Intent Recognition and Out-of-Scope Detection using LLMs in Multi-party Conversations

Galo Castillo-López **Gaël de Chalendar** **Nasredine Semmar**
Université Paris-Saclay, CEA, List, Palaiseau, France
{galo-daniel.castillolopez, gael.de-chalendar, nasredine.semmar}@cea.fr

Abstract

Intent recognition is a fundamental component in task-oriented dialogue systems (TODS). Determining user intents and detecting whether an intent is Out-of-Scope (OOS) is crucial for TODS to provide reliable responses. However, traditional TODS require large amount of annotated data. In this work we propose a hybrid approach to combine BERT and LLMs in zero and few-shot settings to recognize intents and detect OOS utterances. Our approach leverages LLMs generalization power and BERT's computational efficiency in such scenarios. We evaluate our method on multi-party conversation corpora and observe that sharing information from BERT outputs to LLMs leads to system performance improvement.

1 Introduction

Advances in dialogue systems have facilitated their employment to assist users on daily tasks in domains such as banking, health consulting, hospitality and others (Valizadeh and Parde, 2022; Camilleri and Troise, 2023; Casanueva et al., 2020). Task-oriented dialogue systems (TODS) in real-world applications must be able to both recognize user intents and detect Out-of-Scope (OOS) intents to generate reliable responses. Standard methods for intent recognition generally require large amounts of annotated data. However, annotations are scarce in some real-world applications, especially when new intents are introduced into systems. Large Language Models have been shown to be robust at classification tasks in zero and few-shot settings. Nevertheless, inferences from LLMs are computationally costly, thus their extensive use remains unpractical in some scenarios. Previous work have proposed hybrid approaches combining LLMs and smaller language models, by only routing uncertain inferences to LLMs at inference time (Arora et al., 2024). These approaches consist in processing queries in two steps: first through the computationally efficient model, and then through LLMs, if necessary. In doing so, overall computational costs are reduced without compromising prediction quality. However, these methods do not share information among models, and hence LLMs miss potential relevant information from the preceding step.

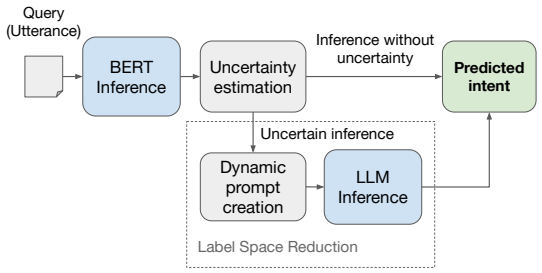

Figure 1: Overview of our Label Space Reduction method.

In this work we propose a hybrid approach that combines small language models and LLMs for intent recognition and OOS detection in multi-party conversations, i.e. dialogues between three or more participants. We route inferences with high uncertainty from fine-tuned BERT models to LLMs, and use the information from the outputs of the fine-tuned models to dynamically generate the prompts at inference time. Such information is employed to reduce the label space on the classification task. Experiments in this study are conducted on three open source LLMs. Our work leverages the efficiency of (relatively) small models and the power of LLMs in zero-shot settings for intent classification and OOS detection. Figure 1 illustrates our proposed method.

2 Related Work

In recent years, intent detection methods have mainly consisted in fine-tuning small models (Larson et al., 2019; Arora et al., 2020; Wang et al.,

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 504–512
Aug 25–27, 2025.

2023). Gautam et al. (2024) studied the use of class names to improve in-scope (IS) intent classification and OOS detection, using BERT and Spherical Variational Autoencoders (Davidson et al., 2018). Vishwanathan et al. (2022) observed that fine-tuning sentence transformers presents largely better IS and OOS performance than traditional methods.

