Exploring the Use of Natural Language Descriptions of Intents for Large Language Models in Zero-shot Intent Classification

Taesuk Hong[1,2], Youbin Ahn[1], Dongkyu Lee[1], Joongbo Shin[1,†], Seungpil Won[1], Janghoon Han[1], Stanley Jungkyu Choi[1], Jungyun Seo[1,2]
[1]LG AI Research, [2]Sogang University
lino.taesuk@gmail.com,
{youbin.ahn, movingkyu, jb.shin, seungpil.won,
janghoon.han, stanleyjk.choi, seojy}@lgresearch.ai

Abstract

In task-oriented dialogue systems, intent classification is crucial for accurately understanding user queries and providing appropriate services. This study explores the use of intent descriptions with large language models for unseen domain intent classification. By examining the effects of description quality, quantity, and input length management, we identify practical guidelines for optimizing performance. Our experiments using FLAN-T5 3B demonstrate that 1) high-quality descriptions for both training and testing significantly improve accuracy, 2) diversity in training descriptions doesn't greatly affect performance, and 3) off-the-shelf rankers selecting around ten intent options reduce input length without compromising performance. We emphasize that high-quality testing descriptions have a greater impact on accuracy than training descriptions. These findings provide practical guidelines for using intent descriptions with large language models to achieve effective and efficient intent classification in low-resource settings.

1 Introduction

In task-oriented dialogue systems, mapping user utterances to a predefined set of intents is crucial and is known as 'intent classification.' This process is essential because it helps determine the service that the user requires, making it the foundational step in fulfilling the user's goal via a chatbot (Bang et al., 2023; Sung et al., 2023; Zhang et al., 2021a, 2022). Due to the vast range of domains where chatbots can be utilized and the limited availability of intent classification data, research on transferring intent classifiers to unseen domains under low-resource conditions is very active (Zhang et al., 2021b; Mueller et al., 2022; Kuo and Chen, 2023).

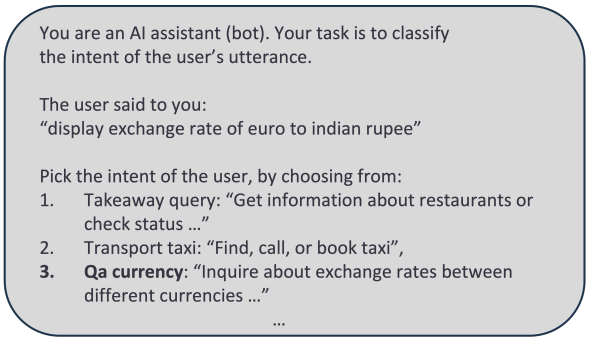

Figure 1: An example of an intent classification input for a large language model that includes intent descriptions. The figure was adopted from Parikh et al. (2023).

Parikh et al. (2023) proposed an in-context learning classification method using large language models to classify intents in unseen domains. They provided detailed intent descriptions as inputs to compensate for the lack of user query examples for each intent. Figure 1 illustrates how descriptions are included in the in-context learning input. Provided descriptions can capture the subtle semantic nuances and exceptions that are challenging to address with intent names alone. However, the paper does not clarify the quality or quantity of descriptions that should be used for training or inference, leaving practitioners without concrete guidelines. This paper aims to provide specific guidelines on the effective and efficient use of intent descriptions during training and testing for intent classification with large language models in unseen domains.

This study specifically explores how to utilize intent descriptions in large language models through the following aspects: 1) *Effect of description quality*: Using Chat-

†Corresponding author

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 458–465
September 18–20, 2024. ©2024 Association for Computational Linguistics

GPT (OpenAI, 2023), the study collects intent descriptions for the CLINC150 (Larson et al., 2019), HWU64 (Liu et al., 2021), and BANKING77 (Casanueva et al., 2020) datasets. Three sets of descriptions are collected: `dependent descriptions`, which consider semantic differences between intents, `independent descriptions`, generated without considering semantic differences, and `cleansed descriptions` manually filtered to address the subtle semantics. The impact of description quality on training and testing is investigated. 2) *Impact of description quantity*: The study examines the effect of increasing the number of descriptions used for training on intent classification accuracy. 3) *Input length Management*: Usage of off-the-shelf rankers, selecting the most probable intent options based on the similarity between the user query and descriptions, is examined to address the input length issue caused by descriptions. The optimal number of intent options to select that balances the trade-off between input length and performance is investigated. The study uses FLAN-T5 3B (Chung et al., 2022) which is an instruction-tuned model of T5 3B model (Raffel et al., 2019).

