From Discrimination to Generation: Low-Resource Intent Detection with Language Model Instruction Tuning

Feng Zhang[1,2,3] **Wei Chen**[1,2,3*] **Fei Ding**[4,5] **Meng Gao**[1,2,3]
Tengjiao Wang[1,2,3] **Jiahui Yao**[2,3] **Jiabin Zheng**[1,2,3]
[1]Key Lab of High Confidence Software Technologies (MOE),
School of Computer Science, Peking University
[2]Research Center for Computational Social Science, Peking University
[3]Institute of Computational Social Science, Peking University (Qingdao)
[4]School of Intelligence Science and Technology, Peking University
[5]Institute for Artificial Intelligence, Peking University
{zhangfeng,dingfei,gaomeng}@stu.pku.edu.cn
{pekingchenwei,tjwang,yaojh,jiabinzheng}@pku.edu.cn

Abstract

Intent detection aims to identify user goals from utterances, and is a ubiquitous step towards the satisfaction of user desired needs in many interaction systems. As dynamic and varied intents arise, models that are capable of identifying new intents promptly are required. However, existing studies usually fine-tune discriminative models on the specific defined intent classes, precluding them from being directly adopted to new intent domains. In this paper, we introduce a generative pre-trained intent model that can recognize new intents from different domains in low-resource scenarios. We reformulate intent detection into a generation task and design descriptive and regularized instructions to guide the model effectively to detect new intents in open domains with no parameter updates. To validate the proposed method, we introduce a new intent detection benchmark, including the Meta-Intent Dataset and three types of representative evaluation settings. We conduct extensive experiments which demonstrate that our method outperforms a range of strong baselines that needs further fine-tuning or domain-specific samples.

1 Introduction

Intent detection aims to identify user intentions or main topic, playing the first and foremost role in task-oriented dialogue systems (Gupta et al., 2019), recommend systems (Qian et al., 2023) and other ubiquitous systems. Many existing studies have conducted considerable exploration (Goo et al., 2018; Wang et al., 2021), where pre-trained language models are fine-tuned under full supervision. However, due to the dynamic change of intents over time, gathering a large number of manually annotated examples and pre-defining a comprehensive intent set are quite challenging. Consequently, low-resource intent detection is proposed to recover new emerging intents with few or even zero labeled samples for relieving the need of substantial domain-specific labeled data.

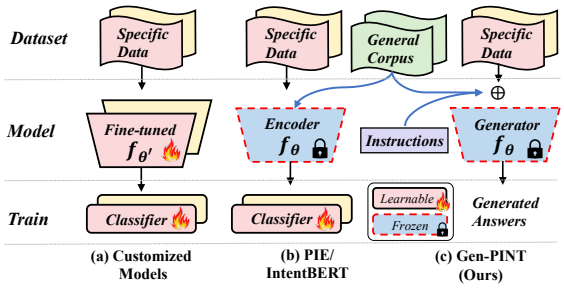

Figure 1: Illustration of different paradigms of intent detection.

Recent studies have been proposed to utilize meta-learning based methods, which transfer knowledge from seen intents to target new intents (Dopierre et al., 2021b; Chen et al., 2022a; Zhang et al., 2023a). Also, Zhang et al. (2023b) propose to directly fine-tune the model on the new intents with context augmentation and self-distillation. As illustrated in Figure 1a, these methods aim to fine-tune a customized model for each dataset. However, they require domain-specific labeled data for each intent detection task, which restricts the model generalization and leads to poor performance when facing large gap domains. Another type of method is designed to learn a unified intent-aware encoder, as shown in Figure 1b. Zhang et al. (2021) propose IntentBERT, an intent utterance encoder which firstly is trained on public source corpora and then is adapted to target domains via non-parametric

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 10167–10183

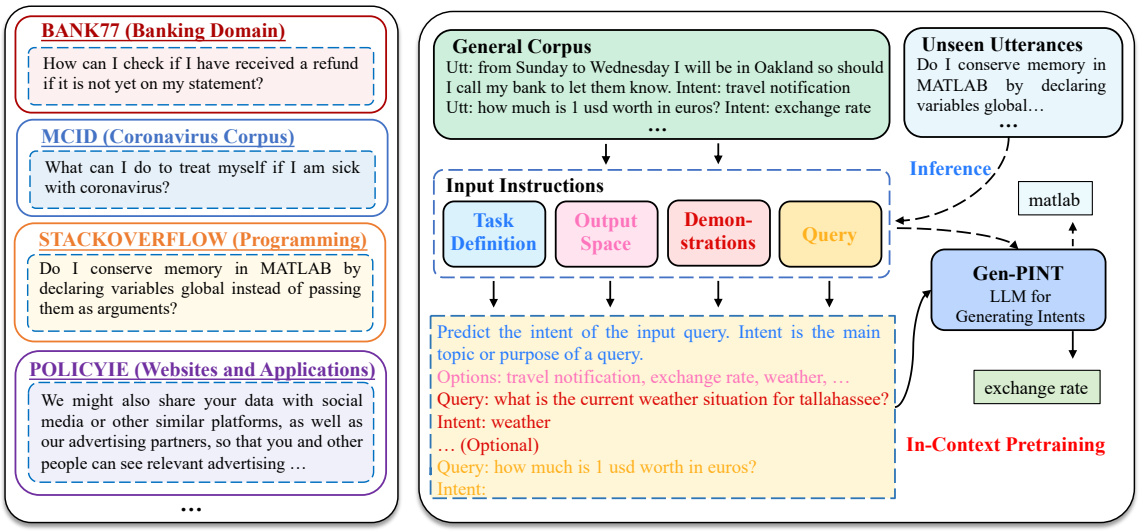

Figure 2: **Left**: Intent detection tasks from different domains. **Right**: Illustration of our framework.

matching or tuning additional classifiers. Sung et al. (2023) propose PIE, which is designed to align embeddings of utterances and pseudo-labels. However, they still need additional training for new tasks according to given intent classes and can not generate meaningful intent classes.

In this paper, we reformulate the intent detection as a language generation paradigm and propose a generative pre-trained intent model (Gen-PINT). Instead of customized networks in closed-world scenarios, Gen-PINT, as a unified model, can recognize new intents by the guidance of easily crafted prompts in open domains, as shown in Figure 1c. Specifically, Gen-PINT is tuned on a large set of general corpora to learn how to detect intents with instructions, and then is evaluated on new and diverse unseen target tasks. For the instruction format, we present four descriptive components to direct it to understand commonalities in intent detection. Task definition is the driver to enable Gen-PINT to comprehend the whole task paradigm, and the output space consists of all intent candidate options as restriction of the answer space. And demonstrations built on several sampled training samples guide the model to recover semantics and increase the diversity of training set. Through instruction tuning on general corpora, Gen-PINT acquires the ability to uncover intents in unlabeled utterances with provided context. Then even for distant target domains, it can generate correct and novel intents in the form of natural language. To evaluate the performance of our method, we introduce a new intent detection benchmark including the Meta-Intent Dataset and three types of evaluation settings. Meta-Intent Dataset, a dataset of intent datasets, consists of 12 intent datasets and covers diverse domains. We conduct comprehensive experiments under three types of settings including any-way any-shot intent detection, episodic intent detection and new intent discovery.

