GPT-RE: In-context Learning for Relation Extraction using Large Language Models

Zhen Wan [1] **Fei Cheng** [1] **Zhuoyuan Mao** [1]
Qianying Liu[1] **Haiyue Song**[1] **Jiwei Li**[2] **Sadao Kurohashi**[1]
[1] Kyoto University, Japan
[2] Zhejiang University, China
{zhenwan, zhuoyuanmao, ying, song}@nlp.ist.i.kyoto-u.ac.jp
{feicheng, kuro}@i.kyoto-u.ac.jp
{jiwei_li}@zju.edu.cn

Abstract

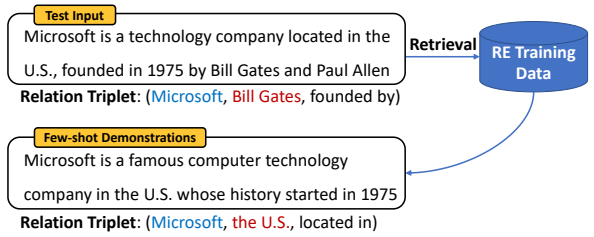
In spite of the potential for ground-breaking achievements offered by large language models (LLMs) (e.g., GPT-3), they still lag significantly behind fully-supervised baselines (e.g., fine-tuned BERT) in relation extraction (RE). This is due to the two major shortcomings of LLMs in RE: (1) low relevance regarding entity and relation in retrieved demonstrations for in-context learning; and (2) the strong inclination to wrongly classify NULL examples into other pre-defined labels.

In this paper, we propose GPT-RE to bridge the gap between LLMs and fully-supervised baselines. GPT-RE successfully addresses the aforementioned issues by (1) incorporating task-specific entity representations in demonstration retrieval; and (2) enriching the demonstrations with gold label-induced reasoning logic. We evaluate GPT-RE on four widely-used RE datasets, and observe that GPT-RE achieves improvements over not only existing GPT-3 baselines, but also fully-supervised baselines. Specifically, GPT-RE achieves SOTA performances on the Semeval and SciERC datasets, and competitive performances on the TACRED and ACE05 datasets.

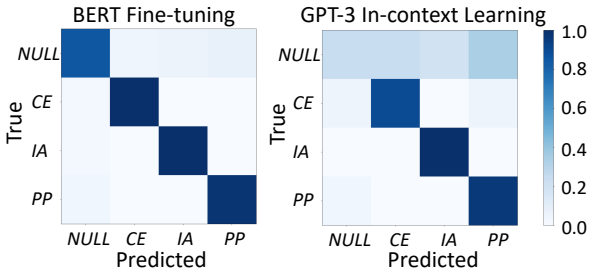
1 Introduction

The emergence of large language models (LLMs) such as GPT-3 (Brown et al., 2020; Thoppilan et al., 2022; Chowdhery et al., 2022; Rae et al., 2021; Hoffmann et al., 2022) represents a significant advancement in natural language processing (NLP). Instead of following a pretraining-and-finetuning pipeline (Devlin et al., 2019; Beltagy et al., 2019; Raffel et al., 2019; Lan et al., 2019; Zhuang et al., 2021), which finetunes a pretrained model on a task-specific dataset in a fully-supervised manner, LLMs employ a new paradigm known as in-context learning (ICL) (Brown et al., 2020; Min et al., 2022a) which formulates an NLP task under the paradigm of language generation and makes predictions by learning from demonstrations. Under the framework of ICL, LLMs achieve remarkable performance rivaling previous fully-supervised methods even with only a limited number of demonstrations provided in the prompt in various tasks such as solving math problems, commonsense reasoning, text classification, fact retrieval, natural language inference, and semantic parsing (Brown et al., 2020; Min et al., 2022b; Zhao et al., 2021; Liu et al., 2022b; Shin et al., 2021).

Despite the overall promising performance of LLMs, the utilization of ICL for relation extraction (RE) is still less optimal. RE seeks to identify a semantic relationship between a given entity pair mentioned in a sentence, which is the central task for knowledge retrieval requiring a deep understanding of natural language. Recent research (Gutiérrez et al., 2022) has sought to apply GPT-3 ICL to biomedical RE, the results are relatively negative and suggest that GPT-3 ICL still signifi-



(a) Retrieval without entity information results in noisy demonstrations.



(b) Confusion matrix on Semeval dataset with three selected relation labels. The NULL examples are overpredicted to other relations by GPT-3. CE: Cause-Effect, IA: Instrument-Agency, PP: Product-Producer.

Figure 1: **Two problems of GPT-3 ICL on RE**.

cantly underperforms fine-tuned models on the full dataset.

The reasons that cause the pitfall of GPT-3 ICL in RE are two folds: (1) The low relevance regarding entity and relation in the retrieved demonstrations for ICL. Demonstrations are selected randomly or via k-nearest neighbor (kNN) search based on sentence representations (Liu et al., 2022b; Gutiérrez et al., 2022). Regrettably, kNN-retrieval based on sentence-level representations is more concerned with the relevance of the overall sentence semantics and not as much with the entities and relations it contains. This leads to low-quality demonstrations retrieved. As shown in Figure 1a, the test input retrieves a semantically similar sentence but is not desired in terms of entities and relations.

