Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again

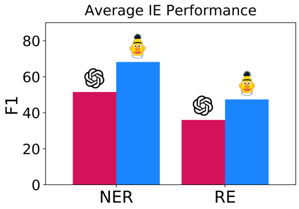
Bernal Jiménez Gutiérrez[1], **Nikolas McNeal**[1], **Clay Washington**[1],
You Chen[2], **Lang Li**[1], **Huan Sun**[1], **Yu Su**[1]
[1]The Ohio State University, [2]Vanderbilt University
{jimenezgutierrez.1,mcneal.121,washington.534,sun.397,su.809}@osu.edu
lang.li@osumc.edu, you.chen@vumc.org

Abstract

Large pre-trained language models (PLMs) such as GPT-3 have shown strong in-context learning capabilities, which are highly appealing for domains such as biomedicine that feature high and diverse demands of language technologies but also high data annotation costs. In this paper, we present the first systematic and comprehensive study to compare the few-shot performance of GPT-3 in-context learning with fine-tuning smaller (i.e., BERT-sized) PLMs on two representative biomedical information extraction (IE) tasks: named entity recognition and relation extraction. We follow the true few-shot setting (Perez et al., 2021) to avoid overestimating models' few-shot performance by model selection over a large validation set. We also optimize GPT-3's performance with known techniques such as contextual calibration and dynamic in-context example retrieval. However, our results show that GPT-3 still significantly underperforms compared to simply fine-tuning a smaller PLM. In addition, GPT-3 in-context learning also yields smaller gains in accuracy when more training data becomes available. More in-depth analyses further reveal issues of in-context learning that may be detrimental to IE tasks in general. Given the high cost of experimenting with GPT-3, we hope our study provides helpful guidance for biomedical researchers and practitioners towards more practical solutions such as fine-tuning small PLMs before better in-context learning is available for biomedical IE.[1]

1 Introduction

Given the overwhelming pace of biomedical research and clinical text production, transforming large amounts of biomedical text into structured information has become increasingly important for researchers and practitioners alike. In recent years, pre-trained language models (PLMs), both general-domain and biomedicine-specific ones, have remarkably boosted performance on biomedical information extraction (IE) tasks (Lee et al., 2019; Peng et al., 2019; Gu et al., 2021; Alsentzer et al., 2019; Beltagy et al., 2019).



	GPT-3 In-Context	PLM Fine-Tuning
No Training	✓	✗
No External Costs	✗	✓
More Data = Better Performance	✗	✓

Figure 1: Main findings: (*Left*) fine-tuning BERT-sized PLMs substantially outperforms GPT-3 in-context learning under true few-shot setting. (*Right*) Feature comparison for consideration of practical applications.

The latest round of PLMs such as GPT-3 (Brown et al., 2020), Megatron-Turing NLG (Smith et al., 2022), the Switch Transformer (Fedus et al., 2022), among others, feature hundreds of billions of parameters and have achieved impressive performance in many NLP tasks using *in-context learning*—a new few-shot learning paradigm first introduced by Brown et al. (2020). In-context learning allows PLMs to use their natural language generation capabilities to solve any task almost like how humans would—by completing a piece of text or *prompt*. This paradigm allows large PLMs to solve various NLP problems without updating their parameters, potentially resulting in massive savings in both data annotation and engineering overhead compared with standard model training. Even more impressively, GPT-3 in-context learning yields competitive performance against fully supervised baselines in many NLP tasks by adding only a handful of demonstrative examples in the prompt (Brown et al., 2020).

The variety of potential biomedical information extraction applications, the high cost of biomedical annotations, and the complexity of model training make in-context learning particularly appealing for

[1] Our source code is available at `https://github.com/dki-lab/few-shot-bioIE`.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4497–4512
December 7-11, 2022 ©2022 Association for Computational Linguistics

biomedical applications. To investigate its practicality, we present the first systematic and comprehensive comparative study of GPT-3 in-context learning and BERT-sized (Devlin et al., 2019) PLM fine-tuning in the few-shot setting on named entity recognition (NER) and relation extraction (RE), two representative and highly valued biomedical IE tasks. For consistency and comprehensiveness, we use all the biomedical NER and RE tasks compiled in the BLURB benchmark (Gu et al., 2021). We operate under the true few-shot setting introduced by Perez et al. (2021) to avoid overestimating models' few-shot performance via model selection over a large validation set.

We optimize GPT-3's in-context learning performance for biomedical information extraction by leveraging multiple recent techniques. Firstly, inspired by studies that show the importance of optimal prompt selection (Perez et al., 2021; Schick and Schütze, 2021; Gao et al., 2021), we formulate a prompt structure which allows us to construct prompt designs and select optimal ones systematically. Secondly, similar to Liu et al. (2022), we introduce a k-nearest neighbor (kNN) module for in-context example retrieval. Finally, for NER, we also use logit biases to ensure that the generated tokens are from the input sentence; for RE, we use contextual calibration (Zhao et al., 2021) to reduce contextual bias.

Even when equipped with these latest techniques, which indeed improve GPT-3's performance as shown in ablation studies, we find that fine-tuning BERT-sized PLMs substantially outperforms GPT-3 in-context learning across all biomedical information extraction datasets when using the same small training set (e.g., 100 labeled examples). We also find that fine-tuning small PLMs yields a more reliable return in terms of data annotation: as training data size increases, fine-tuning performance steadily improves while in-context learning performance lags behind. In-depth analyses further reveal that in-context learning struggles with the `null` *class*, e.g., sentences that contain no named entity (for NER) or entity pairs that hold none of the target relations (for RE), which is likely detrimental to IE tasks in general. In summary, our findings suggest that fine-tuning PLMs is still a more cost-effective option than GPT-3 in-context learning for biomedical IE tasks, at least before qualitatively better methods for in-context learning are discovered.

2 Approach

In this section, we describe the two paradigms we explored under the true few-shot setting for NER and RE: BERT-sized PLM fine-tuning and GPT-3 in-context learning.

2.1 Tasks

We use named entity recognition (NER) and relation extraction (RE) as two representative and highly valued tasks to comprehensively evaluate the potential of GPT-3 in-context learning in biomedical IE.

2.2 True Few-Shot Setting

Recent work has questioned the performance of few-shot learning in very large PLMs like GPT-3 as well as small PLM fine-tuning, arguing that large validation sets have played a strong biasing role in model and prompt selection (Perez et al., 2021). To avoid overestimating the few-shot learning performance of PLMs, we follow their proposed true few-shot setting. In this setting, all model selection decisions are made systematically on the few-shot training set rather than on a large validation set. For our main experiments, we use cross-validation on 100 training examples to choose the prompt structure, the number of few-shot examples per prompt and the fine-tuning hyperparameters.

2.3 BERT-Sized PLM Fine-Tuning

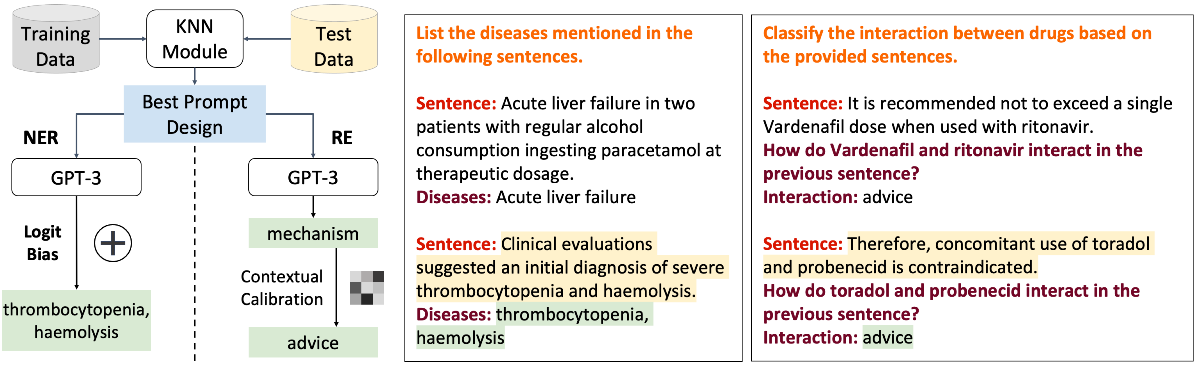
We follow the standard PLM fine-tuning process for NER and RE used in Gu et al. (2021). We use 5-fold cross-validation on the 100 example training set mentioned above to select the best performing values for learning rate, batch size, warm-up ratio, weight decay, and stopping checkpoint for all of our fine-tuning experiments. The hyperparameter values we select from are specified in Appendix C.