Our findings and contributions can be summarized as follows:

- Fine-tuning is required for effective understanding of descriptions in large language models and high-quality descriptions improve classification accuracy for both training and testing.
- Enhancing the quality of test descriptions has a more significant impact on accuracy than improving those used for training.
- Using a ranker to reduce to around ten classification achieves similar performance to using all options.

2 Method for Analysis

2.1 Quality-varied Description Generation and Filtering

To investigate the impact of description quality on intent classification using large language models, three different qualities of descriptions were collected using ChatGPT (gpt-3.5-turbo) via the OpenAI API. Prompts used for the API calls can be found in the Appendix A.

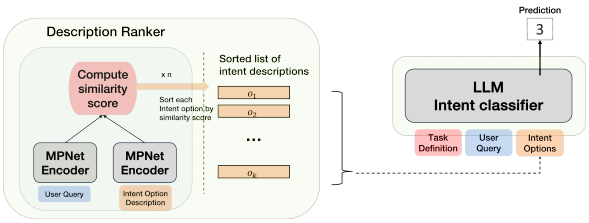

Figure 2: An off-the-shelf ranker scores the similarity between the user query and each description, selecting the top 'k' intent options for intent classification input.

Independent Description Generation The nuanced differences between distinct intents pose challenges for intent classification. For `independent descriptions`, prompts were crafted to include only a single intent and three user query examples specific to that intent, excluding other intents. Consequently, the collected description may lack comparative context, resulting in relatively lower quality. Prompts for each intent was called seven times to collect a total of seven `independent descriptions` per intent.

Dependent Description Generation In contrast, `dependent descriptions` include all possible intents within the prompt to ensure that the generated description uniquely distinguishes itself from others. Thus, these descriptions are considered relatively higher quality. For each intent, seven unique `dependent descriptions` was collected using API call.

Human-Cleansed Description Since the automatically collected descriptions may not fully capture differences between intents, manual review was added. One description per intent was carefully filtered to ensure clear distinction from other intents. This final filtering aimed to produce highest quality descriptions among our control-group for description quality. Henceforth, we will refer to this type of description as a `cleansed description`.

2.2 Description-Based Intent Option Ranker

Including intent descriptions increases the input length proportionate to the number of intent options. Given a model with a maximum length of 1024 tokens, descriptions of just ten

https://openai.com/index/openai-api

words per intent for 100 intents would exceed this limit. To address this, a description-based ranker was used to optimize input length. The off-the-shelf mpnet-base-v2 (Song et al., 2020; Reimers and Gurevych, 2019) model was employed. Figure 2 shows how this ranker integrates into the intent classification architecture. It calculates the similarity between user queries and intent descriptions, sorts intent options by similarity, and passes the top-k intents to the intent classifier. This paper experimentally determines the optimal k to maintain high performance while reducing input length.

2.3 Fine-Tuning Large Language Models for Intent Classification

Consider the user's utterance of i-th instance as u_i and intent options as $\{o_{i1}, o_{i2}, \ldots, o_{in}\}$. Descriptions for each intent are denoted as $\{d_{i1}, d_{i2}, \ldots, d_{in}\}$. All intent options are organized as

$$\begin{gathered}\text{`1. } o_1\text{: } d_1, \\ \text{2. } o_2\text{: } d_2, \\ \ldots \\ n.\ o_n\text{: } d_n\text{'}.\end{gathered}$$

Replacing this option text with a predefined instruction template forms the input $INST$. The training objective for FLAN-T5 and Llama-2-Chat is defined as:

$$L(\theta) = -\sum_{i}^{N} \log p(y_i | INST; \theta), \qquad (1)$$

where y_i is the correct intent index mapped to the i-th instance, N is the total number of instances, and θ represents model parameters. An example of an input as an instruction format, $INST$, can be found in Appendix E.

3 Experiments

3.1 Datasets

We used the publicly recognized intent classification datasets CLINC150, HWU64, and BANKING77. For training, we divided ten domains of the CLINC150 dataset in half and trained on 75 intents from five domains. The remaining 75 intents from the other five domains were reserved for testing. This domain split simulates an unseen domain scenario for the intent classification test. Detailed statistics for the datasets are provided in Appendix F.

https://sbert.net/

Table 1: Rows lower in the table represent higher description quality used during training. Similarly, columns further to the right indicate higher description quality used during testing. The accuracy for CLINC dataset is reported.

