X-SHOT : A SINGLE SYSTEM TO HANDLE FREQUENT, FEW-SHOT AND ZERO-SHOT LABELS IN CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, few-shot and zero-shot learning, which focus on labels with limited annotated instances, have garnered significant attention. Traditional approaches often treat freq-shot (labels with numerous instances), few-shot, and zero-shot learning as distinct challenges, optimizing systems for just one of these scenarios. Yet, in real-world settings, label occurrences vary greatly. Some labels might appear thousands of times, while others might only appear sporadically or not at all. Ideally, a system should accommodate any label, regardless of its training frequency. Notably, while few-shot systems often falter on zero-shot tasks, zero-shot systems don't leverage available annotations when certain downstream labels possess them. For practical deployment, it's crucial that a system can adapt to any label occurrence. We introduce a novel classification challenge: X-Shot , reflecting a real-world context where freq-shot, few-shot, and zero-shot labels emerge without predefined limits. Here, X can span from 0 to $+\infty$. The crux of X-Shot centers on open-domain generalization and devising a system versatile enough to manage various label scenarios. Our solution leverages Instruction Learning, bolstered by data autonomously generated by pre-trained Language Models (PLMs). Our unified system, X-Shot, surpasses preceding state-of-the-art techniques on three benchmark datasets across diverse domains in both single-label and multi-label classifications. This is the first work addressing X-Shot learning, where X remains variable.[1]

1 INTRODUCTION

Over recent years, few-shot and zero-shot learning techniques have seen significant advancements, aiming to address the challenge of training models with scant or even no annotated instances for specific labels (Bragg et al., 2021; Xia et al., 2020). Historically, the fields of frequent-shot, few-shot, and zero-shot learning have been approached as distinct paradigms, with systems optimized uniquely for each setting. Yet, in real-world scenarios, label frequencies can exhibit broad variation, with certain labels occurring prolifically, and others being scarce or completely absent. Given this variability, it becomes imperative to craft learning systems adept at managing labels across the full frequency spectrum. Regrettably, current few-shot systems often fall short when confronted with zero-shot challenges(Zhang et al., 2022; Cui et al., 2022; Zhao et al., 2021). In contrast, zero-shot systems, while adept in their domain, typically overlook the potential benefits of available annotations(Zhang et al., 2019; Obamuyide & Vlachos, 2018; Yin et al., 2019). Thus, mastering the ability to handle all conceivable label occurrences is paramount for systems aiming for practical deployment.

In this paper, we introduce an innovative and inherently more challenging task, termed X-Shot . This task mirrors real-world environments where label frequencies span a continuum, seamlessly incorporating frequent-shot, few-shot, and zero-shot instances, all without a priori constraints. In this paradigm, the variable X is unbounded, ranging freely within the interval $[0, +\infty)$. At the heart of X-Shot lies the objective of attaining open-domain generalization and architecting a system resilient across a plethora of label scenarios.

[1]Data & code will be released upon acceptance.

Tackling $X-\texttt{Shot}$ spawns two core technical conundrums: ($\mathcal{Q}_1$) Amidst the paucity of annotations characteristic of few-shot and zero-shot contexts, how might one identify apt sources of indirect supervision (Yin et al., 2023) to navigate the $X-\texttt{Shot}$ setting? ($\mathcal{Q}_2$) Traditional multi-class classifiers grapple with the heterogeneity of label sizes across tasks, often mandating distinct classification heads tailored to these variations. Here, the challenge is formulating a cohesive system capable of effectively managing labels of diverse sizes.

To address $\mathcal{Q}_1$, we tap into the availability of indirect supervision from instruction tuning datasets, such as Super-Naturalinstructions (Wang et al., 2022). These datasets primarily contain various NLP tasks enriched with textual instructions. Our method involves pretraining our model on these datasets, aiming for robust generalization to the unseen $X-\texttt{Shot}$ task when supplemented with pertinent instructions. For ($\mathcal{Q}_2$), we advocate a triplet-oriented binary classifier. This classifier functions by accepting a triplet of (`instruction`, `input`, `label`), anticipating a binary response (Yes" or No") that confirms the suitability of the `label` for the specified `input` under the given `instruction`. Such a triplet-oriented classifier acts as a cohesive architecture, adept at managing text classification tasks with labels of varied dimensions. By amalgamating solutions for both $\mathcal{Q}_1$ and $\mathcal{Q}_2$, we forge a holistic framework, $X-\texttt{Shot}$. This framework capitalizes on indirect supervision sourced from a diverse set of tasks, incorporating instructions as guidance, and thus presents a unified architecture proficient in handling text classification challenges with both open-shot and open-size labels.

