontinuation of risperidone and continuation of clozapine.
Relations: [[‘risperidone’, ‘hypothermia’]]
Explanation: risperidone caused hypothermia since it was resolved with its discontinuation.<s>

TEXT: Eighty-two patients with various malignancies who received imipenem/

cilastatin 143 times for neutropenic fever between March 1994 and October 1999 in Department of Pediatric Oncology, Gazi University, were identified.
Relations: [[‘cilastatin’, ‘neutropenic fever’], [‘imipenem’, ‘neutropenic fever’]]
Explanation: Patients who received either cilastatin or imipenem were identified with neutropenic fever.<s>

TEXT: This increase when clozapine was switched to risperidone and vice versa is consistent with our previous report of elevated serum triglyceride levels in clozapine-treated patients.
Relations: [[‘clozapine’, ‘elevated serum triglyceride levels’]]
Explanation: There was a report of elevated serum triglyceride levels in clozapine-treated patients.<s>

TEXT: Autopsy findings were consistent with bleomycin and oxygen-induced pulmonary damage.
Relations: [[‘bleomycin’, ‘pulmonary damage’], [‘oxygen’, ‘pulmonary damage’]]
Explanation: Both bleomycin and oxygen caused pulmonary damage in the autopsy findings.<s>

TEXT: CD4 T-lymphocyte depletion, myelosuppression, and subsequent severe infections are the major side effects of fludarabine phosphate therapy.
Relations: [[‘fludarabine phosphate’, ‘CD4 T-lymphocyte depletion’], [‘fludarabine phosphate’, ‘myelosuppression’], [‘fludarabine phosphate’, ‘severe infections’]]
Explanation: Following major side-effects are known of fludarabine phosphate therapy, CD4 T-lymphocyte depletion, myelosuppression, and severe infections.<s>

TEXT: OBJECTIVE: To describe a case of severe skin necrosis resulting from peripheral intravenous administration of low-dose vasopressin in a patient with catecholamine-resistant septic shock.
Relations: [[‘vasopressin’, ‘skin necrosis’]]
Explanation: A case of skin necrosis was described after administration of low-dose vasopressin.<s>

TEXT: In vitro inhibition of hematopoiesis in a patient with systemic sclerosis treated with D-penicillamine.
Relations: [[‘D-penicillamine’, ‘inhibition of hematopoiesis’]]

Explanation: Patient treated with D-penicillamine had in vitro inhibition of hematopoiesis.<s>

TEXT: PURPOSE: We report an unusual paradoxical effect of brimonidine.
Relations: [[‘brimonidine’, ‘paradoxical effect’]]
Explanation: paradoxical effect of brimonidine was reported.<s>

TEXT: Hepatocellular damage following therapeutic intravenous iron sucrose infusion in a child.
Relations: [[‘iron sucrose’, ‘Hepatocellular damage’]]
Explanation: Hepatocellular damage occurred in a child after infusion of iron sucrose.<s>

CoNLL

Examplee Instructional Prefix: List the relations of the types [OrgBased In, Work For, Located In, Live In, Kill] among the entities [PERSON, LOCATION, ORGANIZATION, OTHER] in the given text and provide a reasonable explanation.
TEXT: “If it does not snow, and a lot, within this month we will have no water to submerge 150,000 hectares (370,500 acres) of rice”, said Bruno Pusterla, a top official of the Italian Agricultural Confederation.
Relations: [[‘Bruno Pusterla:Per’, ‘Work For’, ‘Italian Agricultural Confederation:Org’]]
Explanation: Bruno Pusterla is a top official of the Italian Agricultral Confederation.<s>

TEXT: Meanwhile, Shi Liming at the Institute of Zoology of Kunming found that pandas lack variety in their protein heredity, which may serve as one of the major reasons for pandas’ near extinction.
Relations: [[‘Shi Liming:Per’, ‘Work For’, ‘Institute of Zoology:Org’], [‘Institute of Zoology:Org’, ‘OrgBased In’, ‘Kunming:Loc’]]
Explanation: Shi Liming works for the Institute of Zoology, which is an organization based in Kunming.<s>

TEXT: The viewers of “JFK” and “The Men Who Killed Kennedy” never learn about these facts, nor do they ever learn about all of the other massive body of evidence that conclusively proves beyond a reasonable doubt that Oswald was the lone gunman who killed President Kennedy and Officer Tippit and that there was no coverup by Earl Warren or by the Warren Commission.
Relations: [[‘Oswald:Per’, ‘Kill’, ‘President Kennedy:Per’], [‘Oswald:Per’,

‘Kill’, ‘Officer Tippit:Per’]]
Explanation: Oswald was the lone gunman who killed President Kennedy and Officer Tippit.<s>

TEXT: PURCHASE, N.Y .
Relations: [[‘PURCHASE:Loc’, ‘Located In’, ‘N.Y.:Loc’]]
Explanation: PURCHASE is a place located in N.Y..<s>

TEXT: BELGRADE, Yugoslavia (AP)
Relations: [[‘BELGRADE:Loc’, ‘Located In’, ‘Yugoslavia:Loc’], [‘AP:Org’, ‘OrgBased In’, ‘BELGRADE:Loc’], [‘AP:Org’, ‘OrgBased In’, ‘Yugoslavia:Loc’]]
Explanation: City of BELGRADE is located in Yugoslavia and AP is an organization based in BELGRADE, Yugoslavia.<s>

TEXT: Rome is in Lazio province and Naples in Campania.
Relations: [[‘Rome:Loc’, ‘Located In’, ‘Lazio:Loc’], [‘Naples:Loc’, ‘Located In’, ‘Campania:Loc’]]
Explanation: Rome is a place located in Lazio and Naples is a place located in Campania.<s>

TEXT: (By ITAR-TASS correspondent Mikhail Shevtsov)
Relations: [[‘Mikhail Shevtsov:Per’, ‘Work For’, ‘ITAR-TASS:Org’]]
Explanation: Mikhail Shevtsov is a correspondent for the ITAR-TASS.<s>

TEXT: In the communique, the Group of Rio states that the Haitian crisis can be resolved only if unrestricted respect is shown for the Governor’s Island Agreement which calls for the prompt return of Haitian President Jean Bertrand Aristide to the exercise of his constitutional powers in Haiti.
Relations: [[‘Jean Bertrand Aristide:Per’, ‘Live In’, ‘Haiti:Loc’]]
Explanation: Jean Bertrand Aristide was the president of Haiti and therefore lived in Haiti.<s>

