

# QueryForm: A Simple Zero-shot Form Entity Query Framework

Zifeng Wang<sup>1\*</sup>, Zizhao Zhang<sup>2†</sup>, Jacob Devlin<sup>3</sup>, Chen-Yu Lee<sup>2</sup>,  
Guolong Su<sup>3</sup>, Hao Zhang<sup>3</sup>, Jennifer Dy<sup>1</sup>, Vincent Perot<sup>3†</sup>, and Tomas Pfister<sup>2</sup>

<sup>1</sup>Northeastern University <sup>2</sup>Google Cloud AI <sup>3</sup>Google Research

†Core contribution

## Abstract

Zero-shot transfer learning for document understanding is a crucial yet under-investigated scenario to help reduce the high cost involved in annotating document entities. We present a novel query-based framework, QueryForm, that extracts entity values from form-like documents in a zero-shot fashion. QueryForm contains a dual prompting mechanism that composes both the document schema and a specific entity type into a query, which is used to prompt a Transformer model to perform a single entity extraction task. Furthermore, we propose to leverage large-scale query-entity pairs generated from form-like webpages with weak HTML annotations to pre-train QueryForm. By unifying pre-training and fine-tuning into the same query-based framework, QueryForm enables models to learn from structured documents containing various entities and layouts, leading to better generalization to target document types without the need for target-specific training data. QueryForm sets new state-of-the-art average F1 score on both the XFUND (+4.6%~10.1%) and the Payment (+3.2%~9.5%) zero-shot benchmark, with a smaller model size and no additional image input.

## 1 Introduction

Form-like document understanding has become a booming research topic recently thanks to its many real-world applications in industry. Form-like documents refer to documents with rich typesetting formats, such as invoices and receipts in everyday inventory workflow. Automatically extracting and organizing structured information from form-like documents is a valuable yet challenging problem.

Recent methods (Xu et al., 2020b; Garncarek et al., 2021; Lee et al., 2022) often discuss the problem of form-like document understanding, e.g. document entity extraction (DEE), in the supervised

\*Work done while the author was an intern at Google Cloud AI Research. Email: zifengwang@ece.neu.edu

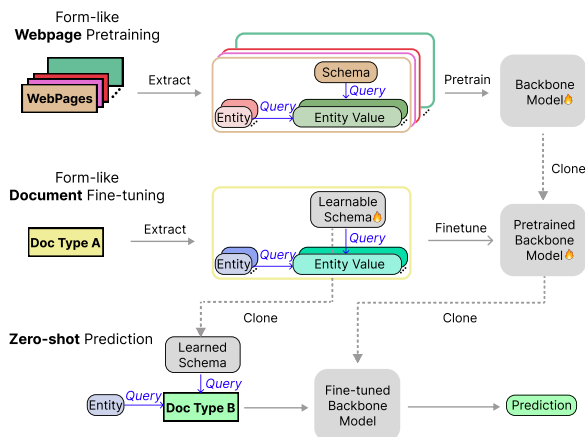


Figure 1: Illustration of the zero-shot transfer learning stages of QueryForm. In the pre-training stage, we extract millions of schemas and entity-value pairs from publicly available webpages to generate a large amount of query-value pairs to teach the backbone model to make query-conditional prediction. During fine-tuning, we extract more accurate entity-value pairs from the available annotated document and directly learn schema information from data. Finally, we evaluate the pre-trained model on a different target document without training data.

setting, assuming the training and test sets are of the same document type. However, in real-world scenarios, there is often the need for generalizing models from seen document types to new unseen document types. Beyond annotation costs, endlessly training specialized models on new types of documents is not scalable in many practical scenarios. Moreover, the methods in the supervised setting pre-define the document schema, *i.e.* the set of entities contained in the document, following the sequence-to-sequence tagging framework via the BOISE labeling format (Ratinov and Roth, 2009). Consequently, the models lack the ability to learn from different documents with diverse schemas.

Thus, it is desirable to have a systematic way to learn knowledge from existing annotated documents of different types to an un-annotated target document type (e.g. invoice in Figure 2, right).

This learning paradigm is usually defined as zero-shot transfer learning in literature (Xu et al., 2021). Beyond this, it is even more desirable to leverage highly-structured form-like documents with rich schema, such as form-like webpages in Figure 2, left. Although webpages do not have explicit human annotations, we believe the diverse schemas and natural “entities”, such as headers and text paragraphs, that exist in webpages can be valuable for document understanding. However, how to effectively utilize these webpages with a high discrepancy from documents like invoice and receipt is an unknown yet challenging problem.

In this work, we propose a novel query-based framework, QueryForm, to learn transferable knowledge from different types of documents for zero-shot entity extraction on the target document type. The workflow of QueryForm is illustrated in Figure 1. Ideally, we would like to prompt the model: *This document has the following [SCHEMA], please extract its [ENTITY] value*, and model is able to accurately predict the corresponding word tokens belong to the queried entity. To this end, we encode both schema and entity information in our query, so that the model is no longer limited by a certain document type and a fixed set of entity types (or classes). Moreover, our query-based design can even benefit further from large-scale datasets with diverse schemas and entity types.

