X-Shot: A Unified System to Handle Frequent, Few-shot and Zero-shot Labels in Classification

Anonymous ACL submission

Abstract

In recent years, few-shot and zero-shot learning, which learn to predict labels with limited annotated instances, have garnered significant attention. Traditional approaches often treat frequent-shot (freq-shot; labels with abundant 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 of them might appear thousands of times, while others might only appear sporadically or not at all. For practical deployment, it is 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 co-occur 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. To solve X-Shot, we propose BinBin (**bin**ary **in**ference **b**ased on **in**struction following) that leverages the *Indirect Supervision* from a large collection of NLP tasks via instruction following, bolstered by *Weak Supervision* provided by large language models. BinBin surpasses preceding state-of-the-art techniques on three benchmark datasets across multiple domains. To our knowledge, this is the first work addressing X-Shot learning, where X remains variable.[1]

1 Introduction

Over recent years, there's been a growing focus in AI on enhancing model performance while minimizing the need for extensive human labeling, which is typically termed as few-shot or zero-shot. Historically, the fields of frequent-shot, few-shot, and zero-shot learning have been approached as distinct paradigms, with systems optimized separately for each setting. Yet, in real-world scenarios, label frequencies can exhibit broad variation, with certain labels occurring prolifically, while 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 and Vlachos, 2018; Yin et al., 2019; Xu et al., 2022). Thus, mastering the ability to handle all conceivable label occurrences is paramount for systems aiming for practical deployment.

In this work, we introduce a more challenging and practically useful task: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.

Tackling X-Shot spawns two core technical conundrums: ($\mathcal{Q}_1$) Amidst the paucity of annotations characteristic of few-shot and zero-shot contexts, how one might identify apt sources of *Indirect Supervision* (Yin et al., 2023) to navigate the X-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-NaturalInstruction (Wang

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

et al., 2022). These datasets primarily contain various NLP tasks enriched with textual instructions. Our method trains the model on these datasets, aiming for robust generalization to the unseen X-Shot task when supplemented with pertinent instructions, especially for the low-shot labels. 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 that manages text classification tasks with labels of varied sizes. By amalgamating solutions for both $\mathcal{Q}_1$ and $\mathcal{Q}_2$, we forge a holistic framework, BinBin (**bin**ary **in**ference **b**ased on **in**struction following).

There are, however, no existing datasets that explicitly cater to this challenge. To evaluate our system, we turn to three representative classification tasks: relation classification, event detection, and argument role identification. We recompile their associated datasets: *FewRel* (Han et al., 2018), *MAVEN* (Wang et al., 2020), and *RAMS* (Ebner et al., 2020) to simultaneously contain frequent-shot, few-shot, and zero-shot instances. Sourced from diverse domains (Wikipedia, news articles, etc.), and featuring vast label counts (ranging from 30 to 78), these datasets pose a formidable challenge to contemporary text classification systems. Moreover, the *MAVEN* dataset uniquely integrates an "None" label, further amplifying the realistic nature of the task. Experiments reveal our system's resilience across datasets, consistently outperforming leading baselines, including GPT-3.5.

Our contributions can be summarized as follows: (i) We introduce X-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 BinBin, harnessing the potential of instruction-following 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 do not 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 (Chang et al., 2008; Qiao et al., 2017). Later works adopted word embeddings from pre-trained language models to represent the meaning of the text or the label (Alcoforado et al., 2022; Wang et al., 2023). Recent works have been enhancing the embedding representations by integrating class hierarchy, class descriptions, 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 (Holtzman et al., 2021; Zhao et al., 2021; Min et al., 2022). The latest trend in zero-shot text classification leverages *Indirect Supervision* from well-annotated NLP tasks such as text entailment (Obamuyide and Vlachos, 2018; Yin et al., 2019). However, these methods often underutilize available annotations for labels.

Indirect Supervision There is a burgeoning interest in *Indirect Supervision* (Yin et al., 2023) in recent years. 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; Xu et al., 2023c), event argument extraction (Sainz et al., 2022), entity typing (Li et al., 2022) and relation extraction (Xia et al., 2021; Lu et al., 2022; Xu et al., 2023b; Zhou et al., 2023). Beyond entailment, knowledge from areas like question answering (Yin et al., 2021), summarization (Lu et al., 2022) and dense retrievers (Xu et al., 2023c) has been incorporated. However, precious *Indirect Supervision* is usually collected from a single source task. 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.

