Exploring Zero and Few-shot Techniques for Intent Classification

Soham Parikh, **Prashil Tumbade**, **Quaizar Vohra**, and **Mitul Tiwari**

ServiceNow, Inc.

{soham.parikh,prashil.tumbade,quaizar.vohra,mitul.tiwari}@servicenow.com

Abstract

Conversational NLU providers often need to scale to thousands of intent-classification models where new customers often face the cold-start problem. Scaling to so many customers puts a constraint on storage space as well. In this paper, we explore four different zero and few-shot intent classification approaches with this low-resource constraint: 1) domain adaptation, 2) data augmentation, 3) zero-shot intent classification using descriptions large language models (LLMs), and 4) parameter-efficient fine-tuning of instruction-finetuned language models. Our results show that all these approaches are effective to different degrees in low-resource settings. Parameter-efficient fine-tuning using T-few recipe (Liu et al., 2022) on Flan-T5 (Chung et al., 2022) yields the best performance even with just one sample per intent. We also show that the zero-shot method of prompting LLMs using intent descriptions is also very competitive.

1 Introduction

Intent classification is the primary natural language understanding task for a virtual agent or a chatbot. Providing intent-utterances for training intent classification models is a laborious process. In this paper, we address this problem by exploring zero and few-shot intent identification using Large Language Models (LLMs) as well as instruction finetuned models. Zero-shot and few-shot intent prediction completely remove or substantially reduce the work to provide intent-utterances, respectively. We demonstrate that the following four approaches work well in practice for zero/few-shot intent classification.

- **Domain adaptation** We use a sentence encoder that is pre-trained with our domain knowledge and show that it performs well in a few-shot setting compared to off-the-shelf sentence encoders.
- **Data Augmentation** By supplementing human-curated training data with LLM-generated data to improve training data.
- **Zero-shot intent classification** High capacity LLMs can be prompted creatively with intent descriptions to do zero-shot classification.
- **Parameter-efficient fine-tuning (PEFT)** Finetuning a small number of parameters added to instruction finetuned LMs using only a few examples

Here is the outline of the rest of the paper. In Section 2 we describe the related work. In Section 3 we detail the datasets used. In Section 4 we describe the four approaches covered in this work for zero/few-shot intent classification. Finally, we conclude with observations in Sections 5 and 6.

2 Related Work

Recent work has successfully used domain adaptation and contrastive learning for few-shot intent classification. One approach is to use embeddings from a BERT model (Devlin et al., 2019) pretrained on domain data to search for utterances belonging to new intents in the domain (Yu et al., 2021). In a similar vein, (Zhang et al., 2021) finetune a BERT model on few-shot data using contrastive learning which learns to discriminate between semantically similar sentences. Our work on domain adaptation differs from these mainly due to our setting which involves serving thousands of customers. For legal reasons, we cannot co-mingle data from these customers to pre-train a single model. Instead, we pre-train a sentence encoder based on an intent taxonomy and out-of-the-box intents, which consist of human generated synthetic data. In this setting, we can only train very lightweight models for each customer, e.g. a dense layer on top of a pre-trained sentence encoder.

Data Augmentation is another widely used technique to solve the problem of data scarcity. Recent

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 744–751
July 10-12, 2023 ©2023 Association for Computational Linguistics

Dataset	Intents	Train Size	Test Size	OOS Samples in Test
MASSIVE	60	11514	2974	No
OOTB-dataset*	27	1363	3099	No
Benchmark01*	9	270	300	Yes
Benchmark02*	13	390	420	Yes
Benchmark03*	31	930	960	Yes

Table 1: Statistics for intent classification datasets used in this paper. Datasets marked with an asterisk (*) are private, internal benchmarking datasets. Train and Test Sizes correspond to the number of utterances in the respective splits. OOS samples in test set indicates whether there are any out-of-scope samples in the test set.

work on data augmentation has focused on using multiple methods to improve model performance (Chen and Yin, 2022). LLMs like GPT-3 (Brown et al., 2020) can be prompted to generate labeled training data for intent classification (Sahu et al., 2022). The quality of generated training data using LLMs is highly dependent on the prompts. In this work, we show various prompt-based approaches that generate diverse data for training and boost the performance of intent classifiers.