TEXT: Moscow ITAR-TASS
Relations: [[‘ITAR-TASS:Org’, ‘OrgBased In’, ‘Moscow:Loc’]]
Explanation: ITAR-TASS is an organization based in Moscow.<s>

TEXT: King rose to prominence after Mrs. Parks ‘ action in December 1955 in Montgomery , Ala. , set the stage for a boycott and subsequent

demonstrations that caught the nation by surprise.
Relations: [[‘Mrs. Parks:Per’, ‘Live In’, ‘Montgomery:Loc’], [‘Mrs.Parks:Per’, ‘Live In’, ‘Ala.:Loc’], [‘Montgomery:Loc’, ‘Located In’, ‘Ala.:Loc’]]
Explanation: Mrs. Parks actions were in Montgomery, Ala., where she lived. It can be derived that Montgomery is located in Ala..<s>

TEXT: Sirhan says he was the lone assassin but can’t remember shooting Kennedy.
Relations: [[‘Sirhan:Per’, ‘Kill’, ‘Kennedy:Per’]]
Explanation: Sirhan was the lone assassin in the Kennedy assassination.<s>

TEXT: In Colorado, 13 inches of snow in Denver Wednesday prompted officials to close Interstate 270 temporarily.
Relations: [[‘Denver:Loc’, ‘Located In’, ‘Colorado:Loc’]]
Explanation: Denver officials closed Interstate 270 in Colorado, consequently we can see that Denver is located in Colorado.<s>

TEXT: Edward Marks, an official with the Montgomery County Democratic Party, argued that if Ms. Toth is not interested in the job, “she should get out”.
Relations: [[‘Edward Marks:Per’, ‘Work For’, ‘Montgomery County Democratic Party:Org’]]
Explanation: Edward Marks is an official that works for the Montgomery County Democratic Party.<s>

NYT

TEXT: Massachusetts ASTON MAGNA Great Barrington; also at Bard College, Annandale-on-Hudson, N.Y., July 1-Aug.
Relations: [[‘Annandale-on-Hudson’, ‘/location/location/contains’, ‘Bard College’]]
Explanation: Annandale-on-Hudson is a location in N.Y. that contains Bard College.<s>

TEXT: It will be the final movie credited to Debra Hill, a film producer and native of Haddonfield, who produced “Halloween” and was considered a pioneering woman in film.
Relations: [[‘Debra Hill:Per’, ‘/people/person/place-of-birth’, ‘Haddonfield:Loc’]]

Explanation: Debra Hill was a film producer born (native of) in Haddonfield.<s>

TEXT: Under pressure from Mr. Kerkorian and other disgruntled shareholders, Mr. Wagoner started talks on Friday in Detroit with Carlos Ghosn, the chief executive of Renault and Nissan.
Relations: [[‘Carlos Ghosn:Per’, ‘/business/person/company’, ‘Renault:Org’]]
Explanation: Carlos Ghosn is a business person (chief executive) associated with Renault and Nissan.<s>

TEXT: Mr. Ferrer still holds commanding leads over the other two Democrats in the race – United States Representative Anthony D. Weiner of Brooklyn and Queens, and City Council Speaker Gifford Miller – and is also ahead of Mayor Michael R. Bloomberg in most polls.
Relations: [[‘Anthony D. Weiner:Per’, ‘/people/person/place-lived’, ‘Brooklyn:Loc’], [‘Anthony D. Weiner:Per’,’/people/person/place-lived’, ‘Queens:Loc’]]
Explanation: Anthony D. Weiner is a person representing Brooklyn and Queens, therefore we can infer he lives in those places.<s>

TEXT: Quebec, Canada’s second most populous province, after Ontario, has not decided to go that far.
Relations: [[‘Ontario:Loc’, ‘/location/administrative-division/country’, ‘Canada:Loc’], [‘Canada:Loc’, ‘/location/location/contains’, ‘Ontario:Loc’], [‘Canada:Loc’, ‘/location/country/administrative-divisions’, ‘Ontario:Loc’]]
Explanation: Ontario is a place located in the administrative divisions of the country Canada. Quebec is Canada’s second most populous province and hence, Canada is a place that contains Quebec.<s>

TEXT: And Abu Izzadeen , who converted to Islam at 17 and heads another successor group to Al Muhajiroun, called Al Ghurabaa, called suicide bombing “martyrdom operations”.
Relations: [[‘Abu Izzadeen:Per’, ‘/people/person/religion’, ‘Islam:Org’]]
Explanation: Since Abu Izzadeen converted to Islam at the age of 17, we can infer that this is a person who belongs to the religion of Islam.<s>

TEXT: And yet, despite the success of its exhibitions, the institute remains something of a strange hybrid: located southeast of Notre-Dame, in a striking building designed by Jean Nouvel, it has operated since 1987 as a partnership between France and 22 Arab countries.

Relations: [[‘Jean Nouvel:Per’, ‘/people/person/nationality’, ‘France:Loc’]]
Explanation: Jean Nouvel was a french designer and we can derive his nationality/citizenship as French or France.<s>

TEXT: They could have done it Sunday, when we were closed,” said Joseph Bastianich, who owns Del Posto with his mother, Lidia Bastianich, and the chef, Mario Batali.
Relations: [[‘Lidia Bastianich:Per’, ‘/people/person/children’, ‘Joseph Bastianich:Per’]]
Explanation: Joseph Bastianich owns Del Posto with his mother Lidia Bastianich.<s>

TEXT: A French court sentenced six Algerian-French men to prison terms of up to 10 years on Tuesday for their role in a 2001 plot to attack the United States Embassy in Paris , closing the books on one of France ‘s most serious terrorist cases.
Relations: [[‘Paris:Loc’, ‘/location/ administrative-division/country’, ‘France:Loc’], [‘France:Loc’, ‘/location/location/contains’, ‘Paris:Loc’], [‘France:Loc’, ‘/location/country/ administrative-divisions’, ‘Paris:Loc’], [‘France:Loc’, ‘/location/country/capital’, ‘Paris:Loc’]]
Explanation: Paris is located in the administrative divisons of the country France. Consequently, France is a place that contains Paris. US embassies are located in the capital of countries, therefore it can be inferred that Paris is the capital of France.<s>