In order to feed this kind of composite query, we propose a *dual prompting* strategy to effectively prompt the backbone model, e.g., a pre-trained Transformer, to make conditional prediction. As its name suggests, the dual prompting strategy consists of an E(ntity)-Prompt and a S(chema)-Prompt. Depending on the annotations we have, we can either generate the prompts from semantic labels, or learn them directly from data. Although similar concepts to dual prompting exist in the vision field (Wang et al., 2022a,b) to solve different problems, the main design in QueryForm is original in DEE. We also propose a query-based pre-training method, QueryWeb, which leverages the highly-accessible and inexhaustible resource - publicly available webpages. During the pre-training stage, the model learns to quickly adapt to various queries composed of different S-Prompts and E-Prompts generated from HTML source of webpages. After the decoupling of entity and schema that are tied to document types, the model can learn more

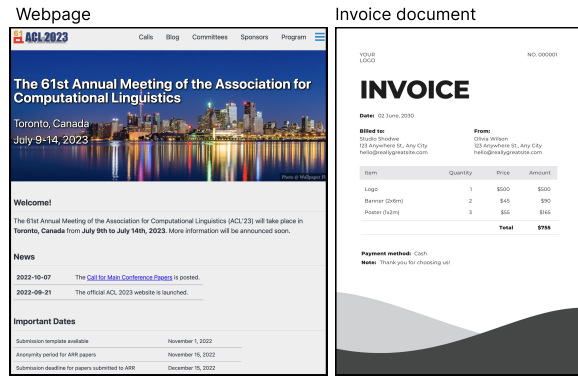


Figure 2: Form-like examples of Webpage and Invoice documents. Webpage appears to have distinct layouts and contents from invoice documents, but they both contain rich entity-value pairs, such as “*page title*-The 61st Annual Meeting of the Association for Computational Linguistics” in Webpage and “*total amount* - \$755” in Invoice.

transferable knowledge - leveraging the rich layout, scale and content information in webpages to make query-conditional predictions.

In summary, our work makes the following contributions:

- We propose QueryForm, a novel yet simple query-based framework for zero-shot document entity extraction. QueryForm provides a new dual prompting mechanism to encode both document schema and entity information to learn transferrable knowledge from source to target document types.
- We demonstrate an effective pre-training approach, QueryWeb, that collects publicly available webpages with various layouts and HTML sources, and pre-trains QueryForm via the dual prompting mechanism. Although webpages show high discrepancy from the target documents, we show this approach consistently improves the zero-shot performance.
- With extensive empirical evaluation, QueryForm sets new state-of-the-art F1 score on both Inventory-Payment and FUNSD-XFUND zero-shot transfer learning benchmarks.

## 2 Related Work

**Document entity extraction (DEE).** Researchers started to study extracting information from documents using rule-based models (Lebourgeois et al.,

1992; O’Gorman, 1993; Simon et al., 1997), or learning-based approaches with hand-engineered features (Marinai et al., 2005; Wei et al., 2013; Schuster et al., 2013). These methods have limited representation power and generalization ability.

More recently, neural models have been the main-stream solution for document entity extraction (DEE). Both RNN-based (Palm et al., 2017; Aggarwal et al., 2021) and CNN-based models (Katti et al., 2018; Zhao et al., 2019; Denk and Reisswig, 2019) have been adopted for DEE task. Nevertheless, motivated by the superior performance of the Transformers (Vaswani et al., 2017) in various NLU tasks (Devlin et al., 2018; Raffel et al., 2020), recent work develops multiple Transformer-based models for DEE. Majumder et al. (2020) extended BERT (Devlin et al., 2018) to learn representations for form-like documents; (Kim et al., 2022) propose a encoder-decoder structure that directly extracts document information from image input; Xu et al. (2020b, 2021) leverage both image and text inputs to capture cross-modality information. Lee et al. (2021, 2022) further introduced GCN (Kipf and Welling, 2016) to encode spatial relationships in addition to the Transformer backbone. However, these methods only consider the usual supervised learning setting, *i.e.*, training and test sets are from the same document type. On the contrary, our work proposes a novel query-based framework and tackle the challenging yet under-investigated zero-shot transfer learning (Xu et al., 2021) setting.

**Pre-training for DEE.** Existing large-scale pre-training techniques in NLP (Devlin et al., 2018; Conneau et al., 2019; Liu et al., 2019) are readily available for serialized document tokens. Multi-modal pre-training (Xu et al., 2020a,b, 2021; Appalaraju et al., 2021) achieves better performance than text-modality alone by incorporating visual information at the cost of more expensive data collection and computation costs. Our work presents a novel pre-training method using text modality alone, which is complementary to models that relies on image modality (Kim et al., 2022) or multiple modalities (Xu et al., 2020b, 2021). Moreover, we leverage publicly available webpages, which contain rich structured information and are much more accessible than documents. Different from the common Mask Language Model (MLM) objective used in pre-training, QueryForm has the same query-conditional objective during both pre-training and fine-tuning, which intuitively strength-

ens the transferability of pre-trained knowledge.

To the best of our knowledge, DQN (Gao et al., 2022) and Donut (Kim et al., 2022) are the closest work to ours in the DEE domain. However, our work is still very different from them from multiple perspectives, including problem setting, query design, and pre-training technique. On the other hand, leveraging webpages to pre-train language models has been explored in prior work. Liu et al. (2019); Brown et al. (2020) extract text corpora from webpages and (Aghajanyan et al., 2021) use HTML source for pre-training. However, to the best of our knowledge, we are the first to leverage both webpages and the corresponding HTML source in a novel query-based pre-training framework to address the challenging zero-shot DEE task. Our framework fully takes advantage of the rich schema and layout information from webpages and utilizes HTML tags as weak entity annotation to align pre-training with the downstream DEE task.