As the usage of conversational agents grows, it is important for them to generalize to new intents. Recent work has focused on performing zero-shot intent detection on unseen intents and domains. Using knowledge from ontologies or attributes (Ferreira et al., 2015; Yazdani and Henderson, 2015) can help in detecting and generalizing to new intents. A more recent approach by (Liu et al., 2019) makes modifications to capsule networks to generalize to unseen domains. Embeddings of intent descriptions have also shown to be quite meaningful in generalizing to new intents and services (Ma et al., 2019). While these methods are effective, they all require training on an initial set of intents. Large Language Models (LLMs) like GPT-3 (Brown et al., 2020) and more recently instruction finetuned models like (Chung et al., 2022) have shown good zero-shot performance on newly seen tasks without any prior training data on those tasks. In this work, we show that these models are also effective for zero-shot intent classification using just intent descriptions.

3 Datasets

We use public and private intent classification datasets to benchmark different approaches. For evaluation on public dataset, we use the English train and test sets from MASSIVE for intent classification. MASSIVE contains utterances directed at a physical device spanning 60 intents and 18 domains. For more details on the MASSIVE dataset (FitzGerald et al., 2022), we encourage readers to refer to their paper. We also use private benchmarking datasets internal to our company. These datasets contain various intents and utterances in the enterprise setting spanning 3 different domains: IT Service Management (ITSM), HR and Customer Service Management (CSM). The utterances are inspired by interactions between humans and chatbots and are typically queries from goal-oriented conversations where the user needs to resolve an issue. Additionally, some of these datasets also contain out-of-scope (OOS) utterances in their test set i.e. utterances that do not belong to any intent, in order to benchmark irrelevance detection of intent classification models. Table 1 shows statistics for different datasets used in our benchmarking.

4 Methodology

In this section, we describe the various methods we evaluate for zero and few-shot learning.

4.1 Domain Adaptation

Domain and task-specific pre-training of language model (Gururangan et al., 2020) has shown to significantly improve classification accuracy in both low and high resource settings. Techniques like contrastive learning (Gao et al., 2021) (Feng et al., 2022) are effective for improving the quality of sentence encoders, specifically in low-resource settings. Inspired by these ideas, we use a sentence encoder trained on our domain-specific data along with public datasets. Starting with the LaBSE checkpoint (Feng et al., 2022) we train it further by converting intent classification, paraphrasing, etc, as sentence similarity tasks. We will refer to this model as ELMSE (enterprise language model based sentence encoder).

For training intent-classification models, we freeze ELMSE weights and use its sentence embeddings as features for a trainable non-linear dense layer for classification. We compare ELMSE against other publicly available sentence encoders, namely LaBSE, Multilingual Universal Sentence

Dataset	Intent Names	Utterance
OOTB - dataset*	UpdateChangeRequest	I could I update CHG1234567
	TroubleshootSlowComputer	My laptop is taking too long to load
	SubmitARequest	I need a new phone
	IdentifyScheduledChanges	What are the upcoming scheduled changes
	CreateProblem	report new critical problem
Benchmark01 - dataset*	GuestWifiAccess	How do I get in the guest wifi
	IdentifyScheduledChanges	Can you pull up the list of scheduled changes
	MyAssignedEquipment	Show me my devices list
	SearchKnowledgeBase	I want information on policies
	RepositoryAccess	How can I access the shared drive
Benchmark02 - dataset*	EscalateITTicket	increase priority of my incident
	LocalAdminAccess	Can I get authorization as local admin on my pc
	RSAToken	RSA login is not working
	EmailSetup	How do I configure outlook on my device
	BillingInvoiceIssue	I was billed twice but have no account
Benchmark03 - dataset*	SubmitARequest	I request a new computer
	RSAToken	I have problem with authentication code
	CreateChangeRequest	I want to request a change
	LocalAdminAccess	How can I login as local admin
	Feedback	I have bad experience

Table 2: Few samples of intents and their respective utterances from the private internal benchmarking datasets.