Named Entity Recognition. For NER, we use the BIO tag token classification formulation and fine-tune a separate model for each entity type.

Relation Extraction. For RE, we mask the object and subject entities in the input sentence and use the `[CLS]` token to classify the relation between them.

2.4 GPT-3 In-Context Learning

In this section, we first describe how we reformulate the NER and RE tasks for in-context learning. We then provide thorough descriptions of our



List the diseases mentioned in the following sentences.

Sentence: Acute liver failure in two patients with regular alcohol consumption ingesting paracetamol at therapeutic dosage.
Diseases: Acute liver failure

Sentence: Clinical evaluations suggested an initial diagnosis of severe thrombocytopenia and haemolysis.
Diseases: thrombocytopenia, haemolysis

Classify the interaction between drugs based on the provided sentences.

Sentence: It is recommended not to exceed a single Vardenafil dose when used with ritonavir.
How do Vardenafil and ritonavir interact in the previous sentence?
Interaction: advice

Sentence: Therefore, concomitant use of toradol and probenecid is contraindicated.
How do toradol and probenecid interact in the previous sentence?
Interaction: advice

Figure 2: Overall architecture for GPT-3 in-context learning for both NER and RE (left). One-shot learning example prompt for NER (middle) and RE (right). Different colors indicate different prompt design elements: orange for overall task instructions, red for sentence introduction and purple for the retrieval message portion. The current input sentence and the completion by GPT-3 are highlighted.

prompt design and in-context example retrieval approaches as well as other recent techniques we use to improve GPT-3's in-context learning performance for biomedical IE.

2.4.1 Task Linearization

As shown in the examples in Figure 2, in order to use in-context learning, we must first reformulate NER and RE as language generation tasks.

For NER, we extract all entity spans from the original sentence and combined them using a separator (entities are only added once), as was done in previous work (Raval et al., 2021). GPT-3 will then be expected to generate a list of entities joined by the chosen separator when conditioned on the current input and its context, as shown in Figure 2 (middle).

For relation extraction, we draw inspiration from Huguet Cabot and Navigli (2021) and transform every example into a prompt as shown in Figure 2 (right). For all our prompt templates shown in Appendix D, we add the subject and object entities, in their verbatim lexical form in the original sentence, to the prompt.

2.4.2 Prompt Design

Given the importance of prompt selection in obtaining strong performance from GPT-3 in-context learning (Perez et al., 2021; Schick and Schütze, 2022, 2021; Gao et al., 2021), we provide a systematic and task-agnostic process for constructing GPT-3 prompts. As shown in the examples in Figure 2, we identify three main parts of a prompt: overall task instructions, a sentence introduction and a retrieval message. The overall task instruction provides broad instructions for the task as concisely as possible. The sentence introduction describes the input text (i.e., scientific article excerpt, tweet, sentence, etc.). Finally, the retrieval message directly precedes the expected completion and is meant to reiterate what is needed for the task. For relation extraction, similar to Schick et al. (2020), we also define a label verbalizer which maps relation categories to plausible natural language phrases to facilitate generation.

For each task, we manually create a set of alternatives for each prompt section and select their best combination. We use leave-one-out cross-validation (LOOCV) to choose the best combination of the prompt alternatives as well as the number of in-context examples included in the prompt. To keep costs reasonable, we compare 8 prompt alternatives for each dataset. A list of all the options for each dataset can be found in Appendix D.

2.4.3 Logit Biases

In order to prevent GPT-3 from generating tokens that are not in the original sentence, we use the *logit bias* option from the OpenAI Completion API.[2] This option allows us to add a fixed value to the final probability of a specified set of tokens, restricting the possible tokens that GPT-3 can generate. Specifically, we add a value of 10 to all tokens present in the original sentence, our chosen separator and the newline token (used to designate the end of the entity list). Additionally, any predicted entities that do not match any span in the original sentence are discarded during post-processing.

2.4.4 Contextual Calibration

During preliminary studies, we found that each set of few-shot in-context examples biased GPT-3

[2] https://beta.openai.com/docs/api-reference/completions

	Task	Train	Dev	Test	Eval. Metric
BC5CDR-disease	NER	4,182	4,244	4,424	F1 entity-level
BC5CDR-chem	NER	5,203	5,347	5,385	F1 entity-level
NCBI-disease	NER	5,134	787	960	F1 entity-level
JNLPBA	NER	46,750	4,551	8,662	F1 entity-level
BC2GM	NER	15,197	3,061	6,325	F1 entity-level
DDI	RE	25,296	2,496	5,716	Micro F1
ChemProt	RE	18,035	11,268	15,745	Micro F1
GAD	RE	4,261	535	534	Micro F1

Table 1: Dataset statistics.

towards certain labels regardless of the test input. Previous work (Zhao et al., 2021) proposes to address these biases by calibrating the output using a linear transformation which equalizes all label probabilities generated by GPT-3 when conditioned on a null prompt (a version of the original prompt in which the test input is replaced by a null value such as "N/A"). This linear transformation is then used to update the output probabilities of the true few-shot prompt, thereby removing the context induced biases. We adopt this approach for RE and create the null prompt by replacing the original sentence as well as the subject and object entities in the retrieval message with "N/A".

2.4.5 Retrieval Module

Several studies (Liu et al., 2022; Rubin et al., 2022; Shin et al., 2021) suggest that choosing few-shot in-context examples for each test example dynamically instead of using a fixed set of in-context examples yields strong improvements for GPT-3 in-context learning. Following Liu et al. (2022), we use a k-nearest neighbor (kNN) retrieval module to select the most similar examples in our training set as the few-shot in-context prompt for each test example. We opt for RoBERTa-large as the encoder for our kNN retrieval module after preliminary experiments showing its advantages over other alternatives including biomedical PLMs (Lee et al., 2019; Gu et al., 2021), sentence-transformer models (Reimers and Gurevych, 2019) and a BM25 baseline (Robertson and Zaragoza, 2009).

3 Experiments

3.1 Datasets

We use all NER and RE datasets exactly as they are used in the BLURB benchmark (Gu et al., 2021) to evaluate biomedical IE. Table 1 lists the datasets and their statistics. For processing and train/dev/test splits, we refer the interested reader to Section 2.3 of Gu et al. (2021).

3.1.1 Named Entity Recognition

BC5CDR. The BioCreative V Chemical-Disease Relation corpus (Li et al., 2016) contains PubMed abstracts with both disease and chemical annotations; we evaluate models on each entity type separately following previous work (Gu et al., 2021).

NCBI-disease. The Natural Center for Biotechnology Information Disease corpus (Doğan et al., 2014) contains disease name and concept annotations for 793 PubMed abstracts.

JNLPBA. The Joint Workshop on Natural Language Processing in Biomedicine and its Applications dataset (Collier and Kim, 2004) contains 2,000 abstracts from MEDLINE selected and annotated by hand for gene related entities.

BC2GM. The Biocreative II Gene Mention corpus (Smith et al., 2008) contains 17,500 sentences from PubMed abstracts labeled for gene entities.

3.1.2 Relation Extraction

DDI. The DDI dataset (Herrero-Zazo et al., 2013) consists of sentences from MEDLINE and DrugBank labeled with drug-drug interactions categorized into 4 true and one vacuous relation.