TEXT: Anheuser-Busch, which has been the exclusive beer sponsor for the Super Bowl since 1989, will do so again for the Super Bowls in 2007 and 2010 on CBS and in 2008 and 2011 on Fox Broadcasting, said Anthony T. Ponturo, vice president for global media and sports marketing at Anheuser-Busch in St.Louis.
Relations: [[‘Anheuser-Busch:Org’, ‘/business/company/place-founded’, ‘St. Louis:Loc’], [‘St. Louis:Loc’, ‘/location/location/contains’, ‘Anheuser-Busch:Org’]]
Explanation: Anheuser-Busch is a business that was founded in St. Louis. Consequently, St. Louis is a place that contains Anheuser-Busch.<s>

TEXT: Somewhat chastened by his retreat in the polls, Mr. Blair acknowledged that Britons had turned against him in part over accusations that he led them into a war in Iraq on dubious legal grounds and on the false premise that Saddam Hussein presented a direct threat because of a supposed arsenal of unconventional weapons that was never found.”

Relations: [[‘Saddam Hussein:Per’, ‘/people/deceased-person/place-of-death’, ‘Iraq:Loc’], [‘Saddam Hussein:Per’, ‘/people/person/place-of-birth’, ‘Iraq:Loc’], [‘Saddam Hussein:Per’, ‘/people/person/nationality’, ‘Iraq:Loc’]]
Explanation: Saddam Hussein was killed in Iraq. His place of birth was also Iraq. We can infer that his nationality was Iraq.<s>

TEXT: Rupert Murdoch and John C. Malone , who have wrangled for two years over Mr. Malone ‘s challenge to Mr. Murdoch ‘s control of the News Corporation , have made peace . Relations: [[‘Rupert Murdoch’, ‘/business/person/company’, ‘News Corporation’], [‘News Corporation’, ‘/business/company/founders’, ‘Rupert Murdoch’]] Explanation: Rupert Murdoch is a business person associated with News Corporation, which was a company founded by Rupert Murdoch.<s>

TEXT: Manhattan, especially the East Village , has long been well stocked with cheap and raucous yakitori places that specialize in skewers and beer.
Relations: [[‘Manhattan:Loc’,’/location/location/contains’, ‘East Village:Loc’], [‘East Village:Loc’, ‘/location/neighborhood/neighborhood-of’, ‘Manhattan:Loc’]]
Explanation: East Village is a neighborhood in Manhattan.<s>

TEXT: HEADING OUT – Sanford I. Weill stepped down as chairman of Citigroup , the worldwide financial supermarket he had meticulously and single-mindedly stitched together through dozens of mergers and acquisitions.
Relations: [[‘Citigroup:Org’, ‘/business/company/advisors’, ‘Sanford I. Weill:Per’]]
Explanation: Citigroup is a business company who was associated with (advised by) Sanford I. Weill.<s>

TEXT: He had decided to use the premiere to publicize the issue; his plan was to invite the neighborhood’s Russian speakers to sign a petition against piracy, a common practice at the area’s Russian-language video outlets, which sell films and music from Russia and by Russian immigrants in the United States.
Relations: [[‘Russian:Per’, ‘/people/ethnicity/geographic-distribution’, ‘Russia:Loc’]]
Explanation: Russian is an ethnicity in United States associated with immigrants who came from the geographic distribution of Russia.<s>

TEXT: In 1995, Cleveland successfully lobbied to have the name Cleveland Browns stay in that city after that venerable franchise's owner, Art Modell, opted to move it to Baltimore.
Relations: [['Cleveland:Loc', '/sports/sports-team-location/teams', 'Cleveland Browns:Org'], ['Cleveland Browns:Org', '/sports/sports-team/location', 'Cleveland:Loc']]
Explanation: Cleveland Browns is the sports franchise located in Cleveland, consequently Cleveland's sports team is Cleveland Browns.<s>

TEXT: Mr. Fields, speaking from vacation in France, added, "That a mogul like Sumner Redstone could make a statement so vicious, so pompous, so petulant as that he didn't want to make a deal with Tom Cruise because of his personal conduct – it tells you more about Sumner Redstone and Viacom, than about Tom Cruise".
Relations: [['Sumner Redstone:Per', '/business/company-shareholder/major-shareholder-of', 'Viacom:Org']]
Explanation: Sumner Redstone is a major shareholder of the company Viacom.<s>

TEXT: It is a room of paintings by Leonard Peltier , a citizen of the Anishinabe and Dakota and Lakota nations who is serving two consecutive life terms in Pennsylvania for the murder of two F.B.I. agents on the Pine Ridge Reservation in South Dakota.
Relations: [['Leonard Peltier:Per', '/people/person/ethnicity', 'Lakota:Per'], ['Lakota:Per', '/people/ethnicity/people', 'Leonard Peltier:Per']]
Explanation: Leonard Peltier is a member of the Lakota native-american tribe and consequently belongs to that ethnic group.<s>

TEXT: INSIDE THE N.B.A. Correction : February 9 , 2006 , Thursday A sports article on the Spotlight page on Sunday about Dick Bavetta , a longtime referee in the National Basketball Association, misstated the number he was approaching to set the record for regular-season games worked.
Relations: [['Dick Bavetta:Per', '/people/person/profession', 'National Basketball Association:Org']]
Explanation: Dick Bavetta is a person who's profession is that of a referee in National Basketball Association.<s>

TEXT: Now the United States Postal Service may be displaying a similar rebellious streak : tomorrow at the huge Sturgis motorcycle rally in the Black Hills of South Dakota, the Postal Service will issue a set of four

```
stamps that depict classic American bikes.
Relations: [['United States Postal Service:Org',
'/business/company/industry', 'Postal Service:Org']]
Explanation: United States Postal Service is a business company in the
industry of providing postal services.<s>
```

D: Learning to Identify *False* False Positives and Negatives

As discussed in the main paper, one common problem across datasets in generative RE is evaluation, given that LMs are flexible in how they might express entities and relations. Prior work in RE has tended rely on standard metrics to quantify performance (precision, recall, micro-F1). These rely on matching *classified* (or in our case, *generated*) labels to reference labels to calculate the number of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs).