		Types of Descriptions Used in Testing		
		without descriptions	`independent descriptions`	`dependent descriptions`
Types of Descriptions Used in Training	without descriptions	**84.28**%±3.95%	84.15%±3.67%	90.55%±2.16%
	`independent descriptions`	81.93%±4.05%	85.64%±3.53%	90.97%±1.45%
	`dependent descriptions`	82.1%±3.56%	**86.99**%±3.09%	**91.75**%±1.91%

3.2 Impact of Description Quality on Intent Classification Training and Testing

Table 1 examines how description quality affects training and testing in intent classification models. When testing without descriptions, model trained without descriptions achieves the highest performance at 84.28%, while the performances of models trained with `independent` and `dependent descriptions` drop by 2.35% and 2.18% absolute points, respectively. However, when models trained with descriptions are tested with descriptions (specifically, `independent descriptions`), scores improve by 1.49% and 2.84% over the model trained without descriptions, respectively. This indicates that descriptions not only help models understand the detailed semantics of intents to improve classification accuracy but that fine-tuning models to understand descriptions enhances their ability to leverage them in testing.

The score improvements of the model trained with `dependent descriptions` over the model trained with `independent descriptions` demonstrate that fine-tuning with higher-quality descriptions optimizes their effective use in classification. This result supports the premise of this research that improving description quality is crucial and should not be left to random selection. In testing, higher-quality descriptions can boost performance, and their influence is more significant than in training. The model trained with `dependent descriptions` starts at 82.1% when tested without descriptions, improves by 4.89% when tested with `independent descriptions`, and achieves an additional 4.76% increase when tested with `dependent descriptions`. The improvement in testing quality has a larger impact

Table 2: The middle row shows models trained using a single type of `dependent description` per intent. In contrast, the top row represents models trained using five different `dependent descriptions` per intent, alternating during training. The bottom row shows models trained with a single manually filtered `cleansed description`.

		Types of Descriptions Used in Testing	
		1 dependent description	1 cleansed description
Types of Descriptions Used in Training	5 dependent descriptions	82.05%±3.15%	85.28%±1.69%
	1 dependent description	82.52%±2.44%	85.71%±1.61%
	1 cleansed description	**83.4**%±3.71%	**85.79**%±1.69%

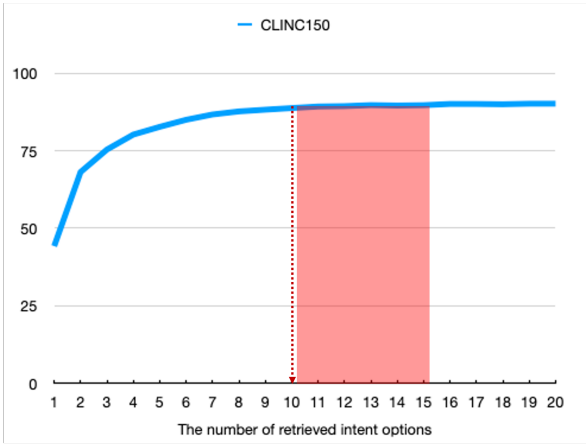

Figure 3: Graph depicts the intent classification accuracy on CLINC dataset converges as k becomes near 10.

than that in training. Notably, a model trained with `independent descriptions` but tested with `dependent descriptions` scored 90.97%, while one trained with `dependent descriptions` but tested with `independent descriptions` scored only 86.99%. This clearly shows that testing is particularly sensitive to description quality.

3.3 Impact of Description Quantity on Intent Classification Training

This experiment evaluates the effect of the quantity and quality of descriptions on model training. The result is shown in Table 2. The results reveal little to no difference between models trained with multiple descriptions and those trained with just a single description. In fact, performance tends to decline with the inclusion of varied descriptions. However, training with higher-quality descriptions – `cleansed descriptions` – resulted in the highest performance. This highlights the importance of training with a higher-quality description, even if only one, rather than relying on multiple descriptions of varying quality.

3.4 Optimizing 'k' for Efficient Intent Classification with Ranker

This experiment investigates the optimal value of 'k' for a ranker, determining the number of intent options to include in the intent classification input. Figure 3 demonstrates the performance trends as k increases. In the CLINC dataset, starting at approximately 44.21% accuracy with k set to 1, performance improves consistently as k increases, peaking at around 90% when k reaches around 13.