## 3 Preliminaries

### 3.1 Problem Formulation

Given serialized words from a form-like document, we formulate the DEE problem as sequence tagging for tokenized words, *i.e.*, for each word, we predict its corresponding entity class. Recent methods (Xu et al., 2020b; Garncarek et al., 2021; Lee et al., 2022) use the BOISE labeling format (Ratinov and Roth, 2009) - classifying the token as  $\{e\text{-Begin, Outside, } e\text{-Inside, } e\text{-Single, } e\text{-End}\}$  of a certain entity  $e \in \mathbf{E}$  to mark the entity span, where  $\mathbf{E}$  is the set of entities of interest. Thus the cardinality of the label space will be  $(4 \times |\mathbf{E}| + 1)$ . In our formulation, we explicitly encode entity in the E-Prompt. Therefore, we are able to use a more succinct and generalizable BOISE labeling with only 5 labels, {Begin, Outside, Inside, Single, End} to mark the span. Our approach decouples the label space with entity types. Following Lee et al. (2022), we then apply the Viterbi algorithm to get the final prediction.

In our work, we focus on the zero-shot DEE setting proposed by Xu et al. (2021), where 1) the training source documents have significant domain gap from the target test documents (e.g. languages or document types), 2) there is no training documents available from the target documents, and 3) source documents include entities contained in the target documents.

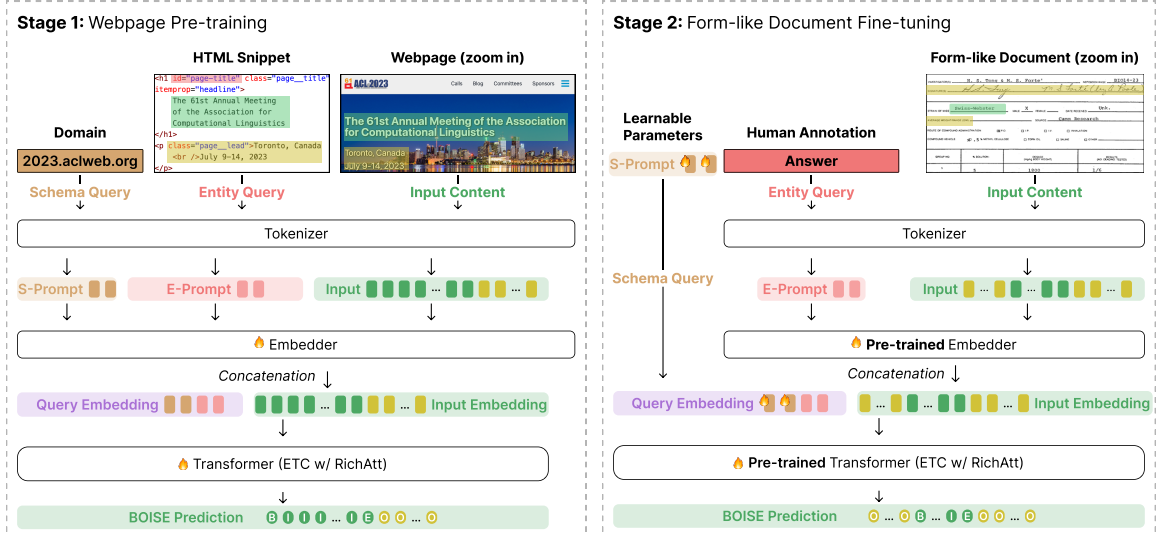


Figure 3: Overview of QueryForm. Our dual prompting design yields a consistent objective in both pre-training and fine-tuning stages. Note that the schema query in pre-training comes from website domains while it is a learnable parameter in fine-tuning. See Section 4 for more details.

### 3.2 Architecture design

Following the setting in earlier work (Majumder et al., 2020; Lee et al., 2022), our method takes the WordPiece (Wu et al., 2016) tokenized outputs from the Optical Character Recognition (OCR) engine in reading order (left-right and top-bottom). By design, our method is compatible with any sequence encoder model as the backbone. We adopt the long-sequence transformer extension, ETC (Ainslie et al., 2020) as our backbone, following the adoption of Lee et al. (2022), which contains Rich Attention as an enhancement of self-attention layers to encode 2D spatial layout information. We find this method (used as our baseline) performs fairly strong in the usual supervised learning setup, however, its performance drops significantly in the zero-shot learning setting.

Note that in practice, one can use QueryForm with OCR engines with different heuristics or other model backbones (Zaheer et al., 2020). The work focuses on how to enrich entity query abilities from forms via our proposed QueryForm.

## 4 Methodology

We propose QueryForm as a general query-based framework for solving the zero-shot DEE problem. QueryForm consists of a novel dual prompting strategy and a specially-designed pre-training approach called QueryWeb. In Figure 3, the model is first pre-trained on a large-scale webpage dataset to learn to make conditional prediction under rela-

tively noisy queries generated from combination of webpage domains (proxy of schema) and HTML tags (proxy of entity). Then, the model is fine-tuned on form-like documents with a unified schema to learn more specialized knowledge, by learning schema information in S-Prompt and further encode more accurate entity-level knowledge in E-Prompt. Finally, we test the model on the target document type in a zero-shot fashion (Figure 1).

### 4.1 Dual Prompting

Given a serialized document represented as a sequence of tokens  $\mathbf{x}$  from the set of all documents  $\mathbf{X}$ , and a set of entities of interest  $\mathbf{E} = \{e_1, \dots, e_m\}$ , the goal is to let the model predict the corresponding label sequence  $\mathbf{y}$ . In our query-based framework, we additionally define  $\mathbf{Q} = \{q_1, \dots, q_m\}$  as the set of queries, where there is a bijection between  $\mathbf{Q}$  and  $\mathbf{E}$ . The model takes an input tuple  $(q_i, \mathbf{x})$  and predicts the conditional output  $\mathbf{y}_{q_i}$  (see BOISE prediction in Figure 3 for example).  $\mathbf{y}_{q_i}$  defines the token spans of the given query  $q$  with 5 classes (*i.e.*, BOISE).