(2) Overpredicting: we observe that LLMs have the strong inclination to wrongly classify NULL examples into other pre-defined labels as shown in Figure 1b. A similar phenomenon has also been observed in other tasks such as NER (Gutiérrez et al., 2022; Blevins et al., 2022). This phenomenon is because it is relatively simple to display an example that meets the criteria of a pre-defined label, yet hard and complex to illustrate an example that does not belong to that label. NULL examples are the collection with various undefined relations in nature. This fact causes the complex distribution of NULL examples, which hinders kNN-retrieval to obtain similar NULL demonstrations. This issue can be alleviated if the representations for retrieval can be supervised with a huge number of NULL, i.e., the supervised setting, or the reasoning logic can be accessed to enhance GPT-3 inference.

In this paper, we propose GPT-RE for the relation extraction task. GPT-RE employs two strategies to resolve the issues above: (1) **entity-aware retrieval** and (2) **gold label-induced reasoning**. For (1) entity-aware retrieval, its core is to use representations that deliberately encode and emphasize entity and relation information rather than sentence-level representations for kNN search. We propose the first encoding method to append the entity-pair prompt to the sentence. The second method is to obtain representations from a RE model fine-tuned on the RE training set, which naturally places emphasis on entities and relations. Both methods contain more RE-specific information than sentence semantics, thus effectively addressing the problem of low relevance.

For (2) gold label-induced reasoning, we propose to incorporate the reasoning steps to the demonstration, a strategy akin to Wei et al. (2022); Wang et al. (2022b); Kojima et al. (2022). But different from previous work, the reasoning process not only explains why a given sentence should be classified under a particular label but also why a NULL example should not be assigned to any of the pre-defined categories. This process of explaining significantly improves the prediction when fewer demonstrations are provided.

We evaluate our proposed method on three popular general domain RE datasets: Semeval 2010 task 8, TACRED and ACE05, and one scientific domain dataset SciERC. We observe that GPT-RE achieves improvements over not only existing GPT-3 baselines, but also fully-supervised baselines. Specifically, GPT-RE achieves SOTA performances on the Semeval and SciERC datasets, and competitive performances on the TACRED and ACE05 datasets.

2 Preliminary Background

2.1 Task Definition

Let $\mathcal{C}$ denote the input context and $e_{\text{sub}} \in \mathcal{C}$, $e_{\text{obj}} \in \mathcal{C}$ denote the pair of subject and object entity. Given a set of pre-defined relation classes $\mathbb{R}$, relation extraction aims to predict the relation $y \in \mathbb{R}$ between the pair of entities $(e_{\text{sub}}, e_{\text{obj}})$ within the context $\mathcal{C}$, or if there is no pre-defined relation between them, predict $y = \text{NULL}$.

2.2 BERT-based Fine-tuning

Current BERT-based fine-tuning methods for RE (Baldini Soares et al., 2019; Zhong and Chen, 2021; Wan et al., 2022) attempts to capture both the context information and the entity information by adding extra marker tokens to highlight the subject and object entities and their types.

Specifically, given an example: "*He* has a sister *Lisa*.", the input tokens are "[CLS] [SUB_PER] *He* [/SUB_PER] has a sister [OBJ_PER] *Lisa* [/OBJ_PER]. [SEP]" where "PER" is the entity type if provided. Denote the n-th hidden representation of the BERT encoder as $\mathbf{h}_n$. Assuming i and j are the indices of two beginning entity markers [SUB_PER] and [OBJ_PER], we define the relation representation as $\mathbf{Rel} = \mathbf{h}_i \oplus \mathbf{h}_j$ where $\oplus$ stands for concatenation of representations in the first dimension. Subsequently, this representation is fed into a feedforward network for predicting the relation probability $p(y \in \mathbb{R} \cup \{\text{NULL}\} \mid \mathbf{Rel})$.