No existing datasets explicitly cater to this challenge. To evaluate our system, we turn to three representative classification tasks: relation classification, ultra-fine entity typing, and situation detection. We reconfigure their associated datasets: *FewRel* (Han et al., 2018), *UFET* (Choi et al., 2018), and *Situation* (Yin et al., 2019) to simultaneously encapsulate frequent-shot, few-shot, and zero-shot instances. Sourced from diverse domains (Wikipedia, crowdsourcing, and more), and featuring vast label counts (ranging from 12 to the thousands), these datasets pose a formidable challenge to contemporary text classification systems. Moreover, both *UFET* and *Situation* function as multi-label classification datasets. The *Situation* dataset uniquely integrates an " None" label, further amplifying the realistic nature of the task. Empirical results reveal our system's resilience across datasets and instruction templates, consistently outclassing leading methods, including GPT, in frequent-shot, few-shot, and zero-shot contexts.

Our contributions can be summarized as follows: (i) We introduce $X-\texttt{Shot}$, a hitherto underexplored, open-domain open-shot text classification problem that mirrors real-world complexities. (ii) We innovate a unique problem setting that reframes any text classification challenge into a binary classification task, adaptable to any number of labels and occurrences. (iii) Our $X-\texttt{Shot}$, harnessing the potential of Instruction Tuning datasets, excels past existing approaches, demonstrating versatility across various domains, label magnitudes, and classification paradigms.

2 Related Work

Few-shot Learning. Few-shot learning refers to machine learning methods that can perform tasks with only a few labeled training examples. This technique has gained traction in NLP for two reasons: (i) labeled data can be expensive to obtain and (ii) extensive training or fine-tuning, particularly with large models, can be both costly and unstable. Ideally, a model would generalize from a handful of examples, capturing the core knowledge. The main challenge lies in effectively using limited labeled samples for broad generalizations. Initially, the approach to few-shot learning was metric-based, focusing on a shared feature space and distance metrics for label predictions (Vinyals et al., 2016; Snell et al., 2017; Sung et al., 2018). Recently, Large Language Models (LLMs) have been recognized as efficient few-shot learners. Fine-tuning these pre-trained LLMs with minimal samples often produces notable results (Brown et al., 2020). Additionally, due to the success of prompting in GPT models, prompt-tuning has been applied to tackle classification problems under few-shot settings (Zhang et al., 2022; Cui et al., 2022; Zhao et al., 2021). However, these methods don't typically manage zero-shot scenarios where certain labels are without annotated data.

Zero-shot Learning. Building on the concept of few-shot learning, we transition to the even more challenging zero-shot learning where no labeled examples are available. Early techniques in this domain employed metrics to align texts and labels in shared spaces. More recent works adopted

word embeddings from pre-trained language models to represent the meaning of the text or the label. The latest work enhanced the embedding representations by integrating class descriptions, class hierarchy, and the word-to-label paths found within ConceptNet(Zhang et al., 2019). Today's LLMs are so adept that they can tackle NLP tasks without any labeled instances, either by reformatting the classification tasks or through in-context learning as seen with the GPT models (Brown et al., 2020; Wei et al., 2022). Similarly, an alternative approach is to calibrate and score outputs from LLM models for the label assignment(Zhao et al., 2021; Holtzman et al., 2021; Min et al., 2022). The most recent trend in zero-shot text classification is to draw on the power of indirect supervision from other well-annotated NLP tasks, like text entailment (Obamuyide & Vlachos, 2018; Yin et al., 2019). Still, these methods don't fully utilize annotations when they exist for labels.

Indirect Supervision There's a burgeoning interest in indirect supervision. Here, easily available signals from relevant tasks are used to aid in learning the target task, especially when task-specific supervision is in short supply. The technique of using entailment for indirect supervision in zero-shot classification was pioneered by (Yin et al., 2019) and has since been adapted for a variety of NLP tasks, including few-shot intent identification (Zhang et al., 2020), event argument extraction (Sainz et al., 2022), and relation extraction (Xia et al., 2021). Beyond entailment, knowledge from areas like question answering (Yin et al., 2021) and summarization (Lu et al., 2022) has been incorporated. Recent studies have demonstrated that modern language models, after fine-tuning on a plethora of instruction-based tasks, can generalize to multiple unseen tasks (Wang et al., 2022; Mishra et al., 2022; Ye et al., 2021). Our work is inspired by the observed efficacy of NLP models when given task instructions and their ability to generalize knowledge across tasks.

3 PROBLEM STATEMENT

X−Shot has the following requirements:

- **Input** t: Versatile text in form, length, and domain.
- **Label space** L: L contains arbitrary size of labels: $\{\cdots, l_i, \cdots\}$ and an optional *None* label (i.e., all labels in L are incorrect for the input). Within L, some labels are zero-shot, some are few-shot, and some are frequent.

Then, the task of X−Shot is to figure out a subset of $L_s \in L$ that are correct for the input t, where $|L_s|$ can be zero (i.e., "*None*"), 1 or >1.

Research questions of X−Shot : i) Given that the above formulation encompasses various text classification problems, how can we move away from constructing individual models for each problem, and instead develop a singular classifier adept at handling diverse classification challenges? ii) Beyond frequently-encountered labels, low-shot labels necessitate additional supervision for effective reasoning. Where can we source this supervision? In the following section, we delve deeper into our approach concerning the universal system and the process of seeking supervision.