ChemProt. ChemProt (Krallinger et al., 2017) is a dataset consisting of 1,820 PubMed abstracts with annotated chemical-protein interactions categorized into 5 true and one vacuous relation.

GAD. The Genetic Association Database corpus (Bravo et al., 2015) consists of scientific excerpts and abstracts distantly annotated with gene-disease associations.

3.2 Compared Methods

In our main experiments, we compare three pre-trained language models, PubMedBERT-base (Gu et al., 2021),[3] BioBERT-large (Lee et al., 2019) and RoBERTa-large (Liu et al., 2019), fine-tuned on 100 training examples, with GPT-3 in-context learning where each test example's in-context prompt was retrieved from the same 100 training examples.[4] Both PubMedBERT and BioBERT were pre-trained on a large corpus of PubMed articles; PubMedBERT was pre-trained from scratch with a biomedical-specific vocabulary while BioBERT was initialized from a BERT checkpoint. We use RoBERTa-large as a strong representative for general-domain PLMs. We refer the interested

[3] We use the base version of PubMedBERT since larger versions are not publicly available.

[4] We used the original `davinci` model for all GPT-3 experiments.

	PubMedBERT-base	BioBERT-large	RoBERTa-large	GPT-3 In-Context
	Precision / Recall / F1	Precision / Recall / F1	Precision / Recall / F1	Precision / Recall / F1
BC5CDR-disease	$67.4_{3.7}$/ $67.5_{1.2}$/ $67.4_{2.4}$	$62.9_{5.0}$/ $69.0_{3.0}$/ $65.8_{4.1}$	$66.9_{1.7}$/ $68.7_{4.7}$/ $\mathbf{67.7}_{1.8}$	$57.9_{2.3}$/ $35.0_{2.9}$/ $43.6_{2.2}$
BC5CDR-chem	$86.1_{1.9}$/ $88.6_{4.8}$/ $\mathbf{87.3}_{1.3}$	$84.8_{2.6}$/ $87.3_{3.3}$/ $86.0_{1.1}$	$82.1_{1.8}$/ $87.3_{1.0}$/ $84.6_{1.3}$	$74.7_{2.5}$/ $71.4_{2.2}$/ $73.0_{0.3}$
NCBI-disease	$68.5_{4.7}$/ $67.6_{2.4}$/ $\mathbf{68.0}_{2.9}$	$59.6_{10.6}$/ $67.0_{6.1}$/ $63.0_{8.7}$	$64.3_{3.7}$/ $68.7_{6.7}$/ $66.4_{5.1}$	$55.2_{6.7}$/ $49.0_{6.1}$/ $51.4_{1.4}$
JNLPBA	$56.9_{2.9}$/ $67.9_{1.7}$/ $61.9_{2.4}$	$57.4_{1.9}$/ $73.7_{1.8}$/ $64.6_{1.8}$	$57.2_{2.9}$/ $75.1_{2.4}$/ $\mathbf{65.0}_{2.7}$	$44.7_{1.0}$/ $52.4_{3.7}$/ $48.3_{2.1}$
BC2GM	$55.4_{0.4}$/ $57.9_{7.2}$/ $\mathbf{56.5}_{3.2}$	$53.6_{0.8}$/ $59.2_{2.0}$/ $56.2_{1.0}$	$49.7_{2.1}$/ $56.3_{5.3}$/ $52.7_{2.2}$	$43.0_{8.2}$/ $40.8_{2.3}$/ $41.4_{2.7}$
NER Average	$66.9_{1.0}$/ $69.9_{0.9}$/ $\mathbf{68.2}_{0.8}$	$63.7_{1.8}$/ $71.3_{0.4}$/ $67.1_{0.9}$	$64.0_{1.6}$/ $71.2_{0.5}$/ $67.2_{0.9}$	$55.1_{3.6}$/ $49.7_{0.6}$/ $51.5_{1.3}$
DDI	$19.9_{2.0}$/ $79.1_{3.0}$/ $31.8_{2.7}$	$17.3_{1.4}$/ $75.4_{1.2}$/ $28.2_{1.9}$	$25.5_{2.2}$/ $77.9_{3.5}$/ $\mathbf{38.4}_{2.6}$	$9.6_{1.1}$/ $48.6_{1.9}$/ $16.1_{1.6}$
ChemProt	$17.9_{2.2}$/ $62.0_{3.9}$/ $27.7_{2.9}$	$19.0_{6.8}$/ $60.6_{8.2}$/ $28.7_{8.7}$	$22.0_{0.3}$/ $69.7_{1.2}$/ $\mathbf{33.4}_{0.4}$	$15.9_{0.8}$/ $68.9_{1.9}$/ $25.9_{1.3}$
GAD	$63.7_{6.6}$/ $57.2_{7.9}$/ $60.2_{7.4}$	$63.2_{5.8}$/ $72.7_{5.7}$/ $67.6_{5.8}$	$64.1_{4.0}$/ $78.5_{11.5}$/ $\mathbf{70.3}_{5.6}$	$51.4_{0.9}$/ $92.3_{4.2}$/ $66.0_{1.8}$
RE Average	$33.8_{2.0}$/ $66.1_{2.8}$/ $39.9_{2.2}$	$33.2_{0.6}$/ $69.6_{2.3}$/ $41.5_{1.4}$	$37.2_{1.8}$/ $75.4_{4.5}$/ $\mathbf{47.4}_{1.9}$	$25.6_{0.1}$/ $70.0_{1.4}$/ $36.0_{0.4}$

Table 2: Comparison of the true few-shot performance of fine-tuned BERT-sized PLMs with GPT-3 in-context learning on biomedical IE datasets from the BLURB benchmark (Gu et al., 2021). We run all experiments on at most 1,000 test examples from each dataset and use 3 different 100-example training sets to account for data variance (standard deviation found in subscripts).

reader to Appendix E for results on the base versions of BioBERT and RoBERTa.

Implementation Details. We choose 100 training examples for our experiments as a reasonable number of annotated examples with which to start training an IE model for a new task.[5] For the RE tasks, we use a balanced set of 100 examples evenly distributed over all relation types. All BERT-sized PLMs are fine-tuned using the HuggingFace Transformers library (Wolf et al., 2020). For our GPT-3 experiments, we use a maximum of 10 and 5 in-context examples for NER and RE respectively to remain within GPT-3's input length limit. Due to the high cost of GPT-3, we evaluate all methods on at most 1,000 test examples from each dataset, using the same subset for all methods. For RE, the test examples are sampled in a stratified fashion to reflect the original test set distribution of relation types. Model and prompt design selection are done following the true few-shot framework we described in §2.2. To account for training data variance, we run all experiments using 3 different 100-example training sets and report the mean and standard deviation.

4 Results & Discussion

4.1 Main Results

Our main experimental results can be found in Table 2. We first note that fine-tuned BERT-sized PLMs outperform GPT-3 in-context learning across all datasets, often by a large margin (on average 15.6-16.7% for NER and 3.9-11.4% for RE in F1). For NER, even though GPT-3's precision already drops by an average of 10 points, recall drops by twice as much. This indicates that entity *under-prediction* is an important factor for GPT-3's poor in-context learning performance. In contrast, GPT-3's precision decreases much more steeply in the RE tasks due in part to the poor performance on the `none` relation class. In §4.4, we explore the reasons behind these issues in greater depth.

Drilling down into the fine-tuning results, we note that BERT-sized PLMs obtain reasonable performance on the NER tasks, considering the extremely small size of the training sets. We obtain strong performance in the mid 80s for the drug extraction task (BC5CDR-chem) due to the high lexical regularity of drug names (e.g., suffixes like "-ate", "-ine" or "-ol" are very frequent). On other biomedical NER datasets such as disease and gene extraction, performance stalls in the high and low 60s, respectively. This performance gap is likely due to the higher lexical diversity present in gene and disease names and is also observed in PLMs fine-tuned on the full training sets, which typically achieve scores in the low or mid 80s compared to low 90s for disease recognition (Gu et al., 2021). It is also worth noting that the base version of PubMedBERT outperforms the larger versions of the general-domain RoBERTa model and biomedicine-specific BioBERT model, suggesting that pre-training on domain-specific text and vocabulary from scratch is especially beneficial for NER, reinforcing the findings in Gu et al. (2021).