LLMs have gained attention in multiple NLP tasks, including intent recognition in dialogue systems (Lin et al., 2024; Wang et al., 2024; Shin et al., 2024). Findings are contradictory, as some works have found that LLMs outperform fine-tuned models (Addlesee et al., 2023) and others have shown the opposite (Zhang et al., 2024b). To the best of our knowledge, the only study on intent recognition and OOS detection with a focus on efficiently using LLMs is proposed in (Arora et al., 2024). They propose a hybrid method that uses sentence transformers and LLMs, which reduced the performance gap to 2% while reducing computing latency up to 50%. However, their approach does not consider sharing information among models. Furthermore, we focus our work in multi-party conversations as such scenarios have been overlooked in previous work across most dialogue system tasks (Ganesh et al., 2023; Castillo-López et al., 2025).

3 Experimental Procedure

Let $D = \{(u_i, y_i) \mid y_i \in \mathcal{Y}_A\}$ be a labeled dataset where u_i denotes the i_{th} utterance labeled with intent y_i, and $\mathcal{Y}_A = \{1, \ldots, m, m+1\}$ denotes the set of m in-scope intents plus the out-of-scope label. Our aim is to build a multiclass classification system that detects whether u_i corresponds to an OOS intent from an unknown distribution or whether u_i can be classified into any of m possible in-scope intents.

3.1 Datasets

We use two multi-party conversations corpora in this work. The first corpus is **MIntRec2.0**, a multimodal dataset of 15K multi-party dialogues from TV shows (Zhang et al., 2024a). Modalities include audio, video and transcripts. In this study we are interested in systems working with text input data, thus we only use the text modality. The second dataset is **MPGT**, which is a collection of 29 multi-party dialogues between users and a receptionist robot in a hospital (Addlesee et al., 2023). The MIntRec2.0 and MPGT datasets contain 30 and 8 in-scope intents, respectively, and both count with OOS utterances. Additional information about the datasets is detailed in Appendix B.

3.2 Methods

We evaluate four different approaches for intent recognition and OOS utterance detection: fine-tuned BERT; LLMs zero-shot classification; Uncertainty-based Query Routing combining BERT and LLMs, following the strategy proposed in (Arora et al., 2024); and our proposed Label Space Reduction method (LSR) using BERT inference outputs and LLMs. We detail such methods below.

3.2.1 Small Language Model Fine-tuning

Fine-tuning. We use the same pre-trained BERT language model as in (Zhang et al., 2024a). Since having large number of examples per intent is challenging in real-world scenarios, especially when introducing new intents into systems, we conduct experiments on ten-shot settings following a similar approach to (Zhang et al., 2024a). In addition, we concatenate each utterance with its 3 preceding utterances to enhance the model performance by introducing context information. A special turn-shift token `<ts>` is included between each pair of concatenated utterances to explicitly indicate change of turns in dialogues. Thus, each input example is a text sequence corresponding to an utterance with its concatenated context. We fine-tune the uncased version of $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019) for multiclass classification over the entire set of m in-scope classes. Fine-tuning is performed over 5 different seeds to compute uncertainty scores from multiple runs at inference time. We detail the hyperparameter set we use on BERT-fine-tuning experiments on Appendix A.1. At inference time, the predicted class is obtained by a majority voting strategy. Note that in this approach, the pre-trained model is fine-tuned without the OOS label, thus a OOS class detection strategy is needed.

Out-of-Scope Detection. In order to detect OOS intents from our fine-tuned models, we quantify model uncertainty from the 5 outputs by computing the standard deviation of the softmax function applied on the logits in the last layer of the models (i.e. the probability estimates). Analysis on the validation sets showed that standard deviations $\sigma = 0.10$ and $\sigma = 0.12$ on the fine-tuned model probabilities provide good performance while maintaining a balance between OOS recall and IS macro F1-score,