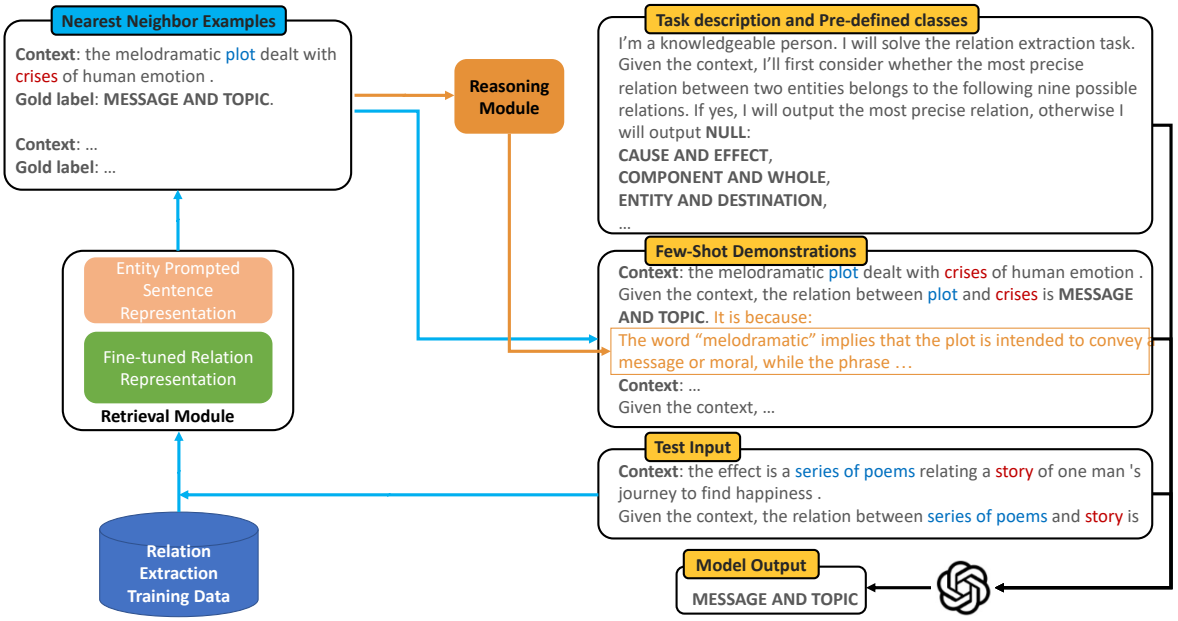


Figure 2: **An illustration of GPT-RE**. The blue line denotes the retrieval process, and the orange line denotes the reasoning process.

3 GPT-RE

GPT-RE is formalized under the ICL framework, using GPT-3 as shown in Figure 2.

3.1 Prompt Construction

We construct a prompt for each given test example, which is fed to the GPT-3 model. Each prompt consists of the following components:

Task Description and Pre-defined Classes We provide a succinct overview of the RE task description and the set of pre-defined classes $\mathbb{R}$, denoted by $\mathcal{O}$. The model is explicitly asked to output the relation, which belongs to the pre-defined classes. Otherwise, the model will output NULL.

Few-shot Demonstrations In the demonstration part, we reformulate each example by first showing the input prompt $x_{demo} = Prompt(\mathcal{C}, e_{\text{sub}}, e_{\text{obj}})$ and the relation label y_{demo}. The input prompt can be further enriched by the reasoning process.

Test Input Similar to the demonstrations, we offer the test input prompt x_{test}, and GPT-3 is expected to generate the corresponding relation y_{test}.

In summary, GPT-RE can be formulated as:

$$p\left(y_{test} | \mathcal{O} \uplus \biguplus_{i=1}^{K} \left(x^i_{demo} \uplus y^i_{demo}\right) \uplus x_{test}\right) \quad (1)$$

where "$\uplus$" denotes the concatenation of two textual pieces, "$\biguplus$" indicates cumulative concatenation, and "K" is the number of demonstrations.

3.2 Entity-aware Demonstration Retrieval

Since ICL demonstrations closer to the test sample in the embedding space result in more consistent and robust performance (Liu et al., 2022b). Recent work (Gutiérrez et al., 2022; Liu et al., 2022b) employs the kNN to retrieve the most similar examples in the training set as the few-shot demonstrations for each test example. As kNN relies on the choice of the embedding space, they propose to obtain sentence representations using PLMs, or other improved sentence representations.

However, using sentence-level representations for kNN retrieval has a severe drawback: relation extraction focuses on pair-wise entities, which diverge from the semantic meaning of the entire sentence, leading to an ambiguous retrieval using sentence embeddings. In this study, we propose two novel methods to provide more robust representations for better retrieval quality.

3.2.1 Entity Prompted Sentence Representation

Given the discrepancy between sentence embedding and relation extraction, the original context is insufficient for demonstration retrieval. Considering the importance of entity information in RE, we propose reconstructing the context by incorporating entity pair information. For example,

Dataset	# Rel.	# Train	# Dev	# Test (# Subset)	NA (%)
Semeval	9	6,507	1,493	2,717 (2,717)	17.40%
TACRED	41	68,124	22,631	15,509 (1,600)	79.40%
SciERC	7	16,872	2,033	4,088 (4,088)	90.16%
ACE05	6	121,368	27,597	24,420 (2,442)	95.60%

Table 1: **Statistics of datasets**. Rel. denotes relation types.

given the context "*He* has a sister *Lisa*," the reconstructed context with the entity prompted will be "The relation between 'He' and 'Lisa' in the context: He has a sister Lisa." This approach addresses the feature of RE as it preserves both the semantic meaning of the entire sentence and the entity pair-centered information during retrieval. In the paper, we employ the latest robust model SimCSE (Gao et al., 2021) for the sentence similarity calculation to select the nearest neighbors between the reconstructed contexts. We formulate the encoding process as $SimCSE(Prompt_{retrival}(\mathcal{C}, e_{\text{sub}}, e_{\text{obj}}))$.

3.2.2 Fine-tuned Relation Representation

Compared to prompt entity information into context sentences, a more straightforward solution is to extract the relation representation from a fine-tuned model for retrieving demonstrations. As shown in Sec. 2.2, the entity markers have explicitly encoded subject and object entities, and the relation representation $\mathbf{Rel}$ is naturally enriched with the entity information.

We believe this approach can potentially compensate for the limitations of GPT-3 in RE. While GPT-3 ICL has a constraint of limited demonstrations, the fine-tuning process is unbundled and can be done on the whole train data. It has two subsequent merits. First, the relation representations are directly fine-tuned to fit the RE task, which could significantly boost the overall retrieval quality. Second, the overpredicting NULL issue will be substantially alleviated because the similar NULL demonstrated can be accurately recognized by the fine-tuned model.

3.3 Gold Label-induced Reasoning

The recent CoT (Wei et al., 2022; Wang et al., 2022b) has reported significant progress in commonsense and numerical reasoning tasks by reasoning prompts that successfully elicits the reasoning logic towards the final output. While in RE, a pair of entities potentially holds multiple possible relations, which could leave the reasoning out of focus.