Unified Discriminative Classifier Previous research, such as the work presented in (Xu et al., 2023a), also attempts to transform classification problems into binary tasks. While this system represents a discriminative classifier approach similar to ours, there are several significant differences. The most notable distinction is that it focuses exclusively on zero-shot learning scenarios, whereas our X-Shot encompasses the entire range of label occurrences. Additionally, it relies solely on the instance itself and therefore is less flexible than ours, while our method utilizes instructions to enrich the context and can be adapted to more diverse tasks. Most importantly, this system benchmarks its performance against generative models, rather than comparing it with state-of-the-art (SOTA) systems specifically designed for classification tasks.

3 Problem Statement

X-Shot has the following components:

- **Input** t: Versatile text in varied forms, lengths, and domains.

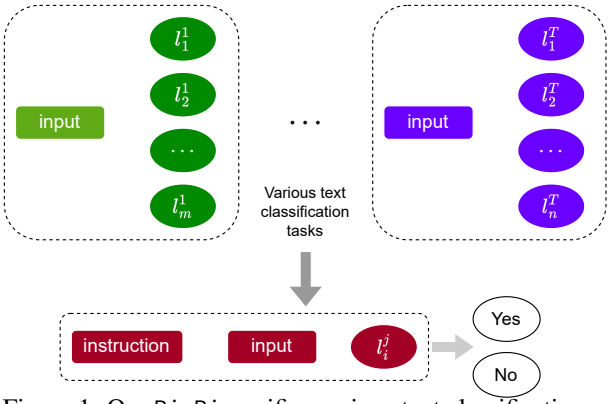

Figure 1: Our BinBin unifies various text classification tasks as an instruction tuning problem. More details in Appendix A.2

- **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, each label can be either zero-shot, few-shot, or more frequent.

Then, the task of X-Shot is to figure out label $L_s \in L$ that is correct for the input t, where $|L_s|$ might be zero (i.e., "*None*").

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 BinBin 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 BinBin architecture

We have devised a broad architecture that seamlessly transitions most classification tasks into a unified, instruction-driven binary classification formation. 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.

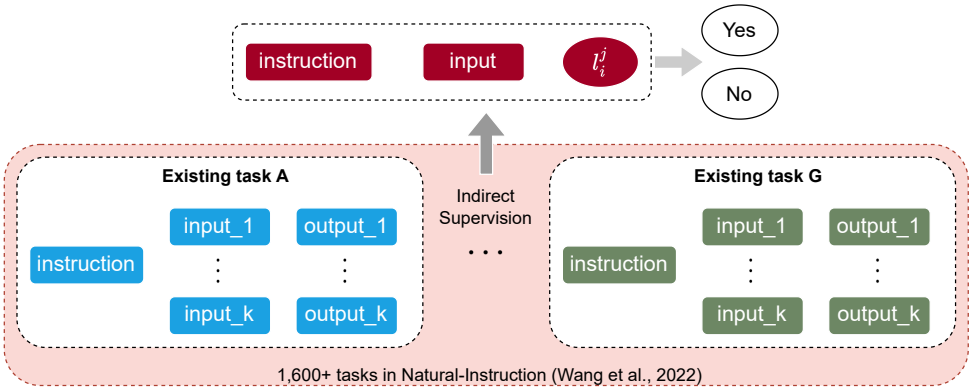

Figure 2: *Indirect Supervision* for `BinBin`. More details in Appendix A.1

The task then becomes determining if the label is appropriate ("Yes") or not ("No") given the input under the instruction. An example of the conversion can be found in Appendix A.2.

`BinBin` can freely support classification tasks with any number of labels. Instead of converting labels into numerical IDs as traditional supervised classifiers do, we retain the actual label names. Optionally, we can also employ sophisticated verbalizers (Schick and 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.

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

4.2 Supervision acquisition for low-shot labels

In this section, we will introduce how we conduct and combine *Indirect Supervision* and *Weak 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), summarization (Lu et al., 2022) and passage retrieval (Xu et al., 2023c). 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 is not a universally adaptable system to address the X`-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 task is accompanied by instructions and numerous input-output examples (example of tasks in Appendix A.1). This dataset offers an invaluable source of *Indirect Supervision* for our target X`-Shot`. As illustrated in Appendix A.1, for every task within the Super-NaturalInstruction dataset, we are 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 (`instruction`, `input`, `gold output`) and one negative triplet (`instruction`, `input`, `random output`) for each example in our training dataset as in Figure 2. 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 positive label l, we add an instruction and craft $|L|$ triplets (`instruction`, t, l) for each label l from the label space L, with the gold label as positive and the remainings as 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.