4 METHODOLOGY

This section outlines our approach to the X−Shot problem. We first explain our process of transforming all classification problems into a unified binary classification framework. Next, we discuss the type of supervision we gather to address this problem with limited annotations.

4.1 SYSTEM ARCHITECTURE FOR X−SHOT

We've devised a broad architecture that seamlessly transitions most classification challenges into a unified, instruction-driven binary classification task. As depicted in Figure 1, for any text classification task with its set of inputs and labels, we model it as (instruction, input, label) triplet. The task then becomes determining if the label is appropriate ("Yes") or not ("No") for the input given an instruction. This new framework is referred to as X−Shot.

X−Shot can capably manage both multi-class and multi-label classification challenges. Instead of converting labels into numerical IDs as traditional supervised classifiers do, we retain the actual

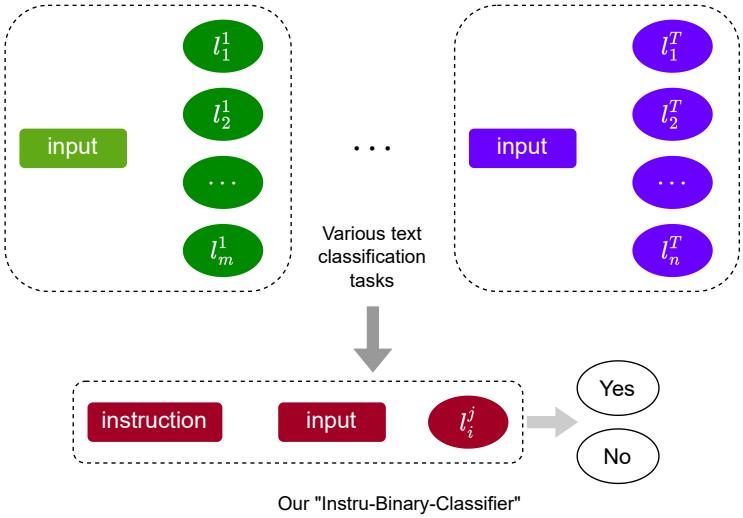

Figure 1: Our $X-$`Shot` unifies various text classification tasks as an instruction tuning problem.

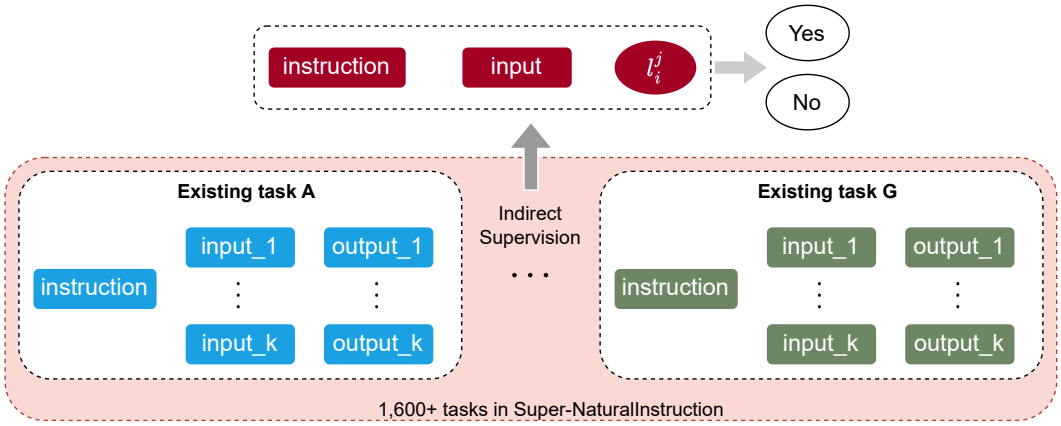

Figure 2: Indirect supervision for $X-$`Shot`.

label names. Optionally, we can also employ sophisticated verbalizers (Schick & Schütze, 2021) to enhance the expression of the label. This ensures a more intuitive understanding of the relationship between inputs and labels, all within the context of task instructions.

$X-$`Shot` paves the way to tackle a variety of low-shot text classification tasks using an instruction-guided approach. Two primary challenges arise: i) Ensuring the model comprehends the instructions, and ii) Guiding the model to identify seldom seen or entirely new labels. We'll delve deeper into our supervision-seeking approaches to address these challenges in the following subsection

4.2 Supervision Acquisition for Low-Shot Labels

In this section, we will introduce how we conduct and combine *Indirect Supervision* and *GPT supervision* to solve $X-$`Shot` .

Indirect Supervision. Previous best-performing systems for low-shot text classification have primarily relied on indirect supervision *from a single source task*. Examples of these source tasks include natural language inference (Yin et al., 2019) and summarization (Lu et al., 2022). This approach presents three main drawbacks: i) the usable supervision from the single source task is finite, and there's often a domain mismatch between the source task and the target classification tasks; ii) typically, instances of the target problems need to be reformatted to align with the specific source tasks to enable zero-shot generalization—a process that's frequently complex; iii) there isn't

a universally adaptable system to address the $X-\texttt{Shot}$ situation, where labels might vary in their visibility or frequency.