Methods	In-scope			In-scope + Out-of-scope		
	ACC	WF1	WP	ACC	F1-OOS	F1
ChatGPT$_{zero}$ ♠	35.27	37.10	48.22	27.68	21.21	28.34
Mixtral 8×7B$_{zero}$	31.87	32.17	**51.35**	31.46	**38.66**	26.97
Llama-3 70B$_{zero}$	36.65	36.87	47.10	25.54	11.64	27.88
DeepSeek-R1 70B$_{zero}$	**41.22**	**43.47**	49.99	**35.79**	35.06	**35.14**
MAG-BERT$_{ten}$ ♠	9.82	11.58	13.34	34.58	**50.57**	3.75
ChatGPT$_{ten}$ ♠	34.53	36.39	49.27	29.72	27.85	28.41
BERT$_{ten}$	10.53	15.20	47.38	**34.64**	49.35	16.68
BERT$_{ten}$ + Mixtral 8×7B$_{zero}$	33.34	34.26	47.83	29.60	31.95	28.65
BERT$_{ten}$ + Llama-3 70B$_{zero}$	37.38	37.92	48.33	25.33	8.67	29.03
BERT$_{ten}$ + DeepSeek-R1 70B$_{zero}$	**41.42**	**43.51**	**49.58**	33.41	28.46	**35.05**
BERT$_{ten}$ + Mixtral 8×7B$_{zero}$ **(LSR)**	35.02	35.76	47.03	28.61	25.98	29.96
BERT$_{ten}$ + Llama-3 70B$_{zero}$ **(LSR)**	39.45	40.00	49.88	26.32	7.36	31.21
BERT$_{ten}$ + DeepSeek-R1 70B$_{zero}$ **(LSR)**	**41.66**	**44.82**	**52.65**	**33.96**	**29.48**	**36.55**
Humans$_{ten}$ ♠	*64.34*	*67.82*	*72.80*	*60.43*	*62.83*	*57.83*

Table 1: Results on the MIntRec 2.0 Corpus. Learning strategies include fine-tuning in *ten*-shot as well as *zero*-shot prompting. Results from (Zhang et al., 2024a) are denoted with ♠. Our results implement a label space reduction approach **(LSR)** leveraging BERT probability outputs. IS evaluation metrics include accuracy (ACC), weighted F1 (WF1) and weighted precision (WP). IS+OOS settings are evaluated on accuracy (ACC), macro F1 (F1), and F1 score on the out-of-scope label (F1-OOS). Scores in **bold** highlight the best performing model per setting, and scores in **blue** highlight the best performances overall.

Methods	In-scope			In-scope + Out-of-scope		
	ACC	WF1	WP	ACC	F1-OOS	F1
Mixtral 8×7B$_{zero}$	65.41	67.38	83.64	56.25	17.65	51.02
Llama-3 70B$_{zero}$	87.22	86.55	86.43	73.12	6.25	55.18
DeepSeek-R1 70B$_{zero}$	**89.47**	**91.17**	**93.65**	**78.75**	**35.90**	**75.17**
BERT$_{ten}$	50.38	58.92	**96.89**	56.25	**40.70**	64.65
BERT$_{ten}$ + Mixtral 8×7B$_{zero}$	72.93	74.65	88.28	61.88	13.33	63.53
BERT$_{ten}$ + Llama-3 70B$_{zero}$	90.23	90.23	90.98	75.00	0.0	62.98
BERT$_{ten}$ + DeepSeek-R1 70B$_{zero}$	**90.98**	**92.30**	94.30	**79.38**	33.33	**73.15**
BERT$_{ten}$ + Mixtral 8×7B$_{zero}$ **(LSR)**	72.18	80.49	**95.28**	69.38	40.54	68.57
BERT$_{ten}$ + Llama-3 70B$_{zero}$ **(LSR)**	89.47	90.73	92.43	75.62	12.12	67.21
BERT$_{ten}$ + DeepSeek-R1 70B$_{zero}$ **(LSR)**	**91.73**	**92.70**	93.97	**81.88**	**45.00**	**73.08**