Weak Supervision for zero-shot labels. In addition to *Indirect 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-3.5 (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 from *Maven* designed to generate text and event trigger for a zero-shot event type label:

```
event type: Competition
event trigger: tournament
sentence: The final tournament was Played in two
stages: the group stage and the knockout stage.

event type: Motion
event trigger: throwing
sentence: Simultaneously, Sayhood gained a lock
on Rodriguez, throwing him onto the defensive.

event type: Manufacturing
```

In this approach, upon exposing GPT-3.5 to event and event statement examples associated with the event type labels "Competition" and "Motion", we introduce the zero-shot label "Manufacturing." Subsequently, GPT-3.5 generates an event trigger along with an event statement, serving as a weakly supervised instance for this unseen label.

Training strategy. We first train the RoBERTa-large model (Liu et al., 2019) on the transformed binary Super-NaturalInstruction dataset, then fine-tune on the augmented instances of downstream X-Shot tasks. The model used will be consistent in all experiments and baselines.

5 Experiments

5.1 Experimental setting

Datasets. There are no existing datasets that can exactly align with X-Shot. In this work, we standardize datasets that can cover (i) multiple domains, (ii) various sizes of labels, and (iii) out-of-domain label scenarios. Therefore, we recompile: *FewRel* (Han et al., 2018), *MAVEN* (Wang et al., 2020), and *RAMS* (Ebner et al., 2020), referring to *relation classification*, *event detection*, and *argument role identification* problems respectively. Next, we elaborate on the details of reorganizing each of them.

	domain	#freq	#few	#zero
FewRel$_{X\text{-Shot}}$	Wikipedia	26	26	26
MAVEN$_{X\text{-Shot}}$	Wikipedia	23	23	23+1
RAMS$_{X\text{-Shot}}$	News articles	10	10	10

Table 1: Statistics of dataset labels.

We rename each resulting dataset as "[]$_{X\text{-Shot}}$." Details can be seen in Table 1.

- **FewRel**$_{X\text{-Shot}}$: *FewRel* is a well-established relation classification dataset while each instance provides a relation statement, two entities from the statement, and their corresponding relation label. Since the test set of *FewRel* is not available, we include 78 relations from its *train* and *dev* 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 new $train$, and 200 instances for each label in the new dev and new $test$.
- **MAVEN**$_{X\text{-Shot}}$: The standard event detection task in *MAVEN* includes two steps: detecting the event trigger and predicting the event label from the trigger. In this work, we will focus on the second step, where we assume the event trigger is known and aim to predict the corresponding event label. The annotation of the original test set is not publicly available. To make *MAVEN* align with our setting, we reorganize its $train$ and dev sets as follows: since the event label distribution is significantly imbalanced, we adopt 69 of them who have 400+ instances plus the "*None*" label as our label set. Labels are divided into 23/23/23+1 as freq/few/zero-shot labels with "*None*" belonging to the zero-shot group. We put 300/5/0 instances for each freq/few/zero label in the new $train$, and 100 instances for each label in the new dev and $test$.
- **RAMS**$_{X\text{-Shot}}$: *RAMS* tackles the task of identifying semantic role labels given the sentence marked with event triggers and argument terms. There are 30 labels that have more than 100 instances; we split them into 10/10/10 for each label group. Similarly, we put 300/5/0 instances for each freq/few/zero label in the new $train$, and 50 instances for each label in the new dev and $test$.

Baselines. For baselines, we compare our system with the current SOTA multi-way classification model (for traditional frequent label setting), the

most advanced few-shot/zero-shot learning methods, and the in-context learning with GPT-3.5.

- **Multi-way classification (MWC, (Soares et al., 2019))** . This methodology is the prior SOTA approach for relation classification. We employ this strategy for all three datasets, given that they all contain term features (entity, event trigger, argument, etc.) within their inputs.
- **In-context learning with GPT-3.5 (GPT-3.5).** We create a prompt that includes three demonstrations, two positive and one negative, and each comes with the input, prediction, and a True/False label that indicates whether the prediction is correct. The template can be seen in Appendix A.3.
- ***Indirect Supervision* from NLI (NLI; Li et al. 2022)**. The prior SOTA approach for addressing a zero-shot or few-shot classification with *Indirect Supervision* from merely the NLI source task. This paradigm uses the input text as the premise and transforms the label into a hypothesis sentence.
- **Prototypical Prompt learning (PPL; Cui et al. 2022)** The prior SOTA system for few-shot classification. For each of the dataset, we select 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 available instances until we reach the number. For zero-shot labels, we simply put the label itself as the text for the training.