In this work, we leverage indirect supervision from an extensive assortment of NLP tasks. The Super-NaturalInstruction dataset (Wang et al., 2022) encompasses over 1,600 tasks across 76 categories. Each of these tasks is accompanied by instructions and numerous input-output examples. As depicted in Figure 2, this dataset offers an invaluable source of indirect supervision for our target $X-\texttt{Shot}$. For every task within the Super-NaturalInstruction dataset, we're presented with the associated instruction as well as (input, gold output) pairs. For each instance selected, we will randomly pick one output from the task label space that is different from the gold output, whether the task is generation or classification. As a result, we obtain one positive triplet (task instruction, input, gold output) and one negative triplet (task instruction, input, random output) for each example in our training dataset. Our indirect supervision stems from this dataset training. When evaluated on benchmark classification tasks, we convert every sample into triplets similarly, complemented by a human-written instruction. For an instance with text t and L_s positive labels, we add an instruction and craft $|L|$ triplets (task instruction, t, l) for each label l from the label space L, with L_s of them are positive and the remaining are negative.

Through this indirect supervision, minor alterations—be it a word or a few words—can pivot the class completely. By enabling the model to distinguish the positive and negative classes from marginally tweaked inputs, we ensure the model establishes more distinct decision boundaries.

GPT Supervision for zero-shot labels. In addition to Instruction Supervision, we aim to enhance our model's performance on zero-shot labels. Given that we cannot procure annotated instances for these labels, how can we enhance the model's understanding of these labels without human intervention or labeling? This is where we leverage the capabilities of GPT (Brown et al., 2020) to produce weakly labeled examples. For generating instances related to zero-shot labels, we utilize in-context learning. This involves a random selection of demonstrations from either few-shot or frequently labeled data. Below is a sample prompt designed to generate entity typing text for a zero-shot type label:

```
entity type: paper
entity: New York Times
sentence:  I enjoy reading articles in The New York Times to stay
updated on current events and global news

entity type: gathering
entity: concert
sentence: The concert was captivating, with the musicians' stellar
performance earning an encore request from the audience.

entity type: star
```

In this approach, upon exposing GPT to entity and entity statement examples associated with the entity type labels "paper" and "gathering", we introduce the zero-shot label "star". Subsequently, GPT generates an entity along with an entity statement, serving as a weakly supervised instance for this previously unseen label.

Training strategy. We first train the RoBERTa (Liu et al., 2019) model on the transformed binary Super-NaturalInstruction dataset, then fine-tune on the augmented instances of downstream $X-\texttt{Shot}$ tasks.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTING

Datasets. Our objective is to choose datasets that can cover (i) multiple domains, (ii) various sizes of labels, and (iii) both single-label and multi-label scenarios. Therefore, we evaluate on three mainstream classification datasets: *FewRel* (Han et al., 2018), *UFET* (Choi et al., 2018), and

Situation (Yin et al., 2019), referring to relation exaction, entity typing, and situation identification problems respectively. All of them are considered the benchmark dataset in the text classification field. The number of labels in these datasets varies from 12 to 230 labels, making the classification task very challenging. In addition, an extra "*None*" label in one of the datasets makes the problem setting more realistic.

While the original datasets provide a foundation, they don't align with our needs because: i) some maintain consistent instance counts across all labels, whereas others display varied label coverage distributions; and ii) they aren't tailored for binary classification. To better accommodate the X-Shot scenario, we modify each dataset. This results in three distinct label groups: *freq-shot* labels, *few-shot* labels, and *zero-shot* labels. We'll delve into the specifics of this augmentation in the subsequent section.

• **FewRel (Han et al., 2018)** is a well-established relation classification dataset containing relation statements extracted by aligning terms from Wikipedia to the knowledge base facts in Wikidata. *FewRel* uses 64/16/20 relations for training, dev, and test sets, while each relation has 700 instances. Each instance in *FewRel* provides a relation statement, two entities from the sentence, and their corresponding relation label. Even though *FewRel* includes a large number of labels, the original experimental setting is to evaluate few-shot learning for a limited number of relations (Soares et al., 2019; Dong et al., 2020; Wang et al., 2020). Previous approaches usually perform an N-way K-shot learning, while N is usually 5 or 10 while K is usually 1 or 5.

To align with the objectives of X-Shot , our evaluation framework adopts a comprehensive setup, ensuring the inclusion of a diverse set of labels. Since the test set is not available for *FewRel*, we include 78 relations and divide them into 26/26/26 as freq/few/zero-shot labels. We put 500/5/0 instances for each freq/few/zero label in the training set, and 200 instances for each label in the dev/test set.

• **UFET (Choi et al., 2018)** is a human-labeled entity typing dataset with more than 5000 instances and 2519 unique labels. Each instance in *UFET* consists of an entity statement, the target entity, and the list of possible types of the entity. In contrast to *FewRel*, *UFET* is a free-form multi-label dataset while one instance can be labeled with several types of roles based on the context. UEFT has been studied as a multi-label classification problem in previous studies (Choi et al., 2018; Zhang et al., 2021).