To encode entity information into the query, we can use the entity name as query, *i.e.*,  $q_i = e_i$ . We denote by  $t$  the tokenizer,  $f_\theta$  the input embedding layer and  $p_\phi$  the rest of the language model. Then we can get the token-wise BOISE prediction:

$$\hat{\mathbf{y}}_{q_i} = p_\phi(f_\theta([t(e_i); t(\mathbf{x})])), \quad (1)$$

where “[ ; ]” is the concatenation operation along the token length dimension. Note that although  $e$

itself is not learnable, we can still learn its embedding  $f_\theta(t(e))$  by optimizing  $\theta$ .

We name the query directly generated from the entity name as E-Prompt. However, in QueryForm, our novel pre-training stage requires learning from large amount of webpages, which contain diverse categories of schema. Therefore, the model naturally requires more informative queries that also encodes the schema information. To this end, we propose the S-Prompt to capture schema information. In pre-training, we can generate S-Prompt in a similar way as we do for E-Prompt, please see Section 4.2 for more details. During fine-tuning, the schema for form-like documents is often very different from that of webpages. Thus, we let the model learn the schema representation directly from the data, so that it can align well to the S-Prompts used in pre-training. We denote S-Prompt by  $s$ , learnable vectors in the token embedding space during fine-tuning to capture schema information implicitly from the data. According to the assumption in Section 3.1, the documents we used in fine-tuning includes the target entities of interest. Intuitively, the schema information from fine-tuning documents should be transferrable to the target test document type. So we directly reuse the learned S-Prompt when testing the target documents. In this case,  $\mathbf{q}_i = (s, e_i)$  and the prediction becomes:

$$\hat{\mathbf{y}}_{\mathbf{q}_i} = p_\phi([s; f_\theta([t(e_i); t(\mathbf{x})])]). \quad (2)$$

Finally, the model is trained with the objective:

$$\min_{\theta, \phi, s} \sum_{\mathbf{x}} \sum_{i=1}^m \mathcal{L}(\mathbf{y}_{\mathbf{q}_i}, \hat{\mathbf{y}}_{\mathbf{q}_i}), \quad (3)$$

where  $\mathcal{L}$  is the cross-entropy loss.

## 4.2 QueryWeb: Webpage-based Pre-training

Distinct from recent work that focuses on multi-modal pre-training, our proposed pre-training approach provides a new perspective with two core ideas: (1) Aligning pre-training and fine-tuning objectives. (2) Utilizing easy-accessible and informative webpages.

Recall that, in the fine-tuning stage, QueryForm is trained with a moderately sized set of queries composed of E-Prompts generated from human-annotated entities and a learnable S-Prompt that encodes schema information. It is reasonable to believe that if we can pre-train the model with an extremely large set of queries composed of different

```
http://www.example.com
<div class="product">
<span id="name">Bath Mat</span>
<span id="price">$13.99</span>
</div>
```

Figure 4: An example of HTML snippets, with two entities *product/name* and *product/price*.

E-Prompt and S-Prompt generated from weakly-annotated documents, the model will perform better than the Masked Language Model (MLM) (Devlin et al., 2018) pre-training alone. Here, we present our simple webpage-based pre-training technique as well as our data collection recipes to empower the pre-training.

**Dual prompting based pre-training.** We directly extract schema and entity information from rich HTML structure of various webpages and use them to generate the S-Prompt and the E-Prompt, respectively. With a slight abuse of notation, we denote S-Prompt by  $\tilde{s}$ , where  $\tilde{\cdot}$  indicates that S-Prompt is no longer a learnable parameter. Different from the fine-tuning stage with a single set of entities under a unified schema, we can group the webpages by schema:  $\{(\tilde{s}_1, \mathbf{E}_1, \mathbf{X}_1), \dots, (\tilde{s}_n, \mathbf{E}_n, \mathbf{X}_n)\}$ , where each schema  $\tilde{s}_j$  corresponds to a set of entities  $\mathbf{E}_j$  and a set of webpages  $\mathbf{X}_j$ . Similarly, the model takes the query-document tuple  $(\mathbf{q}_{ij}, \mathbf{x})$ , where  $\mathbf{q}_{ji} = (\tilde{s}_j, e_{ji})$ ,  $e_{ji} \in \mathbf{E}_j$  and  $\mathbf{x} \in \mathbf{X}_j$ , and outputs the following conditional prediction  $\hat{\mathbf{y}}_{\mathbf{q}_{ji}}$ :

$$\hat{\mathbf{y}}_{\mathbf{q}_{ji}} = p_\phi(f_\theta([t(\tilde{s}_j); t(e_j); t(\mathbf{x})])). \quad (4)$$

Equation 4 is analogous to equation 2, however,  $s_j$  here is directly sourced from webpage data, which makes it different from the learnable  $s$  in equation 2. Then we have the following pre-training objective:

$$\min_{\theta, \phi} \sum_{j=1}^n \sum_{\mathbf{x} \in \mathbf{X}_j} \sum_{i=1}^{|\mathbf{E}_j|} \mathcal{L}(\mathbf{y}_{\mathbf{q}_{ji}}, \hat{\mathbf{y}}_{\mathbf{q}_{ji}}). \quad (5)$$

The pre-training format is highly aligned with the fine-tuning in equation 3, so the model learns consistently during both stages to make query-conditional predictions.

**Data collection recipe.** How to extract schema and entity information from any webpage is another contribution of this paper. Consider the HTML snippets from Figure 4. First, it naturally contains two entities, and the combination of HTML tags defines what the entity is about, or its “entity

Task	Pre-training	Train	Test
Zero-shot Zero-shot	QueryWeb	FUNSD Inventory	XFUND Payment
Few-shot	QueryWeb + Inventory	Payment X-shot	Payment

Table 1: Experiment design of two zero-shot transfer learning and one few-shot learning tasks.