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 `BinBin`, we draft three distinct instruction prompts and present the average outcomes to demonstrate the system's stability. All template are available in Appendix A.4.

• **Weak supervision.** We use the "text-davinci-003" GPT-3.5 completion model to augment zero-shot instances. Temperature is set 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*.

• **Prediction threshold.** Both NLI baseline and our method necessitate a threshold for assigning label predictions. We use the probability of the positive class the model produces for this purpose. For *FewRel* and *RAMS*, the label with the highest score is chosen. In *MAVEN*, we introduce a threshold parameter, t. If the label receiving the highest probability does not exceed this probability threshold, we assign the label as "*None*". We experiment with various values of t, ranging from 0.5 to 1, and select the optimal one based on dev.

5.2 Results

Table 2 reports the main comparison between our `BinBin` system and those baselines. Our model consistently outperforms all baselines by a significant margin in the "all" and "zero" dimensions, while occasionally showing slightly lower but on-par performance with the baselines in "freq" and "few". Analyzing these baselines, we notice that most are ill-suited for the X`-Shot` problem setting, particularly in zero-shot scenarios where annotations are absent. MWC is entirely determined by the number of label-wise training examples; therefore, its performance, although pretty high for "freq", drops quickly to be 0.0 for "zero". In a similar vein, the few-shot prompting (PPL) baseline encounters difficulties with unseen class instances, underscoring the limitations of classification models in the X`-Shot` context. NLI, representing the SOTA in low-shot learning settings, is the only model adept at managing all three label sets. Nonetheless, when pitted against `BinBin`, NLI's performance remains subpar in few-shot and zero-shot situations. This indicates that, despite its competency in handling sparse or non-existent annotations, NLI's capacity for reasoning and exploiting limited supervision is inferior to our system.

As one of the most advanced closed-source LLMs, GPT-3.5 shows limited effectiveness in this task, with its performance across three label sets appearing strikingly similar. Although GPT-like models demonstrate robust capabilities in in-context learning, they *fall short in utilizing rich annotations when available* and often *struggle in scenarios with a vast decision space*. This highlights the flexibility of our `BinBin` in handling classification

Models	FewRel$_{X\text{-Shot}}$				RAMS$_{X\text{-Shot}}$				MAVEN$_{X\text{-Shot}}$			
	all	freq	few	zero	all	freq	few	zero	all	freq	few	zero
MWC (Soares et al., 2019)	49.82	94.23	55.23	0.0	34.47	**78.40**	25.00	0.0	42.43	85.17	43.96	0.0
NLI (Li et al., 2022)	63.46	**95.35**	48.81	46.22	43.07	71.40	20.40	37.40	56.31	**85.65**	39.83	44.00
PPL (Cui et al., 2022)	53.23	95.15	**63.54**	0.0	27.13	65.00	16.20	0.20	46.84	85.04	**55.52**	0.0
GPT-3.5	18.24	18.22	25.33	11.17	18.19	21.21	15.15	18.19	21.43	15.15	12.12	37.50
BinBin	**68.48**	94.06	58.04	**53.34**	**54.70**	77.00	**29.00**	**58.07**	**64.96**	84.32	46.64	**63.97**

Table 2: Main results on three benchmarks

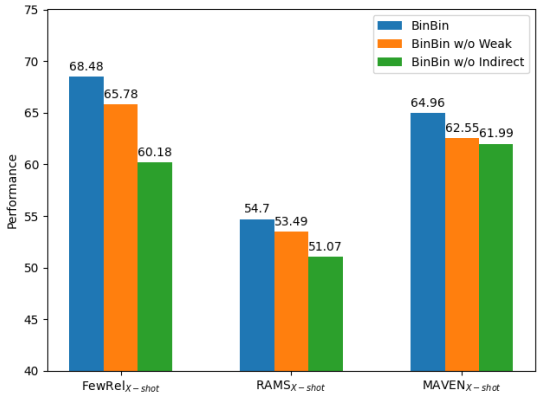

Figure 3: Ablation study of BinBin

	all	freq	few	zero
FewRel$_{X\text{-Shot}}$	63.34	89.04	**60.95**	40.04
RAMS$_{X\text{-Shot}}$	51.64	**78.74**	**30.13**	40.07
MAVEN$_{X\text{-Shot}}$	63.83	**85.68**	**47.48**	58.57

Table 3: Results of training BinBin after deleting top-10 similar tasks from Natural-Instruction. Bold numbers indicate enhanced performance compared to the pre-deletion state.

labels of different sizes and #examples.