Given the higher complexity of the task, it is not surprising that performance deteriorates for all evaluated methods on RE tasks (especially for DDI and ChemProt since they contain more relation types

[5] A smaller training size (e.g., 10) would likely work in GPT-3's favor but is less representative of practical applications: a serious practitioner would likely annotate 100 (compared to 10) examples if it would bring significant gains.

and higher class imbalance). In contrast with the NER task and previous work using larger training sets (Gu et al., 2021), RoBERTa-large notably outperforms PubMedBERT-base and BioBERT-large in the RE task. This suggests that, in the low-resource setting, larger-scale general-domain pre-training offsets the advantage of domain-specific pre-training in tasks which require more advanced syntactic and semantic understanding such as RE.

4.2 Ablation Studies for GPT-3

In Tables 3 and 4, we present ablation studies demonstrating the effectiveness of the techniques used to improve GPT-3 performance. These studies are done on a subset of 250 validation examples from one representative dataset for each task. We follow the LOOCV process discussed in §2.4.2 and use the same experimental setup as the main experiments with the exception of using only one 100-example training set instead of three.

We ablate the kNN module for both tasks, replacing it with a module which randomly assigns examples from the training set to each test example's in-context prompt. As we can see in both Table 3 and 4, removing the kNN module reduces GPT-3 in-context learning performance. Performance drops more steeply for RE than NER, indicating that NER is more resilient to different in-context examples. This is to be expected given that there are only a limited number of completions to choose from in the RE task and thus having similar examples (with likely the same class label as the test example) would favorably bias GPT-3 towards predicting that class label. For NER, conversely, the diversity of entities is large and so it is rare that a training sentence would have similar completions to a given test example in the low-resource setting.

	F1	Precision	Recall
Best Model	46.3	42.5	50.9
– **kNN Module**	45.3	42.7	48.2
– **Logit Biases**	42.6	66.7	31.3
– **Both**	38.7	60.2	28.5

Table 3: NER ablation study on BC5CDR-disease.

	F1	Precision	Recall
Best Model	26.1	16.1	68.0
– **kNN Module**	18.6	11.5	48.0
– **Calibration**	23.6	14.6	62.0
– **Both**	16.9	10.9	38.0

Table 4: RE ablation study on DDI.

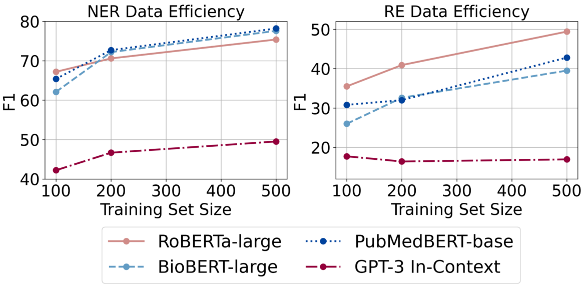


Figure 3: Data efficiency curves for BC5CDR disease NER dataset (left) and DDI RE dataset (right).

In our NER-specific ablation study, we find that removing the logit bias option leads to a large drop in performance even though precision improves. This boost in precision is due to our post-processing which removes predicted entities that are not in the original sentence and eliminates false positives. However, since invalid entities are generated instead of the valid spans which could be correct, recall drops. When ablating the kNN module and removing the logit bias option, we see an even greater drop, indicating that they are complementary. As for our RE-specific ablation study, removing the calibration module results in a drop in both precision and recall, with or without the kNN module, verifying its effectiveness.

4.3 Data Efficiency

In practice, choosing an optimal machine learning model requires considering not only a model's overall performance but also, crucially, its *data efficiency*, i.e., how performance improves w.r.t the amount of labeled data added. Previous work shows that GPT-3 in-context learning performance improves as dataset size increases when using kNN retrieval (Liu et al., 2022). Thus, we explore whether adding more training examples to sample from leads to performance improvements via more relevant in-context examples. In this experiment, we expand the training dataset to 200 and 500 training examples for one representative dataset from each task: BC5CDR-disease and DDI. For the BERT-sized PLMs, we carry out the same cross-validation procedure for model selection as in the main experiments. For GPT-3, we utilize the same optimal prompt design obtained from the main experiments to keep costs manageable. As shown in Figure 3, for NER, we find that performance for in-context learning improves at a similar rate as the small PLMs, keeping the large gap between them constant. On the other hand, for RE, GPT-3's performance quickly falls behind. This behavior can be partially explained by the fact that `none` relation examples are more challenging to retrieve by leveraging simple lexical features than their positive class counterparts.[6] Overall, GPT-3 in-context learning does not seem to yield a high return for more data annotation to compensate for its lower few-shot performance, so fine-tuning BERT-sized PLMs is likely still a better choice in the medium to high-data regime.

4.4 Detailed Error Analysis

In this section, our in-depth analysis reveals the difficulty of in-context learning in handling the `null` class,[7] such as sentences that contain no entities (for NER) and entity pairs that hold none of the target relations (for RE). Such issues do not seem to be specific to biomedical applications but are likely detrimental for IE tasks in general.

4.4.1 NER Error Analysis

When applying an NER (or similar span extraction tasks such as slot filling) model in practice on an input sentence, it may, more often than not, contain no relevant entity at all (what we call `null` class examples). For example, up to 50% sentences in the BC5CDR-disease dataset contain no disease. However, existing work on GPT-3 in-context learning has ignored this issue. For instance, Zhao et al. (2021) chose to remove all examples that contain no relevant slots from their slot filling experiment. Unfortunately, as we will show, such `null` class examples turn out to be a major culprit of in-context learning's poor performance.

Original BC5CDR-disease			
	F1	**Precision**	**Recall**
GPT-3 In-Context	43.6	57.9	35.0
RoBERTa-large	67.7	66.9	68.7
Modified BC5CDR-disease			
	F1	**Precision**	**Recall**
GPT-3 In-Context	59.8	60.3	59.3
RoBERTa-large	70.4	68.0	72.9

Table 5: Evaluation on modified BC5CDR-disease where sentences with no disease entity are removed.

To explore the effect of such `null` examples, we compare GPT-3 in-context learning with fine-tuned RoBERTa-large on a modified BC5CDR-disease dataset in which all sentences containing no disease entities are removed. As shown in Table 5, recall for GPT-3 improves by around 24%, compared to only 4% for RoBERTa-large, indicating that including `null` examples in a prompt biases GPT-3 much more strongly to predict few entities than adding them to the fine-tuning data.

Number of Entities	$\mathbb{P}$(**`null`**) **2-Shot**	$\mathbb{P}$(**`null`**) **3-Shot**	**Absolute** Δ	**% Increase**
Zero (`null`)	19.4	49.1	29.7	153%
One or More	15.8	40.9	25.1	159%

Table 6: We compare the `null` token probability assigned by GPT-3 to examples with zero and non-zero entities in the BC5CDR-disease training dataset. We run GPT-3 on 2-shot and 3-shot prompts (the 3-shot prompts contain one extra `null` example to examine its effect). We present the average over 3 randomly chosen prompts.

We hypothesize that this bias is due, at least in part, to the fact that GPT-3 in-context learning must infer that relevant entities should only be predicted if they are present in the given sentence, in contrast with smaller PLMs using the token-classification formulation. In order to examine this hypothesis more closely, we simplify our experimental setting to isolate the effect that an additional `null` example has on GPT-3's predictions. We run GPT-3 on the BC5CDR-disease training dataset without the k-NN retrieval module, instead using the same randomly chosen two-shot prompt (containing an example with no entities and one with at least one) across all examples. We then add one more random example without entities to every prompt and compare the probability of a `null` prediction in each setting.[8] As shown in Table 6, we find that, while adding the second `null` example increases the `null` probability slightly more for zero entity examples than ones with entities in absolute terms, accounting for the lower initial `null` probability that is assigned to examples with one entity or more reverses this effect. The absence of a significantly larger increase on the `null` probability for examples with zero entities over others suggests that GPT-3 struggles to infer the appropriate prediction constraint for this task and rather increases the `null` probability somewhat uniformly across examples.