Prior to the introduction of LLMs for generative RE, Taillé et al. (2020) attempted to unify evaluation and provide useful guidelines around issues associated with prior methods and how different evaluation strategies rendered an accurate comparison infeasible. They broadly recommended the use of a *strict* evaluation scheme where for a relation triplet to be considered correct, the head and tail entity surface forms must be an exact match, as well as their corresponding types (when available). While this provides a standardized framework for traditional models where entities and and relations are hard *classification* labels, in a generative setting we often find that LLMs, under varying levels of supervision, produce relation triplets (or pairs) that do not correspond exactly to their reference counterparts, but are nonetheless correct upon manual review. Consider the following example from CoNLL in Figure 2

```
Text: On Friday, U.S. Ambassador Vernon A. Walters... fuselage.
Gold Reference: [(Vernon A. Walters, 'Live In', U.S.)]
Generated Relations: [[Vernon A. Walters, 'Works For', U.S.]]
```

In this example, one can reasonably infer that Vernon A. Walter is a U.S. Ambassador. Therefore, by definition a U.S. diplomat to another country cannot live inside the U.S., but such a person must work for the U.S. (commonsense dictates that a diplomat would work for a specific country).

To achieve a more accurate characterization of how LLMs perform on generative RE tasks, we hired human annotators on Amazon Mechanical Turk[5] to manually re-assess all ostensible FPs and FNs from each of our datasets. To control for quality and recruit annotators we ran pilot experiments on 50 instances of pre-annotated data.[6] We required AMT workers to have an overall approval rating of $>95\%$ irrespective of geographic region. Based on these initial set of results we hired a total of 9 workers who reliably followed our

[5]We set the payrate to average at $15/hour using time estimates informed by pilot experiments.
[6]These instances used in pilot experiments were annotated by a graduate student familiar with this research.

instructions. Recruited workers were paid periodic bonuses (equivalent to one hour of pay) based on the quality of their annotations.

To identify potentially faulty "false positives", we provided annotators with the input text along with the relation identified as a FP, and ask the following question: "Can the given given relation be reasonably derived from the text?". Similarly, to identify erroneous "false negatives", we provide annotators with the input text, the full set of generated labels, the *ostensible* FN from the reference set, and ask: "Can the reference relation triplet (or pair) be inferred from the generated set of relations?". Each instance was annotated by three different AMT workers, and we considered a potential FP/FN to be inaccurate only when **all** annotators agree on a label.[7] We provide specific examples of FPs and FNs in Tables 8 and 7. We summarize the dataset-specific findings in Table 6.

In light of these findings, we make a first effort in using simple, learned models to classify false-positives/negatives in generative RE. We experiment with fine-tuned BERT (Devlin et al., 2019) classifier to classify "false positives" and "false negatives" as being accurate designations (or not). For FPs, we concatenate the input with a generated relation pair/triplet (*potential* FP) and classify using the `[CLS]` token -

```
[CLS] Input Text [SEP] Potential FP
```

Similarly, for FNs we concatenate the input text with a *potential FN* and the full set of generated labels, and classify using the `[CLS]` token -

```
[CLS] Input Text [SEP] Potential FN [SEP] Generated Labels
```

We analyze the effectiveness of this approach in Figure 4 using the AUC-ROC. We find that this approach is most effectiveness in identifying potential potential false positives for CoNLL (AUC 0.88), while being least effective at identifying false negatives for CoNLL (AUC 0.73). This suggests that learning to identify erroneous "false positives" and "false negatives" may be a promising avenue to facilitate accurate automated evaluation of generative LLMs for RE.

D.1 List of out-of-domain relation-types generated by Flan during Few-Shot Prompting with CoNLL

Assassinates, Purpose, isPartOf, Mother, Spouse, President, date, killed, Summer, Works_at, Sentenced_To_Death, Source, Statue, Secretary, Born, Year, Born_in, Day, Place, Number_Of_Passengers, Callers, Governor, Hometown, has_a_leader, is_a_member_of, Nickname, is_part_of, Office, Rank, Works_For, WorkedFor, Worked_For, Killed_By, Piano, Term, Sentence, Person, Movie, Said, Brother, Date_of_Death, Type, Death_Penalty, assassination_date, Worked_for, capital, Killed, Killing, Occupation, Crime,

[7] We observe a high degree of agreement among the annotators with a Fliess *kappa* of 0.83

Years_in_use, Org, Education, Order_to_ignore, Assassination, Location, Officer, language, former_name, Total_acres, Age, Cause, Chairman, worked_for, Son, Staff_name, departure, Capsule_name, Operator, Spin-off, Owner, located_in, theory, Birth_Place, on_duty_with, City, Top_Leader, Director, structure, Known_as, former_chief_executive, Works_for, Native_name, Percentage, department, Component, reminds_someone_of, Sex, Bank, Appointed_By, Activity, Title, has_a_river_name, Size, Office_Space, Part, Kingdom, Attached_to, Death_Place, Years_on_the_Supreme_Court, Assassin, location, Newspaper, City,, island, Employee, Friend, Native_Son, Speaker, Visitor, Date, Aircraft, channel, Sale_to, Creditor, Client, Nationality, Flight_Status, assassinater, on_behalf_of, Shot_By.

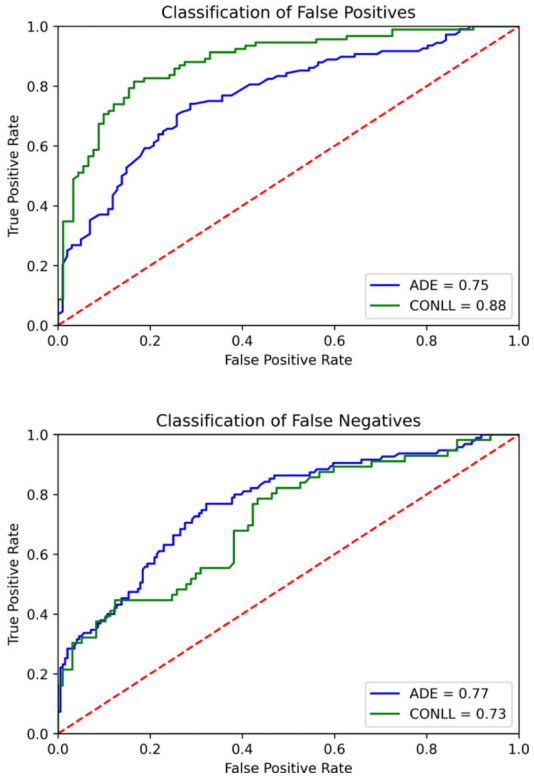

Figure 4:
AUC plots for FPs and FNs.