The main contributions of this paper are: (1) We reformulate intent detection into the language generation paradigm and propose a unified framework for intent detection, Gen-PINT, which generates correct intents by following specialized natural language instructions. (2) We present an intent detection benchmark, including the Meta-Intent Dataset which contains 12 datasets from different domains and three types of evaluation settings in low-resource intent detection. (3) We conduct extensive evaluations on various experimental settings. Experimental results show that our method consistently outperforms strong baselines.

2 Related Work

2.1 Intent Detection

Intent detection aims to identify user intentions and is an important part in ubiquitous systems, such as accomplishing a helpful conversation in chatbots (Ouyang et al., 2022; Song et al., 2022), building commercial recommend systems (Qian et al., 2023; Wang et al., 2023a) and collecting information from search engines (Chen et al., 2023). Existing studies (Gong et al., 2021, 2022) have achieved promising performance by training a supervised model on large amounts of labeled examples. However, collecting sufficient labeled samples and re-training

models for new intents are challenging. Recent studies (Krone et al., 2020; Dopierre et al., 2021a; Chen et al., 2022a; Liu et al., 2023) have adopted few-shot learning methods to recognize novel intents. Mi et al. (2022) introduce a comprehensive instruction that exploits pre-trained language models with extra task-specific instructions. Parikh et al. (2023) leverage label description and intent filtering in zero-shot setting, and tune instruction-tuned language models with parameter-efficient fine-tuning in the few-shot setting. Zhang et al. (2023b) propose to directly fine-tune pre-trained language models with a few labeled examples. However, these methods train a new model for each dataset, which is not only costly in deployment but also inevitably hard to scale to new intents from different domains.

2.2 Instruction Tuning

Instruction tuning (Wang et al., 2022; Iyer et al., 2022; Longpre et al., 2023), is becoming a novel paradigm that unlock the generalized knowledge of large language models by leveraging natural language instructions. During inference, it only conditions a language model with domain-specific instructions and several optional demonstrations. Recently, some existing studies have explored question answering (Min et al., 2022; Chen et al., 2022b), information extraction (Wang et al., 2023b; Sainz et al., 2023), and so on (Chung et al., 2022; Muennighoff et al., 2023). However, specialized intent language model with instruction tuning is still under exploration. Some studies (Song et al., 2023; Raedt et al., 2023) conduct intent discovery evaluation on ChatGPT (OpenAI, 2023), while they suffer two limitations, the risk of test data contamination (Li and Flanigan, 2023) and data privacy restrictions.

2.3 Pre-trained Intent Language Models

Recent studies (Wu et al., 2020; Henderson et al., 2020) have explored continual pre-training on discriminative models with dialogue corpora to improve generalization on new intent detection tasks. Zhang et al. (2021) propose IntentBERT, a BERT model tuned on public intent datasets. Sung et al. (2023) introduce a pre-trained intent-aware encoder by using contrastive learning to align embeddings of utterances with their pseudo labels. They focus on tuning discriminative language models while neglecting powerful generative language models with instructions. One of the shining points of generative models is that generated text may provide supervising view for the novel target task. Here, we design detailed instructions and propose a unified generative intent model to generalize beyond the tuning data and recognize new intents with specialized instructions.

3 Approach

We introduce Gen-PINT: a **Gen**erative **P**re-trained **INT**ent detection model via instruction tuning. As illustrated in Figure 2, Gen-PINT is a generative model with instructions. Different from previous methods, Gen-PINT is trained on a collection of general intent detection corpora and is directly evaluated on unseen target datasets without parameter updates. At test phase, leveraging several demonstrated samples and task instructions, it learns to recognize semantic information in utterances and then generates the corresponding intent label from given options as the predicted result.

3.1 Input-Output Schema

Different from typical discriminative classification methods, we reformulate the intent detection into text-to-text format and solve it through in-context learning on generative language models. Figure 2 shows four main properties of the input: task definition, output space, demonstrations and query. Output is the intent label.

Task Definition describes the intent detection task. This section contains task explanation, and target format. Task explanation shows the main topic of the task and target format indicates the definition of generated output. By aligning to the pre-trained tasks, the detailed guidelines lead the language model to understand how to perform intent detection tasks and how to output the expected content, which more properly utilizes the knowledge stored in language models. To ensure the fine-tuned model follows the provided guidelines rather than overfitting and memorizing specific inputs, we paraphrase the task instructions. By doing this, the tuned model is more robust to different task definition inputs and remains better generalization ability. Table 9 lists the used task definition.

Output Space is the label options for each task. The output space adds constraints for output content and guides the model to conduct multiple choice questions. Output space is flexible and varies for each task. It is easy to refine output space by just replacing different label options, rather than

re-train a classification model. In training and inference phase, we shuffle the label option order to obtain more stable performance.

Demonstrations are concatenated query-output pairs. Following previous literature (Brown et al., 2020), we randomly sample k training samples and concatenate them together to build up inputs. The demonstrations are optional and we also conduct experiments without demonstrations. We observe empirically that training with context brings more improvement due to the increased diversity of the training corpora.

Query is the test utterance. For each test sample, it is appended by task definition, output space and demonstrations as an single input to the model.

3.2 Training and Inference

The model is trained on a collection of general intent detection corpora. Formally, given a training sample $(x_i, y_i) \in \mathcal{D}_{train}$, where x_i is the query and y_i is the corresponding label, we first randomly sample k examples, and then combine them to construct demonstrations:

$$D_i = \pi(x_1, y_1) \oplus \pi(x_2, y_2) \oplus ... \oplus \pi(x_k, y_k), \quad (1)$$

where π is the template of transforming the query-label pair and $\oplus$ denotes the concatenation operation. Then we fine-tune a pre-trained language model and force the model to learn from task definition, demonstrations and generate predictions from output space. The in-context fine-tuning objective $\mathcal{L}$ is calculated as:

$$\mathcal{L} = \sum_{(x_i, y_i) \in \mathcal{D}_{train}} -\log p(y_i | I_i, c_i, D_i, x_i, \boldsymbol{\theta}), \quad (2)$$

where I_i and c_i is the corresponding task definition and output space, and $\boldsymbol{\theta}$ is parameters of the model.

At test phase, for each test sample $x_* \in \mathcal{D}_{test}$ from a unseen target task, we directly concatenate its task definition I_*, output space c_*, demonstrations and query to form the input and then obtain generated predictions. Note that there are not any updating of parameters in test phase. The fine-tuned model Gen-PINT can be used to any intent detection tasks by simply specializing the output space and demonstrations.