Few-shot K	model	Massive	Benchmark01	Benchmark02	Benchmark03
3	LaBSE	46 (1.7)	59 (2.9)	52 (2.7)	58 (3.1)
	MUSE3	53 (2.8)	64 (3.8)	62 (2.7)	64 (1.3)
	GTR-3b	**59 (1.4)**	76 (1.4)	**70 (3.3)**	**78 (2.2)**
	ELMSE	57 (2.3)	**77 (2.4)**	63 (4.6)	74 (1.7)
5	LaBSE	58 (1.7)	65 (3.3)	59 (1.7)	67 (1.8)
	MUSE3	61 (0.9)	70 (2.2)	66 (1.4)	70 (1.7)
	GTR-3b	**66 (1.2)**	78 (1.0)	**73 (1.7)**	**84 (1.0)**
	ELMSE	63 (1.1)	**80 (1.7)**	67 (2.6)	79 (1.2)

Table 3: Results for domain adaptation on 3 internal datasets along with MASSIVE comparing LaBSE, MUSE, ELMSE, and GTR-3B models. The metric reported here is in-scope accuracy averaged over 5 different selections of few shot data. Numbers inside parenthesis indicate standard deviation across the 5 selections

Encoder (MUSE) (Yang et al., 2020) and GTR-3B. ELMSE is comparable in size to LaBSE and MUSE while almost 30 times smaller than GTR-3b. We simulate few-shot setting by randomly selecting K utterances per intent from full datasets. We use K=3,5,8,10,15,20 as well as the full dataset. We report results on 4 datasets from Table 1. Since OOTB-dataset was used for pretraining ELMSE, we exclude it from few-shot evaluation.

4.1.1 Results for Domain Adaptation

Table 3 reports in-scope accuracy and standard deviation averaged of 5 random seeds for 3-shot and 5-shot classification. The results demonstrate that domain adaptation is a very effective approach with improvements of greater than 5 percent in most cases when compared with models of similar size. These results carry over as we increase the number of few-shot utterances to more than 5 as shown in Figure 1. The plots also show that the gap between ELMSE and LaBSE is much larger in a few-shot setting and reduces as K increases. Moreover,

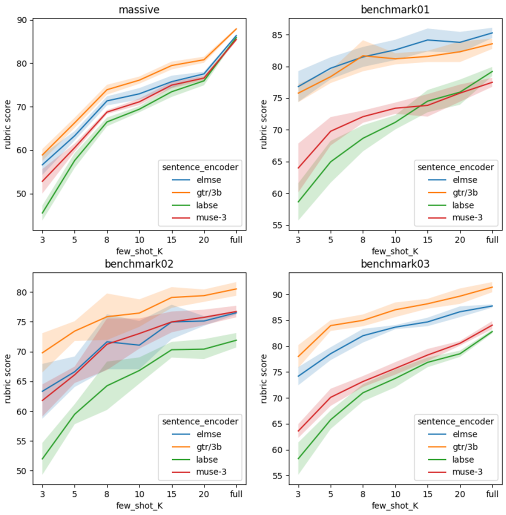


Figure 1: Comparison of ELMSE which is domain adapted with sentence encoders which are not domain adapted

ELMSE is only 2-3% worse than GTR-3b which is 30 times larger model.

4.2 Data Augmentation

We use data augmentation to generate labeled data for training starting with a seed set of 5 utterances per intent. In this section, we explore different ways of prompting GPT-3 and T5 (Raffel et al., 2020). For evaluating the generated utterances, we use them for training the same type of lightweight classifier as described in 4.1 using ELMSE as the sentence encoder. This section describes different prompt-based approaches for data generation.

GPT-3 + Paraphrase : Following (Sahu et al., 2022), we ask GPT-3 to generate 20 paraphrases of utterances from the same intent taken from the seed set. To encourage diverse generations, we set high temperature and top_p values.