[6] See §4.4.2 for more discussion.

[7] It is named after the null hypothesis for its similar nature.

[8] We measure `null` prediction probability instead of performance since entity-level F1 would not capture any information about examples with no entities.

Label	Example	Model	Correct
Effect	Concurrent use of **phenothiazines** may antagonize the anorectic effect of **diethylpropion**.	RoBERTa-large	✓
	Concurrent use of **phenothiazines** may antagonize the anorectic effect of **diethylpropion**.	GPT-3	✓
None	Other strong inhibitors of CYP3A4 (e.g., itraconazole, **clarithromycin**, nefazodone, troleandomycin, ritonavir, **nelfinavir**) would be expected to behave similarly.	RoBERTa-large	✓
	Other strong inhibitors of CYP3A4 (e.g., itraconazole, **clarithromycin**, nefazodone, troleandomycin, ritonavir, **nelfinavir**) would be expected to behave similarly.	GPT-3	✗ (Mechanism)

Table 7: We compare LIME-based saliency scores for two DDI examples predicted by GPT-3 in-context learning and RoBERTa-large. Masking out words highlighted in blue changes the model's current prediction (the color's intensity indicates the effect of removing each word on the final prediction). The drugs shown in **bold** are the head and tail entities for the relation being queried. The second example shows that GPT-3 in-context learning is more prone to spurious surface-level signals and thus suffers in correctly predicting the `none` class.

4.4.2 RE Error Analysis

We similarly examine the effect of the `null` class for RE, which is denoted as the `none` relation in the DDI dataset analyzed. As seen in Table 2, GPT-3 in-context learning achieves high recall but low precision on RE datasets that have multiple relation types such as DDI and ChemProt. Based on the confusion matrices derived from LOOCV (Appendix F.1), the `none` relation in DDI is rarely predicted by GPT-3. This bias against the `none` class greatly degrades the model's precision given that the DDI dataset is, rightfully so, heavily skewed towards this class.

In order to further understand this bias, we use LIME (Ribeiro et al., 2016)[9] to analyze the predictions for both GPT-3 and RoBERTa on an `effect` example and a `none` example.[10] The first example in Table 7 was labeled correctly by both models by relying on "anorectic effect" as a relevant signal. For `none` examples, however, correct predictions often require the use of more implicit structural understanding rather than reliance on surface level signals, as can be seen in the second example in Table 7. In this `none` example, we note that RoBERTa-large's prediction is strongly affected by the phrase "*of CYP3A4 (e.g.,*" which helps express that the drugs within the parenthesis are examples of the same drug class and therefore do not interact with each other. This suggests that RoBERTa correctly leverages the linguistic structure of the sentence. On the other hand, GPT-3's incorrect `mechanism` prediction appears to be supported by the phrase "*expected to behave similarly*", which is not relevant to the relation between the drugs being queried. This suggest that GPT-3 in-context learning is more prone to spurious surface-level signals and thus suffers in predicting the `none` class.

4.4.3 General Limitation or Domain Shift?

Our analysis suggests that GPT-3's in-context learning faces a broader issue concerning the higher complexity of `null` examples compared to positive examples. However, given that there is little work thoroughly studying GPT-3 for general domain IE, we leave it for future work to determine to what extent our findings stem from this `null` class limitation, the biomedical domain shift, or some other unforeseen reasons.

5 Related Work

In-Context Learning. GPT-3 in-context learning (Brown et al., 2020) has been found to be competitive against supervised baselines in a broad range of tasks including text classification, natural language inference, machine translation, question answering, table-to-text generation and semantic parsing (Brown et al., 2020; Zhao et al., 2021; Liu et al., 2022; Shin et al., 2021). Many techniques have been introduced to bolster its performance by removing biases through calibration (Zhao et al., 2021; Malkin et al., 2022) as well as by optimizing prompt retrieval (Liu et al., 2022; Rubin et al., 2022; Shin et al., 2021), prompt ordering (Lu et al., 2022) and prompt design (Perez et al., 2021).

Previous work exploring GPT-3's in-context learning performance for information extraction tasks is limited. Zhao et al. (2021) evaluate smaller GPT-3 models on a modified slot filling task in which all examples have at least one entity of interest. Additionally, Epure and Hennequin (2021) evaluate the in-context learning performance of GPT-2 on open-domain NER datasets, modified to keep a specific ratio of empty to non-empty examples. Our prompt design for biomedical NER draws heavily from both of these works.

[9] Our LIME process is described in Appendix F.3.

[10] Other similar examples are discussed in Appendix F.2.

As far as we know, our work is among the first to comprehensively evaluate GPT-3's in-context learning performance on IE tasks.

Prompt Design. Apart from work on in-context learning, several other research directions study how to reformulate NLP tasks as language generation tasks. Schick and Schütze (2021) reformulate text classification and natural language inference tasks using a diverse set of manually constructed cloze-style templates as prompts to improve few-shot learning in smaller pretrained language models. Gao et al. (2021) explore a similar setting but leverage an external language model to generate such templates. Both of these demonstrate the importance of using a variety of prompt designs.

In a related direction, Huguet Cabot and Navigli (2021) achieve state-of-the-art performance on relation extraction benchmarks by reformulating it as an end-to-end sequence-to-sequence task. In the biomedical domain, several works (Raval et al., 2021; Phan et al., 2021; Parmar et al., 2022) follow the multi-task sequence-to-sequence paradigm introduced by Raffel et al. (2020) and outperform previous methods on many tasks such as side effect extraction, NER, RE, natural language inference and question answering. Our prompt design is heavily inspired by many of these efforts to reformulate IE tasks as sequence-to-sequence tasks.

True Few-Shot Learning. Perez et al. (2021) argue that previous work overestimates the few-shot learning abilities of PLMs by using large validation sets for model and prompt selection. This setting has been adopted by many works in this direction in an effort to more accurately estimate few-shot performance (Logan IV et al., 2022; Schick and Schütze, 2022; Lu et al., 2022).

Biomedical In-Context Learning. Previous work evaluating GPT-3's in-context learning abilities on biomedical NLP tasks suggests that using the GPT-3 API directly yields poor performance in the biomedical domain (Moradi et al., 2021). Their work provides experimental results on 5 biomedical NLP datasets on distinct tasks including relation extraction. In our study, we aim to provide a comprehensive and in-depth evaluation on biomedical IE by using an established multi-dataset biomedical NLP benchmark and leverage recent in-context learning techniques to obtain the highest possible performance to our knowledge and ability. However, our results ultimately provide more evidence for the inadequacy of GPT-3 in-context learning for biomedical IE tasks, which cannot be easily overcome with existing techniques. Interestingly, a concurrent work (Agrawal et al., 2022) finds that GPT-3 perform well on a different set of *clinical* IE tasks, including one on biomedical evidence extraction that is clinical in nature. More work is needed to ascertain the cause of this surprising gap in IE performance between the clinical and biomedical domains for in-context learning.

6 Conclusions

In this work, we explored the potential of GPT-3 in-context learning for the high impact task of biomedical information extraction (IE). Given that such a paradigm would provide significant advantages for biomedical IE applications, we spent considerable efforts exploring available techniques that have been proven effective for other in-context learning settings. We showed, however, that current techniques do not enable GPT-3 in-context learning to surpass BERT-sized PLM fine-tuning on a range of benchmark datasets for biomedical NER and RE. Additionally, we discussed some potentially general limitations of in-context learning in biomedical IE to be explored in future work: its difficulty in handling the `null` class, such as entity-less NER examples and vacuous relation examples for RE. Apart from posing this question for further study, we hope our work provides helpful guidance to biomedical researchers and practitioners towards more promising and cost-effective tools for low-resource IE such as small PLM fine-tuning or perhaps even directly fine-tuning GPT-3.