These results indicate that using a description ranker with the top k around 10 intent options provides near-optimal performance. The CLINC dataset, with 75 intent options and descriptions of 10 to 20 tokens each, requires around 1,200 to 1,300 tokens in total. By retrieving only the top 10 descriptions, the required input length drops to 300 to 400 tokens, reducing the input size by roughly 75%. This demonstrates that the approach proposed in this study significantly optimizes instruction-tuned models, enhancing their efficiency by minimizing the input length required for classification. For the HWU and BANKING dataset, the similar trend is shown and it can be found in Appendix B.

4 Conclusions

This paper thoroughly explored the impact of intent description quality and quantity on zero-shot intent classification using large language models while addressing the challenges of increased input length. The results show that fine-tuned models with descriptions are more effective for intent classification with descriptions. Additionally, higher-quality descriptions for both training and testing enhance performance, particularly during testing. Using an off-the-shelf ranker to reduce input length by selecting the top ten intent options minimizes input length without significant trade-offs in performance. Overall, this study provides practical guidelines for leveraging intent descriptions with large language models to address intent classification in low-resource settings.

References

Namo Bang, Jeehyun Lee, and Myoung-Wan Koo. 2023. Task-optimized adapters for an end-to-end task-oriented dialogue system. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7355–7369, Toronto, Canada. Association for Computational Linguistics.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew 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. *Preprint*, arXiv:2210.11416.

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.

Hui-Chi Kuo and Yun-Nung Chen. 2023. Zero-shot prompting for implicit intent prediction and recommendation with commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 249–258, Toronto, Canada. Association for Computational Linguistics.

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 *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 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. *Benchmarking Natural Language Understanding Services for Building Conversational Agents*, pages 165–183. Springer Singapore, Singapore.

Tingting Ma, Qianhui Wu, Zhiwei Yu, Tiejun Zhao, and Chin-Yew Lin. 2022. On the effectiveness of sentence encoding for intent detection meta-learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3806–3818, Seattle, United States. Association for Computational Linguistics.

Aaron Mueller, Jason Krone, Salvatore Romeo, Saab Mansour, Elman Mansimov, Yi Zhang, and Dan Roth. 2022. Label semantic aware pre-training for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8318–8334, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2023. Chatgpt (gpt-3.5-turbo). Accessed via https://chat.openai.com.

Soham Parikh, Mitul Tiwari, Prashil Tumbade, and Quaizar Vohra. 2023. Exploring zero and few-shot techniques for intent classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 744–751, Toronto, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

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*. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10433–10442, Singapore. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto.

2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Haode Zhang, Haowen Liang, Yuwei Zhang, Li-Ming Zhan, Xiao-Ming Wu, Xiaolei Lu, and Albert Lam. 2022. Fine-tuning pre-trained language models for few-shot intent detection: Supervised pre-training and isotropization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 532–542, Seattle, United States. Association for Computational Linguistics.

Haode Zhang, Yuwei Zhang, Li-Ming Zhan, Jiaxin Chen, Guangyuan Shi, Xiao-Ming Wu, and Albert Y.S. Lam. 2021a. Effectiveness of pre-training for few-shot intent classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1114–1120, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianguo Zhang, Trung Bui, Seunghyun Yoon, Xiang Chen, Zhiwei Liu, Congying Xia, Quan Hung Tran, Walter Chang, and Philip Yu. 2021b. 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.

Appendix

A Prompt for Description Generation using ChatGPT

Prompt for Independent Description Generation The independent description generation prompt is illustrated as follows:

Independent-Description Generation Prompt

Intent Name: {intent name}
Few-Shot Queries: {q1}, {q2}, {q3}
Instruction:
The above is the list of intents and their examples. Now, I want you to create unique descriptions for the intent. Make the description of the intent, 'intent name'. Here, make the description that encompasses the provided few-shot queries. Also, don't use the given use cases examples of intent for the description. Make the descriptions no longer than 10 words. I want you to return the result as following format of json:

List({
"{intent}": "description"
})
DO NOT return any words other except for the requested format of the result.

Prompt for Dependent Description Generation Dependent description generation prompt has the following format:

Dependent-Description Generation Prompt Example

Intent Name: {intent name}
Few-Shot Queries: {q1}, {q2}, {q3}
...
Intent Name: {intent name}
Few-Shot Queries: {q1}, {q2}, {q3}

Instruction:
The above is the list of intents and their examples. Now, I want you to create unique descriptions for each intent. This time, please make the description of the intent, '{intent}'. Here, the most important thing is that each description of intents is distinct and separate to each other. Don't make one description of intent to be inclusive to another. For example, if you have an intent, 'find restaurant', 'restaurant', don't make the description of each of them to be 'Find a available restaurant' and 'every acts related to restaurant' so that the former one is inclusive to the latter one. Also, don't use the given use cases examples of intent for the description. Make the descriptions longer than 10 words. Generate as long as possible. I want you to return the result as following format of json:
List({
"{intent}": "description"
})
DO NOT return any words other except for the requested format of the result.