Table 3:

Average micro metrics over 5 seeds for the test sets (10-folds for ADE).

Model	Data	P	R	F-1
Few-Shot In-Context Prompting GPT-3	ADE	80.85	84.54	82.66
	CoNLL	78.31	74.82	76.53
	NYT	66.63	70.58	68.55
Vanilla Fine-Tune Flan-T5-Large	ADE	89.11	77.93	83.15
	CoNLL	78.81	72.05	75.28

Model	Data	P	R	F-1
	NYT	91.82	90.25	91.03
Fine-Tune Flan on GPT-3-generated CoT	ADE	91.74	92.60	92.17
	CoNLL	81.22	80.31	80.76
	NYT	95.49	94.97	95.23
Fine-Tune Flan w/CoT Explanations and Reference labels generated from GPT	CoNLL	76.41	75.85	76.13

Table 4:

Hyperparameters and compute time for the fully fine-tuned Flan models (corresponding to main results table 2).

Model	Data	Batch Size	Warm-up	Learning Rate	Time/Epoch (minutes)	Max Epochs
Vanilla Fine-Tune Flan-T5-Large	ADE	8	10%	3e-5	36	6
	CoNLL	4	12%	3e-5	22	10
	NYT	4	12%	2e-5	99	4
Fine-Tune Flan on GPT-3-generated CoT	ADE	8	10%	3e-5	38	6
	CoNLL	4	12%	3e-5	28	10
	NYT	4	12%	2e-5	107	4
Fine-Tune Flan w/CoT Explanations and Reference labels generated from GPT	ADE	8	10%	3e-5	37	6
	CoNLL	4	12%	3e-5	28	10
	NYT	4	12%	2e-5	109	4

Table 5:

Summary of costs incurred by prompting and using GPT-3 as a labeler for RE.

Experiment	Data	Cost (US$)
Evaluation of Few-Shot In-Context Prompting	ADE	64.91
	CoNLL	19.24
	NYT	238.70
Generation of CoT Explanations (Training Set)	ADE	93.96
	CoNLL	44.20
	NYT	983.86
Generation of Target Labels + CoT Explanations	CoNLL	86.41

Table 6:

Number of inaccurate false positives (FNs) and false negatives (FNs) identified during automated evaluation in GPT-3 labelled outputs under the in-context few-shot prompting setting.

Data	Inaccure FPs / Total FPs	Inaccurate FNs / Total FNs
ADE	108 / 209	136 / 417
CoNLL	92 / 183	56 / 152

Table 7:

Sample of inaccurate false *negatives* identified by human annotators for each dataset. Examples from CoNLL and ADE were generated from GPT-3, while those in NYT were generated by Flan.

Dataset	Input	Detected FN (From the Gold Reference Set)	Full Set of Generated Relations
CoNLL04	• The three reactors, all at the Savannah River Plant, in Aiken, S.C., have been shut down since last April undergoing changes to make them safer.	[Savannah River Plant, Located_In, S.C.]	[Savannah River Plant, Located_In, Aiken] [Aiken, Located_In, S.C.]
	• They will also be cleaning the car Oswald drove on the day Kennedy was shot and the ambulance that took Oswald to hospital after he was shot by Jack Ruby.	[‘Jack Ruby’, **‘Kill’**, ‘Oswald’]	[Jack Ruby, **Shoot**, Oswald]
ADE	• We report the case of an 11 -year-old female treated for mediastinal t-cell lymphoma who presented renal failure following the second cycle of high-dose methotrexate (hdmtx).	[‘hdmtx’, ‘renal failure’]	[‘methotrexate’, ‘renal failure’]
	• Four days after the initial injection of 3.6 mg of goserelin acetate, severe dyspnea developed due to worsening pleuritis carcinomatosa, which was considered as a flare-up.	[‘goserelin acetate’, ‘flare’]	[‘goserelin acetate’, ‘dyspnea’]
NYT	• This time, the president chose Father Leon to replace one of those clergymen, the Rev. Franklin Graham, who was filling in for his father, the Rev. Billy Graham, who was ill in 2001.	[‘Franklin Graham’, **‘people/ people/children’**, ‘Billy Graham’]	[‘Billy Graham’, **‘/people/ person/children’**, ‘Franklin Graham’]
	• The moves by Citigroup and Deutsche Bank are only the latest illustrations of the difficulty of retaining jobs in New York City and rebuilding the business district in Lower Manhattan.	[‘Citigroup’, ‘business/place/ founded’, ‘New York City’] [‘Lower Manhattan’, **‘/location/location/ contains’**, ‘New York City’]	[‘New York City’, **‘/location/ location/contains’**, ‘Lower Manhattan’]

Table 8:

Sample of inaccurate false *positives* (FPs) identified by human annotators for each dataset. Examples from CoNLL and ADE were generated from GPT-3, while those in NYT were generated by Flan.

Dataset	Input	Detected FPs (Generated)	Full Set of True Relations
CoNLL04	• Illinois Gov. James Thompson signed legislation last month scrapping Chicago's central school board next July, to be replaced by parentrun councils empowered to set budgets and hire and fire principals.	[James Thompson, **Work_For**, Illinois], [Chicago, **Located_In**, Illinois]	[[James Thompson, **Live_In**, Illinois]]
	• On Friday, U.S. Ambassador Vernon A. Walters displayed photographs of one Libyan jet showing shapes resembling missile pods on its wings and fuselage.	[Vernon A. Walters, **Work_For**, U.S.]	[[Vernon A. Walters, **Live_In**, U.S.]
ADE	• Background: how to best treat psychotic patients who have had past clozapine-induced agranulocytosis or granulocytopenia remains a problem.	[clozapine, agranulocytosis]	[[clozapine, granulocytopenia]]
	• Acute renal failure is a rare complication following the administration of intravenous immunoglobulin (ivig).	[ivig, acute renal failure]	[[immunoglobulin, acute renal failure]]
NYT	• Many people in Turkey have lost hopes in joining Europe and they are looking for other horizons , " said Onur Oymen , an opposition politician whose party is staunchly secular.	[Turkey, **location/ administrative_division/ country**, Europe]	[[Europe, **LOC_CONTAINS**, Turkey]]
	• To make his case , Dr. von Hagens invited two journalists to Dalian for a tour of his facility , which he said was the first center in China to preserve bodies.	[Dalian, **location/ administrative_division/ country**, China]	[[China, **location/ location/contains**, Dalian]]

References

Brown Tom, Mann Benjamin, Ryder Nick, Subbiah Melanie, Jared D Kaplan Prafulla Dhariwal, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, Agarwal Sandhini, Herbert-Voss Ariel, Krueger Gretchen, Henighan Tom, Child Rewon, Ramesh Aditya, Ziegler Daniel, Wu Jeffrey, Winter Clemens, Hesse Chris, Chen Mark, Sigler Eric, Litwin Mateusz, Gray Scott, Chess Benjamin, Clark Jack, Berner Christopher, Sam McCandlish Alec Radford, Sutskever Ilya, and Amodei Dario. 2020a. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.