3.3 Intent Detection Benchmark

We introduce a benchmark of intent detection in low-resource scenarios, consisting of the Meta-Intent Dataset, 12 representative and diverse intent datasets, and three types of evaluation settings. When collecting Meta-Intent Dataset, diversity is carefully considered. We include corpora from various domains, such as the common intent domains, e.g., travel, flight, banking, weather, and alarm etc., and also include some specific domains that have a large gap compared to common domains, like privacy policies, covid-19 and software etc. These diverse corpora from different domains can be used to evaluate model generalization. Table 8 in Appendix A reports detailed dataset statistics. All task instances follow the same text-to-text schema described in Section 3.1. For each dataset split, we follow the original split if it has other we divide the datasets into training/validation/test set with the ratio of 8:1:1.

The benchmark also offers comprehensive and standardized evaluation rules for the performance of models in both zero-shot and few-shot situations for intent detection tasks. Specifically, we conduct evaluations from three perspectives, consisting of any-way any-shot intent detection, episodic intent detection and new intent discovery. In any-way any-shot intent detection, strict zero/few-shot experiments are conducted where the number of unseen classes is larger and there is no validation set. For episodic evaluation, we employ meta-learning strategy to simulate few-shot scenarios, i.e., the popular N-way K-shot setting. In new intent discovery, we test the model ability of discovering new intents. The three types of evaluation make it easier to compare different approaches more comprehensively and will help create stronger and more efficient models.

4 Experimental Setup

4.1 Datasets

We introduce Meta-Intent Dataset to evaluate our proposed method in zero/few-shot settings. We select the most common intent datasets, Clinc (Larson et al., 2019), HWU64 (Liu et al., 2019a), Facebook (Schuster et al., 2019) and MTOP (Li et al., 2021) to make up the training set. The remaining eight datasets, Bank77 (Casanueva et al., 2020), MCID (Arora et al., 2020a), HINT3 (Arora et al., 2020b), SNIPS (Coucke et al., 2018), ATIS (Hemphill et al., 1990), StackOverflow (Xu et al., 2015), PolicyIE (Ahmad et al., 2021) and StackExchange (Braun et al., 2017), are used as test set in zero/few-shot settings. Detailed split and statistics of datasets can be found in Appendix A.

Method	Bank77	MCID	HINT3	SNIPS	ATIS	StackO.	PolicyIE	StackE.	Avg.
IntentBert	43.33	56.44	37.42	80.86	47.82	20.92	37.58	55.00	47.42
IntentBert 1-shot	52.30	64.53	50.27	85.60	62.56	24.60	45.92	71.43	57.15
IntentBert 2-shot	56.52	67.96	56.13	87.14	67.43	27.60	46.21	75.00	60.50
PIE	54.98	63.61	47.89	86.96	78.05	63.65	57.58	83.92	67.08
PIE 1-shot	49.19	54.15	46.40	84.17	32.36	34.31	42.07	73.57	52.03
PIE 2-shot	58.50	66.01	56.52	**91.05**	49.92	45.28	42.80	80.19	61.28
BERT-A	44.91	55.28	46.10	78.28	55.04	65.37	45.57	69.16	57.46
0-shot	22.03	29.34	20.35	47.89	41.12	42.90	58.48	51.19	39.16
ICL 1-shot	49.98	55.98	45.35	81.34	61.50	77.60	64.60	87.97	65.54
PT	53.31	70.89	54.29	85.20	85.26	78.81	62.73	82.50	71.62
PT 1-shot	56.46	76.44	56.92	86.71	83.79	80.98	66.12	88.45	74.48
Gen-PINT (Ours)	54.73	68.42	55.98	84.14	86.81	77.89	71.46	81.79	72.65
Gen-PINT 1-shot (Ours)	**60.29**	**77.19**	**62.68**	86.57	**86.96**	**82.30**	**72.53**	**90.48**	**77.38**

Table 1: The average accuracy results in zero-shot and few-shot settings.

4.2 Baselines

We compare Gen-PINT with a series of baselines under three scenarios, i.e., strict any-way any-shot intent detection, episodic intent detection and new intent discovery. In any-way any-shot intent detection, baselines contains discriminative models and generative language models. For both baselines and our method, we test them under zero-shot and few-shot setting respectively. **(1) IntentBERT**: Zhang et al. (2021) propose to continually train BERT on public intent datasets and then detect novel intents by adding specific classifiers. **(2) PIE**: It is a pre-trained intent-aware encoder through key phrases identification and contrastive learning on public intent datasets (Sung et al., 2023). **(3) BERT-A**: Comi et al. (2023) first train the model for natural language inference and then perform zero-shot intent recognition. **(4) 0-shot**: We run zero-shot inference on raw pre-trained generative language models (Mi et al., 2022). **(5) ICL 1-shot**: We follow the in-context learning by conditioning on several training samples (Brown et al., 2020). **(6) Prompt Tuning (PT)**: We train the language model on the same training tasks without demonstrations in context.

In episodic intent detection, we compare with three strong baselines: **(1) PN** (Snell et al., 2017) averages the corresponding support samples to calculate the prototype for each class, and uses the negative Euclidean distance between queries and prototypes to perform classification. **(2) CNet** (Chen et al., 2022a) aims to generate better sample representations by designing a task-level and an instance-level unsupervised contrastive losses. **(3) DE** (Liu et al., 2023) proposes to estimate the novel class distribution by leveraging support samples and the nearest unlabeled query samples.

In new intent discovery, we compare with recent promising baselines: **(1) MTP** (Zhang et al., 2022) leverages the multi-task pre-training strategy to learn semantic utterance representations. MTP-CLNN is a variant that integrates contrastive loss. **(2) IDAS** (Raedt et al., 2023) formulates the intent discovery task into abstractive summarization to retain core elements of user utterances.

4.3 Evaluation Metrics

In any-way any-shot intent detection, we use accuracy and macro-F1 as evaluation metric to evaluate balanced and imbalanced classification tasks. For both zero-shot and few-shot settings, we report results on all test samples in test set as described in Table 8. In few-shot setting, i.e., k-shot, we random sample k samples from training set for each class to construct training samples. We repeat this procedure 5 times using five different seeds and report average results. And we follow strict few-shot setting which means that there is no validation set to be used. In episodic evaluation, we follow Liu et al. (2023) and report accuracy in 5-way 1/5-shot settings. For new intent discovery, following Zhang et al. (2022), we report the Normalized Mutual Information (NMI) and Cluster Accuracy (ACC) according to the Hungarian algorithm (Kuhn, 2010).

Method	Bank77	MCID	HINT3	SNIPS	ATIS	StackO.	PolicyIE	StackE.	Avg.
IntentBert	39.74	53.05	34.40	79.74	29.01	18.62	36.99	47.49	42.38
IntentBert 1-shot	51.18	63.54	47.83	84.82	47.60	23.13	43.86	59.13	52.64
IntentBert 2-shot	54.26	66.56	53.32	86.84	52.33	26.98	45.15	65.32	56.35
PIE	53.06	63.58	45.60	85.62	33.81	62.91	53.45	68.76	58.35
PIE 1-shot	48.15	52.62	47.46	82.22	36.33	32.50	39.79	65.23	50.54
PIE 2-shot	57.95	65.53	58.12	**90.91**	45.95	45.31	42.43	70.48	59.59
BERT-A	35.48	48.59	33.04	78.17	35.21	64.28	33.72	41.15	46.21
0-shot	29.61	41.09	25.96	59.45	27.54	49.95	56.45	59.59	43.71
ICL 1-shot	52.42	61.70	48.65	81.69	50.85	78.48	63.04	83.54	65.05
PT	55.15	73.43	53.81	84.99	46.19	79.06	61.02	80.36	66.75
PT 1-shot	58.10	78.24	58.63	86.01	59.01	82.03	65.03	82.28	71.17
Gen-PINT (Ours)	54.68	70.21	53.35	83.57	49.39	78.54	69.82	80.99	67.57
Gen-PINT 1-shot (Ours)	**60.55**	**78.47**	**62.31**	85.79	**67.06**	**82.46**	**71.54**	**87.74**	**74.49**

Table 2: The average macro-F1 scores in zero-shot and few-shot settings.