For our approach, we adopted the most frequent 230 entity types and split them into 30/100/100 as the freq/few/zero-shot labels since the remaining entity type labels occur less than 20 times in the dataset. These 230 entities cover around 90% of the dataset. Since *UFET* is a multi-label dataset, it will be difficult to assign a specific number of instances per label. Therefore, we put all instances without zero-shot labels (around 70%) as the training set and the remaining as the dev/test set.

• **Situation (Yin et al., 2019)** is an event-typing classification dataset including 5,956 labeled instances. There are two kinds of situations here: i) 8 "need" situations where a specific kind of aid is needed, such as food or water supply. ii) 3 "issue" situations where an issue, such as a crime, is happening. Similar to *UFET*, this dataset is also a multi-label dataset. However, one thing that makes it different from the other datasets is that there is one special situation, "*None*", which means that none of the 11 situations fit. This dataset was used as a benchmark dataset for zero-shot classification in the previous study. The methodology is to convert it into a Natural Language Inference (NLI) binary classification problem, which is adopted as one of our baselines. If none of the labels are positive (receiving a probability higher than a threshold), then the "*None*" label is assigned.

To create a dataset with varying label occurrences, we separate the 11+1 situations as 4/4/4 freq/few/zero-shot labels, while the "*None*" label belongs to the zero-shot group. Similar to *UFET*, we treat instances without zero-shot labels (around 60%) as training instances and the remaining as the dev/test set.

For *UFET* and *Situation* datasets, even though we cannot assign a specific number of instances for each label in the training set due to the multi-label setting, we always limit the number of occurrences of few-shot labels to around 5 times in the training set in order to be consistent with *FewRel*. More dataset details are in Table 1.

Table 1: Dataset statistics

	domain	#freq	#few	#zero
FewRel	Wikipedia	26	26	26
UFET	crowdsourcing	30	100	100
Situation	/	4	4	3+1

Baselines. For baselines, we compare our system with the current state-of-the-art multi-way classification model, the in-context learning with GPT, and the most advanced few-shot/zero-shot learning methods in the literature.

• **Multi-way classification (MWC, (Soares et al., 2019))** . In this baseline, we treat it as a traditional multi-way classification problem with a special Marker scheme called "Entity Marker"(Soares et al., 2019). Entity Marker introduces extra entity token markers to the model besides the entity terms and feeds the concatenation of start entity tokens into the classification head. For each statement containing entities, we put $< Ei >$ and $< /Ei >$ as the start and end entity tokens for each entity i. One example is as follows:

<E1> LONDON <\E1> is the capital of <E2> UK <\E2>

This methodology stands as the leading approach for extracting relation representations, especially in the realm of entity relation classification (Soares et al., 2019). We employ this strategy for both the *FewRel* and *UFET* datasets, given that they contain entities within their inputs. However, for the *Situation* dataset, given the absence of predefined entity spans in the situational statements, we continue to use the [cls] token as input for the classification head without integrating any Entity Markers.

• **Indirect Supervision from NLI (NLI, (Li et al., 2022))**. The previously established best approach for addressing a zero-shot classification challenge was to reframe it as an NLI task. This technique eliminates the need for specific annotations related to the label space or any label-specific data. A sequence classification task can be adapted into a text entailment problem by using the original statement as the premise and transforming the label into a hypothesis. Our method distinguishes itself from this NLI-centric technique in two significant ways: first, we broaden the range of indirect supervision sources from just NLI to encompass a diverse set of NLP tasks; second, we implement an instruction tuning schema rather than adopting a pairwise classification framework.

• **In-context learning with GPT (GPT-3.5).** For in-context learning, we create a prompt that includes three demonstrations, two positive and one negative, and each comes with the sentence, optional entity (entities), the relation/entity type/ situation term, and the label that indicates whether the term is correct. Then, we provide the same features for the instance we want to predict but let the GPT complete the label part. A template can be seen in Appendix A.1.

• **Prototypical Prompt learning (PPL, (Cui et al., 2022))** The most popular approach for addressing classification challenges within the few-shot framework is through the practice of prompt learning in recent years. It combines the strength of LLMs and a well-designed verbalizer that maps the model output to the pre-defined labels. This baseline utilized the prototypical verbalizer (ProtoVerb) that is built directly from training data N-shot setting converts classification into a sequence mask problem that, in each training iteration, the model puts N sentences from each label into a prompt and has the label token been replaced by the [mask] vector. For *FewRel*, *UFET*, and *Situation*, we select 500, 100, and 500 instances during training for prototype learning. Since we want to be consistent with the freq, few, and zero-shot learning approach, for freq and few shot labels, we keep selecting instances from the limited instances until we reach the number. For zero-shot labels, we simply put the label itself as the text for the training and test on the original test set.

Implementation details We elaborate on our implementation details at different stages here.

• **Indirect Supervision**. Consistent with the original experimental setup, we select 100 random instances from each task for training when compiling the indirect supervision dataset from Super-NaturalInstruction. Our prefix template follows the previous benchmark strategy, incorporating only the instruction and two positive examples—provided this inclusion doesn't surpass the word limit. When adjusting classification tasks to fit X-Shot, we draft three distinct instruction prompts and present the average outcomes to demonstrate the system's stability. Further details about each template are available in Appendix A.2.