Brown Tom B., Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared, Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, Agarwal Sandhini, Herbert-Voss Ariel, Krueger Gretchen, Henighan TJ, Child Rewon, Ramesh Aditya, Ziegler Daniel M., Wu Jeff, Winter Clemens, Hesse Christopher, Chen Mark, Sigler Eric, Litwin Mateusz, Gray Scott, Chess Benjamin, Clark Jack, Berner Christopher, Sam McCandlish Alec Radford, Sutskever Ilya, and Amodei Dario. 2020b. Language models are few-shot learners. ArXiv, abs/2005.14165.

Chen Mingda, Du Jingfei, Pasunuru Ramakanth, Mihaylov Todor, Iyer Srini, Stoyanov Veselin, and Kozareva Zornitsa. 2022. Improving in-context few-shot learning via self-supervised training. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3558–3573, Seattle, United States. Association for Computational Linguistics.

Chung Hyung Won, Hou Le, Longpre Shayne, Zoph Barret, Tay Yi, Fedus William, Li Yunxuan, Wang Xuezhi, Dehghani Mostafa, Brahma Siddhartha, Webson Albert, Shixiang Shane Gu Zhuyun

Dai, Suzgun Mirac, Chen Xinyun, Chowdhery Aakanksha, Alex Castro-Ros Marie Pellat, Robinson Kevin, Valter Dasha, Narang Sharan, Mishra Gaurav, Yu Adams, Zhao Vincent, Huang Yanping, Dai Andrew, Yu Hongkun, Petrov Slav, Chi Ed H., Dean Jeff, Devlin Jacob, Roberts Adam, Zhou Denny, Le Quoc V., and Wei Jason. 2022. Scaling instruction-finetuned language models.

Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eberts Markus and Ulges Adrian. 2019a. Span-based joint entity and relation extraction with transformer pre-training. ArXiv, abs/1909.07755.

Eberts Markus and Ulges Adrian. 2019b. Span-based joint entity and relation extraction with transformer pre-training. CoRR, abs/1909.07755.

Eberts Markus and Ulges Adrian. 2021. An end-to-end model for entity-level relation extraction using multi-instance learning. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3650–3660, Online. Association for Computational Linguistics.

Gurulingappa Harsha, Abdul Mateen Rajput Angus Roberts, Fluck Juliane, Hofmann-Apitius Martin, and Toldo Luca. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of biomedical informatics, 45 5:885–92. [PubMed: 22554702]

Cabot Pere-Lluís Huguet and Navigli Roberto. 2021. REBEL: Relation extraction by end-to-end language generation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lafferty John D., McCallum Andrew, and Pereira Fernando C. N.. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lewis Mike, Liu Yinhan, Goyal Naman, Ghazvininejad Marjan, Mohamed Abdelrahman, Levy Omer, Stoyanov Veselin, and Zettlemoyer Luke. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Liu Jiachang, Shen Dinghan, Zhang Yizhe, Dolan Bill, Carin Lawrence, and Chen Weizhu. 2022. What makes good in-context examples for GPT-3? In Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Lu Yao, Bartolo Max, Moore Alastair, Riedel Sebastian, and Stenetorp Pontus. 2022a. Fantastically ordered prompts and where to find them: Overcoming fewshot prompt order sensitivity. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Lu Yaojie, Liu Qing, Dai Dai, Xiao Xinyan, Lin Hongyu, Han Xianpei, Sun Le, and Wu Hua. 2022b. Unified structure generation for universal information extraction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Min Sewon, Lewis Mike, Zettlemoyer Luke, and Hajishirzi Hannaneh. 2022. MetaICL: Learning to learn in context. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Nayak Tapas and Ng Hwee Tou. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In AAAI Conference on Artificial Intelligence.

Paolini Giovanni, Athiwaratkun Ben, Krone Jason, Ma Jie, Achille Alessandro, ANUBHAI RISHITA, Santos Cicero Nogueira dos, Xiang Bing, and Soatto Stefano. 2021. Structured prediction

as translation between augmented natural languages. In International Conference on Learning Representations.

Radford Alec and Narasimhan Karthik. 2018. Improving language understanding by generative pretraining.

Radford Alec, Wu Jeff, Child Rewon, Luan David, Amodei Dario, and Sutskever Ilya. 2019. Language models are unsupervised multitask learners.

Riedel Sebastian, Yao Limin, and McCallum Andrew. 2010. Modeling relations and their mentions without labeled text. In ECML/PKDD.

Roth Dan and Yih Wen-tau. 2004. A linear programming formulation for global inference in natural language tasks. In Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004, pages 1–8, Boston, Massachusetts, USA. Association for Computational Linguistics.

Taboureau Olivier, Sonny Kim Nielsen Karine Audouze, Weinhold Nils, Daniel Edsgärd Francisco S. Roque, Kouskoumvekaki Irene, Bora Alina, Curpan Ramona, Thomas Skøt Jensen Søren Brunak, and Oprea Tudor I.. 2010. ChemProt: a disease chemical biology database. Nucleic Acids Research, 39:D367–D372. [PubMed: 20935044]

Taillé Bruno, Guigue Vincent, Scoutheeten Geoffrey, and Gallinari Patrick. 2020. Let's Stop Incorrect Comparisons in End-to-end Relation Extraction! In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3689–3701, Online. Association for Computational Linguistics.

Tsochantaridis Ioannis, Hofmann Thomas, Joachims Thorsten, and Altun Yasemin. 2004. Support vector machine learning for interdependent and structured output spaces. In Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04, page 104, New York, NY, USA. Association for Computing Machinery.