In this section, we propose to let GPT-3 induce

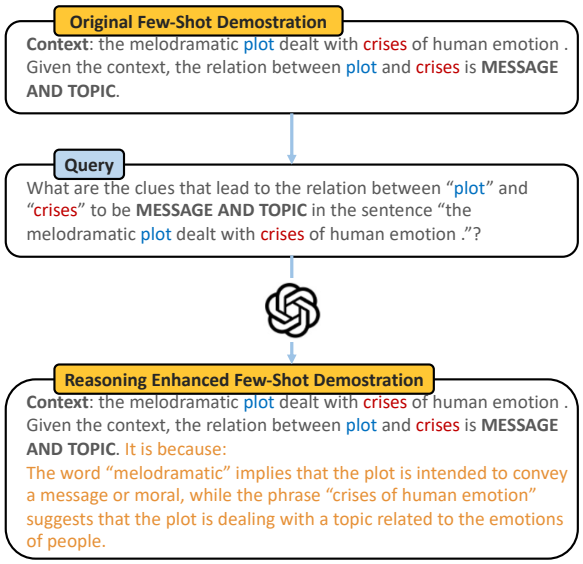


Figure 3: **An illustration of adding reasoning**.

the reasoning logic for each demonstration by the corresponding gold RE label. As shown in Figure 3, given a selected example, we first generate a query prompt based on the example and subsequently ask GPT-3 to generate clues on the labeled relation between the pair of entities in the context. Finally, we augment the demonstration by incorporating the generated clues with the original example.

4 Experiment Setup

4.1 Datasets

We evaluate our proposed method on three popular general domain RE datasets and one scientific domain dataset. Table 1 lists the datasets and their statistics.

Semeval 2010 task 8 Hendrickx et al. (2010) focuses on semantic relations (e.g., "cause and effect") between pairs of nominals and contains 10,717 annotated examples covering nine relations collected from general domain resources.

TACRED Zhang et al. (2017) is a large-scale relation extraction dataset with 106,264 examples built over newswire and web text. It spans 41 relations labels, which hold between persons, locations, organizations, dates, and so on (e.g., "siblings," "dates of birth," "subsidiaries," etc.).

SciERC Luan et al. (2018) collects AI paper abstracts and annotated relations, especially for scientific knowledge graph construction.

Methods	Retriever	Semeval	TACRED	SciERC	ACE05
Fine-tuning Baselines					
Cohen et al. (2020)		**91.90**	-	-	-
Wang et al. (2022a)		-	♣**76.80**	-	-
PURE (Zhong and Chen, 2021)		89.90	69.72	68.45	**70.09**
GPT-3 Baselines (Best k-shot)					
GPT-Random	-	70.04 (30)	32.49 (15)	17.92 (25)	9.04 (25)
GPT-Sent	SimCSE	79.94 (30)	33.45 (15)	20.96 (25)	6.31 (25)
Ours (Best k-shot)					
GPT-RE_SimCSE	SimCSE	81.02 (30)	37.44 (15)	26.46 (25)	8.67 (25)
GPT-RE_SimCSE*	SimCSE	77.49 (15)	31.58 (10)	-	-
+ Reasoning	SimCSE	79.88 (15)	33.18 (10)	-	-
GPT-RE_FT	RE fine-tuning	**91.90** (25)	72.14 (15)	**69.00** (30)	68.73 (25)
GPT-RE_FT*	RE fine-tuning	91.11 (15)	70.38 (10)	-	-
+ Reasoning	RE fine-tuning	91.82 (15)	70.97 (10)	-	-

Table 2: **Main Results on four RE datasets**. All results are given by Micro-F1. * denotes the same k-shot for the comparison with + Reasoning. Due to the costly GPT-3 expense, we conducted Reasoning experiments on the two relatively smaller datasets Semeval and TACRED. ♣ denotes that this performance is not comparable as it evaluates on the entire test set. The underline denotes the results outperforming the fine-tuning baseline PURE.

ACE05 contains the entity, relation, and event annotations from 511 documents in total collected from multiple domains including newswire, broadcast, discussion forums, etc.

Due to the cost of running the model in the API with GPT-3, in our main results, we sample a subset from the original test set for two datasets: ACE05 and TACRED as shown in Table 1. We will release all of these subsets and the corresponding sampling codes for reproducibility and further study.

4.2 Baseline Methods

Fine-tuning baseline In our experiment, we choose PURE (Zhong and Chen, 2021) as our fine-tuning baseline model. We follow their single-sentence to keep consistency among datasets as Semeval and TACRED are both sentence-level RE datasets. For the PLMs, we also follow PURE by using *scibert-scivocab-uncased* (Beltagy et al., 2019) as the base encoder for SciERC and *bert-base-uncased* (Devlin et al., 2019) for the remaining three general domain datasets.

GPT-3 baseline For all GPT-3 baselines and our proposed methods, we select "`text-davinci-003`" and use the identical prompt construction, as defined in Section 3.1. Other hyperparameters are listed in Appendix A. We compare our proposed method with two categories of GPT-3 baselines from previous work.