B Optimizing 'k' for Efficient Intent Classification with Intent Option Ranker in HWU and BANKING datasets

For HWU, initial accuracy is 26.51 points with k set to 1, rising to almost maximum of 80 points around k equals 15. Lastly, in the BANKING dataset, the accuracy begins at 50.83 points and reaches around 80 points near the optimal k value of 10.

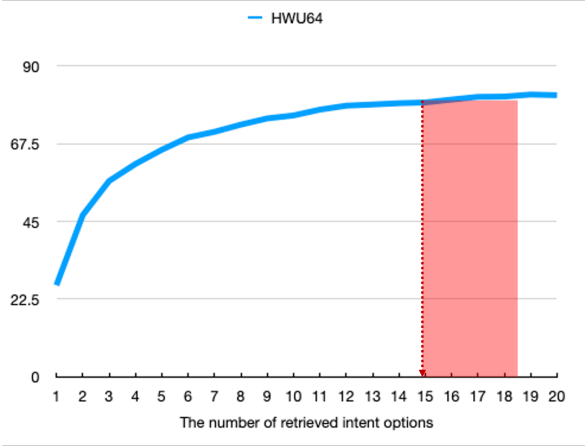

Figure 4: The graph shows that intent classification accuracy on the HWU dataset converges as k approaches near 15.

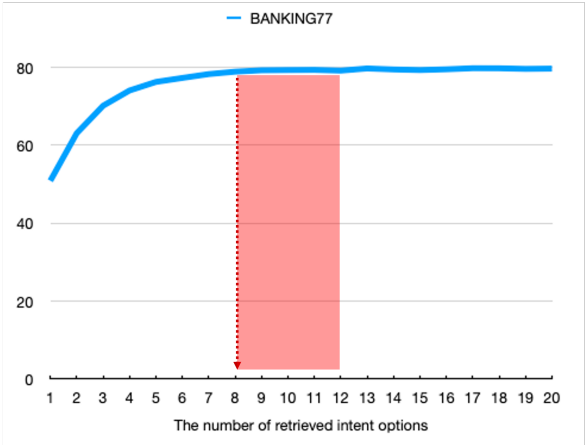

Figure 5: The Graph depicts the intent classification accuracy on BANKING dataset converges as k approaches near 10.

C Baseline Experiments

In Table 3, we compare our model, labeled as FLAN-T5-ranker (ours), with state-of-the-art models presented by Sung et al. (2023), which performed intent classification on the CLINC, HWU, and BANKING datasets. Please refer to the original paper for details on the baselines: L-BERT$_{\text{TAPT}}$, L-SBERT$_{\text{Paraphrase}}$, and L-PIE.

Table 3: Out-domain intent classification accuracy compared to state-of-the-art models and baselines. Zero and few-shot accuracy results are reported as percentages. The datasets had 50, 25, and 27 intent options for CLINC, HWU, and BANKING datasets respectively. We followed the same configuration but trained on CLINC data only outside the 50 test set intents.

	CLINC$_{N=50}$		HWU$_{N=25}$		BANKING$_{N=27}$	
	K=0	K=1	K=0	K=1	K=0	K=1
L-BERT$_{TAPT}$(Gururanganet al., 2020)	79.5	86.5	63.1	69.4	70.1	78.5
L-SBERT$_{Paraphrase}$(Maet al., 2022)	84.5	90.9	67.5	75.5	77.4	82.8
L-PIE (Sung et al., 2023)	86.5	91.8	70.6	77.4	77.6	82.9
FLAN-T5 (ours)	**97.58**	**97.62**	**87.92**	**87.22**	**84.72**	**85.52**
FLAN-T5-ranker (ours)	96.46	96.26	86.23	85.92	84.88	85.34
Llama-2-Alpaca (ours)	96.38	96.91	85.92	86.07	83.61	84.10
Llama-2-Alpaca-ranker (ours)	96.15	96.24	85.71	86.01	83.95	84.44

Using the FLAN-T5 3B model fine-tuned with `dependent descriptions` and tested with the top-10 ranked `cleansed descriptions` per option, our zero-shot approach outperformed L-PIE by 9.96, 15.63, and 7.28 points for the CLINC, HWU, and BANKING datasets, respectively. When trained on one sample per intent (one-shot learning), our model showed improvements of 4.44, 8.61, and 1.54 points over L-PIE for those datasets. The significant gap between our model and the state-of-the-art may be attributed to size differences, but these results demonstrate the objectivity of our findings and the model's superior performance over existing models.