Table 2: Results on the MPGT Corpus. Learning strategies include fine-tuning in *ten*-shot as well as *zero*-shot prompting. Our results implement a label space reduction approach **(LSR)** leveraging BERT probability outputs. IS evaluation metrics include accuracy (ACC), weighted F1 (WF1) and weighted precision (WP). IS+OOS settings are evaluated on accuracy (ACC), macro F1 (F1), and F1 score on the out-of-scope label (F1-OOS). Scores in **bold** highlight the best performing model per setting, and scores in **blue** highlight the best performances overall.

on the MIntRec2.0 and MPGT datasets, respectively.

3.2.2 Large Language Models

Large Language Models have been shown to excel at various classification tasks in zero-shot settings. We use three mid-sized instruct-tuned versions of open source LLMs: Mixtral8×7B, LLaMA-3 70B, and DeepSeek-R1 70B (distilled). More details about the used LLMs can be found in Appendix A.2. Our experiments on all LLMs are conducted on zero-shot prompting and use the same prompt template. The prompts we use describe the classification task; list the possible intents; define an OOS label; provide context from preceding utterances; define the expected output format; and include the utterance to classify. The prompt template is shown in Figure 3. We investigate how LLMs alone perform in our classification task, as well as in combination with BERT, as described in further sections. In contrast to the OOS detection strategy used on our fine-tuned models, we do not need to add an additional step at inference time as our prompts already instructs either recognizing intents or detecting OOS samples. In other words, the OOS detections are directly obtained from the LLMs.

3.2.3 Uncertainty-based Query Routing

Following the uncertainty-based query routing strategy proposed in (Arora et al., 2024), we combine BERT and LLMs by dispatching uncertain inferences made by BERT to LLMs. By doing so, only examples with high uncertainty are handled by LLMs, and costs due to the use of LLMs are

reduced. We use the output probabilities by the 5 fine-tuned models and compute their standard deviation to quantify the uncertainty of the prediction, as explained in 3.2.1. Prompts used in these experiments are the same as in the only-LLMs approach, where models are instructed to classify utterances into any of the IS intents or determine whether the utterance is OOS.

3.2.4 Label Space Reduction

We propose leveraging the outputs from the fine-tuned language models, and using such information to dynamically create prompts for LLM inference on routed queries. Our method extends the strategy described in Section 3.2.3. Instead of including all labels on the LLM prompts, we consider the intents with the highest probabilities (i.e. estimates from the softmax function on the final layer logits) outputted by the fine-tuned models. The intent set selection is conducted as follows. For every routed utterance u_i, we retrieve subset K_i of top-ranked intents whose cumulative sum of probabilities is at least P. The subset K_i is the smallest subset of intents defined as $K_i = \{y_1, y_2, \ldots, y_k\} \subseteq \mathcal{Y}_S$ such that $\sum_{j=1}^{k} p_i(y_j) \geq P$, where $\mathcal{Y}_S$ is the full set of in-scope intents, $p_i(y_j)$ is the softmax probability of label j for inference on u_i, and P is a hyper-parameter that controls the label space reduction (LSR). Lower values of P result in higher space reduction. Therefore, the amount of intents included on the routed LLM inferences vary among examples. We found on the validation sets that $P = 0.85$ achieves average hit rates slightly above 90% on the intent subsets on both datasets, while reducing the label spaces on average by ≈80% and ≈50% on the MIntRec2.0 and MPGT corpora, respectively. This suggests that our approach retrieves pertinent labels after label filtering.

3.3 Evaluation

Method evaluations are performed in IS and IS+OOS scenarios. In-scope evaluation does not consider test examples belonging to the OOS label, whereas IS+OOS considers all labels including the OOS label. We follow previous work (Zhou et al., 2024; Zhang et al., 2024a; Chen et al., 2024) and adopt three metrics for IS evaluation: Accuracy (ACC), Weighted F1 (WF1), and Weighted Precision (WP). Similarly, we use three commonly used metrics for IS+OOS evaluation: Accuracy (ACC) and F1-score (F1) over all classes, as well as F1-score over the OOS label (F1-OSS).