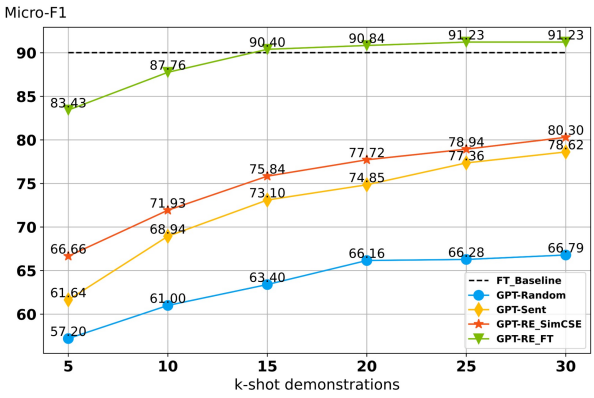
(1) GPT-Random For each test input, we randomly choose few-shot demonstrations from the training data. Unlike the vanilla random selection, we add extra constraints to make the label distribution of selected demonstrations more uniform. Our preliminary experiments suggest that this is a stronger baseline than the vanilla random.

(2) GPT-Sent Gutiérrez et al. (2022) uses the [CLS] of RoBERTa-large as the representation in retrieval, Liu et al. (2022b) fine-tunes RoBERTa-large on two natural language inference (NLI) datasets: SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) to enhance the quality of sentence representations. In this work, our implementation adopted SimCSE (Gao et al., 2021) for sentence embedding, which has been demonstrated to be the state-of-the-art method for sentence similarity tasks. In our experiment, we utilize the version: `sup-simcse-bert-base-uncased`.

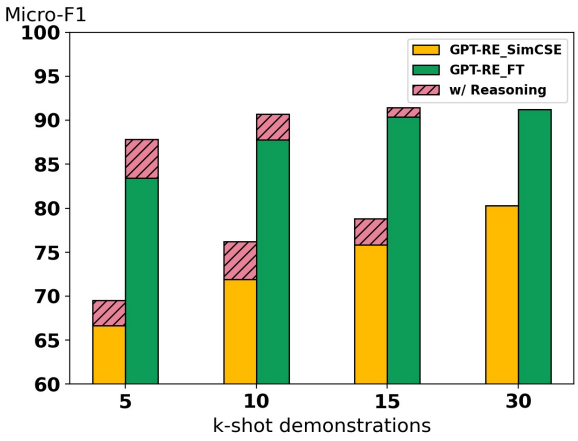
5 Experimental Results

5.1 Main Results

We compare our main experiment results with previous methods in Table 2. From the table, we can observe that: (1) both *GPT-RE_SimCSE* and *GPT-RE_FT* outperform the retrieval-based *GPT-Sent*, indicating that it is necessary to capture the task-specific information into sentence encoding for



(a) **The comparison on retrieval modules**



(b) **Reasoning with fewer demonstrations**.

Figure 4: **Ablation study on the retrieval and reasoning components on Semeval.** We sampled a subset from the test data with 300 examples. We show the 'w/o reasoning' results with $k = 30$ for comparison.

selecting proper demonstrations; (2) *GPT-RE_FT* succeeds to outperform the fine-tuning baseline on three datasets by $+2.00$, $+2.42$, $+0.55$ Micro-F1. It suggests that GPT-3 has the potential to beat fine-tuning when the retriever has prior task knowledge. *GPT-RE_FT* eventually achieves SOTA results on Semeval and SciERC. (3) reasoning module improves *GPT-RE_SimCSE* by around 2% Micro-F1, indicating that gold label-induced reasoning successfully enriches the knowledge of demonstrations. Meanwhile, the high-quality demonstrations obtained by *GPT-RE_FT* offset the effort of enriching reasoning into demonstrations, which shows relatively trivial improvements. Since reasoning aims at enriching demonstrations, this feature potentially works better with fewer demonstrations, as shown in Section 5.3.

5.2 Ablation Study on Entity-aware Retrieval

We first implement the ablation experiments of the retrieval component with the setting of increasing

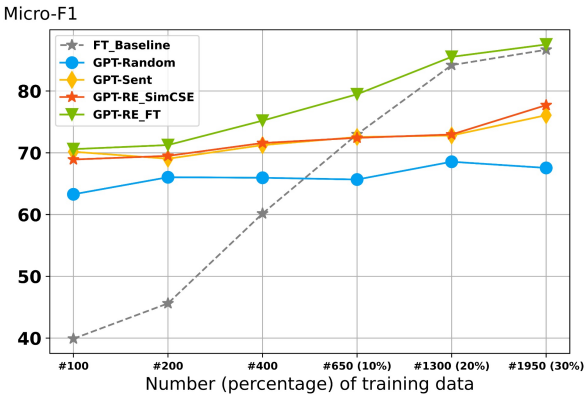
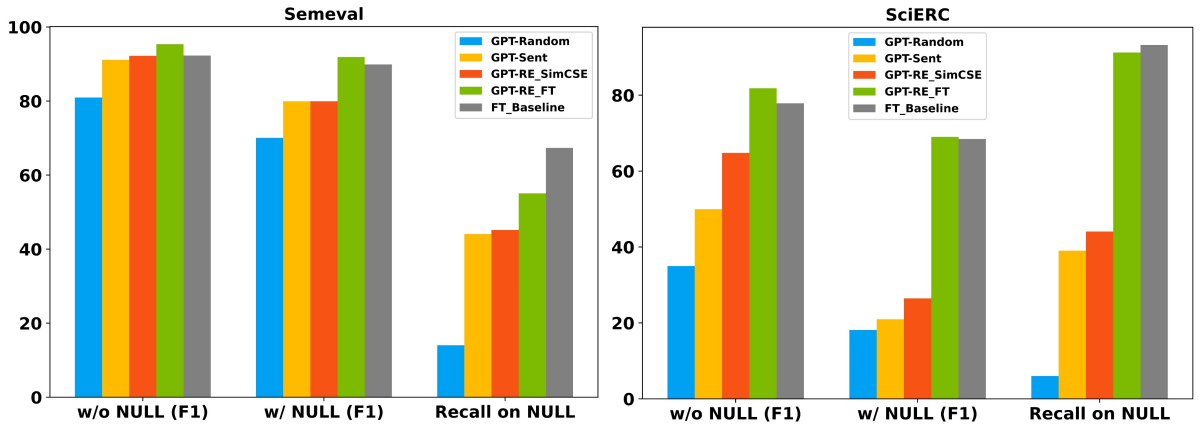