Dataset	Schema	Lang.	Entity	Example
FUNSD	1	1	4	199
XFUND	1	7	4	1393
Payment	1	1	7	10k
Inventory	1	1	7 or 28	24k
QueryWeb	87K	1	2.6M	1.2M
QueryWeb-ML	113K	>50	11.5M	13M

Table 2: Detailed statistics of used datasets.

type”. Therefore, we have "Bath Mat" is of entity *product/name*, and "\$13.99" is of entity *product/price*. Second, the schema of the webpage = {*product/name*, *product/price*}, and the schema is usually shared by a series of similar webpages under the same domain. Therefore, we can extract the domain name "www.example.com" as the schema information. Both the schema information and entity types generated from webpages are then respectively encoded by our dual prompting mechanism. In practice, the schema and entity information automatically generated from webpages are often noisy. However, in the experiments, our model is still able to learn structured information from noisy queries and obtain significantly better entity extraction performance on the target form-like documents. Moreover, in order to represent webpages in a manner that generalizes to form-like documents, the webpage representation consists only of the visible text tokens and corresponding  $x/y$  coordinates.<sup>1</sup>

## 5 Experiments

### 5.1 Datasets and Experiment Design

We use 3 publicly available datasets and 2 in-house datasets that we collected to design and conduct extensive experiments to validate our method. Table 2 summarizes the datasets.

**FUNSD** (Jaume et al., 2019) is a form understanding benchmark consisting of 199 annotated forms in English with 4-entity types: *header*, *question*, *answer*, and *other*.

**XFUND** (Xu et al., 2021) is a multilingual

<sup>1</sup>Visible text and coordinates are generated by rendering each page with Headless Chrome: <https://developer.chrome.com/blog/headless-chrome/>.

form understanding benchmark by extending the FUNSD dataset. The XFUND benchmark has 7 different languages with 1,393 fully annotated forms, where each language includes 199 forms with the same set of 4 entity types as FUNSD.

**Payment** (Majumder et al., 2020) consists of around 10K documents and 7 entity types from human annotators. The corpus is collected from different vendors with various layout templates. In the few-shot learning experiments, we create multiple subsets by randomly subsampling documents from its training set.

**Inventory** is collected by us that contains inventory-related purchase documents in English (e.g., utility bills), containing a few document types different from the Payment dataset. We prepare the dataset consists of  $\sim 24k$  documents in two annotated versions. The first version, *Inventory-7*, is annotated at word-level with the same 7 entity types from Payment. The second version, *Inventory-28*, is annotated at word-level with 21 additional entities types, including common entity types such as shipping address, supplier name.

**QueryWeb** is collected by us from publicly available webpages in English from the Internet with the acquisition procedure stated in Section 4.2.

**QueryWeb-ML** is a multilingual (ML) version of QueryWeb with more than 50 languages (at 99% percentile). We collect the dataset to validate the effectiveness for multilingual pre-training for zero-shot generalizability across different languages.

### 5.2 Experimental Details

We use the BERT-multilingual vocabulary (Devlin et al., 2018) to tokenize the serialized OCR words. We have two variants of QueryForm: a 6-layer ETC with 512 hidden size and 8 attention heads and a 12-layer ETC with 768 hidden size and 12 attention heads. For both S-Prompt and E-Prompt generated from dataset annotations, we use a maximum token length of 32 with zero padding. For learnable S-Prompt used in the fine-tuning stage, we treat its token length as a hyperparameter to search.

Our method uses the proposed QueryWeb pre-training approach on the 2 large-scale webpage-based datasets. Other comparing methods use MLM pre-training with the corresponding datasets mentioned in their papers, including  $\sim 0.7k$  unlabeled form documents for ETC+RichAtt (Ainslie et al., 2020), IIT-CDIP dataset (Lewis et al., 2006) with 7M documents for FormNet (Lee et al., 2022),

Method	Pre-training		Image	# Layers (Model Size)	FUNSD	XFUND							
	Method	Size				ZH	JA	ES	FR	IT	DE	PT	Avg.
XLM-RoBERTa	MLM	2.5TB		12L(270M)	66.70	41.44	30.23	30.55	37.10	27.67	28.86	39.36	38.24
InfoXLM	MLM	2.5TB		12L(270M)	68.52	44.08	36.03	31.02	40.21	28.8	35.87	45.02	41.19
LayoutXLM	MLM	30M	✓	12L(345M)	79.40	60.19	47.15	45.65	57.57	48.46	52.52	53.90	55.61
XLM-RoBERTa	MLM	2.5TB		24L(550M)	70.74	52.05	39.39	36.27	46.72	33.98	41.80	49.97	46.37
InfoXLM	MLM	2.5TB		24L(550M)	73.25	55.36	41.32	36.89	49.09	35.98	43.63	51.26	48.35
LayoutXLM	MLM	30M	✓	24L(625M)	<b>82.25</b>	<b>68.96</b>	51.90	49.76	61.35	55.17	59.05	60.77	61.15
QueryForm	QW	1.2M		6L(82M)	75.92	59.30	44.43	47.75	67.63	55.23	70.75	62.79	58.27
QueryForm	QW-ML	13M		12L(185M)	80.84	67.68	<b>53.30</b>	<b>55.80</b>	<b>72.91</b>	<b>67.30</b>	<b>74.25</b>	<b>69.30</b>	<b>65.79</b>

Table 3: Comparison between QueryForm and competing methods on FUNSD-XFUND zero-shot benchmark.