5.3 Analyses

In addition to reporting the main results, we further analyze our system in the following dimensions: ($\mathcal{Q}_1$) the individual contribution of our *Indirect Supervision* and *Weak Supervision*; ($\mathcal{Q}_2$) why does "zero" show better performance than "few" in RAMS$_{X\text{-Shot}}$ and MAVEN$_{X\text{-Shot}}$? ($\mathcal{Q}_3$) Given that our *Indirect Supervision* is derived from a diverse range of NLP tasks in Natural-Instruction (Wang et al., 2022), is there a possibility of task leakage? ($\mathcal{Q}_4$) When selecting source tasks for *Indirect Supervision* in instruction-following, which configuration is more effective: having more (diverse) tasks or having more (task-wise) instances? ($\mathcal{Q}_5$) The efficiency of our system. ($\mathcal{Q}_6$) The mistakes our system makes.

($\mathcal{Q}_1$) Ablation study. Figure 3 depicts the ablation study, where either *Indirect Supervision* of *Weak Supervision* is discarded from our system BinBin. Our findings reveal that both supervision sources fulfill complementary roles in the X-Shot task. Encouragingly, while their combined usage yields the best results, each type of supervision, on its own, still significantly surpasses the baselines. This underscores the efficiency of our system.

($\mathcal{Q}_2$) Why do zero-shot labels outperform few-shot labels in the *MAVEN*$_{X\text{-Shot}}$ and *RAMS*$_{X\text{-Shot}}$ benchmarks? We observe that this phenomenon applies not only to our system, but also to baselines "NLI" and "GPT-3.5". We suspect two reasons: i) Some zero-shot labels in *RAMS*$_{X\text{-Shot}}$ seem easier upon visual inspection; ii) In *MAVEN*$_{X\text{-Shot}}$, "None" is treated as a zero-shot label in the test set, contributing notably due to threshold tuning.

($\mathcal{Q}_3$) Influence of Task Type Overlap. Although the Natural-Instruction task repository doesn't directly contain our target datasets, we still remove 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 (Reimers and Gurevych, 2019) embeddings of the 757 training task definitions in the Natural-Instruction dataset and each X-Shot test dataset's instruction.

From Table 3, we can observe that: i) The main decreases when the top-10 similar tasks are deleted happen to zero-shot labels. Recall that we only provided *Weak Supervision* for them; this phenomenon indicates that pretraining on similar source tasks can help diminish the impact of noise in the weakly supervised data. ii) Despite slight decreases in "all", our results still surpass baselines in Table 2, underscoring the value of diverse training tasks. This is further supported by subsequent analysis.

($\mathcal{Q}_4$) Number of Tasks vs Number of Instances. Balancing the number of tasks and the number of instances per task is pivotal in curating instruction-

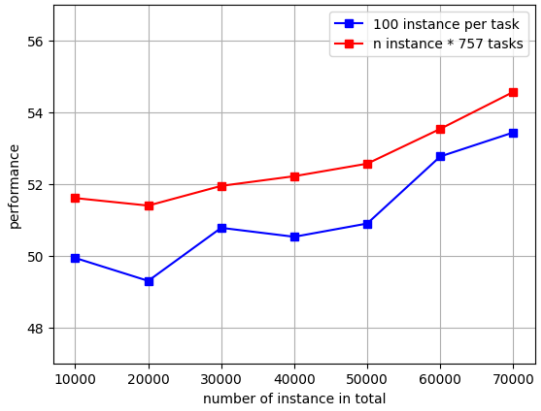

Figure 4: #instances vs. #tasks

following datasets (Lou et al., 2023). 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. From Figure 4, it's evident that both task count and instance count boost performance. While increasing either is beneficial, having more (diverse) 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.

($\mathcal{Q}_5$) Efficiency Analysis. Efficiency concerns center around the inference stage, where our system, like "NLI," converts varied-label classification problems into a binary inference task. Training of our system takes more time due to pretraining on Natural-Instruction, but during testing, both systems are equally efficient as they make binary decisions for each label candidate.