Figure 5: **Low-resource Scenario on Semeval**. We limit the percentage of training data for both fine-tuning and GPT-RE.

k-shot demonstrations (Figure 4a). We find that: (1) compared to *GPT-Random*, all the retrieval-based models have higher F1 scores and large gradients of the performance curves. It means that GPT-3 can learn from high-quality demonstrations more effectively; (2) after adding entity information to the SimCSE retrieval, *GPT-RE_SimCSE* achieves better performance throughout all K shots, indicating that entity-aware sentence representation can capture the feature of RE and provide more proper demonstrations; (3) finally, the fine-tuned relation representation retriever *GPT-RE_FT* significantly outperforms all retrieval-based methods and beats the fine-tuning baseline when $k > 15$. Note that even with $k = 5$ demonstrations, *GPT-RE_FT* still works better than *GPT-RE_SimCSE* with $k = 30$ ($80.30 \rightarrow 83.43(+3.13)$), which indicates that the quality of demonstrations shows much more important than the number of demonstrations.

5.3 Ablation Study on Reasoning Enhancing

We then check the influence of our proposed reasoning-enhanced demonstration, as shown in Figure 4a. Due to the limited amount of input tokens of GPT-3, we have to set the $k \leq 15$ for the tokens of reasoning, leading to a trade-off between adding reasoning and adding more demonstrations. From the result, we find that: (1) with reasoning-enhanced demonstrations, GPT-3 always achieves better scores across all the k-shot settings of both *GPT-RE_SimCSE* and *GPT-RE_FT*, indicating that the reasoning induced from ground truth relation labels can effectively unlock the reasoning ability of GPT-3 and improve the ICL with a deeper understanding of demonstrations. Specifically, for *GPT-RE_FT*, the performance improve-



(a) General domain with 17.4% NULL examples

(b) Scientific domain with 90.16% NULL examples

Figure 6: **Analysis on the effects of NULL examples**. w/o NULL refers to the classification setting that NULL examples are excluded from the train and test data. w/ NULL refers to the original extraction setting. We use the full test set for the evaluation

ment becomes less significant when more demonstrations provided, which is feasible as with more high-quality demonstrations available, GPT-3 can already learn the internal reasoning behind each demonstration; (2) since the reasoning enhancement works better with fewer demonstrations, we expect this method can be an effective solution to low-shot relation extraction (Han et al., 2018; Geng et al., 2020; Liu et al., 2022a), which aims at recognizing novel relations with very few or no examples, and we leave this for future work.

5.4 Low-resource Scenario

We conduct the experiment for observing the low-resource performance in the general domain Semeval task. As shown in Figure 5, we observe that: (1) all the GPT-3 based results work better than fine-tuning in when the training examples are less than # 650 (10%). It indicates that in the general domain RE, GPT-3 benefits from its abundant prior knowledge to understand the relations; (2) *GPT-RE_SimCSE* starts to show a substantial difference to *GPT-Sent* after the training size surpasses 30%. We believe fewer training candidates could limit the effects of retrieval; (3) *GPT-RE_FT* achieves an upper bound performance in all settings, even when the fine-tuned model shows poor performance with hundreds of training data (from #100 to #400). This emphasizes the impressive effectiveness of fine-tuned relation representations for capturing higher-quality demonstrations. The observation in the low-resource setting is very different from Gutiérrez et al. (2022). We assume the difference could be caused by the domain and NULL proportion of the task.

6 Analysis

6.1 Analysis on the effects of NULL Examples

To analyze the influence of NULL class, we compare the effectiveness of each method for solving this issue on two datasets: general domain semeval with 17.4% NULL examples and scientific domain SciERC with 90.16% NULL examples. As shown in the Figure 6, (1) by comparing the performance on Semeval and SciERC, a larger percentage of NULL examples results in more significant performance drop showing the negative influence of over-predicting NULL examples; (2) by comparing w/o NULL and w/ NULL, our *GPT-RE_FT* shows the most robustness to the influence of NULL examples, indicating that the RE fine-tuned representations in retrieval can release the overpredicting issue of GPT-3 by providing higher-quality demonstrations; (3) however, even with entity-aware representations, all GPT-3 methods still underperform the fine-tuning baseline on NULL examples, this is due to the confusing definition of NULL, in many cases, there is a certain relation between entities in the context, but out of the distribution of pre-defined classes. In these cases, GPT-3 tends to overpredict as the relation information may be covered in its prior knowledge. We think this ability of GPT-3 can be useful in more open fields, such as open RE (Banko and Etzioni, 2008) which has no pre-defined relation classes.