Limitations

While we have uncovered a large performance gap between current GPT-3 in-context learning techniques and standard fine-tuning in the true few-shot setting, there are several important limitations that are worth discussing. Our limited budget restricted our study to a small set of prompt styles and number of examples in the prompt. Although our experiments suggest otherwise, it is possible that having a larger prompt design search space or using more examples per prompt could narrow the gap between small PLM fine-tuning and GPT-3 in-context learning. Additionally, it is still unclear to what degree using larger validation sets, at the cost of compromising the few-shot assumption, for prompt selection could improve GPT-3's in-context learning performance. Perhaps more notably, the kNN retrieval module used in this study relies on whole sentence embeddings, as commonly done in the existing literature. However, intuitively, tasks like relation extraction require a more focused view around the target entity pair. We speculate that developing a better retrieval module that is able to incorporate such task-specific inductive biases will likely be beneficial for in-context learning, but we leave it for future work. Finally, it is important to note that while contextual calibration (Zhao et al., 2021) is shown to work well in some text classification tasks, it is unclear whether other more recent methods such as that by Malkin et al. (2022) could better address GPT-3's text generation biases or if more task-specific calibration mechanisms are necessary for IE tasks.

Acknowledgements

The authors would like to thank colleagues from the OSU NLP group for their thoughtful comments. This research was supported in part by OSU TDAI, NSF OAC 2118240, NSF OAC 2112606, NSF IIS 1815674, NSF CAREER 1942980, and Ohio Supercomputer Center (Center, 1987).

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are zero-shot clinical information extractors.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. 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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16(1):55.

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.

Ohio Supercomputer Center. 1987. Ohio Supercomputer Center.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the Bio-entity Recognition Task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.

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.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Elena V. Epure and Romain Hennequin. 2021. A Realistic Study of Auto-regressive Language Models for Named Entity Typing and Recognition. *CoRR*, abs/2108.11857.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3(1).

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics*, 46(5):914–920.

Pere-Lluís Huguet Cabot and Roberto Navigli. 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.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López, Umesh Nandal, et al. 2017. Overview of the biocreative VI chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)*, 2016:baw068.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Robert Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting down on prompts and parameters: Simple few-shot learning with language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835, Dublin, Ireland. 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.

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*.

Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. 2021. GPT-3 Models are Poor Few-Shot Learners in the Biomedical Domain. *CoRR*, abs/2109.02555.

Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, Murad Mohammad, and Chitta Baral. 2022. In-BoXBART: Get instructions into biomedical multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 112–128, Seattle, United States. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True Few-Shot Learning with Language Models.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. SciFive: a text-to-text transformer model for biomedical literature. *CoRR*, abs/2106.03598.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Shivam Raval, Hooman Sedghamiz, Enrico Santus, Tuka Alhanai, Mohammad Ghassemi, and Emmanuele Chersoni. 2021. Exploring a Unified Sequence-To-Sequence Transformer for Medical Product Safety Monitoring in Social Media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3534–3546, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

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.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically Identifying Words That Can Serve as Labels for Few-Shot Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2022. True few-shot learning with Prompts—A real-world perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731.

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.

Larry Smith, Lorraine K Tanabe, Rie Johnson Nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A Struble, Richard J Povinelli, Andreas Vlachos, William A Baumgartner, Jr, Lawrence Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W John Wilbur. 2008. Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(S2):S2.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zheng, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. *CoRR*, abs/2201.11990.

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.

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*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Experimental Setup Details

A.1 Named Entity Recognition

We follow the BIO tag formulation for NER and use standard fine-tuning process for PLMs used in Gu et al. (2021). Given a sentence containing n tokens $X = [x_1, ..., x_n]$, an NER system attempts to predict a tag for each token: $Y = [y_1, ..., y_n]$, which can then be translated into a set of k entities. An encoder H is used to obtain a contextualized representation for the sentence X: $H(X) = [\vec{v_1}, ..., \vec{v_n}]$. Each embedding $\vec{v_i}$ is then used to predict y_i using a linear layer. The encoder H and the linear layer are then fine-tuned using a standard cross entropy objective. We use NLTK (Bird et al., 2009) to tokenize all NER sentences.

A.2 Relation Extraction

For RE we use the simplest formulation in standard fine-tuning, the subject and object entities for each relation are replaced in the original sentence by new special tokens [ENT1] and [ENT2]. An encoder H is then used as in NER to obtain a contextualized representation $H(X) = [\vec{v_1}, ..., \vec{v_n}]$ of the now masked sentence. As is standard for text classification tasks, the [CLS] token embedding is then used to predict each relation type. As with the NER task, a standard cross entropy loss is used to fine-tune the encoder and linear layer.

B Computational Cost

For our experiments, we used 4 NVIDIA GeForce RTX 2080 Ti GPUs. The number of parameters for each model we used as well as the total GPU hours and costs associated with using GPT-3 are listed in Table 8.

	# of Parameters (millions)	Total GPU Hours	Total Cost
RoBERTa-large	354	338	-
PubMedBERT-base	100	138	-
BioBERT-large	345	344	-
GPT-3 (davinci)	175,000	-	~$1,500

Table 8: Total GPU Hours and GPT-3 costs associated with our experiments.

C Fine-Tuning Hyperparameters

We run 5-fold cross validation for each 100 sample training subset to choose between all hyperparameters listed in Table 9.

	Learning Rate	Batch Size	Warmup Ratio	Weight Decay	Early Stopping Checkpoint
Search Space	1e-5 2e-5 3e-5 5e-5	16 32	0.0 0.06	0.0 0.01 0.1	5 10 15 20 25

Table 9: Hyperparameter search grid used with k-fold cross-validation to obtain the optimal hyperparameters for all PLM fine-tuning experiments.

D Prompt Designs

We run leave-one-out cross validation for each 100 sample training subset to choose between all choices listed in Table 10. Prompt design selections were completely independent for each training subset to maintain the true few-shot learning setting.

E Base Models

As expected, the base models added to Table 11 underperform their large counterparts on almost all datasets. Consistent with previous work (Gu et al., 2021), benefiting from the biomedical-specific vocabulary, PubMedBERT-base handily outperforms other base models on the NER task (as well as some large models on several tasks). However, on the RE tasks, RoBERTa models perform the best. Since RE tasks requires more holistic understanding of the whole sentence, this suggests that RoBERTa provides more general linguistic knowledge than other PLMs specific to biomedicine.

F DDI Error Analysis

F.1 Confusion Matrices

Figure 4 shows the error distribution for both GPT-3 and RoBERTa-large in a 100 example training set for the DDI relation extraction dataset. We obtain these by combining all folds from 5-fold and leave-one-out cross-validation for RoBERTa-large and GPT-3 respectively. From the figure, we can see that GPT-3 in-context learning rarely predicts the `none` class which indicates two drugs bare no relation to each other. We note that RoBERTa-large also suffers from a larger error rate for the `none` class than other classes, indicating that this class is challenging for both models, however, the gap is

NER					
	Overall Instructions	**Sentence Introduction**	**Retrieval Message**	**# of In-Context Examples**	**Label Verbalizer**
BC5CDR-disease	""	Sentence:	Diseases:	5	N/A
	List the diseases mentioned in the following sentences.	Scientific Article Excerpt:		10	
BC5CDR-chemical	""	Sentence:	Drugs:	5	
	List the drugs mentioned in the following sentences.	Scientific Article Excerpt:		10	
NCBI-disease	""	Sentence:	Diseases:	5	
	List the diseases mentioned in the following sentences.	Scientific Article Excerpt:		10	
JNLPBA	""	Sentence:	Genes:	5	
	List the genes mentioned in the following sentences.	Scientific Article Excerpt:		10	
BC2GM	""	Sentence:	Genes:	5	
	List the genes mentioned in the following sentences.	Scientific Article Excerpt:		10	