GPT-3 + Intent Descriptions : We describe intents in the prompt and ask GPT-3 to generate 20 utterances for a particular intent. We find that describing all intents prevents hallucinations in the generations.

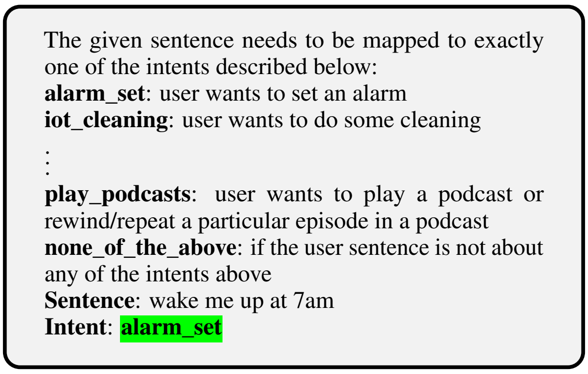
Parrot T5 Paraphrasing : We use the Parrot Paraphrase approach based on T5 (Damodaran, 2021) to generate 20 diverse paraphrased utterances given seed set. Table 4 shows a few generations from our prompt-based approaches.

4.2.1 Experimental Setup and Results

To evaluate the quality of generated utterances, we use them to train intent classifiers. We evaluate the performance of augmented dataset from each approach as mentioned in Section 4.2.1 by training ELMSE classifier model for intent classification task. We evaluate on 4 datasets and compared against ELMSE few-shot baseline where K is set to 5. We report the in-scope accuracy and standard deviation averaged over 3 different random seeds. Table 5 shows the result for all approaches using the data augmentation. Unless mentioned explicitly, we do not add the seed set to the training mix.

We find that using paraphrases from GPT-3 and Parrot T5 Paraphraser give better results compared to ELMSE Baseline even without the seed set. GPT-3 Augmentations using Intent Descriptions does not perform well but when combined with ELMSE Baseline seed set gives better results. Moreover, given a good quality seed-set, we see that data augmentation using LLMs can boost the performance of intent classification in few-shot setting.

4.3 Prompting Zero-shot Prediction


The given sentence needs to be mapped to exactly one of the intents described below:
alarm_set: user wants to set an alarm
iot_cleaning: user wants to do some cleaning
:
:
play_podcasts: user wants to play a podcast or rewind/repeat a particular episode in a podcast
none_of_the_above: if the user sentence is not about any of the intents above
Sentence: wake me up at 7am
Intent: **alarm_set**

We use intent names and descriptions for prompting language models to perform zero-shot prediction. The intent descriptions are prefaced with instructions to predict the correct intent and the test utterance is specified at the end. The output is expected to be the correct intent label. Figure 4.3 shows an example prompt from the MASSIVE dataset and the output from LLMs. For evaluation, we check for the presence of intent names in the LLM completion text as opposed to an exact match and report the first intent predicted in the completion. This is done to account for hallucinations. If no intent names are present in the completion text, we mark it as an "out-of-scope prediction". We create intent descriptions for 4 datasets: 3 internal benchmarking datasets and the open-source MASSIVE dataset. We benchmark 2 language models using this type of prompt: GPT-3 (175B parameters) and Flan-T5-XXL (11B parameters), an instruction fine-tuned model.

Filtering Intents for LLMs Many conversational agents have a lot of intents, sometimes more than 50. As the number of intents increases, the prompt size increases which incurs higher latency and cost (in case of token-based pricing models). To restrict the length of the prompt, we use sentence similarity to retrieve the top-5 intents and only use those 5 intents in the language model prompt. Using sentence similarity needs a few training examples which makes this a few-shot approach. With just 5 examples per intent, we get more than 0.85 recall for Top-5. To restrict the size of the prompt, we do not pass the training examples to the LLMs.