Wang Chenguang, Liu Xiao, Chen Zui, Hong Haoyun, Tang Jie, and Song Dawn. 2022. DeepStruct: Pretraining of language models for structure prediction. In Findings of the Association for Computational Linguistics: ACL 2022, pages 803–823, Dublin, Ireland. Association for Computational Linguistics.

Wang Jue and Lu Wei. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1706–1721, Online. Association for Computational Linguistics.

Wang Shuohang, Liu Yang, Xu Yichong, Zhu Chenguang, and Zeng Michael. 2021. Want to reduce labeling cost? GPT-3 can help. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wei Jason, Bosma Maarten, Zhao Vincent, Guu Kelvin, Adams Wei Yu Brian Lester, Du Nan, Dai Andrew M., and Le Quoc V.. 2022a. Finetuned language models are zero-shot learners. ArXiv, abs/2109.01652.

Wei Jason, Wang Xuezhi, Schuurmans Dale, Bosma Maarten, Chi Ed, Le Quoc, and Zhou Denny. 2022b. Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903.

Xu Hanwei, Chen Yujun, Du Yulun, Shao Nan, Wang Yanggang, Li Haiyu, and Yang Zhilin. 2022. Zeroprompt: Scaling prompt-based pretraining to 1, 000 tasks improves zero-shot generalization. CoRR, abs/2201.06910.

Yao Yuan, Du Jiaju, Lin Yankai, Li Peng, Liu Zhiyuan, Zhou Jie, and Sun Maosong. 2021. CodRED: A cross-document relation extraction dataset for acquiring knowledge in the wild. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4452–4472, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yao Yuan, Ye Deming, Li Peng, Han Xu, Lin Yankai, Liu Zhenghao, Liu Zhiyuan, Huang Lixin, Zhou Jie, and Sun Maosong. 2019. DocRED: A large-scale document-level relation extraction dataset. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Ye Qinyuan, Lin Bill Yuchen, and Ren Xiang. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zeng Daojian, Zhang Haoran, and Liu Qianying. 2020. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. ArXiv, abs/1911.10438.

Zeng Xiangrong, Zeng Daojian, He Shizhu, Liu Kang, and Zhao Jun. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 506–514, Melbourne, Australia. Association for Computational Linguistics.

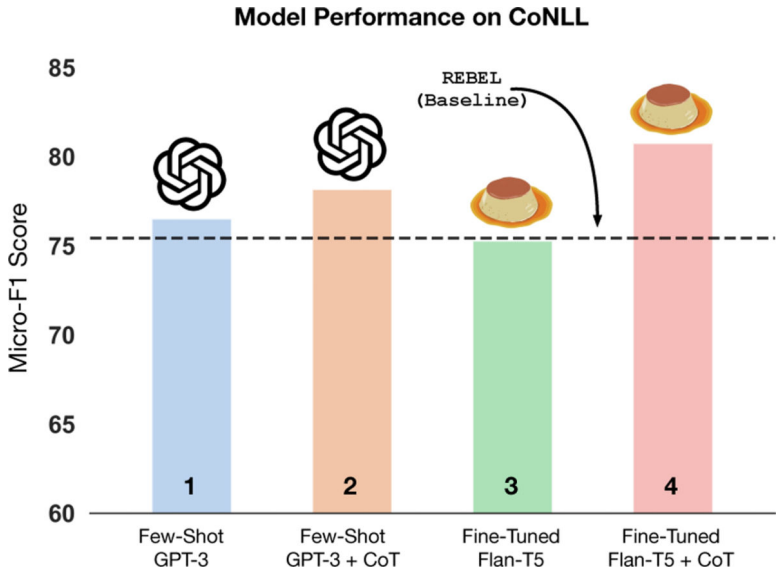

Figure 1:

RE performance of LLMs on the CoNLL dataset. **1** *Few-shot* GPT-3 slightly outperforms the existing *fully supervised* SOTA method (Huguet Cabot and Navigli 2021; dotted horizontal line). **2** Eliciting CoT reasoning from GPT-3 further improves few-shot performance. **3** Fine-tuning Flan-T5 (large) is competitive with, but no better than, existing supervised methods, but **4** supervising Flan-T5 with CoT reasoning elicited from GPT-3 substantially outperforms all other models.

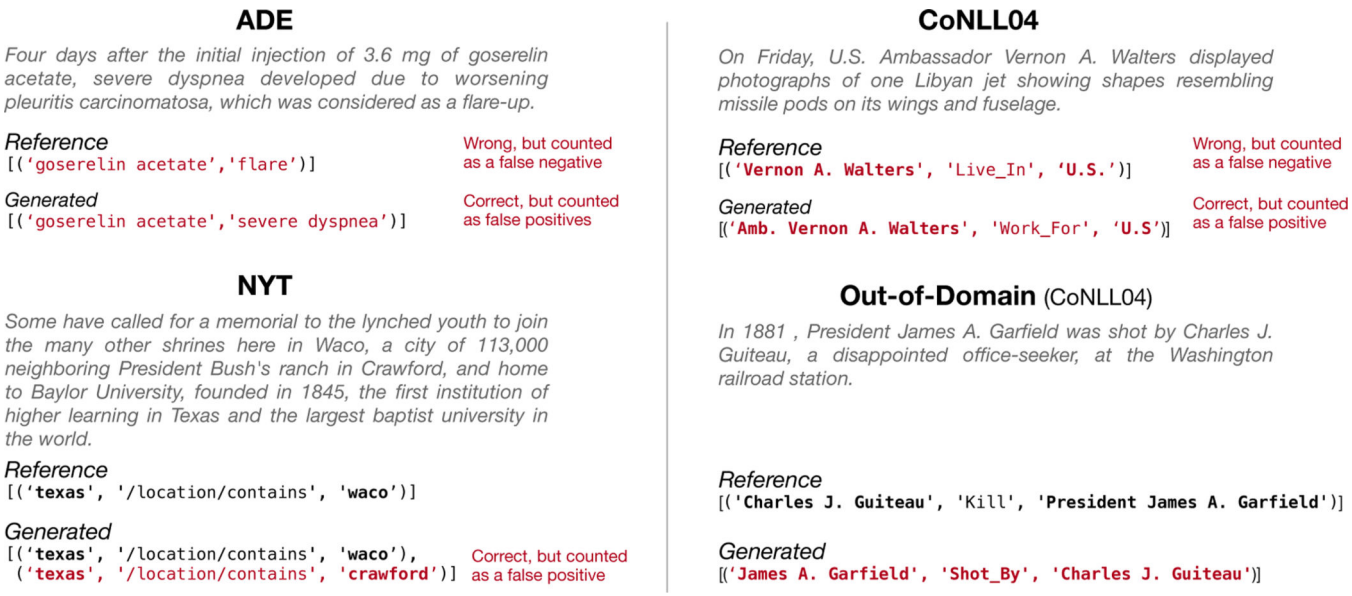

Figure 2:

Examples of misclassified FPs and FNs from GPT-3 (generated under few-shot in-context prompting scheme) under traditional evaluation of generative output. In each instance, the entity-type of subject and object was correctly identified.