Method	Pre-training		Model Size	Supervised Upper-bound	Source → Target	
	Method	Size			I7 → Payment	I28 → Payment
ETC+RichAtt	MLM	0.7M	83M	94.33	81.54	78.37
FormNet	MLM	7M	82M	<b>95.70</b>	79.77	77.88
FormNet	MLM	7M	157M	95.61	86.04	82.42
QueryForm	QW	1.2M	82M	94.60	<b>88.15</b>	<b>89.23</b>

Table 4: Comparison between QueryForm and previous state-of-the-arts on Inventory-Payment zero-shot benchmark. I-7 and I-28 are abbreviations of Inventory-7 and Inventory-28, respectively. QueryForm has much better generalization ability indicated by its stronger zero-shot performance and smaller gap with its supervised upper-bound (trained and tested both on Payment).

Pre-training	E-P	S-P	Source → Target	
			I7 → Payment	I28 → Payment
MLM	✓		81.95	85.71
QueryWeb	✓		85.33	87.41
MLM	✓	✓	84.91	86.41
QueryWeb	✓	✓	<b>88.15</b>	<b>89.23</b>

Table 5: Ablation study of QueryForm on Inventory-Payment benchmark. E-P and S-P are abbreviations of E-Prompt and S-Prompt, respectively.

30M multilingual documents for LayoutXLM (Xu et al., 2021), and 2.5TB multilingual Common-Crawl data for XLM-RoBERTa (Conneau et al., 2019) and InfoXLM (Chi et al., 2020).

We use micro-F1 to evaluate the performance on XFUND related experiments, following Xu et al. (2021); macro-F1 to evaluate Payment related experiments, following Majumder et al. (2020); Lee et al. (2022). We report the mean of best experiment results across three runs with difference seeds. Please see Appendix B for more experimental details.

### 5.3 Zero-shot Transfer Learning Results

To evaluate QueryForm, we introduce two zero-shot transfer learning tasks and one few-shot learning task, as shown in Table 1. We follow the official train-test split for all public available datasets by default, unless specified explicitly. For zero-shot on Payment test set, we pre-train on QueryWeb and fine-tune on Inventory.

**FUNSD-XFUND.** In Table 3, we compare QueryForm with recent zero-shot transfer learning meth-

Method	Payment 1-shot	Payment 10-shot
ETC+RichAtt <sup>2</sup>	59.62	88.21
FormNet <sup>2</sup> (157M)	55.67	86.25
QueryForm	<b>89.26</b>	<b>90.53</b>

Table 6: Comparison between QueryForm and comparing methods further fine-tuned on few-shot Payment training sets.

ods, include XLM-RoBERTa (Conneau et al., 2019), InfoXLM (Chi et al., 2020), and the current state-of-the-art LayoutXLM (Xu et al., 2021), which are all MLM pre-trained on multilingual text or document datasets of different sizes (details in Section 5.2). QueryForm outperforms all comparing methods even with much smaller model size and no image modality. In particular, when we pre-train QueryForm on the multilingual QueryWeb-ML, QueryForm obtains a significant boost on all languages. Although XLM-RoBERTa and InfoXLM are MLM pre-trained on 2.5TB multilingual data, and LayoutXLM especially collected 30M visually rich documents for MLM pre-training, the stronger transferability of QueryForm for zero-shot DEE indicates that our pre-training method is more effective than MLM for this specific task. Since the 12 layer (185M) QueryForm has outperformed previous state-of-the-arts by a large margin with a much smaller model size, we leave further scaling up as future work.

**Inventory-Payment.** Table 4 shows the zero-shot transfer learning result on the Inventory-Payment benchmark. We compare QueryForm against the

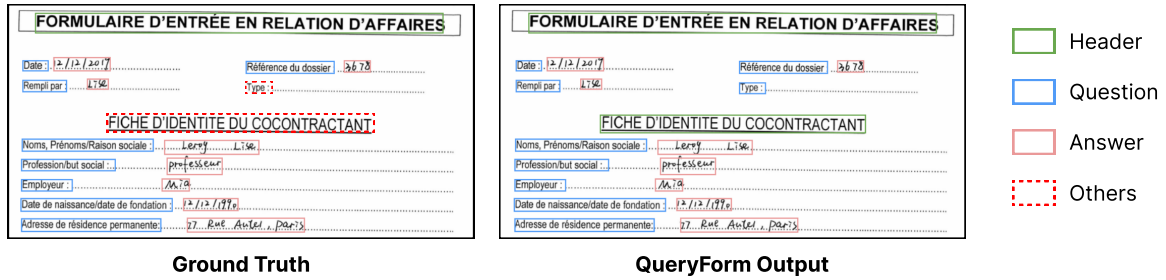


Figure 5: Visualization example from XFUND (French). QueryForm labels entities with ambiguous “others” annotation from ground truth as one of the other three entity types with concrete meanings.

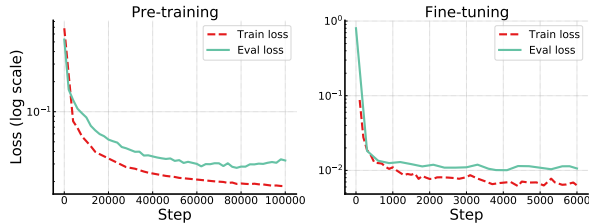


Figure 6: Loss visualization of pre-training on QueryWeb (Left) and fine-tuning on Inventory (Right).

current state-of-the-art on Payment, FormNet (Lee et al., 2022), and our baseline ETC+RichAtt (see Section 3.2). QueryForm outperforms competing methods by a significant margin. Although FormNet obtains the best supervised upper-bound result on Payment, the lower zero-shot results indicate that knowledge transfer from different types of documents is still very challenging.