Method	Bank77		MCID		HINT3		StackOverflow	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
PN (Snell et al., 2017)	86.47	94.08	67.29	83.52	71.01	85.57	64.57	81.09
CNet (Chen et al., 2022a)	91.18	**96.40**	69.14	84.43	70.19	86.55	66.03	81.81
DE (Liu et al., 2023)	90.51	95.76	72.23	85.54	76.31	**87.37**	70.71	83.26
Gen-PINT (Ours)	**92.72**	94.85	**81.49**	**86.64**	**81.77**	86.88	**83.51**	**87.17**

Table 3: The 5-way 1-shot and 5-shot average accuracy on the Bank77, MCID, HINT3 and StackOverflow datasets.

4.4 Parameter Settings

We use FlanT5-Large (Chung et al., 2022) as our backbone model. In training phase, we append 8 samples for each sample as context. We optimize the model parameters using AdamW (Loshchilov and Hutter, 2019) with learning rate of 1×10^{-5}. For fair comparison, we leverage the same training set, as described in Table 8, to re-train baselines. Experiments are conducted on RTX 3090 with 24G GPU memory.

5 Experimental Results

5.1 Any-Way Any-Shot Intent Detection

Tables 1 and 2 report accuracy and macro-F1 scores on eight unseen intent datasets both in zero-shot and few-shot settings. We can make the following observations. For all methods, increasing the number of labeled samples in target domains brings further improvements, especially for in-context learning of raw LLMs, where labeled samples give direct signals to guide models to generate labels. However, the improvements of discriminative models are relative small because they rely on the generalized features of utterances obtained from intent encoder while the limited labeled utterance features may cause biased class decision boundaries. Generative methods that model the semantic relationship between utterances and intent labels achieve better improvement.

Gen-PINT consistently outperforms a series of strong baselines on most datasets and achieves the best results on averaged scores. Specifically, in zero-shot setting, Gen-PINT achieves competitive performance comparing with discriminative and generative models. In few-shot setting, Gen-PINT performs better by conditioning on few target labeled samples, where brings about 11.84% and 9.44% on improvements accuracy and F1 respectively than raw LLMs. Comparing with discriminative mothods, gains are particularly significant in StackOverflow, PolicyIE and MCID datasets, where the test domains are very different from the training corpus. This demonstrates that Gen-PINT is able to infer the semantics of new intents even when the test corpus is from novel domains.

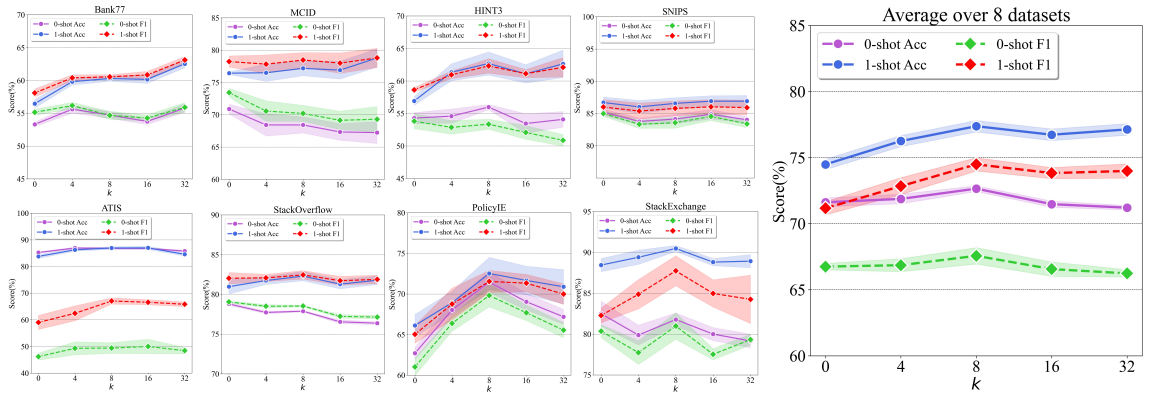

Figure 3: Ablation studies of the number of appended training samples. $k = 0$ is equivalent to the prompt tuning method.

5.2 Episodic Intent Detection

Recently, many papers (Liu et al., 2019b; Chen et al., 2022a; Liu et al., 2023) detect new intents in the few-shot setting via meta-learning paradigm, i.e., episodic strategy. Specifically, these methods split each dataset into training/validation/test set respectively and there is no class overlap among these splits. Then they meta-train models on the training set and select the best one with validation set to evaluate test set, achieving promising performance due to the same distribution of training and test set. We compare Gen-PINT with these models on the 5-way 1/5-shot settings and more detailed experimental settings can be found at Appendix B. Note that baselines train a specific model for each dataset using the training data from the same domain with test data, while our method directly evaluate on the test samples without seeing any training data. Table 3 presents average accuracy from five different dataset splits on Bank77, MCID, HINT3 and StackOverflow. From Table 3, we can observe that Gen-PINT without assessing to any samples in these target datasets, still outperforms the strong baselines trained on each dataset, which further demonstrates the superiority and generalization of our proposed method.

5.3 New Intent Discovery

New intent discovery aims to identify novel unknown intents from a set of unlabeled utterances without any predefined knowledge, which is challenging and practical. Current studies (Zhang et al., 2022; Kumar et al., 2022) usually leverage two-step methods. First, they learn semantic representations of utterances and then cluster them to uncover novel intents. However, these methods just align a serial number to each cluster and they can not generate a meaningful intent name for each cluster, which lacks interpretability and practicality.

Model	MCID		StackOverflow	
	NMI	ACC	NMI	ACC
MTP	72.40	68.94	63.85	66.18
MTP-CLNN	79.95	79.14	78.71	81.43
IDAS	-	-	81.26	83.82
Gen-PINT	77.10	77.08	78.83	83.08

Table 4: Performance of intent discovery on MCID and StackOverflow datasets.

On the contrary, benefiting from the nature of generative models, our proposed method can directly generate novel descriptive intents for utterances. IDAS (Raedt et al., 2023) is the latest work that formulates the intent discovery task into abstractive summarization to retain core elements. It prompts `text-davinci-003` model (Ouyang et al., 2022) to generate intent labels. To evaluate the ability to discover unknown intents, we conduct experiments on MCID and StackOverflow datasets in unsupervised setting. Specifically, we replace the GPT-3 in IDAS with xl-sized Gen-PINT. We conduct comprehensive comparison with both cluster-based and generation-based methods. From Table 4, we can observe that our model achieves competitive results. Note that in StackOverflow dataset, our model performs better than MTP-CLNN and even achieves similar performance with IDAS that leverages GPT-3 (175B). Experimental details can be found in Appendix D.