RE					
	Overall Instructions	**Sentence Introduction**	**Retrieval Message**	**# of In-Context Examples**	**Label Verbalizer**
DDI	""	Sentence:	Drug 1: <DRUG1> Drug 2: <DRUG2> Interaction:	5	DDI-effect > effect DDI-false > none DDI-advise > advice DDI-mechanism > mechanism DDI-int > other
	Classify the interaction between drugs based on the provided scientific article excerpts.	Scientific Article Excerpt:	How do <DRUG1> and <DRUG2>interact according to the previous sentence? Which word best describes their interaction: advice, effect, mechanism, other or none? Interaction:		
ChemProt	""	Sentence:	Drug: <DRUG> Gene: <GENE> Effect:	5	false > none CPR:3 > activator CPR:4 > inhibitor CPR:5 > agonist CPR:6 > antagonist CPR:9 > substrate
	Classify the effect drugs have on the genes mentioned in the following scientific article excerpts.	Scientific Article Excerpt:	What effect does the drug <DRUG>have on gene <GENE>according to the previous sentence? Choose from the following: none, activator, inhibitor, agonist, antagonist or substrate. Effect:		
GAD	""	Sentence:	Gene: <GENE> Disease: <DISEASE> Interaction:	5	0 > no 1 > yes
	Determine if there is any interaction between the diseases and genes mentioned in the provided scientific article excerpts.	Scientific Article Excerpt:	Based on the previous sentence, is there any interaction between gene <GENE>and disease <DISEASE>?		

Table 10: For each element in our proposed prompt design (overall task instruction, sentence introduction and retrieval message), we list every option used for each dataset. For our main experiments, we used LOOCV on 100 training examples to select among 8 combinations of our 3 design elements and the number of in-context examples added to the prompt for each task. We also list the label verbalization used for each relation extraction dataset.

	PubMedBERT-base	BioBERT-base	RoBERTa-base	BioBERT-large	RoBERTa-large	GPT-3 In-Context
	Precision / Recall / F1	Precision / Recall / F1	Precision / Recall / F1	Precision / Recall / F1	Precision / Recall / F1	Precision / Recall / F1
BC5CDR-disease	$67.4_{3.7}$/ $67.5_{1.2}$/ $67.4_{2.4}$	$60.6_{5.1}$/ $66.1_{5.7}$/ $63.0_{2.0}$	$60.4_{2.8}$/ $61.9_{4.4}$/ $61.2_{3.6}$	$62.9_{5.0}$/ $69.0_{3.0}$/ $65.8_{4.1}$	$66.9_{1.7}$/ $68.7_{4.7}$/ $\mathbf{67.7}_{1.8}$	$57.9_{2.3}$/ $35.0_{2.9}$/ $43.6_{2.2}$
BC5CDR-chem	$86.1_{1.9}$/ $88.6_{4.8}$/ $\mathbf{87.3}_{1.3}$	$77.8_{3.3}$/ $85.1_{4.6}$/ $81.2_{0.5}$	$74.6_{3.8}$/ $84.1_{2.1}$/ $79.0_{1.2}$	$84.8_{2.6}$/ $87.3_{3.3}$/ $86.0_{1.1}$	$82.1_{1.8}$/ $87.3_{1.0}$/ $84.6_{1.3}$	$74.7_{2.5}$/ $71.4_{2.2}$/ $73.0_{0.3}$
NCBI-disease	$68.5_{4.7}$/ $67.6_{2.4}$/ $\mathbf{68.0}_{2.9}$	$58.8_{5.4}$/ $65.9_{2.7}$/ $62.1_{4.0}$	$60.6_{3.2}$/ $61.9_{4.6}$/ $61.2_{3.5}$	$59.6_{10.6}$/ $67.0_{6.1}$/ $63.0_{8.7}$	$64.3_{3.7}$/ $68.7_{6.7}$/ $66.4_{5.1}$	$55.2_{6.7}$/ $49.0_{6.1}$/ $51.4_{1.4}$
JNLPBA	$56.9_{2.9}$/ $67.9_{1.7}$/ $61.9_{2.4}$	$49.1_{0.2}$/ $66.7_{1.9}$/ $56.6_{0.8}$	$54.6_{2.7}$/ $71.4_{2.6}$/ $61.9_{2.7}$	$57.4_{1.9}$/ $73.7_{1.8}$/ $64.6_{1.8}$	$57.2_{2.9}$/ $75.1_{2.4}$/ $\mathbf{65.0}_{2.7}$	$44.7_{1.0}$/ $52.4_{3.7}$/ $48.3_{2.1}$
BC2GM	$55.4_{0.4}$/ $57.9_{7.2}$/ $\mathbf{56.5}_{3.2}$	$46.4_{2.5}$/ $57.3_{1.0}$/ $51.3_{1.9}$	$46.2_{3.0}$/ $53.7_{0.4}$/ $49.7_{1.6}$	$53.6_{0.8}$/ $59.2_{2.0}$/ $56.2_{1.0}$	$49.7_{2.1}$/ $56.3_{5.3}$/ $52.7_{2.2}$	$43.0_{8.2}$/ $40.8_{2.3}$/ $41.4_{2.7}$
NER Average	$66.9_{1.0}$/ $69.9_{0.9}$/ $\mathbf{68.2}_{0.8}$	$58.6_{0.9}$/ $68.2_{1.6}$/ $62.8_{1.0}$	$59.3_{2.8}$/ $66.6_{1.7}$/ $62.6_{2.2}$	$63.7_{1.8}$/ $71.3_{0.4}$/ $67.1_{0.9}$	$64.0_{1.6}$/ $71.2_{0.5}$/ $67.2_{0.9}$	$55.1_{3.6}$/ $49.7_{0.6}$/ $51.5_{1.3}$
DDI	$19.9_{2.0}$/ $79.1_{3.0}$/ $31.8_{2.7}$	$18.9_{0.6}$/ $78.3_{4.4}$/ $30.5_{0.9}$	$19.6_{1.3}$/ $68.8_{3.9}$/ $30.5_{1.6}$	$17.3_{1.4}$/ $75.4_{1.2}$/ $28.2_{1.9}$	$25.5_{2.2}$/ $77.9_{3.5}$/ $\mathbf{38.4}_{2.6}$	$9.6_{1.1}$/ $48.6_{1.9}$/ $16.1_{1.6}$
ChemProt	$17.9_{2.2}$/ $62.0_{3.9}$/ $27.7_{2.9}$	$18.7_{0.9}$/ $59.4_{5.0}$/ $28.4_{0.9}$	$18.1_{0.7}$/ $56.8_{1.6}$/ $27.5_{0.7}$	$19.0_{6.8}$/ $60.6_{8.2}$/ $28.7_{8.7}$	$22.0_{0.3}$/ $69.7_{1.2}$/ $\mathbf{33.4}_{0.4}$	$15.9_{0.8}$/ $68.9_{1.9}$/ $25.9_{1.3}$
GAD	$63.7_{6.6}$/ $57.2_{7.9}$/ $60.2_{7.4}$	$60.5_{5.0}$/ $62.8_{14.3}$/ $61.2_{8.1}$	$60.2_{1.4}$/ $71.2_{20.1}$/ $64.4_{9.1}$	$63.2_{5.8}$/ $72.7_{5.7}$/ $67.6_{5.8}$	$64.1_{4.0}$/ $78.5_{11.5}$/ $\mathbf{70.3}_{5.6}$	$51.4_{0.9}$/ $92.3_{4.2}$/ $66.0_{1.8}$
RE Average	$33.8_{2.0}$/ $66.1_{2.8}$/ $39.9_{2.2}$	$32.7_{1.7}$/ $66.8_{5.1}$/ $40.0_{2.7}$	$35.1_{4.6}$/ $68.0_{10.7}$/ $43.5_{7.7}$	$33.2_{0.6}$/ $69.6_{2.3}$/ $41.5_{1.4}$	$37.2_{1.8}$/ $75.4_{4.5}$/ $\mathbf{47.4}_{1.9}$	$25.6_{0.1}$/ $70.0_{1.4}$/ $36.0_{0.4}$

Table 11: Main experimental results from Table 2 with additional results from BioBERT and RoBERTa base models for appropriate comparison.