Our model without the ranker, labeled FLAN-T5 (ours), shows slightly better performance than the version using a ranker, but the difference is minimal.

We also trained another well-known instruction-tuned model, Meta's Llama-2-Chat 7B (Touvron et al., 2023). This model was initially instruction-tuned with the Stanford Alpaca dataset (Taori et al., 2023) and further fine-tuned using intent classification data. Our model, referred to as Llama-2-Alpaca-ranker (ours), achieved accuracy comparable to our state-of-the-art FLAN-T5 model. Notably, our proposed method of using a ranker did not negatively impact performance and even provided slight improvements on the BANKING dataset. This confirms that using a ranker can not only reduce the burden of handling long inputs but also maintain effective performance in zero-shot intent classification.

D Training Detail

We use the *HuggingFace* implementation for fine-tuning FLAN-T5 models. In training FLAN-T5 model, AdamW optimizer with the learning rate $2e-5$ is used in training. The learning rate is gradually decayed during training with a cosine scheduler. The model is trained for 2 epochs and the batch size is 64. Every FLAN-T5 model performance reported in this work is the model of the final epoch. We run experiments with 4 NVIDIA A100 GPUs.

E Instruction Input Example

The Figure 6 shows an example of an input to the model for fine-tuning intent classification task. We manually crafted ten instruction templates following the FLAN v2 format (Chung et al., 2022) for the intent classification task. The input consists of a section that instructs the model to classify the given intent, a section with the user query, and another with the intent options.

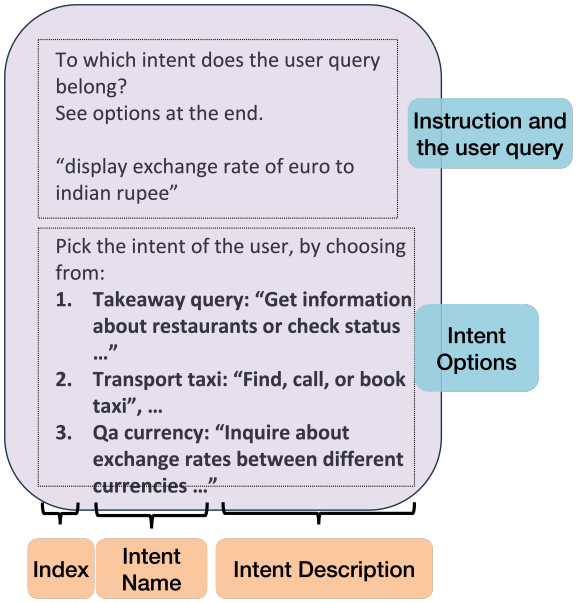

Figure 6: An input example of an instruction format.

F Dataset Statistics

The Table 4 provides statistics for the training and testing datasets of CLINC, HWU, and BANKING. For the CLINC and HWU datasets, the domains were split in half for different seeds, while for the BANKING dataset, all intents were split in half. The numbers below the dashed line represent the number of instances for each seed. The 'Seen domain' column corresponds to the training data, and the 'Unseen domain' column corresponds to the testing instances.

Table 4: The statistics for the training and testing datasets of CLINC, HWU, and BANKING.

	Seen domain	**Unseen domain**		
seed	**CLINC**	**CLINC**	**HWU**	**BANKING**
42	credit cards, banking, auto and commute, meta, utility	home, travel, work, kitchen and dining, small talk	music, recommendation, news, email, general, iot, transport, qa, datetime	banking
	7,500	2,250	669	1,560
52	auto and commute, banking, work, utility, kitchen and dining	home, meta, travel, credit cards, small talk	music, cooking, iot, play, transport, qa, datetime, social, weather	banking
	7,500	2,250	524	1,560
62	meta, kitchen and dining, credit cards, utility, work	home, travel, auto and commute, banking, small talk	alarm, music, audio, recommendation, general, play, lists, qa, cooking	banking
	7,500	2,250	621	1,560

https://huggingface.co/