5.4 Ablation Studies

Instruction regularization We conduct ablation studies to examine the contribution of several com-

Method	Bank77	MCID	HINT3	SNIPS	ATIS	StackO.	PolicyIE	StackE.	Avg.
				Accuracy					
FlanT5-XL	43.74	58.28	40.55	84.00	76.24	46.28	73.54	82.50	63.14
FlanT5-XXL	50.24	67.73	50.12	82.48	58.34	73.66	**74.32**	**83.92**	67.60
Gen-PINT	**54.73**	**68.42**	**55.98**	**84.14**	**86.81**	**77.89**	71.46	81.79	**72.65**
				Macro-F1					
FlanT5-XL	46.30	60.31	41.13	85.79	47.65	52.26	66.18	78.89	59.81
FlanT5-XXL	54.54	69.27	47.30	**86.61**	48.32	78.26	67.94	**81.09**	66.67
Gen-PINT	**54.68**	**70.21**	**53.35**	83.57	**49.39**	**78.54**	**69.82**	80.99	**67.57**

Table 5: Experimental results on raw FlanT5-XL, FlanT5-XXL and large-sized Gen-PINT.

Model	Zero-Shot		Few-Shot	
	Acc	F1	Acc	F1
Gen-PINT	**72.65**	**67.57**	**77.38**	**74.49**
w/o Shuffle	72.04	67.13	76.37	73.53
w/o Paraphrase	71.40	66.51	75.62	72.83
w/o Demons.	71.62	66.75	74.48	71.17

Table 6: Ablation results. The average accuracy and F1 scores of all datasets are reported.

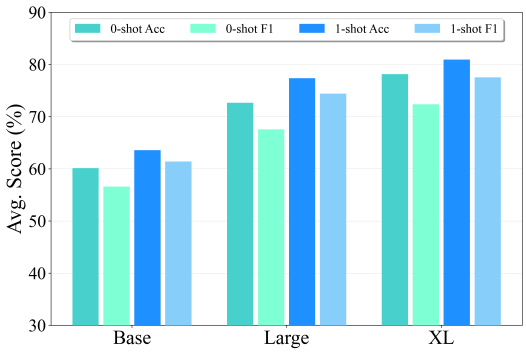

Figure 4: Performance of Gen-PINT with different size. Average accuracy and F1 scores are reported.

ponents in instructions. We use *w/o Shuffle* when we remove label random shuffle regularization, and *w/o Paraphrase* when we remove task definition paraphrase. *w/o Demons.* means that we remove appended demonstrations, which is the prompt tuning baseline. As presented in Table 6, label shuffle and task instruction paraphrase boosts better performance in zero-shot and few-shot evaluation. Also, demonstrations brings more improvements, where all three strategies help to construct diverse training samples and demonstrations may give more clear signals to guide the model.

Number of training samples In the training phase, we conduct experiments with varying number of appended samples, i.e., $k \in \{0, 4, 8, 16, 32\}$. In the test phase, we conduct both zero-shot (0-shot) and few-shot (1-shot) evaluation. Figure 3 shows accuracy and F1 scores on all datasets. We can observe that increasing the number of appended samples brings promising improvements for most datasets especially in few-shot setting. Also, we can see that increasing k may bring light negative influence to zero-shot inference for some datasets due to the gap between training and test input format. Considering both situations, we find that the performance reaches to a saturate value when k equals 8.

Comparison to stronger LLMs In Table 5, we compare our model with FlanT5-XL and FlanT5-XXL raw baselines which consist of 3B and 11B parameters respectively. We can observe that scaling up language model size yields superior performance, while it inevitably brings significant increased computational costs and resources. Comparing with these baselines, our large-sized Gen-PINT achieves competitive or better performance and is more efficient on computing and hardware.

Model scaling Recent studies has reported that more parameters yields better generalization (Brown et al., 2020). We scale Gen-PINT with Base (250M), Large (780M) and XL (3B) versions. Figure 4 reports the average results on all datasets and more details can be found at Figure 5 in Appendix C. We can see that for average performance, more parameters brings superior improvements while this is accompanied by increasing computational costs and deployed resources. Also, we find that scaling model size brings more improvements for zero-shot evaluation than the few-shot setting. In general, large-sized Gen-PINT has exhibited strong generalization ability and simultaneously has relative low computational load in zero-shot

and few-shot intent detection.

6 Conclusion

In this paper, we reformulate intent detection into a generation task and introduce Gen-PINT, a new generative pre-trained model for low-resource intent detection. By conditioning on several training samples and instructions, Gen-PINT, as a unified and end-to-end model, is able to recover semantic of inputs and generate intents for utterances from different domains. We further build a new benchmark for intent detection with a diverse collections of datasets and conduct comprehensive evaluation under three settings. Gen-PINT outperforms a series of strong baselines including discriminative methods, in-context learning and prompt tuning. Also, Gen-PINT performs better than recent dominated meta-learning methods and it can generate new descriptive intents that are beyond the reach of discriminative models.

Limitations

We propose a well-generalized generative pre-trained intent model (Gen-PINT) for low-resource intent detection, and introduce a benchmark to conduct comprehensive evaluations. For a new intent task, Gen-PINT can recognize new emerging intents from different domains with easily crafted instructions rather than any further fine-tuning. Instructions used in this paper consist of task definition, output space, demonstrations and query, which have achieved promising performance. However, we do not discuss the target domain knowledge due to space limits, like the description and definition of intent labels. Integrating domain knowledge can provide more information to excite model performance and we leave this for future exploration.

Acknowledgements

This research is supported by the National Science and Technology Major Project (No. 2021ZD0111202).

References

Wasi Uddin Ahmad, Jianfeng Chi, Tu Le, Thomas Norton, Yuan Tian, and Kai-Wei Chang. 2021. Intent classification and slot filling for privacy policies. In *ACL*, pages 4402–4417.

Abhinav Arora, Akshat Shrivastava, Mrinal Mohit, Lorena Sainz-Maza Lecanda, and Ahmed Aly. 2020a. Cross-lingual transfer learning for intent detection of covid-19 utterances.

Gaurav Arora, Chirag Jain, Manas Chaturvedi, and Krupal Modi. 2020b. HINT3: raising the bar for intent detection in the wild. In *The First Workshop on Insights from Negative Results in NLP*, pages 100–105.

Daniel Braun, Adrian Hernandez-Mendez, Florian Matthes, and Manfred Langen. 2017. Evaluating natural language understanding services for conversational question answering systems. In *The 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185.

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. In *NeurIPS*.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *The 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45.

Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. 2022a. Contrastnet: A contrastive learning framework for few-shot text classification. In *AAAI*, pages 10492–10500.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022b. Meta-learning via language model in-context tuning. In *ACL*, pages 719–730.