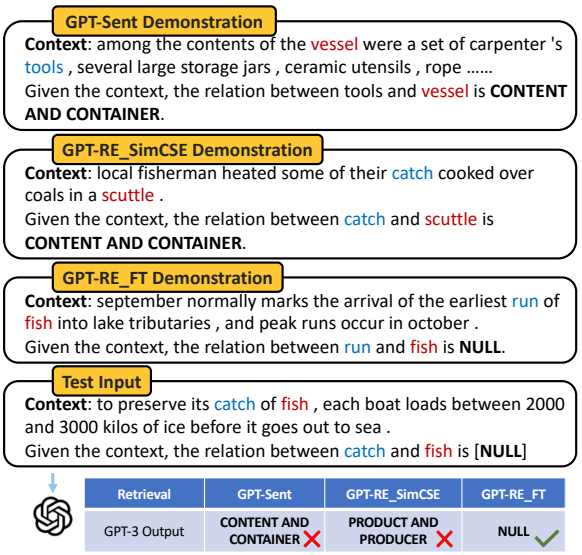


Figure 7: **A case study of demonstration quality on Semeval.** [NULL] denotes the gold label.

6.2 Case Study of Demonstration Quality

We select one typical test example to better illustrate the amendment of our entity-aware demonstration retrieval. As shown in Figure 7, given the NULL Example, we show the most similar demonstration in retrieval based on three methods. The *GPT-Sent* retrieved demonstration focuses on the semantic meaning of "CONTENT AND CONTAINER" which is shared in the test context, but not revealed in the target entity pair. This mismatch confirms the problem of lacking entity information in retrieval. Instead, *GPT-RE_SimCSE* retrieves a much more relevant demonstration that shows the same semantic relation between "catch" and "fish" but still faces a minor mismatch as the gold label is between "catch" and "scuttle." Finally, *GPT-RE_FT* demonstration shares a similar structure with the test input regarding the pair of entities, which is the key clue for predicting the relation between entities. This result shows a level-by-level enhancement with more entity information provided in retrieval. We also show some other case examples in Appendix D.

7 Related Work

In-context Learning Recent work shows that ICL of GPT-3 (Brown et al., 2020) can perform numerous tasks when provided a few examples in a natural language prompt. Existing work focus on various aspects to effectively utilize the advantages of GPT-3, from prompt design (Perez et al., 2021) for proper input to coherence calibration (Malkin et al., 2022) for tackling the diverse generated output. Another research path locates in the demonstration part, including ordered prompts (Lu et al., 2022) and retrieval-based demonstrations (Rubin et al., 2022; Liu et al., 2022b; Shin et al., 2021).

To the best of our knowledge, there is no previous work exploring the potential of GPT-3 on general domain RE tasks. A recent work attempts to leverage GPT-3 in biomedical information extraction (NER and RE), and reveals issues of ICL that may be detrimental to IE tasks in general. Our work succeeds in overcoming these issues to some extent and confirms the potential of GPT-3 in both general and scientific domain RE.

Retrieval-based Demonstrations Several studies have demonstrated that dynamically selecting few-shot demonstrations for each test example, instead of utilizing a fixed set, leads to significant improvement in GPT-3 ICL (Liu et al., 2022b; Shin et al., 2021; Rubin et al., 2022). They also show that nearest neighbor in-context examples yield much better results than the farthest ones. This leads to the significance of better retrieval modules for demonstrations. Existing attempts rely on sentence representations in retrieval, including the sentence encoders of PLMs such as BERT (Devlin et al., 2019), RoBERTa (Zhuang et al., 2021) KATE (Liu et al., 2022b) , SimCSE (Gao et al., 2021), Sentence-BERT (Reimers and Gurevych, 2019; Wolf et al., 2020). Unlike these sentence representations, we propose to fine-tune PLMs on our target RE tasks to produce more task-specific and robust representations for retrieval.

8 Conclusions

This work explores the potential of GPT-3 ICL on RE. Given the difficulties in utilizing GPT-3 for RE, we propose GPT-RE to solve these issues and bridge the performance gap to the fine-tuning baselines via two strategies: (1) entity-aware demonstration retrieval emphasizes entity and relation information for improving the accuracy of searching demonstrations; (2) gold label-induced reasoning enriches the reasoning evidence of each demonstration. The experimental results show that GPT-RE significantly outperforms the fine-tuning baseline on three datasets and achieves SOTA on Semeval and SciERC. We implement detailed studies to explore how GPT-3 overcomes the difficulties such as NULL example influence.

Limitations

Despite the overall positive results, GPT-RE still faces two shortcomings: (1) the issue of overpredicting has been significantly alleviated but not completely solved, and the NULL recall still lags behind full-supervised baselines, especially on the datasets containing a large proportion of NULL examples such as ACE05 ("95.60%"); (2) Though the entity-aware retriever optimizes the representations of PLMs such as SimCSE and BERT, it is widely considered that LLMs can generate more robust representations than small PLMs. Future work can replace representations generated by smaller PLMs with GPT-3 itself. However, due to the access limitation to the representations of GPT-3, we can nearly confirm this proposal up to now.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Michele Banko and Oren Etzioni. 2008. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08: HLT*, pages 28–36, Columbus, Ohio. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Prompting language models for linguistic structure. *CoRR*, abs/2211.07830.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Amir D. N. Cohen, Shachar Rosenman, and Yoav Goldberg. 2020. Relation extraction as two-way span-prediction. *CoRR*, abs/2010.04829.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoqing Geng, Xiwen Chen, Kenny Q. Zhu, Libin Shen, and Yinggong Zhao. 2020. MICK: A meta-learning framework for few-shot relation classification with small training data. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 415–424. ACM.