QueryForm is expected to take advantage of larger number of queries though they are less relevant. To validate, we compare fine-tuning datasets with 7 and 28 annotated entities. As can be seen, supervised methods like FormNet and ETC+RichAtt suffer performance drop when seeing additional entities not existed in target dataset, while QueryForm gains further performance improvement.

**Ablation study.** We conduct ablation study of QueryForm. From the results in Table 5, we can see that both our dual prompting strategy and QueryWeb pre-training contribute to the zero-shot F1 score individually, and synergistically improve the performance when working together.

#### 5.4 Few-shot Learning on Payment

In practice, it is reasonable to believe that a few annotated document from the target document type can make models quickly adapt. Therefore, we design a few-shot learning step based on the best performing model obtained on the Inventory dataset.

Table 6 shows the 1- and 10-shot results on Payment. To make sure the training is stable on low data regime and comparison is fair, we conduct

hyperparameter search (e.g., learning rates, # of freezing layers) for all methods and select the best performing ones to present. When fine-tuning on the extreme Payment 1-shot, both FormNet and ETC-RichAtt overfit the single document from Payment severely, while QueryForm maintains high performance<sup>2</sup>. When extending to 10-shot, all methods improves, and QueryForm still perform the best. Although FormNet is state-of-the-art in supervised learning setting, it underperforms other methods on low data regime. We hypothesize GCN requires more data to learn layout features.

#### 5.5 Result analysis

**Prediction visualization.** Figure 5 demonstrates an example output of QueryForm. QueryForm infers entities that are annotated as “others” in ground truth as one of the other three entity types with concrete meanings. For example, “Type” is a question in the form, however, without corresponding answer. Although human annotators might find it ambiguous and mark it as “others”, QueryForm successfully recognize it as a “question”.

**Loss visualization.** Figure 6 shows the loss curves of pre-training on QueryWeb (Left) and fine-tuning on Inventory (Right). According to the pre-training loss curve, we observe that the loss converges well despite the fact that the weak supervision extracted from webpages is often noisy. Moreover, according to the fine-tuning curve, we observe that the loss converges very fast, thanks to the knowledge learned during pre-training. The observations indicate that our framework successfully extracts useful information from the weak supervision and leverages the learned information to facilitate fine-tuning on form-like documents.

<sup>2</sup>For compared methods trained by us, although 10-shot leads to improvement, 1-shot degrades with the same parameter-search methods used for QueryForm. More advanced training strategies are not considered here.



## 6 Conclusion

This paper presents QueryForm, a novel framework to address the challenging zero-shot document entity extraction problem. The dual prompting design in QueryForm offers a refreshing view to unify the pre-training and fine-tuning objectives, allowing us to leverage large-scale form-like webpages with HTML tags as weak annotations. QueryForm sets new state-of-the-art results on multiple zero-shot DEE benchmarks. We believe QueryForm serves as a flexible framework for document understanding tasks, and multiple interesting directions could be further explored within the framework, such as prompt design, richer pre-training sources, etc.

### Ethical and Broader Impact

We have read the ACL Code of Ethics, and ensure that our work is conformant to this code. As a novel framework for zero-shot document entity extraction, QueryForm has a great potential to boost the performance of existing DEE systems. However, we would still like to discuss the limitations and risks to avoid any misuse of QueryForm.

Although our proposed QueryWeb pre-training approach can effectively achieve knowledge transfer from publicly available webpages to form-like documents, it inevitably carries the bias and fairness problems (Mehrabani et al., 2021) to the downstream task. Therefore, in real-world applications, we should have more strict rules to filter and clean up the webpages, and thoroughly check the bias and fairness issues of the pre-trained model.

### Limitations

In addition to the bias and fairness concerns that we discussed in the Ethical and Broader Impact section, we discuss the possible limitations of our method in this section.

As a query-based DEE framework, QueryForm may be prone to specific prompting based adversarial attacks (Xu et al., 2022), which may further pose potential security concerns for safety-critical documents. Thus, it is important to test the robustness of QueryForm against adversarial attacks and design defense schemes to further strengthen our method in the future.

Our work focuses on the closed-world setting that source documents include entities contained in the target documents, following (Xu et al., 2021), without further investigating the possible open-world (Shu et al., 2018) setting with unseen test

entities. However, as a query-based framework that makes conditional prediction with no pre-defined set of entities, QueryForm actually supports the prediction of unseen entities at test time and we would like to leave it as an interesting future research direction.

### Acknowledgements

We greatly thank Chun-Liang Li, Harr Chen, Evan Huang, Nan Hua for their valuable feedback.