Setup For GPT-3, we set the temperature to 0 and max tokens for generation to 20. We use the

Approach	Prompt	Generated Text
GPT3 Paraphrase Aug.	**Task**: Create diverse utterances by paraphrasing the following utterances: *schedule alarm to wake me up after 3 hours* *alarm for ten am* *wake me up on friday at five in the morning i need to catch the train* *alarm me at eight am* *please set alarm for today* **Create 20 utterances:**	Set an alarm for 10 o'clock. Wake me up on Friday at 5am so I can make the train. ⋮ Set a timer to wake me up in three hours
GPT3 Aug. Using Intent Descr.	A virtual assistant serves multiple intents. Below are the description of the intents: **alarm_set**: user wants to set an alarm **iot_cleaning**: user wants to do some cleaning ⋮ **play_podcasts**: user wants to play a podcast or rewind/repeat a particular episode in a podcast Generate **20 utterances for alarm_set** intent:	Can you set an alarm for next week? I need to set an alarm for a specific time I want to set an alarm for a certain day ⋮ I'd like to set an alarm for a certain hour

Table 4: Example prompts used in generating text for the corresponding approaches

Approach	MASSIVE	Benchmark01	Benchmark02	Benchmark03
ELMSE Baseline	63 (1.1)	80 (1.7)	67 (2.6)	79 (1.2)
GPT-3 w/ Paraphrase Aug.	**63 (0.5)**	**84 (0.4)**	71 (0.3)	**81 (0.5)**
GPT-3 w/ Intent Descriptions	51 (0.5)	76 (0.4)	69 (0.5)	76 (0.2)
Parrot T5	58 (0.4)	81 (0.2)	**73 (0.4)**	81 (0.4)
Seed Set + GPT-3 w/ Intent Descriptions	63 (0.8)	84 (0.4)	71 (0.3)	78 (0.9)
Seed Set + Parrot T5	63 (0.6)	79 (0.4)	68 (2.2)	76 (0.6)

Table 5: Results for Data Augmentation on 3 internal datasets along with MASSIVE comparing the performance on multiple prompt-based approaches. We report the average in-scope accuracy and standard deviation averaged over 3 different random seeds

Dataset	LLM Intents	Model	In-Scope Accuracy	Out-of-scope Recall
MASSIVE (60 intents)	5	Flan-T5-XXL	68.6	-
		GPT-3	**69.2**	-
	60	Flan-T5-XXL	73.3	-
		GPT-3	**73.9**	-
OOTB-dataset (27 intents)	5	Flan-T5-XXL	**83.7**	-
		GPT-3	83.4	-
	27	Flan-T5-XXL	**86.3**	-
		GPT-3	84.9	-
Benchmark01 (9 intents)	5	Flan-T5-XXL	**86.5**	0.43
		GPT-3	84.6	**0.97**
	9	Flan-T5-XXL	86.5	0.48
		GPT-3	**89.3**	**0.67**
Benchmark02 (13 intents)	5	Flan-T5-XXL	**69.7**	0.65
		GPT-3	60.6	**0.87**
	13	Flan-T5-XXL	**69**	**0.7**
		GPT-3	61.3	0.67

Table 6: Results for zero-shot prediction on 3 internal datasets along with MASSIVE with GPT-3 and Flan-T5-XXL. In-scope accuracy is the accuracy computed for test samples that belong to the intents in the dataset. Out-of-scope recall is the fraction of out-of-scope test samples which were correctly identified as irrelevant by the model i.e., not belonging to any of the intents

default setting generation settings for the Flan-T5-XXL model and do not restrict the number of tokens to be generated. The results with filtering are averaged over 3 runs using different random seeds for sampling the 5 samples per intent.