($\mathcal{Q}_6$) Error Analysis. We collect the most typical errors as follows:

• **Multiple labels make sense** In datasets with a large number of labels, it is often feasible for more than one label to appropriately fit the context. Sometimes, the model's interpretation may align more accurately with certain perspectives than the original data. Consider the example from *RAMS* dataset: "Many high - ranking figures in companies tied to Skolkovo have also donated to the Clinton Foundation" While the ground truth label for the argument "Clinton Foundation" is "recipient", the model strongly suggests "beneficiary"—a label that is equally justifiable.

• **Bias towards more frequent labels** It's quite common for multiple labels to have overlapping semantic meanings. In such scenarios, the model tends to favor the labels it encounters more frequently. For example, consider a sentence from the *FewRel* dataset: "The Spanish - Andorran border runs 64 km between the south of Andorra and northern Spain (by the autonomous community of Catalonia) in the Pyrenees Mountains.". Here, the entities are "Catalonia" and "autonomous community". Although the gold relation for the two entities is "instance of", the model assigns the highest probability to "part of"—a frequent group label. This suggests that not only does the label share semantic similarities with others, but its frequent occurrence also biases the prediction, especially when many labels lead to potential confusion.

• **identifying reciprocal or inverse relationships** This issue arises when the model struggles to differentiate between roles that are directly related to each other but represent opposite positions in a given context, such as in a "receiver" and "giver" scenario while both roles are part of the same transaction, but the model confuses who is who. For instance, in a sentence from *RAMS* "She was shouting , 'I am a terrorist,' and reportedly threatened to blow herself up . ..... he could n't believe that the decapitated child 's head being carried by the woman was real." where "she" is a "killer". However, the model incorrectly labels "she" as a "victim", demonstrating the difficulty in accurately discerning reciprocal roles.

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 (or freq-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 BinBin to navigate this intricate challenge, leveraging instruction *Indirect Supervision* and PLMs' *Weak Supervision*. Our approach consistently outperforms the latest methods across three benchmark datasets crossing multiple domains and diverse label occurrence.

Limitation

One of the primary limitations of our model is its efficiency, particularly when handling datasets with a large number of labels when converting the original task into a binary task. This results in extended training times and increased computational efforts. It is important to note that this limitation is not an isolated challenge for our model; it aligns with the experiences reported in previous state-of-the-art models. Future work can focus on optimizing the training process to enhance efficiency without compromising the model's performance.

References

Alexandre Alcoforado, Thomas Palmeira Ferraz, Rodrigo Gerber, Enzo Bustos, André Seidel Oliveira, Bruno Miguel Veloso, Fábio Levy Siqueira, and Anna Helena Reali Costa. 2022. Zeroberto: Leveraging zero-shot text classification by topic modeling. In *Computational Processing of the Portuguese Language - 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21-23, 2022, Proceedings*, volume 13208 of *Lecture Notes in Computer Science*, pages 125–136. Springer.

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. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, pages 830–835. AAAI Press.

Ganqu Cui, Shengding Hu, Ning Ding, Longtao Huang, and Zhiyuan Liu. 2022. Prototypical verbalizer for prompt-based few-shot tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7014–7024. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8057–8077. Association for Computational Linguistics.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4803–4809. Association for Computational Linguistics.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7038–7051. Association for Computational Linguistics.

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

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

Renze Lou, Kai Zhang, Jian Xie, Yuxuan Sun, Janice Ahn, Hanzi Xu, Yu Su, and Wenpeng Yin. 2023. MUFFIN: curating multi-faceted instructions for improving instruction-following. *CoRR*, abs/2312.02436.

Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. 2022. Summarization as indirect supervision for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6575–6594. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5316–5330. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3470–3487. Association for Computational Linguistics.

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

Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. 2017. Visually aligned word embeddings for improving zero-shot learning. *CoRR*, abs/1707.05427.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. 2022. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2439–2455. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269. Association for Computational Linguistics.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2895–2905. Association for Computational Linguistics.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. 2018. 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*, pages 1199–1208. Computer Vision Foundation / IEEE Computer Society.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3630–3638.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A massive general domain event detection dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1652–1671. Association for Computational Linguistics.

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. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5085–5109. Association for Computational Linguistics.

Yue Wang, Lijun Wu, Juntao Li, Xiaobo Liang, and Min Zhang. 2023. Are the BERT family zero-shot learners? A study on their potential and limitations. *Artif. Intell.*, 322:103953.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. 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.