Zhiyu Chen, Jason Ingyu Choi, Besnik Fetahu, Oleg Rokhlenko, and Shervin Malmasi. 2023. Generate-then-retrieve: Intent-aware FAQ retrieval in product search. In *ACL*, pages 763–771.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Daniele Comi, Dimitrios Christofidellis, Pier Francesco Piazza, and Matteo Manica. 2023. Zero-shot-bert-adapters: a zero-shot pipeline for unknown intent detection. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 650–663.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*.

Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021a. Protaugment: Intent detection meta-learning through unsupervised diverse paraphrasing. In *ACL/IJCNLP*, pages 2454–2466.

Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021b. Protaugment: Unsupervised diverse short-texts paraphrasing for intent detection meta-learning. In *ACL*, pages 2454–2466.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910.

Yantao Gong, Cao Liu, Fan Yang, Xunliang Cai, Guanglu Wan, Jiansong Chen, Weipeng Zhang, and Houfeng Wang. 2022. Confidence calibration for intent detection via hyperspherical space and rebalanced accuracy-uncertainty loss. In *AAAI*, pages 10690–10698.

Yantao Gong, Cao Liu, Jiazhen Yuan, Fan Yang, Xunliang Cai, Guanglu Wan, Jiansong Chen, Ruiyao Niu, and Houfeng Wang. 2021. Density-based dynamic curriculum learning for intent detection. In *CIKM*, pages 3034–3037.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, Chih-Li Huo, Tsung-Chieh Chen, Keng-Wei Hsu, and Yun-Nung Chen. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *NAACL*, pages 753–757.

Arshit Gupta, John Hewitt, and Katrin Kirchhoff. 2019. Simple, fast, accurate intent classification and slot labeling for goal-oriented dialogue systems. In *The 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial*, pages 46–55.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley*.

Matthew Henderson, Iñigo Casanueva, Nikola Mrksic, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulic. 2020. Convert: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 2161–2174.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2022. OPT-IML: scaling language model instruction meta learning through the lens of generalization. *CoRR*, abs/2212.12017.

Jason Krone, Yi Zhang, and Mona Diab. 2020. Learning to classify intents and slot labels given a handful of examples. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 96–108.

Harold W. Kuhn. 2010. The hungarian method for the assignment problem. In *50 Years of Integer Programming 1958-2008*, pages 29–47.

Rajat Kumar, Mayur Patidar, Vaibhav Varshney, Lovekesh Vig, and Gautam Shroff. 2022. Intent detection and discovery from user logs via deep semi-supervised contrastive clustering. In *NAACL*, pages 1836–1853.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *EMNLP*, pages 1311–1316.

Changmao Li and Jeffrey Flanigan. 2023. Task contamination: Language models may not be few-shot anymore. *arXiv preprint arXiv:2312.16337*.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *EACL*, pages 2950–2962.

Han Liu, Feng Zhang, Xiaotong Zhang, Siyang Zhao, Fenglong Ma, Xiao-Ming Wu, Hongyang Chen, Hong Yu, and Xianchao Zhang. 2023. Boosting few-shot text classification via distribution estimation. In *AAAI*, pages 13219–13227.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019a. Benchmarking natural language understanding services for building conversational agents. In *IWSDS*, volume 714, pages 165–183.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. 2019b. Learning to propagate labels: Transductive propagation network for few-shot learning. In *ICLR*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *ICML*, pages 22631–22648.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Fei Mi, Yasheng Wang, and Yitong Li. 2022. CINS: comprehensive instruction for few-shot learning in task-oriented dialog systems. In *AAAI*, pages 11076–11084.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Metaicl: Learning to learn in context. In *NAACL*, pages 2791–2809.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *ACL*, pages 15991–16111.

OpenAI. 2023. `https://chat.openai.com.chat`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Soham Parikh, Mitul Tiwari, Prashil Tumbade, and Quaizar Vohra. 2023. Exploring zero and few-shot techniques for intent classification. In *Annual Meeting of the Association for Computational Linguistics: Industry Track, ACL*, pages 744–751.

Tieyun Qian, Yile Liang, Qing Li, Xuan Ma, Ke Sun, and Zhiyong Peng. 2023. Intent disentanglement and feature self-supervision for novel recommendation. *IEEE Trans. Knowl. Data Eng.*, 35(10):9864–9877.

Maarten De Raedt, Fréderic Godin, Thomas Demeester, and Chris Develder. 2023. IDAS: intent discovery with abstractive summarization. *CoRR*, abs/2305.19783.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction. *CoRR*.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *NAACL*, pages 3795–3805.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *NeurIPS*, pages 4077–4087.

Mengxiao Song, Bowen Yu, Quangang Li, Yubin Wang, Tingwen Liu, and Hongbo Xu. 2022. Enhancing joint multiple intent detection and slot filling with global intent-slot co-occurrence. In *EMNLP*, pages 7967–7977.

Xiaoshuai Song, Keqing He, Pei Wang, Guanting Dong, Yutao Mou, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. 2023. Large language models meet open-world intent discovery and recognition: An evaluation of chatgpt. In *EMNLP*, pages 10291–10304.

Mujeen Sung, James Gung, Elman Mansimov, Nikolaos Pappas, Raphael Shu, Salvatore Romeo, Yi Zhang, and Vittorio Castelli. 2023. Pre-training intent-aware encoders for zero- and few-shot intent classification. *CoRR*, abs/2305.14827.

Jixuan Wang, Kai Wei, Martin Radfar, Weiwei Zhang, and Clement Chung. 2021. Encoding syntactic knowledge in transformer encoder for intent detection and slot filling. In *AAAI*, pages 13943–13951.

Xiangmeng Wang, Qian Li, Dianer Yu, Peng Cui, Zhichao Wang, and Guandong Xu. 2023a. Causal disentanglement for semantic-aware intent learning in recommendation. *IEEE Trans. Knowl. Data Eng.*, 35(10):9836–9849.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023b. Instructuie: Multi-task instruction tuning for unified information extraction. *CoRR*.

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 *EMNLP*, pages 5085–5109.

Chien-Sheng Wu, Steven C. H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: pre-trained natural language understanding for task-oriented dialogue. In *EMNLP*, pages 917–929.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. In *The 1st Workshop on Vector Space Modeling for Natural Language Processing, VS@NAACL-HLT*, pages 62–69.

Feng Zhang, Wei Chen, Fei Ding, and Tengjiao Wang. 2023a. Dual class knowledge propagation network for multi-label few-shot intent detection. In *ACL*, pages 8605–8618.

Haode Zhang, Haowen Liang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Y. S. Lam. 2023b. Revisit few-shot intent classification with plms: Direct fine-tuning vs. continual pre-training. In *Findings of the Association for Computational Linguistics: ACL*, pages 11105–11121.

Haode Zhang, Yuwei Zhang, Li-Ming Zhan, Jiaxin Chen, Guangyuan Shi, Xiao-Ming Wu, and Albert

Y. S. Lam. 2021. Effectiveness of pre-training for few-shot intent classification. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 1114–1120.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Y. S. Lam. 2022. New intent discovery with pre-training and contrastive learning. In *ACL*, pages 256–269.