• **GPT-3.5 for D_{weak} collection.** We utilize the "text-davinci-003" GPT completion model for augmenting zero-shot instances. We configure the temperature to 1.6 to ensure more varied outputs and cap the maximum token output from GPT-3.5 at 80. However, GPT-3.5 doesn't always maximize this limit. For each zero-shot label, we generate 5 instances to serve as weak supervision.

Table 2: Main results on three benchmarks

Models	FewRel				UFET				Situation			
	test	freq	few	zero	test	freq	few	zero	test	freq	few	zero
MWC (Soares et al., 2019)	49.82	94.23	55.23	0	11.69	44.88	13.41	0	28.16	43.00	34.47	7.00
NLI (Li et al., 2022)	63.46	**95.35**	48.81	46.22	38.28	53.26	34.44	**37.62**	42.12	**53.56**	34.02	38.77
PPL (Cui et al., 2022)	53.23	95.15	63.54	0	3.28	10.63	3.48	0.89	25.37	22.83	26.78	26.48
GPT 3.5	18.24	18.22	25.33	11.17	19.87	31.05	16.02	20.37	**57.53**	51.87	**59.95**	**60.78**
X-Shot	**68.48**	94.06	58.04	**53.34**	**38.46**	**55.69**	**34.74**	37.00	44.46	52.82	33.51	47.04

• **Prediction threshold.** Both NLI baseline and our method necessitate a threshold for assigning label predictions. We use the probability of the positive class produced by the model for this purpose. For *FewRel*, the label with the highest score is chosen. In *UFET* and *Situation*, we introduce a threshold parameter, *t*. Any label exceeding this probability threshold, *t*, is considered in the final prediction. We experiment with various values of *t*, ranging from 0.5 to 1, and select the optimal one. For *Situation*, there's a unique label "*None*" which signifies that none of the predefined situations are applicable. If no situation surpasses the threshold *t*, the label "*None*" is assigned.

5.2 Results

The primary results are displayed in Table 2. Our model generally surpasses the baselines. While traditional multi-way classification excels with ample annotations, its performance falters in few-shot and especially zero-shot situations. Similarly, the few-shot prompting baseline struggles when encountering unseen instance texts, highlighting the constraints of classification models in the X-Shot context.

The in-context learning method shines in *Situation* with its limited 12 labels and simpler nature. However, X-Shot still exceeds all other baselines significantly. Also, when it comes to the other two datasets where we have hundreds of labels, the model can no longer make wise decisions.

While the NLI-based indirect supervision—a prevalent method that transforms the zero-shot task into an existing NLI problem—delivers impressive results across various settings, our method proves to be even more potent. This underscores the superior robustness of the instruction-learning approach in the context of the X-Shot setting.

5.2.1 Analysis

Error Analysis. To analyze the error patterns, we pick *Situation* dataset as the example and collect the most typical errors as follows:

• **Bias toward more frequent labels** Under our multi-label classification problem setting where the number of labels can be up to 230, it would be very common for multiple labels to have similar semantic meanings. Even with the situation dataset where we have the least number of labels (11+1), we can still find similar labels, such as "terrorism" versus "crime/violence". For example, one input sentence from the Situation dataset is "@-@ Maiduguri hit @-@ with boko haram squeezed out of captured territory, security analysts have predicted a rise in bomb attacks in towns and cities, including to disrupt elections in three weeks ' time." Even though the gold label is "terrorism", it gives 0.99 probability for "crime/violence" considering it does have a similar meaning and as a frequent label it has been seen multiple times. We can see that the frequency of the label being seen can be an important factor, especially since we have a massive amount of labels that can easily confuse the model.

• **Misled by Textual Cues** Occasionally, the input sentence includes terms directly related to one of the labels, even if the context doesn't correspond to that label.

For example, one input sentence is "the two dead adults were either villagers or rescuers searching for those missing, xinhua added" include, which mentions the term "rescuers". The model strongly favors the "search/rescue" label, while none of the labels fit and the gold label is actually " None".

• **Ambiguous labels** It's common for people to have disagreements on the annotations. Sometimes the model makes a more appropriate judgment than the data provides to some people's perspectives. One such example is "so much untreated sewage has been pumped into sierra leone 's rivers and coastal waters that much of the water itself is contaminated with the cholera bacteria , unicef said .". The ground truth label for this input is "utilities, energy, or sanitation". However, the model also strongly suggested "medical assistance", a fair choice given the mention of "cholera bacteria".

Why do few-shot labels outperform zero-shot labels at times? We observe that within the UFET and Situation benchmarks, the performance of few-shot labels is slightly worse than zero-shot labels. We hypothesize that this outcome is attributed to the robustness of LLMs when endowed with extensive pretrained knowledge, wherein both scenarios of no fine-tuning and fine-tuning with ample data manifest resilience. Conversely, minimal fine-tuning tends to induce overfitting.