Congying Xia, Wenpeng Yin, Yihao Feng, and Philip S. Yu. 2021. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. In *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*, pages 1351–1360. Association for Computational Linguistics.

Haike Xu, Zongyu Lin, Jing Zhou, Yanan Zheng, and Zhilin Yang. 2023a. A universal discriminator for zero-shot generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10559–10575. Association for Computational Linguistics.

Hanzi Xu, Slobodan Vucetic, and Wenpeng Yin. 2022. Openstance: Real-world zero-shot stance detection. *CoRR*, abs/2210.14299.

Jiashu Xu, Mingyu Derek Ma, and Muhao Chen. 2023b. Can NLI provide proper indirect supervision for low-resource biomedical relation extraction? In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 2450–2467, Toronto, Canada. Association for Computational Linguistics.

Nan Xu, Fei Wang, Mingtao Dong, and Muhao Chen. 2023c. Dense retrieval as indirect supervision for large-space decision making. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15021–15033, Singapore. Association for Computational Linguistics.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. Crossfit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7163–7189. Association for Computational Linguistics.

Wenpeng Yin, Muhao Chen, Ben Zhou, Qiang Ning, Kai-Wei Chang, and Dan Roth. 2023. Indirectly supervised natural language processing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 32–40. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3912–3921. Association for Computational Linguistics.

Wenpeng Yin, Dragomir R. Radev, and Caiming Xiong. 2021. Docnli: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4913–4922. Association for Computational Linguistics.

Haoxing Zhang, Xiaofeng Zhang, Haibo Huang, and Lei Yu. 2022. Prompt-based meta-learning for few-shot text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1342–1357. Association for Computational Linguistics.

Jian-Guo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, Philip S. Yu, Richard Socher, and Caiming Xiong. 2020. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5064–5082. Association for Computational Linguistics.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating semantic knowledge to tackle zero-shot text classification. In *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)*, pages 1031–1040. Association for Computational Linguistics.

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

Wenxuan Zhou, Sheng Zhang, Tristan Naumann, Muhao Chen, and Hoifung Poon. 2023. Continual contrastive finetuning improves low-resource relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13249–13263, Toronto, Canada. Association for Computational Linguistics.

A Appendix

A.1 Super-NaturalInstruction to `BinBin`

We convert Super-NaturalInstruction (Wang et al., 2022) into our binary schema for the *Indirect Supervision*. Super-NaturalInstruction is a benchmark In-context learning dataset with 757 train tasks and 119 test tasks. Each task includes a definition, positive examples, negative examples, and thousands of instances. A task example from Super-NaturalInstruction is presented in Figure 6. We select 100 instances from each task and convert them into `BinBin` schema for *Indirect Supervision* training as shown in Figure 7.

A.2 X`-Shot` data to `BinBin`

As discussed in Section 4.1, each X`-Shot` instance is converted into the unified binary format to align with `BinBin`. A detailed example from *FewRel* is illustrated in Figure 5.

A.3 In-context Learning template

For the in-context learning baseline, we provide 3 demonstrations, 2 positive ones and 1 negative one, and let GPT-3.5 complete the label of the test instance. The template is as follows for *FewRel*:

Original Instance:

Sentence: "3D Friends (stylized as 3D FRIENDS) is an American indie rock band from Austin , Texas
Entity 1: 3D Friends
Entity 2: indie rock
Relation: genre

Unified Schema:

Instruction Template

Input:

Definition: Given a sentence about two entities, return a relation between the two entities that can be referred from the sentence.

Positive Example 1 -

Sentence: *Mount Storer* ( ) is a jagged peak in the *Tula Mountains* , 4 nautical miles (7 km) east - northeast of Mount Harvey.
Entity 1: *Mount Harvey*
Entity 2: *Tula Mountains*
Relation: mountain range

Positive Example 2 -

Sentence: On the east side of the square stands the impressive mansion of *Dundas House*, built by *Sir William Chambers* for Sir Lawrence Dundas between 1772 and 1774
Entity 1: *Sir William Chambers*
Entity 2: *Dundas House*
Relation: notable work

Now complete the following example -

Instance for Prediction

Input: sentence: "3D Friends (stylized as 3D FRIENDS) is an American indie rock band from Austin , Texas
Entity 1: 3D Friends
Entity 2: indie rock
Relation: genre / company

label: Yes No

Figure 5: Classification to binary `BinBin`

Definition

In this task, you will be shown a short story with a beginning, two potential middles, and an ending. Your job is to choose the middle statement that makes the story coherent / plausible by writing \"1\" or \"2\" in the output. If both sentences are plausible, pick the one that makes most sense.