A Dataset Details

We elaborate the details of Meta-Intent Dataset, a dataset consisting of 12 intent datasets.

- **Clinc** (Larson et al., 2019) is composed of 22,500 utterances covering 150 intents over 10 general domains. The distribution of utterances is uniform across each intent.
- **HWU64** (Liu et al., 2019a) covers 25,716 original examples. We clean the examples and remain 25,606 examples.
- **MTOP** (Li et al., 2021) is a multilingual task-oriented semantic parsing dataset consisting of 100k annotated utterances in 6 languages across 11 domains. Here we use 22,288 English utterances.
- **Facebook** (Schuster et al., 2019) contains utterances in English from the weather, alarm, and reminder domains. For the training set, we remove duplicated samples.
- **Bank77** (Casanueva et al., 2020) is a single-domain intent dataset consisting of 13,083 annotated customer queries over 77 fine-grained intents, where there may have partially overlap among some intent categories.
- **MCID**(Arora et al., 2020a) consists of 1,745 queries which covers 16 covid-19 specific intents.
- **HINT3** (Arora et al., 2020b) contains 2,011 samples across 51 intents from three domains.
- **SNIPS** (Coucke et al., 2018) is composed of 14,484 utterances collected from Snips personal voice assistant.
- **ATIS** (Hemphill et al., 1990) (Airline Travel Information Systems) is a dataset consisting of 5,871 samples about making flight reservations over 21 intents.
- **StackOverflow** (Xu et al., 2015) contains 20,000 samples collected from the stackoverflow platform about computer technology.
- **PolicyIE** (Ahmad et al., 2021) is a challenging dataset about privacy policies of websites and mobile applications. Here we use 1,989 examples over four privacy policies.

Dataset	#samples	#train/val/test (total) classes
Bank77	13083	25 / 25 / 27 (77)
MCID	1745	5 / 5 / 6 (16)
HINT3	2011	15 / 15 / 21 (51)
StackO.	20000	6 / 6 / 8 (20)

Table 7: Dataset statistics under the episodic setting.

- **StackExchange** (Braun et al., 2017) is composed of 251 samples, where 89 from *ask ubuntu* and 162 from *Web Applications* platforms.

Table 8 reports dataset statistics, where # denotes the number of samples in the specific subset. From Table 8, we can see that our benchmark covers diverse domains, including travel, bank, transport, software, privacy policies and so on, which provides a more comprehensive comparison of various methods to evaluate their effectiveness and robustness. For the dataset with official split, we directly follow it. And for other datasets, we split them into training set, validation set and test set respectively based on the ratio of 8:1:1.

B Episodic Evaluation Details

Meta-learning methods have attracted increasing attentions predominantly to solve few-shot learning problems and they usually adapt the episodic strategy. Each episode usually is 5-way 1-shot or 5-way 5-shot, which means that there are 5 new intents and each intent contains 1 or 5 labeled samples. The baseline models firstly meta-train on the training set and then meta-test on the test set, where training classes and test classes do not have overlaps. We compare Gen-PINT with three strong baselines on Bank77, MCID, HINT3 and StackOverflow in the episodic setting.

Following Liu et al. (2023), we use accuracy to evaluate the performance. All reported results are averaged from five different dataset splits. Table 7 shows detailed dataset splits. At test phase, we present the average accuracy on 600 episodes samples from test set, where each class has five query samples to predict in each episode. Note that our method does not assess any training samples, which is different from other baselines. From Table 3, we can observe that our model achieves better performance in most cases without any domain-specific training samples.

Dataset	Classes	#Train	#Val	#Test	Domains
Clinc (Larson et al., 2019)	150	13500	4500	4500	travel, banking, utility, work, credit cards, meta, home, auto commute, small talk, kitchen dining
HWU64 (Liu et al., 2019a)	68	15318	5144	5144	general, lists, play, takeaway, alarm, social, iot, news, email, cooking, calendar, audio, qa, datetime, weather, transport, music, recommendation
MTOP (Li et al., 2021)	113	15667	2235	4386	alarm, calling, weather, timer, music, news, people, recipes, reminder, messaging, event
Facebook (Schuster et al., 2019)	12	23788	4181	8621	reminder, alarm, weather
Bank77 (Casanueva et al., 2020)	77	7849	2617	2617	bank
MCID (Arora et al., 2020a)	16	1047	349	349	medical
HINT3 (Arora et al., 2020b)	51	1205	403	403	mattress products retail, fitness supplements retail, online gaming
SNIPS (Coucke et al., 2018)	7	13084	700	700	personal voice assistant
ATIS (Hemphill et al., 1990)	21	4478	500	893	flight information
StackOverflow (Xu et al., 2015)	20	12000	4000	4000	software
PolicyIE (Ahmad et al., 2021)	4	1304	329	356	privacy policies
StackExchange (Braun et al., 2017)	12	83	-	168	ask ubuntu, web applications

Table 8: Detailed statistics of the Meta-Intent Dataset.

Instructions
Given options, please tell me the intent of the query.\nOptions:\n
You will be provided with utterance intent detection queries. Each utterance has one intent. You need to recognize the utterance intent. Classify each query into an intent category.\nIntent categories:\n
Predict the intent of the input query. Intent is the main topic or purpose of a query.\nOptions:\n
You will receive queries for detecting the intent of an utterance. Each utterance has a single intent, and your task is to identify the intent category for each query.\nOptions:\n
Predict the intent of the input query. Intent is the main topic or purpose of a query.\nIntent categories:\n
Predict the intent of the input query. Intent is the main topic or purpose of a query.\nYou need to select the most suitable intent from:\n

Table 9: Different task definitions used in experiments.

C Scale Model Size

We scale Gen-PINT to Base (250M), Large (780M) and XL (3B) sized models. Figure 5 shows the accuracy and macro-F1 performance on 8 datasets under the zero-shot (0-shot) and few-shot (1-shot) settings. We can make the following observations. Firstly, from the model size perspective, the improvements from base to large version is more significant than the improvements obtained from large to xl version. Taking both deploy efficiency and performance improvements into consideration, large-sized Gen-PINT has achieved a good balanced point. Secondly, the enhancement brought by increasing model size is more notable on zero-shot scenario than few-shot setting. In other words, conditioning a few samples can further reduce the performance gap caused by model size.

During the experiments, we find that format mismatch is an important cause of generation errors in the base-sized model, such as "top up by card" for "topping up by card" intent. Selecting the most similar intent with the generated answer significantly enhance performance. By using SimCSE (Gao et al., 2021), the average F1 of our base-sized model is 61.51/66.48 (Gen-PINT/Gen-PINT 1-shot). We can observe that Gen-PINT continues to outperform strong baselines. Our focus lies more on its generalization and universal application ability, with post-processing methods left for future exploration.