Influence of Task Type Overlap. The Super-NaturalInstruction dataset doesn't directly include our target datasets. We removed the top 10 tasks closest to each test dataset to assess the impact of similar tasks. The measurement is based on cosine similarity between Sentence-BERT embeddings of the 757 task definitions in the Super-NaturalInstruction dataset and each test dataset's instruction.

Table 3: Results of retraining the model after deleting top-10 similar tasks

	test	**freq**	**few**	**zero**
FewRel	63.34	89.04	60.95	40.04
UFET	38.05	53.39	34.20	37.30
Situation	41.96	49.76	36.96	39.15

Comparing results in Table 3 with those in Table 2, there's a minor performance decline for FewRel and Situation datasets. However, UFET's performance remains stable. This suggests that similar tasks in the Super-NaturalInstruction dataset can be beneficial. Even with slight decreases, results still surpass baseline levels, underscoring the value of diverse training tasks. This is further supported by subsequent analysis.

Number of Tasks vs Number of Instances. Balancing the number of tasks and the number of instances per task is pivotal in data collection. We wonder, by keeping the total instance count constant, should we have more tasks or more instances per task? We try [100,200,..,700] for the varying number of tasks, each with 100 instances.

In total, we have [10,000, 20,000, ... 70,000] instances. Accordingly, for the varying number of instances per task, we have datasets with [10,000/757, 20,000/757, ... 70,000/757] number of instances. The overall instances remain the same in each step.

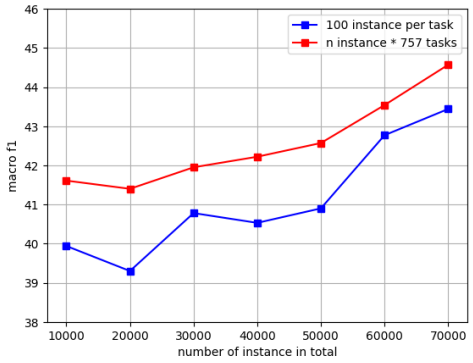

Figure 3: #instances vs. #tasks

From Figure 3, it's evident that both task count and instance count boost performance. While increasing either is beneficial, having more tasks has a greater impact than adding more instances to each task. Given these insights, future work should focus on diversifying the types of tasks exposed to the model, considering data constraints.

6 Conclusion

This work introduces X-Shot , a challenging text classification framework where labels range from non-existent to frequent. X-Shot reflects realistic scenarios where we encounter frequent-shot, few-shot, and zero-shot labels simultaneously. Our innovative approach recasts any text classification issue into a binary task, handling varying label amounts and frequencies. We introduce X-Shotto navigate this intricate challenge, leveraging instruction learning and PLMs' weak supervision. Our approach consistently outperforms the latest methods across three benchmark datasets in both single and multi-label contexts.

REFERENCES

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. FLEX: unifying evaluation for few-shot NLP. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 15787–15800, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8493eeaccb772c0878f99d60a0bd2bb3-Abstract.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. Ultra-fine entity typing. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 87–96. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1009. URL https://aclanthology.org/P18-1009/.

Ganqu Cui, Shengding Hu, Ning Ding, Longtao Huang, and Zhiyuan Liu. Prototypical verbalizer for prompt-based few-shot tuning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 7014–7024. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.483. URL https://doi.org/10.18653/v1/2022.acl-long.483.

Bowen Dong, Yuan Yao, Ruobing Xie, Tianyu Gao, Xu Han, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. Meta-information guided meta-learning for few-shot relation classification. In Donia Scott, Núria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pp. 1594–1605. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.140. URL https://doi.org/10.18653/v1/2020.coling-main.140.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 4803–4809. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1514. URL https://doi.org/10.18653/v1/d18-1514.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 7038–7051. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.564. URL https://doi.org/10.18653/v1/2021.emnlp-main.564.

Bangzheng Li, Wenpeng Yin, and Muhao Chen. Ultra-fine entity typing with indirect supervision from natural language inference. *Trans. Assoc. Comput. Linguistics*, 10:607–622, 2022. doi: 10.1162/tacl_a_00479. URL https://doi.org/10.1162/tacl_a_00479.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. Summarization as indirect supervision for relation extraction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 6575–6594. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.490. URL https://doi.org/10.18653/v1/2022.findings-emnlp.490.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Noisy channel language model prompting for few-shot text classification. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 5316–5330. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.365. URL https://doi.org/10.18653/v1/2022.acl-long.365.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 3470–3487. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.244. URL https://doi.org/10.18653/v1/2022.acl-long.244.

Abiola Obamuyide and Andreas Vlachos. Zero-shot relation classification as textual entailment. In *Proceedings of the first workshop on fact extraction and VERification (FEVER)*, pp. 72–78, 2018.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 2439–2455. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-naacl.187. URL https://doi.org/10.18653/v1/2022.findings-naacl.187.

Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 255–269. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.20. URL https://doi.org/10.18653/v1/2021.eacl-main.20.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4077–4087, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 2895–2905. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1279. URL https://doi.org/10.18653/v1/p19-1279.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1199–1208. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00131. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Sung_Learning_to_Compare_CVPR_2018_paper.html.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg,

Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3630–3638, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html.