Results Table 6 reports the accuracy for in-scope intents and the recall for out-of-scope samples where applicable (samples that do not belong to any of the intents in the dataset). We find that prompting language models with intent descriptions for zero-shot intent classification performs better than few-shot learning using a classifier (Tables 3 and 5). Since this only needs intent descriptions, this approach can generalize to new intents as well. Using the same prompt, Flan-T5-XXL is competitive with GPT-3 in terms of in-scope accuracy and is often better when presented a smaller number of intents in the prompt. While the in-scope accuracy is comparable, GPT-3 clearly outperforms Flan-T5-XXL in terms of the out-of-scope recall, indicating

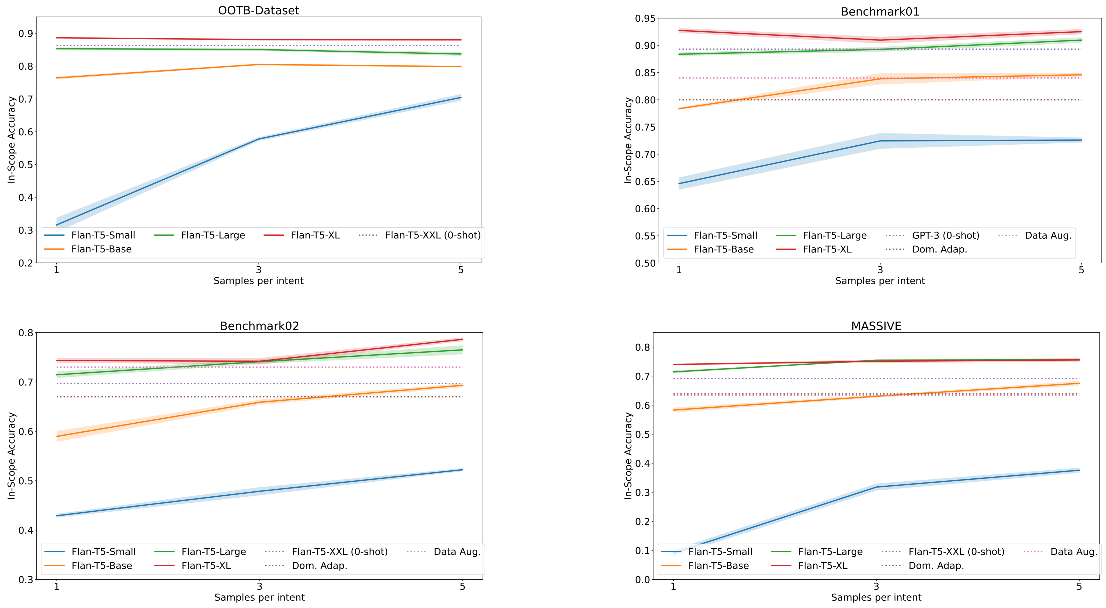


Figure 2: Plots comparing in-scope accuracy of different Flan-T5 models using Parameter-efficient FineTuning (PEFT) with the T-Few recipe. The dotted lines show the best results on each dataset from previously described methods. The shaded regions show the standard deviation

that it is better at detecting irrelevant samples. We attribute the strong performance of Flan-T5-XXL (even though it is 16x smaller) to the multi-task instruction finetuning on over 1800 datasets.

For the 3 internal datasets, we also find that using more intents in the prompt works better only up to a certain extent but have excluded the results for the brevity of this paper. While the intent retrieval method does not give perfect Top-5 recall, it helps in keeping the prompt short and hence provides lesser chances for the language models to give a output a wrong label name. Moreover, filtering can also improve the out-of-scope recall as in the case of Benchmark02 dataset.

4.4 Parameter-Efficient FineTuning (PEFT)

Taking inspiration from the T-Few recipe (Liu et al., 2022), we add and finetune IA3 adapters from scratch in Flan-T5 models in a few-shot setting which is similar to 4.1. We pick K=1,3,5 utterances per intent. Since the Flan-T5 models are instruction fine-tuned, we use the same prompt from 4.3 and provide the intent name as the target string. For MASSIVE and OOTB-dataset, we restrict the number of intents in the prompt to 15 at training time to prevent out-of-memory exceptions. At inference time, we provide all intents in the prompt. We use all 3 loss functions (language modeling, unlikelihood and length normalized losses) and the same hyperparameters as mentioned in the T-Few paper. For more details about the T-Few recipe, we encourage readers to refer to their paper.