4 Results

Table 1 shows the results of our experiments on the MIntRec2.0 corpus. We observe that the best overall results in all in-scope performance metrics are obtained by our method on DeepSeek-R1. Reducing the label space results in an increase of ≈3% on the weighted precision. We also observe that when comparing the same BERT+LLM combinations, with and without label space reduction, better in-scope performance is obtained when reducing the label space in most metrics. BERT and MAG-BERT present the best overall performance on OOS evaluation. Nevertheless, their generalization on in-scope intents are the lowest compared to the other approaches. Additionally, as our method routes utterances with high uncertainty –i.e. potential OOS intents– to LLMs, it is expected to see a decrease on the F1-OSS score (in particular, a decrease on the OOS recall). It is also noteworthy that the classification task is complex even for humans, according to the results reported by Zhang et al. (2024a). We believe that such complexity might be due to a high number of intents and the presence of overlapping intents on annotations. An example of a difficult instance to classify by LLMs is displayed in Figure 2. We observe that similar to the example shown in Figure 2, multiple other instances from the MIntRec2.0 corpus semantically overlap with more than one intent.

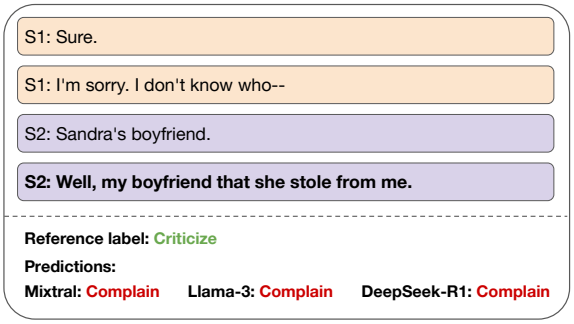

Figure 2: Example of an instance difficult to classify by LLMs from the MIntRec2.0 corpus.

Results on the MPGT corpus are found in Table 2, which show that DeepSeek-R1 with label space reduction obtains the best overall results in all metrics, except on in-scope WP and IS+OOS F1. In contrast to the results we observed on the MIntRec2.0 corpus, our method outperforms the fine-tuned BERT model alone on the F1-OOS score by ≈5%. In fact, all LLMs show to enhance their OOS detection when reducing the label space. We

argue that LLMs struggle to detect OOS intents (more than smaller language models) when there is a higher number of IS intents, as suggested in (Wang et al., 2024). In line with the results on Table 1, Llama-3 is the worst OOS detector in all settings. An increase between 8% and 12% on the IS+OOS accuracy is observed when reducing the label space on Mixtral and DeepSeek-R1. We also conduct additional analysis on the impact of the label space reduction hyperparameter P on the MPGT corpus in Appendix C.

5 Computational Efficiency Analysis

Table 3 shows a computational efficiency comparison between the use of BERT, LLMs, and the proposed label space reduction approach on the MPGT corpus. Note that our analysis considers all the 5 runs on the BERT inferences, which are performed to estimate inference uncertainty. We observe that our proposed method reduces the computational costs in more than 40% when combining BERT with Llama-3 70B and Mixtral 8×7B. To perform fair comparisons among methods and models, we employ the same computational resources on all inferences in this analysis.

Methods	Avg. latency	Latency ratio
Mixtral $8\times7B_{zero}$	1.925	
$BERT_{ten}$	0.065	0.034
$BERT_{ten}$ + Mixtral $8\times7B_{zero}$ (LSR)	1.100	0.571
Llama-3 $70B_{zero}$	4.039	
$BERT_{ten}$	0.065	0.016
$BERT_{ten}$ + Llama-3 $70B_{zero}$ (LSR)	2.236	0.553

Table 3: Method efficiency comparison on the MPGT corpus. Comparison is based on average latency per inference (seconds) and the latency ratio with respect to the zero-shot LLM inference method (without label space reduction).