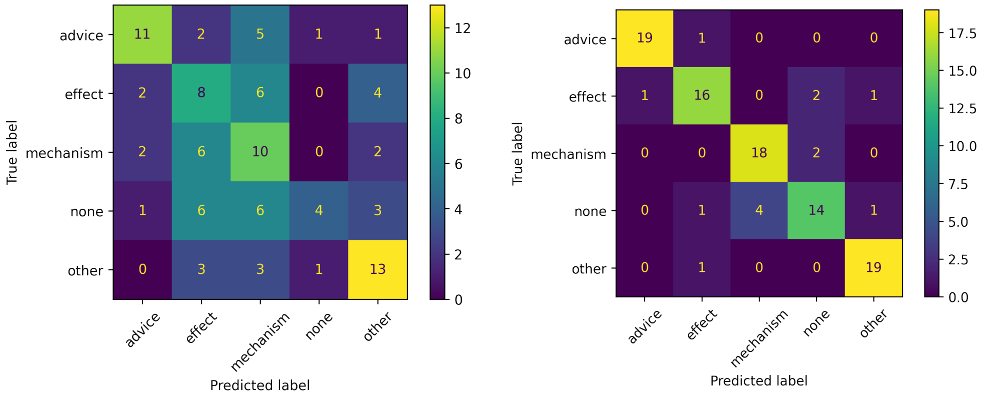


Figure 4: Confusion matrices on 100 validation examples from DDI for GPT-3 (left) and RoBERTa-large (right).

much smaller for RoBERTa than GPT-3 in-context learning.

F.2 Qualitative Analysis

In Table 12, we present 3 positive and 3 none DDI examples respectively to help illustrate the more challenging nature of the none class as well as RoBERTa-large's superior ability to attend to more relevant implicit signals. In all three positive examples, the saliency scores attributed by LIME for RoBERTa and GPT-3 agree closely, suggesting that both models are able to leverage relevant surface level signals. The feature attribution for the none examples, however, suggests that GPT-3 continues leveraging surface level signals when more complex sentence level information is needed which RoBERTa seems to extract and use more effectively.

The first none example shows GPT-3's prediction is affected by several irrelevant features such as other drugs in the drug list ("*channel*", "*quinidine*" and "*carbamazepine*"), the initial phrase explaining that specific studies have not been performed and the word "*metabolized*". In contrast, RoBERTa is unaffected by the removal of drugs from the drug list and is correctly affected by important signals such as the removal of "*CYP3A4 (eg.*", similar to the example in Table 7. For the second none example, GPT-3's incorrect prediction is most strongly affected by the words "*binding*", "*diuretic*" and "*gastrointestinal*" while for RoBERTa the effect of removing words is more uniformly distributed over the phrase "*binding thiazide diuretics and reducing diuretic absorption from the gastrointenstinal tract*". This indicates that RoBERTa's prediction relies on broader phrase level information rather than word level signals. In the last example, we note that removing the phrase "*with L-tryptophan*" from the sentence would create an interaction between the drugs being queried by yielding the phrase "*Using these medicines may increase the chance of side effects*". The fact that RoBERTa's prediction is strongly affected by the removal of this phrase indicates that its decision boundary uses more complex linguistic signals than GPT-3 which leverages single words such as "*inhibitors*", "*Using*" and "*increase*" to arrive at its prediction.

F.3 LIME Details

We choose LIME (Ribeiro et al., 2016) to perform our RE error analysis because it enables us to obtain faithful local explanations for GPT-3 in-context learning which are directly comparable with the ones from RoBERTa or other small PLMs. We

Label	Example	Model	Correct
Advice	Concomitant use of **bromocriptine mesylate** with other **ergot alkaloids** is not recommended.	RoBERTa-large	✓
	Concomitant use of **bromocriptine mesylate** with other **ergot alkaloids** is not recommended.	GPT-3	✓
Advice	Consequently, it is recommended not to exceed a single 2.5 mg **Vardenafil** dose in a 72-hour period when used in combination with **ritonavir**.	RoBERTa-large	✓
	Consequently, it is recommended not to exceed a single 2.5 mg **Vardenafil** dose in a 72-hour period when used in combination with **ritonavir**.	GPT-3	✓
Effect	However, reports suggest that **NSAIDs** may diminish the antihypertensive effect of **ACE inhibitors**.	RoBERTa-large	✓
	However, reports suggest that **NSAIDs** may diminish the antihypertensive effect of **ACE inhibitors**.	GPT-3	✓
None	Although specific studies have not been performed, coadministration with drugs that are mainly metabolized by CYP3A4 (eg, calcium channel blockers, dapsone, **disopyramide**, quinine, amiodarone, quinidine, warfarin, **tacrolimus**, cyclosporine, ergot derivatives, pimozide, carbamazepine, fentanyl, alfentanyl, alprazolam, and triazolam) may have elevated plasma concentrations when coadministered with saquinavir;	RoBERTa-large	✓
	Although specific studies have not been performed, coadministration with drugs that are mainly metabolized by CYP3A4 (eg, calcium channel blockers, dapsone, **disopyramide**, quinine, amiodarone, quinidine, warfarin, **tacrolimus**, cyclosporine, ergot derivatives, pimozide, carbamazepine, fentanyl, alfentanyl, alprazolam, and triazolam) may have elevated plasma concentrations when coadministered with saquinavir;	GPT-3	✗ (Other)
None	- Cholestyramine and colestipol resins: **Cholestytamine** and **colestipol resins** have the potential of binding thiazide diuretics and reducing diuretic absorption from the gastrointestinal tract	RoBERTa-large	✓
	- Cholestyramine and colestipol resins: **Cholestytamine** and **colestipol resins** have the potential of binding thiazide diuretics and reducing diuretic absorption from the gastrointestinal tract	GPT-3	✗ (Mechanism)
None	Monoamine oxidase (MAO) inhibitors such as **isocarboxazid** (e.g., Marplan), phenelzine (e.g., Nardil), procarbazine (e.g., Matulane), selegiline (e.g., Eldepryl), and tranylcypromine (e.g., **Parnate**): Using these medicines with L-tryptophan may increase the chance of side effects.	RoBERTa-large	✓
	Monoamine oxidase (MAO) inhibitors such as **isocarboxazid** (e.g., Marplan), phenelzine (e.g., Nardil), procarbazine (e.g., Matulane), selegiline (e.g., Eldepryl), and tranylcypromine (e.g., **Parnate**): Using these medicines with L-tryptophan may increase the chance of side effects.	GPT-3	✗ (Effect)

Table 12: LIME-based saliency scores for more DDI examples. We present 3 examples with true drug-drug interactions predicted correctly by both models and 3 none examples predicted correctly by RoBERTa-large but incorrectly by GPT-3 in-context learning. As in Table 7, masking out words highlighted in blue changes the model's current prediction and the color's intensity indicates the strength of the effect on the final prediction. The drugs shown in **bold** are the head and tail entities for the relation being queried.

use a modified version of the original LIME implementation[11] (Ribeiro et al., 2016) to carry out our analysis in Appendix F.2 and §4.4.2. Due to resource constraints, we modify the token removal method in the original implementation from randomly masking out tokens to a sliding window of 3 tokens. This allows us to look at how phrase removal changes predictions while still using a reasonable number of neighbor examples. Since we use this tool for analyzing relation extraction only, we do not remove the entities that are being queried. For GPT-3 in-context learning, we keep the few-shot prompts constant and use BLANK as the replacement token given that GPT-3 does not have a mask token. We do not observe a large difference in the saliency scores when this mask token was changed. In our visualizations, the saliency score for each word is normalized by the largest score found for that example in order to make effects more apparent.

[11] https://github.com/marcotcr/lime