Bernal Jiménez Gutiérrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical ie? think again. *CoRR*, abs/2203.08410.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *CoRR*, abs/2205.11916.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

Fangchao Liu, Hongyu Lin, Xianpei Han, Boxi Cao, and Le Sun. 2022a. Pre-training to match for unified low-shot relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5785–5795, Dublin, Ireland. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022b. 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.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot 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.

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.

Nikolay Malkin, Zhen Wang, and Nebojsa Jojic. 2022. Coherence boosting: When your pretrained language model is not paying enough attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8214–8236, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Rethinking the role of demonstrations: What makes in-context learning work?

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? *CoRR*, abs/2202.12837.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11054–11070.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris

Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Zhen Wan, Qianying Liu, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi, and Jiwei Li. 2022. Rescue implicit and long-tail cases: Nearest neighbor relation extraction. *CoRR*, abs/2210.11800.

Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. 2022a. 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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *CoRR*, abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

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.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Hyperparameter	In Experiment
Engine	text-davinci-003
Temperature	0.0
Max_tokens	256
Top_p	1
Frequency_penalty	0.0
Presence_penalty	0.0
Best_of	1
Logprob	1

Table 3: GPT-3 Hyperparamters.

Dataset	Lower bound	Upper bound
Semeval	5	30
TACRED	5	15
SciERC	5	30
ACE05	5	25

Table 4: Search range for each dataset.

A Hyperparameters

B GPT-3 Hyperparameters

We use the GPT-3 API during the experiments and set the hyperparameters as in Table 3. Since the "Temperature" is set to be 0.0, denoting the stable output of GPT-3, we report the result of the single run for all experiments. Due to the input length limitation of GPT-3 and the various average lengths of contexts from each dataset, we set different search ranges for the number of demonstrations of each dataset as shown in Table 4.

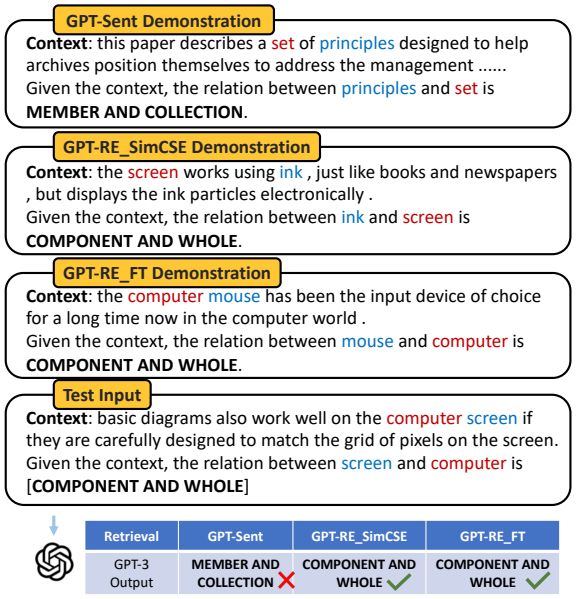
C Other Hyperparameters

For the fine-tuning baseline model PURE and the sentence embedding model SimCSE, we follow all hyperparameters from their papers. We used 2 NVIDIA RTX3090 for training.

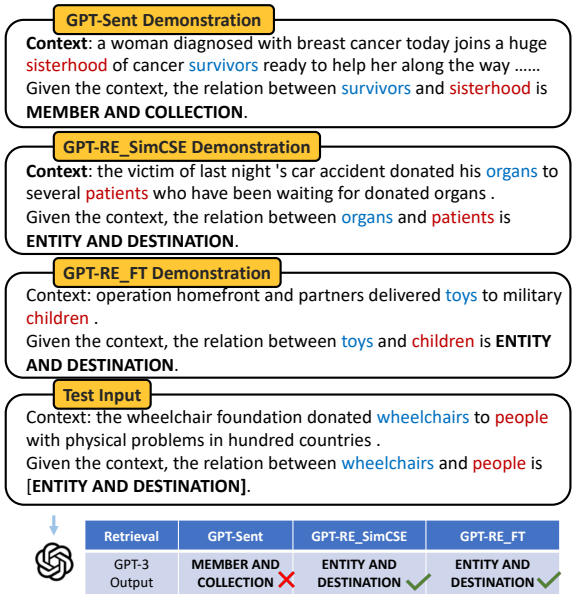
D Case Study

To verify the effectiveness of our entity-aware demonstration retrieval, we provide more cases.

For Figure 8a, *GPT-Sent* retrieves a demonstration that shares the same semantic meaning of "design" with the test input. However, the entity pair is irrelevant to the concept "design" resulting in a noisy demonstration. Instead, *GPT-RE_SimCSE* retrieves a more relative demonstration with closer pair of entities sharing the same relation label. Furthermore, *GPT-RE_FT* retrieves the demonstration containing both the closing entity pair and the same linguistic structure between entities. This case emphasizes level-by-level improvement using our proposed methods. Figure 8b shows a similar phenomenon.



(a) **[COMPONENT AND WHOLE]** denotes the gold label



(b) **[ENTITY AND DESTINATION]** denotes the gold label.

Figure 8: **More casees.**