6 Conclusions

We investigated how (relatively) small language models such as BERT can be combined with LLMs in zero-shot scenarios to reduce computational costs on intent recognition tasks without compromising predictive quality. Our results on MPCs are in line with previous works in dyadic dialogues, suggesting that uncertainty-based routing lead to performance gains. Our work also demonstrates that sharing information among models such as probability estimates to reduce the label space outperforms methods without shared information. Future work may consider exploring other plausible label selection strategies. Additionally, other sources of information to be leveraged in LLM prompts from small models (i.e. BERT) may be investigated in future studies: the actual probability estimates, uncertainty patterns, model's internal representations, etc. Finally, although our experiments are conducted on multi-party corpora, our proposed method could also be applied on dyadic scenarios. We believe that our findings show promising directions towards robust and efficient intent recognition systems in real-world applications.

Ethical Considerations

In developing our hybrid approach for intent detection using BERT and LLMs, we considered several ethical implications to ensure responsible practices. Despite the use LLMs, which are capable of generating potential unsafe content, they are solely employed as text classifiers into sets of defined classes. Therefore, the risk of misuse or producing harmful content available for end users is minimized. However, it is important for any implementation of the proposed methods to be aware of potential biases inherent in those models. In addition, all our experiments use publicly available corpora, which have been curated prior to our work to prevent malicious actions. Overall, the contributions presented in this study are designed for constructive and ethical use, with no direct association with harmful social consequences.

Acknowledgments

We warmly thank our anonymous reviewers for their time and valuable feedback. This publication was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-de-France Regional Council. This work has been partially funded by the EU project CORTEX2 (under grant agreement: N° 101070192).

References

Angus Addlesee, Weronika Sieińska, Nancie Gunson, Daniel Hernández Garcia, Christian Dondrup, and Oliver Lemon. 2023. Multi-party goal tracking with llms: Comparing pre-training, fine-tuning, and prompt engineering. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Gaurav Arora, Chirag Jain, Manas Chaturvedi, and Krupal Modi. 2020. HINT3: Raising the bar for intent detection in the wild. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 100–105, Online. Association for Computational Linguistics.

Gaurav Arora, Shreya Jain, and Srujana Merugu. 2024. Intent detection in the age of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1559–1570, Miami, Florida, US. Association for Computational Linguistics.

Mark Anthony Camilleri and Ciro Troise. 2023. Chatbot recommender systems in tourism: A systematic review and a benefit-cost analysis. In *Proceedings of the 2023 8th international conference on machine learning technologies*, pages 151–156.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*.

Galo Castillo-López, Gaël de Chalendar, and Nasredine Semmar. 2025. A survey of recent advances on turn-taking modeling in spoken dialogue systems. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 254–271.

Zhanpeng Chen, Zhihong Zhu, Xianwei Zhuang, Zhiqi Huang, and Yuexian Zou. 2024. Dual-oriented disentangled network with counterfactual intervention for multimodal intent detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17554–17567.

Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. 2018. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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 (long and short papers)*, pages 4171–4186.

Ananya Ganesh, Martha Palmer, and Katharina von der Wense. 2023. A survey of challenges and methods in the computational modeling of multi-party dialog. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 140–154.

Chandan Gautam, Sethupathy Parameswaran, Aditya Kane, Yuan Fang, Savitha Ramasamy, Suresh Sundaram, Sunil Kumar Sahu, and Xiaoli Li. 2024. Class name guided out-of-scope intent classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9100–9112, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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.

I-Fan Lin, Faegheh Hasibi, and Suzan Verberne. 2024. Generate then refine: Data augmentation for zero-shot intent detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13138–13146, Miami, Florida, USA. Association for Computational Linguistics.