Figure 2 compares the results of PEFT against the best results from previously described methods. Flan-T5-XL (3B parameters) consistently outperforms all other methods with just 1 training example per intent. With a few more examples, Flan-T5-Large (770M parameters) also outperforms all other methods except Flan-T5-XXL on the OOTB dataset. This shows that we can train significantly smaller models which are easier to deploy and also outperform LLMs like GPT-3 with just a few parameters using intent descriptions and a handful of examples.

5 Observations

Comparing results across the 4 approaches, we notice that all 4 approaches are effective in low resource settings. We find that domain adaptation is a cheap option in terms of size of the models but it still requires 5-10 training utterances per intent for getting accuracy above 70%. Data Augmentation using paraphrasing further helps in most cases by 2-4 percentage points. However, expanding to new domains requires sentence-pairs data for training the sentence encoder which can involve days of human labeling. Zero shot classification using intent descriptions with LLMs and instruction finetuned models performs even better than domain adaptation with data augmentation and doesn't require any utterances to be configured per intent. However a good description for each intent is required. Additionally, these models can be expensive to operationalize. Inference on Flan-T5-XXL requires using A100 GPUs. GPT-3 is not open-source and based on a pricing model which can be expensive to scale to thousands of customers. Parameter efficient fine-tuning (PEFT) of instruction finetuned models like Flan-T5-XL and Flan-T5-Large offers the best performance across all methods and often by a large margin. Moreover, these models are only a fraction of the size of GPT-3 and Flan-T5-XXL and much easier to operationalize at scale with far lesser compute resources.

6 Conclusion

In this paper, we addressed the task of zero and few-shot intent identification using Large Language Models (LLMs). We presented four approaches, namely domain adaptation, data augmentation, zero-shot prediction with prompting, and parameter-efficient fine-tuning. Our experimental results demonstrate that LLMs and larger instruction fine-tuned language models are very effective in zero-shot setting with in-context prompting. Smaller instruction finetuned models with adapters are even better when adapter-finetuned on just 1 or 3 examples per intent. We hope these results are useful for practical deployment of conversational agents in low-resource settings as well as aiding non-practitioners in building their intent classification models. In the future, we plan to extend this work by domain adapting smaller instruction fine-tuned models in a multi-task setting and exploring their zero-shot capabilities.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Derek Chen and Claire Yin. 2022. Data augmentation for intent classification. *CoRR*, abs/2206.05790.

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.

Prithiviraj Damodaran. 2021. Parrot: Paraphrase generation for nlu.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Emmanuel Ferreira, Bassam Jabaian, and Fabrice Lefèvre. 2015. Online adaptative zero-shot learning spoken language understanding using word-embedding. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5321–5325.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gökhan Tür, and Prem Natarajan. 2022. MASSIVE: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *CoRR*, abs/2204.08582.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Online. Association for Computational Linguistics.

Han Liu, Xiaotong Zhang, Lu Fan, Xuandi Fu, Qimai Li, Xiao-Ming Wu, and Albert Y.S. Lam. 2019. Reconstructing capsule networks for zero-shot intent classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4799–4809, Hong Kong, China. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Muqeeth Mohammed, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*.

Yue Ma, Zengfeng Zeng, Dawei Zhu, Xuan Li, Yiying Yang, Xiaoyuan Yao, Kaijie Zhou, and Jianping Shen. 2019. An end-to-end dialogue state tracking system with machine reading comprehension and wide & deep classification. *CoRR*, abs/1912.09297.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 87–94. Association for Computational Linguistics.

Majid Yazdani and James Henderson. 2015. A model of zero-shot learning of spoken language understanding. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 244–249, Lisbon, Portugal. Association for Computational Linguistics.

Dian Yu, Luheng He, Yuan Zhang, Xinya Du, Panupong Pasupat, and Qi Li. 2021. Few-shot intent classification and slot filling with retrieved examples. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 734–749, Online. Association for Computational Linguistics.

Jianguo Zhang, Trung Bui, Seunghyun Yoon, Xiang Chen, Zhiwei Liu, Congying Xia, Quan Hung Tran, Walter Chang, and Philip Yu. 2021. Few-shot intent detection via contrastive pre-training and fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1906–1912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.