Positive Examples

Input: Beginning: John was on the trail running. Middle 1: John accelerated the speed and broke his leg accidentally. Middle 2: John was chased by a bear. Ending: He ran even faster until he got to his car safely.
Output: 2
Explanation: When someone breaks his/her leg, it is difficult to run. Therefore, we choose 2 in this case.

Negative Examples

Input: Beginning: Jon decided to steal a police car. Middle 1: Jon crashed the police car into a telephone poll. Middle 2: Jon wasn't caught. Ending: Jon went to prison for three years.
Output: Jon crashed the police car into a telephone poll.
Explanation: You should not answer with the chosen sentence. You should only answer with 1 or 2

Instances

Input: Beginning: Today I was cooking hamburgers inside. Middle 1: I burned my hand. Middle 2: I burned my feet. Ending: Now I have a blister.
Output: 1
.....

Figure 6: Super-Naturalinstructions task example

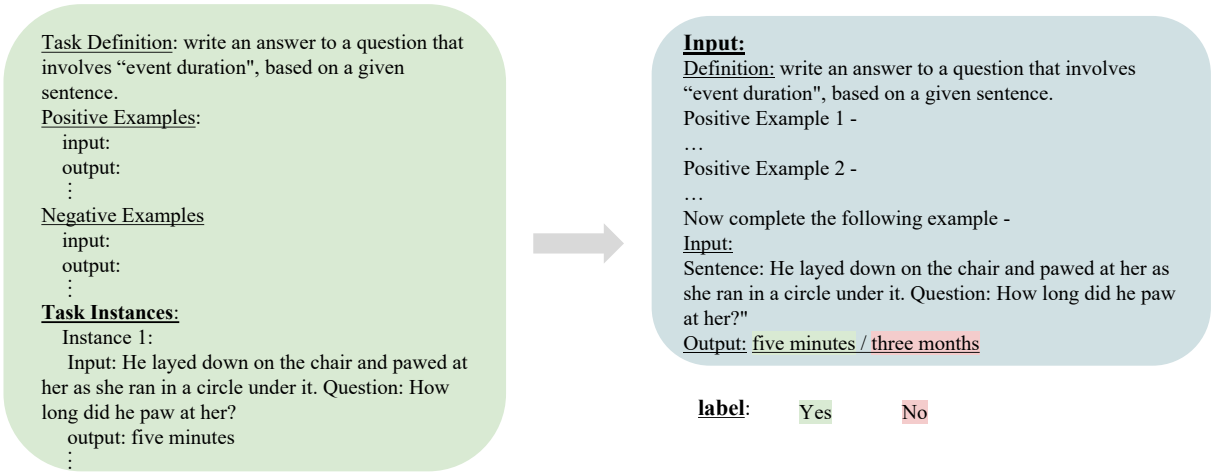

Figure 7: Super-Naturalinstructions to binary BinBin

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.4 BinBin Task Instuctions

To prove the robustness of our model, we create 3 versions of the task instructions for each of the datasets (*FewRel*, *MAVEN*, *RAMS*) 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.

RAMS
Instruction A: Your task is to identify the role of a specified argument within a given sentence, in relation to an identified event trigger.
Instruction B: Identify the role of the argument given the event trigger within the sentence.
Instruction C: Identify the role of the argument given the event trigger within the sentence.

MAVEN

Instruction A: `Given the sentence and the identified trigger word, determine the most appropriate event category for this trigger.`

Instruction B: `Identify the event type in the sentence associated with the trigger word.`

Instruction C: `Classify the event represented by the trigger word in the context of the following sentence.`

A.5 ACL ethics code discussion

• **Scientific artifacts usage** The existing Scientific artifacts included in this work are the RoBERTa model (Liu et al., 2019) and 3 NLP classification datasets. The model and datasets used in this work are publicly available for research purposes and do not contain any sensitive information. Our use of existing Scientific artifacts is consistent with their intended usage.

The license, copyright information, and terms of use information regarding `BinBin`, the asset we proposed, will be specified once the code is released.

• **Computational experiments** The number of parameters in the RoBERTa-large model is 355M. Our system is trained on NVIDIA RTX A5000 GPUs and takes 20 hours on average for a task on a single GPU. We incorporate packages mainly from huggingface for the modeling.