Joongbo Shin, Youbin Ahn, Seungpil Won, and Stanley Jungkyu Choi. 2024. Learning to adapt large language models to one-shot in-context intent classification on unseen domains. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 182–197.

Mina Valizadeh and Natalie Parde. 2022. The ai doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660.

Asha Vishwanathan, Rajeev Warrier, Gautham Vadakkekara Suresh, and Chandra Shekhar Kandpal. 2022. Multi-tenant optimization for few-shot task-oriented FAQ retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 188–197, Abu Dhabi, UAE. Association for Computational Linguistics.

Pei Wang, Keqing He, Yutao Mou, Xiaoshuai Song, Yanan Wu, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. 2023. APP: Adaptive prototypical pseudo-labeling for few-shot OOD detection. In

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 3926–3939, Singapore. Association for Computational Linguistics.

Pei Wang, Keqing He, Yejie Wang, Xiaoshuai Song, Yutao Mou, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. 2024. Beyond the known: Investigating LLMs performance on out-of-domain intent detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2354–2364, Torino, Italia. ELRA and ICCL.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Wenrui Li, Yanting Chen, and 1 others. 2024a. Mintrec2. 0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. *arXiv preprint arXiv:2403.10943*.

Tianyi Zhang, Atta Norouzian, Aanchan Mohan, and Frederick Ducatelle. 2024b. A new approach for fine-tuning sentence transformers for intent classification and out-of-scope detection tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 910–919, Miami, Florida, US. Association for Computational Linguistics.

Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. 2024. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17114–17122. Number: 15.

A Model Information

In this appendix we provide model implementation details of our experiments.

A.1 BERT Implementation Details

All BERT fine-tuning runs were conducted on a single NVIDIA A100 GPU of 40GB. The average execution time for all fine-tuning experiments was less than 30 minutes to complete. We used the `BertForSequenceClassification` class from Hugging Face's Transformers library (Wolf et al., 2020) for sequence classification tasks. $\text{BERT}_{\text{BASE}}$ uncased is used in all the experiments. Table 4 shows the hyperparameter configuration we employ.

hyperparameter	value
eval_monitor	macro F1-score
train_batch_size	16
eval_batch_size	16
test_batch_size	16
wait_patience	3
num_train_epochs	40
warmup_proportion	0.1
lr	1e-5

Table 4: Set of hyperparameters used on BERT fine-tuning experiments.

A.2 Large Language Models

Our experiments on LLMs use mid-sized instruct versions of models. Specifically, we use `Mixtral-8x7B-Instruct-v0.1` (Jiang et al., 2024), `Meta-Llama-3-70B-Instruct` (Grattafiori et al., 2024), and `DeepSeek-R1-Distill-Llama-70B` (DeepSeek-AI, 2025).

A.3 Prompt Template

Figure 3 shows the prompt template we use on all LLM experiments.

B Corpora Details

B.1 MPGT Annotations

In this work, we assume that every utterance corresponds to a single intent, either in-scope or out-of-scope. Thus, the intent recognition task can be defined as a multi-class classification problem. However, the MPGT corpus is built under the assumption that an utterance might belong to none, one, or many intents, i.e. multi-label classification. Hence, we adapted the corpus for multi-class intent classification through manual data curation and multiple strategies. These strategies consisted in combining co-occurring intents, grouping original labels and co-occurring combinations, and assigning the OOS label to rare/irrelevant intents. Figure 4 shows the final label distribution after our adaptation. Our adapted multi-class version of the MPGT corpus for intent recognition is made available online[1].

B.2 Dataset Statistics

Table 5 shows statistics of the datasets we use in this work.