D New Intent Discovery

New intent discovery is a useful and challenging method in practical open domain scenario, where it is hard for experts to pre-define all intent classes especially in the fast-emerging new domains. Most existing studies (Zhang et al., 2022; Kumar et al., 2022) mainly focus on cluster methods to uncover new intents, while they cannot produce descriptive intent labels. In contrast, generative models are able to produce novel intent names. IDAS (Raedt et al., 2023) is proposed to formulate the intent discovery task into abstractive summarization to retain core elements. To evaluate the ability to discover unknown intents, we replace the GPT-3 in IDAS with xl-sized Gen-PINT.

During the process of prototypical label generation, we change the direct format in IDAS and construct set-level similar query context to generate the common intent behind them. As shown in Table 10, the top instructions are used in the direct intent generation, called instance-level instructions and the down instructions are used in the similar context called set-level instructions. Specifically, we choose top-k sentences that are closest to the prototypical center as candidates and then prompt our model to discover the same latent intent behind them. This is inspired by the fact that humans struggle to recognize the core intent from a single case, yet it is natural for people to summarize the common intent from multiple similar examples.

Following Zhang et al. (2022), we report the Normalized Mutual Information (NMI) and Cluster Accuracy (ACC) according to the Hungarian algorithm (Kuhn, 2010). From Table 4, we can observe that our model achieves competitive results. Note that in StackOverflow dataset, our model performs better than MTP-CLNN and even achieves similar performance with IDAS that leverages GPT-3 (175B). Table 11 shows several examples in the MCID dataset. We can observe that Gen-PINT generates abstractive intents. Also, we believe that there is still ample room to design different methods to better leverage our proposed model, which provides a promising starting point for further exploration.

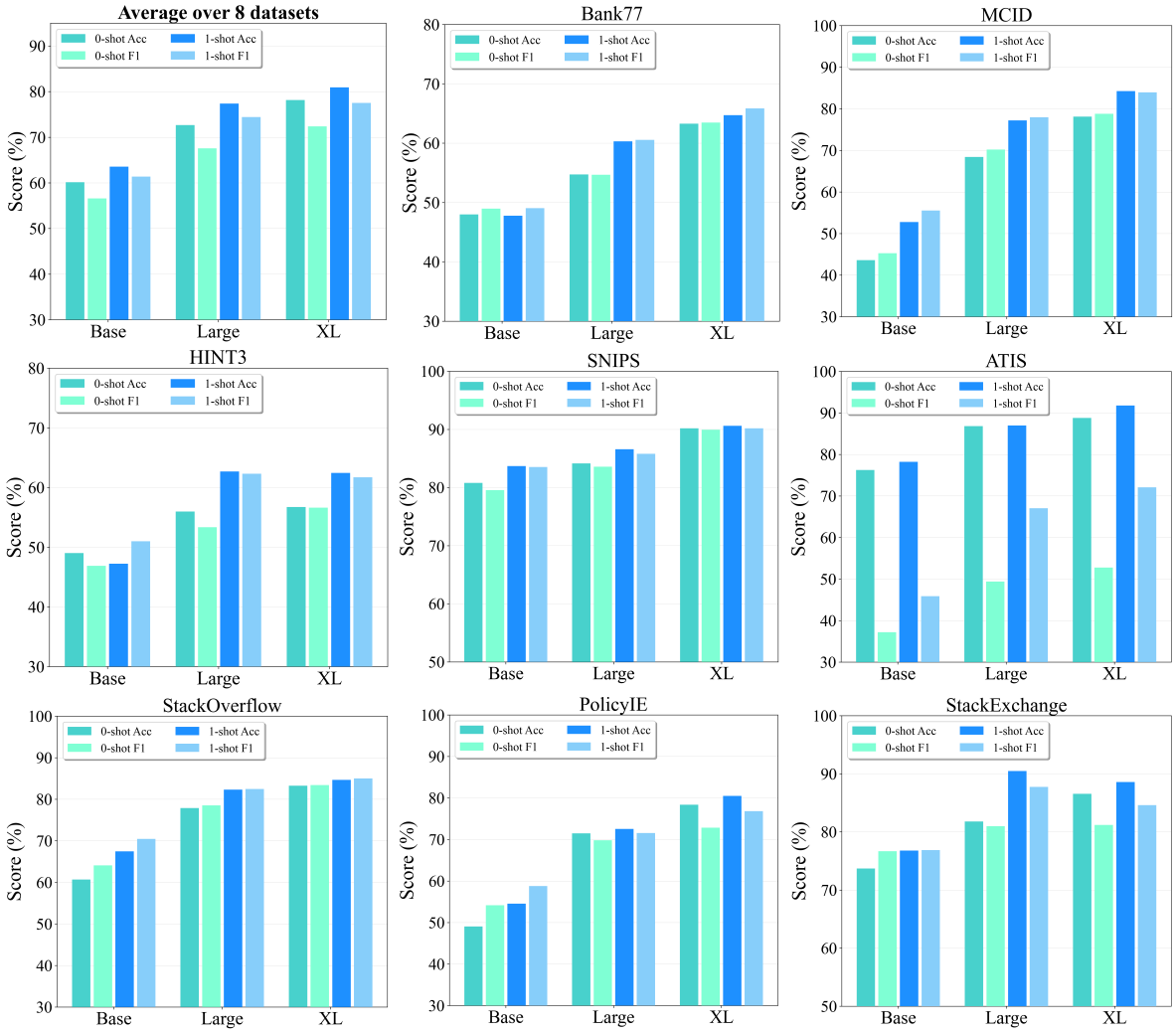

Figure 5: Performance of models with different size on all datasets.

Intent is the main topic or purpose of a query. Please summarize and then tell me the intent of it.\n Sentence: {user_message}\nIntent:
The intent refers to the primary subject or objective of a question. Please explain the intent behind it.\n Sentence: {user_message}\nIntent:
Intent is the main topic or purpose of a query. The following sentences have the same intent. Please summarize and then tell me the intent of them.\n Sentence1: {candidate1}\nSentence2: {candidate2}\nSentence3: {user_message}\nIntent:
The intent refers to the primary subject or objective of a question. Both of the subsequent sentences share the same intent. Please explain the intent behind them.\n Sentence1: {candidate1}\nSentence2: {candidate2}\nSentence3: {user_message}\nIntent:

Table 10: Instructions used in intent discovery.

Intent Discovery Examples
Query: tell me about the different symptoms of corona virus, what symptoms are most common, list all of the corona virus symptoms. **Label:** what_are_symptoms **Generation:** what symptoms are there
Query: what is the latest news on the situation in india, latest news from new york, are there any updates in texas. **Label:** news_and_press **Generation:** latest news
Query: popular coronavirus myths please, is it a myth that coronavirus dies in hot weather, what are some myths floating around about covid. **Label:** myths **Generation:** myths about covid
Query: can i hear updated info on the corona virus, what is there to know about the corona virus, i need more information about covid 19. **Label:** what_is_corona **Generation:** covid 19
Query: can my dog get infected with covid, is my dog immune to covid 19, can i take my pets ivermectin to prevent or treat my covid 19 infection. **Label:** can_i_get_from feces_animal_pets **Generation:** can my dog get infected with covid

Table 11: Intent discovery examples in the MCID dataset.