Yingyao Wang, Junwei Bao, Guangyi Liu, Youzheng Wu, Xiaodong He, Bowen Zhou, and Tiejun Zhao. Learning to decouple relations: Few-shot relation classification with entity-guided attention and confusion-aware training. In Donia Scott, Núria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pp. 5799–5809. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.510. URL https://doi.org/10.18653/v1/2020.coling-main.510.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 5085–5109. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.340. URL https://doi.org/10.18653/v1/2022.emnlp-main.340.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Congying Xia, Chenwei Zhang, Jiawei Zhang, Tingting Liang, Hao Peng, and Philip S. Yu. Low-shot learning in natural language processing. In *2nd IEEE International Conference on Cognitive Machine Intelligence, CogMI 2020, Atlanta, GA, USA, October 28-31, 2020*, pp. 185–189. IEEE, 2020. doi: 10.1109/CogMI50398.2020.00031. URL https://doi.org/10.1109/CogMI50398.2020.00031.

Congying Xia, Wenpeng Yin, Yihao Feng, and Philip S. Yu. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 1351–1360. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.106. URL https://doi.org/10.18653/v1/2021.naacl-main.106.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. Crossfit: A few-shot learning challenge for cross-task generalization in NLP. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 7163–7189. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.572. URL https://doi.org/10.18653/v1/2021.emnlp-main.572.

Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 3912–3921. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1404. URL https://doi.org/10.18653/v1/D19-1404.

Wenpeng Yin, Dragomir R. Radev, and Caiming Xiong. Docnli: A large-scale dataset for document-level natural language inference. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pp. 4913–4922. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.435. URL https://doi.org/10.18653/v1/2021.findings-acl.435.

Wenpeng Yin, Muhao Chen, Ben Zhou, Qiang Ning, Kai-Wei Chang, and Dan Roth. Indirectly supervised natural language processing. In Yun-Nung Vivian Chen, Margot Mieskes, and Siva Reddy (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 32–40. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-tutorials.5. URL https://doi.org/10.18653/v1/2023.acl-tutorials.5.

Haoxing Zhang, Xiaofeng Zhang, Haibo Huang, and Lei Yu. Prompt-based meta-learning for few-shot text classification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 1342–1357. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.87. URL https://doi.org/10.18653/v1/2022.emnlp-main.87.

Jian-Guo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip S. Yu, Richard Socher, and Caiming Xiong. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 5064–5082. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.411. URL https://doi.org/10.18653/v1/2020.emnlp-main.411.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. Integrating semantic knowledge to tackle zero-shot text classification. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 1031–1040. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1108. URL https://doi.org/10.18653/v1/n19-1108.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. Learning with different amounts of annotation: From zero to many labels. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 7620–7632. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.601. URL https://doi.org/10.18653/v1/2021.emnlp-main.601.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021. URL http://proceedings.mlr.press/v139/zhao21c.html.

A APPENDIX

A.1 IN-CONTEXT LEARNING TEMPLATE

For the in-context learning baseline, we provide 3 demonstrations, 2 positive ones and 1 negative one, and let GPT complete the label of the test instance. The template is as follows:

Sentence: Pan was appointed director of the National Academy (Zhejiang Academy of Fine Arts) by the Kuomintang Ministers
Entity 1: Chen Lifu
Entity 2: Kuomintang
Relation: member of political party
Label: Yes

Sentence: Aldo Protti (July 19 ,1920 – August 10 , 1995) was an Italian baritone opera singer
Entity 1: Aldo Protti
Entity 2: baritone
Relation: voice type
Label: Yes

Sentence: Part of DirectXDirect3D is used to render three – dimensional graphics in applications
Entity 1: DirectX
Entity 2: Direct3D
Relation: movement
Label: No

Sentence: The Suzuki GS500 is an entry level motorcycle manufactured and marketed by the Suzuki Motor Corporation.
Entity 1: Suzuki GS500
Entity 2: Suzuki Motor Corporation
Relation: winner
Label:

A.2 TASK INSTUCTIONS

To prove the robustness of our model, we create 3 versions of the task instructions for each of the datasets (*FewRel*,*UFET*,*Situation*) as follows:

FewRel
Instruction A: Given a sentence about two entities, return a relation between the two entities that can be inferred from the sentence.
Instruction B: Your task is to identify a relationship between two entities mentioned in a given sentence.
Instruction C: Identify the relationship between two entities in a given sentence that can be inferred from the sentence.

UEFT
Instruction A: Given a sentence about an entity, return the type of entity that can be inferred from the sentence.
Instruction B: The task is to identify the type of an entity mentioned in the sentence based on the information provided in the sentence.
Instruction C: Determine the type of entity mentioned in the given sentence by analyzing the context of the sentence.

Situation
Instruction A: Given a sentence about a situation, return the type of the situation that can be inferred from the sentence.
Instruction B: The task is to identify the situation mentioned in the sentence based on the information provided in the sentence.
Instruction C: Determine the situation mentioned in the given sentence by analyzing the context of the sentence.