[1]https://github.com/gaalocastillo/mpgt_multiclass

```
**Task description**
You are an out-of-domain intent detector, and your task is to detect whether the
    intent of the last utterance belongs to the intents supported by the system,
    from dialogues of multiple participants. If they do, return the corresponding
    intent label, otherwise return UNK.

**Authorized categories**
The supported intents are:
intent_1, intent_2, intent_3, ... intent_N

**Out-of-domain label**
- UNK

**Previous utterances in the dialogue**
You have the following utterance history from multiple participants to understand
    the context of the dialogue. Each utterance is on a line and starts by "-":
- previous_utterance_1
- previous_utterance_2
- previous_utterance_3

**Expected output format**
Your response should only be a JSON object with the following structure:
{"intent": "intent_label"}
Do not write anything else.

**Task**
The utterance to classify is shown below:
utterance_to_classify

Result:
```

Figure 3: Prompt template used on all LLM experiments. Highlighted text in blue varies among dataset examples.

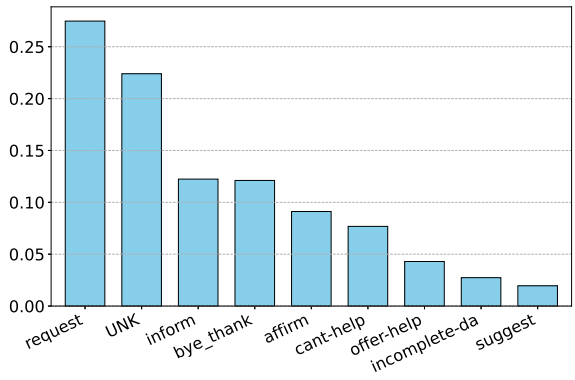

Figure 4: Distribution of the intents in the adapted multi-class version of the MPGT corpus, including the OOS label (UNK).

	#dial.	#utt.	#utt. fs.	#int.	%OOS
MintRec2.0	1.2K	15K	211	30	38%
MPGT	29	768	80	8	22%

Table 5: Dataset statistics: number of dialogues (# dial.), number of utterances (# utt.), number of utterances used on few-shot fine-tuning (# utt. fs.), number of intent categories (# int.), and proportion of OOS utterances (%OOS).

	MIntRec2.0			MPGT		
	train	dev	test	train	dev	test
#dial.	871	125	249	20	4	5
#utt.	9.9K	1.8K	3.2K	517	91	160

Table 6: Number of dialogues (# dial.) and utterances (# utt.) per subset split.

B.3 Subset Splits

Table 6 describes the subset splits we use for training, development and test. Note that our few-shot fine-tuning on BERT does not use all the training sets but only the selected few-shot utterances detailed in Table 5 in Appendix B.2.

C Hyperparameter P

Our proposed method relies on the hyperparameter P, which controls the label space reduction. Lower values of P result in higher reduction, therefore less intents included in the LLM prompts. The main results of this paper, presented in Tables 1 and 2, consider $P = 0.85$. We developed addi-

tional analysis on distinct values of P on the MPGT corpus and the BERT+DeepSeek method. Figure 5 suggests that low label space reduction ($P = 0.95$ and $P = 0.99$) presents better OOS precision and IS-OOS F1-score. Nevertheless, such improvement occurs at cost of missing OOS examples, as a decrease on the OOS recall is observed.

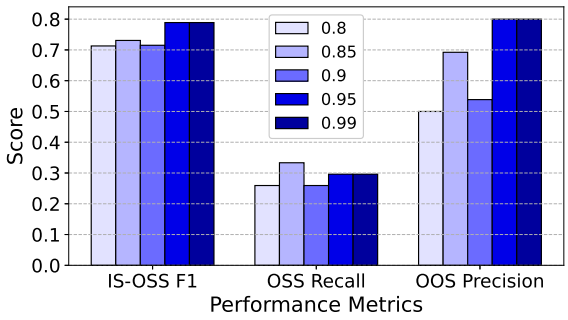

Figure 5: Performance metrics at distinct values of the hyperparameter P of our proposed method on the MPGT Corpus.