### References

- Milan Aggarwal, Hires Gupta, Mausoom Sarkar, and Balaji Krishnamurthy. 2021. Form2seq: A framework for higher-order form structure extraction. *arXiv preprint arXiv:2107.04419*.
- Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2021. Htlm: Hyper-text pre-training and prompting of language models. *arXiv preprint arXiv:2107.06955*.
- Joshua Ainslie, Santiago Ontanon, Chris Alberti, Václav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. Etc: Encoding long and structured inputs in transformers. *arXiv preprint arXiv:2004.08483*.
- Srikanth Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *ICCV*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*.
- Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. InfoLM: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Timo I Denk and Christian Reisswig. 2019. Bert-grid: Contextualized embedding for 2d document representation and understanding. *arXiv preprint arXiv:1909.04948*.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Mingfei Gao, Le Xue, Chetan Ramaiah, Chen Xing, Ran Xu, and Caiming Xiong. 2022. Docquerynet: Value retrieval with arbitrary queries for form-like documents. In *COLING*.
- Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. 2021. Lambert: layout-aware language modeling for information extraction. In *ICDAR*.
- Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *ICDARW*.
- Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. *arXiv preprint arXiv:1809.08799*.
- Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *ECCV*.
- Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Frank Lebourgeois, Zbigniew Bublinski, and Hubert Emptoz. 1992. A fast and efficient method for extracting text paragraphs and graphics from unconstrained documents. In *ICPR*.
- Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022. Formnet: Structural encoding beyond sequential modeling in form document information extraction. In *ACL*.
- Chen-Yu Lee, Chun-Liang Li, Chu Wang, Renshen Wang, Yasuhisa Fujii, Siyang Qin, Ashok Popat, and Tomas Pfister. 2021. Rope: reading order equivariant positional encoding for graph-based document information extraction. In *ACL*.
- David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. Building a test collection for complex document information processing. In *SIGIR*.
- 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. *arXiv preprint arXiv:1907.11692*.
- Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. 2020. Representation learning for information extraction from form-like documents. In *ACL*.
- Simone Marinai, Marco Gori, and Giovanni Soda. 2005. Artificial neural networks for document analysis and recognition. *TPAMI*.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.
- Lawrence O’Gorman. 1993. The document spectrum for page layout analysis. *TPAMI*.
- Rasmus Berg Palm, Ole Winther, and Florian Laws. 2017. Cloudscan-a configuration-free invoice analysis system using recurrent neural networks. In *ICDAR*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.
- Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*.
- Daniel Schuster, Klemens Muthmann, Daniel Esser, Alexander Schill, Michael Berger, Christoph Weidling, Kamil Aliyev, and Andreas Hofmeier. 2013. Intellix—end-user trained information extraction for document archiving. In *ICDAR*.
- Lei Shu, Hu Xu, and Bing Liu. 2018. Unseen class discovery in open-world classification. *arXiv preprint arXiv:1801.05609*.
- Anikó Simon, J-C Pret, and A Peter Johnson. 1997. A fast algorithm for bottom-up document layout analysis. *TPAMI*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS*.
- Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. 2022a. Dual-prompt: Complementary prompting for rehearsal-free continual learning. *ECCV*.
- Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022b. Learning to prompt for continual learning. In *CVPR*.
- Hao Wei, Micheal Baechler, Fouad Slimane, and Rolf Ingold. 2013. Evaluation of svm, mlp and gmm classifiers for layout analysis of historical documents. In *ICDAR*.

- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Lei Xu, Yangyi Chen, Ganqu Cui, Hongcheng Gao, and Zhiyuan Liu. 2022. Exploring the universal vulnerability of prompt-based learning paradigm. *arXiv preprint arXiv:2204.05239*.
- Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.
- Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020b. Layoutlm: Pre-training of text and layout for document image understanding. In *SIGKDD*.
- Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*.
- Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *NeurIPS*.
- Xiaohui Zhao, Endi Niu, Zhuo Wu, and Xiaoguang Wang. 2019. Cutie: Learning to understand documents with convolutional universal text information extractor. *arXiv preprint arXiv:1903.12363*.

## A Additional Dataset Information

**Licensing information.** We provide the licensing information of publicly available dataset as follows:

- **FUNSD** is licensed under the license specified in <https://guillaumejaume.github.io/FUNSD/work/>, where the use of the FUNSD Dataset is solely for non-commercial, research and educational purposes.
- **XFUND** is licensed under the Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license.
- **Payment** has no licensing information, however, we use it accordingly following [Majumder et al. \(2020\)](#); [Lee et al. \(2022\)](#).

The use of existing artifacts is consistent with their intended use for research and educational purposes. **Collection of in-house datasets.** For datasets that we collect, both Inventory and QueryWeb are sufficiently anonymized to remove personal information. Moreover, our internal system filters offensive content when collecting public available webpages. The possible limitation of our current data collection methodology for QueryWeb is that we do not enforce strict check on the bias and fairness issues, as discussed in Section 6. Our method mainly leverages layout and HTML-related information from webpages instead of the semantic content. No images are used. Additionally, it is useful to further improve our data collection method in the future to avoid potential bias and fairness issues.

## B Additional Experiment Information

For QueryWeb pre-training, we use Adam optimizer with a batch size of 512. We set the learning rate to 0.0002 with a warm-up proportion of 0.01 and a linear learning rate decay of 100k steps. We fine-tune all models using the Adam optimizer with a batch size of 128 and a learning rate of 0.0004. No warm-up or learning rate decay is used. Pre-training takes approximately 48 hours on 8x8 TPU v3. Fine-tuning takes approximately 4 hours on the largest corpus, Inventory, on both TPUs and Tesla V100 GPUs. We implement our method based on the same codebase used in [Lee et al. \(2022\)](#).

## ACL 2023 Responsible NLP Checklist

---

### A For every submission:

- A1. Did you describe the limitations of your work?  
*Left blank.*
- A2. Did you discuss any potential risks of your work?  
*Left blank.*
- A3. Do the abstract and introduction summarize the paper’s main claims?  
*Left blank.*
- A4. Have you used AI writing assistants when working on this paper?  
*Left blank.*

### B Did you use or create scientific artifacts?

*Left blank.*

- B1. Did you cite the creators of artifacts you used?  
*Left blank.*
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?  
*Left blank.*
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?  
*Left blank.*
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?  
*Left blank.*
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?  
*Left blank.*
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.  
*Left blank.*

### C Did you run computational experiments?

*Left blank.*

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?  
*Left blank.*

---

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

*Left blank.*

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

*Left blank.*

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

*Left blank.*

**D  Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

*Left blank.*

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

*Left blank.*

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

*Left blank.*

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

*Left blank.*

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

*Left blank.*