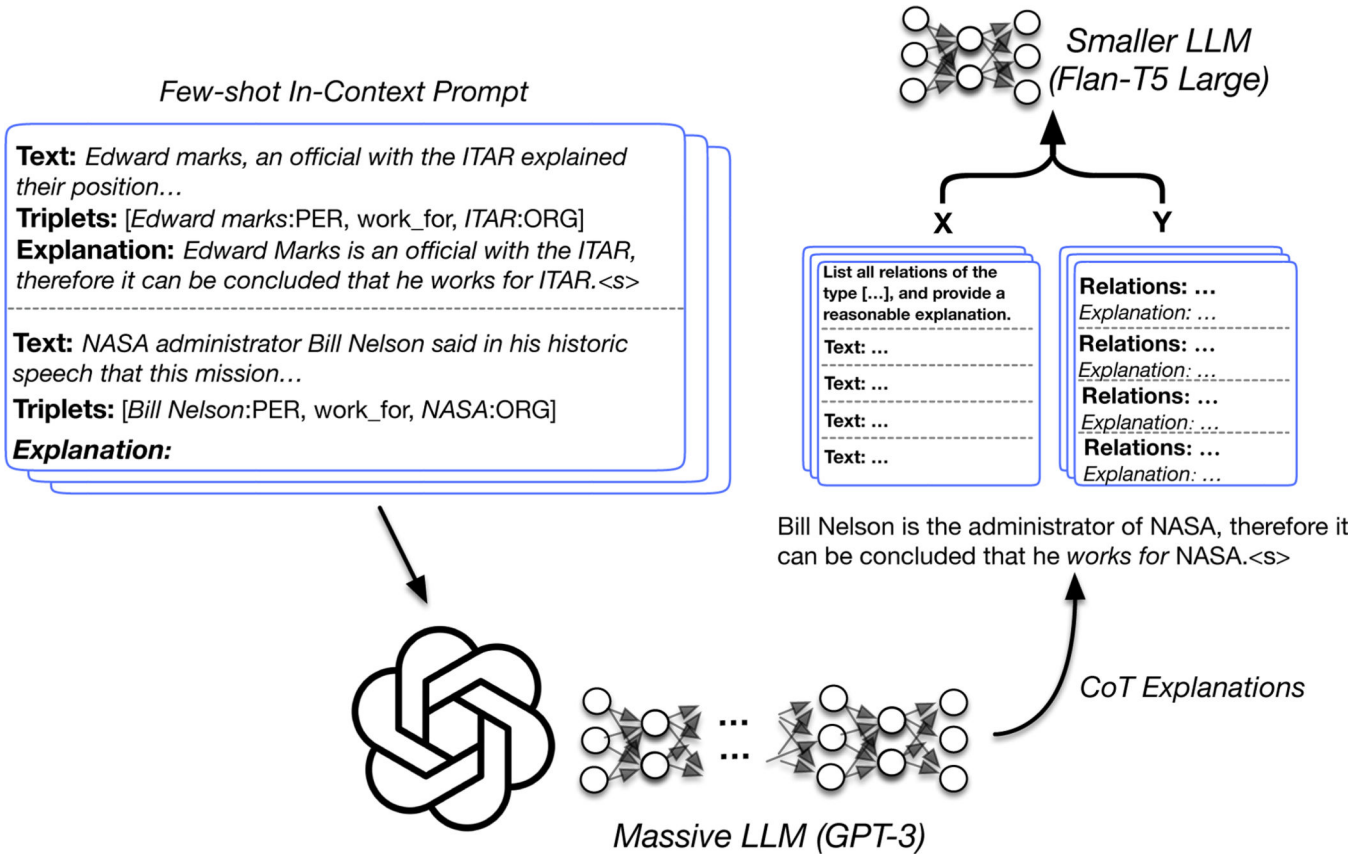

Figure 3:

We propose fine-tuning Flan-T5 (large) for relation extraction (RE) using standard supervision *and Chain-of-Thought (CoT) reasoning* elicited from GPT-3 for RE. This yields SOTA performance across all datasets considered, often by substantial margin (~5 points absolute gain in F1).

Table 1:

Dataset statistics. Train, validation and test indicate the number of relation triplets in each dataset.

	Entity Types	Relation Types	# of relation triplets		
			Train	Val	Test
ADE	2	1	4,272	–	–
CoNLL04	4	5	922	231	288
NYT	4	24	56,196	5,000	5,000
DocRED	6	96	3,008	300	700

Table 2:

Comparison of (micro-F1) performance with recent generative (except SpERT) approaches in RE. Relation triplets/pairs are considered correct only if both of the corresponding entity types are correctly generated.

	Method	Params	CONLL	ADE	NYT
1. Fully supervised	a. SpERT* (Eberts and Ulges, 2019b)	110M	71.54	79.22	–
	b. TANL (Paolini et al., 2021)	220M	71.48	80.61	90.83
	c. TANL (MT) (Paolini et al., 2021)	220M	72.66	80.00	90.52
	d. REBEL (Huguet Cabot and Navigli, 2021)	460M	75.44	82.21	92.00
	e. Flan T5 (Large) (Chung et al., 2022)	760M	75.28	83.15	91.03
	f. + **GPT-3-generated *CoT***	760M	**80.76**	**92.17**	**95.23**
2. Few-shot	a. In-Context GPT-3 (Brown et al., 2020a)	175B	76.53	82.66	61.79
	b. + *CoT*	175B	78.18	–	–
	c. Flan T5 (Large) w/CoT Explanations *and* reference labels generated from GPT-3	760M	76.13	–	–

Proc Conf Assoc Comput Linguist Meet. Author manuscript; available in PMC 2023 September 06.