Exploring Description-Augmented Dataless Intent Classification

Anonymous ACL submission

Abstract

In this work, we introduce several schemes to leverage description-augmented embedding similarity for dataless intent classification using current state-of-the-art (SOTA) text embedding models. We report results of our methods on three commonly used intent classification datasets and compare against previous works of a similar nature. Our work shows promising results for dataless classification scaling to a large number of unseen intents, yielding competitive results to, and in some situations outperforming strong zero-shot baselines, all without training on labelled or task-specific data. Furthermore, we provide qualitative error analysis of the shortfalls of this methodology to help guide future research in this area.

1 Introduction

Task-oriented dialogue systems (TODS) by design, aid the user in accomplishing tasks within specific domains, and can have a wide range of applications from shopping (Yan et al., 2017) to healthcare (Wei et al., 2018; Valizadeh and Parde, 2022). Modular TODS (Wen et al., 2017) will typically contain an intent classification component (Louvan and Magnini, 2020; Chen et al., 2019; Su et al., 2022) used by the dialogue manager to determine the appropriate task the user intends to complete. In recent years, neural-based models using supervised training have reached state-of-the-art on many natural language processing tasks, including intent classification. However, supervised learning methods require human-labelled data for a predefined set of intents, which may be time-consuming and labour-intensive to acquire (Xia et al., 2018), and may have poor scalability if new intents are added, or task definition changed. An early approach to tackle this problem is *dataless intent classification* (Chang et al., 2008; Song and Roth, 2014) which aimed to leverage the pairwise similarities between semantic representations of utterances and intent classes to perform classification without reliance on human-labelled data. However, this approach relies heavily on the quality of semantic representations (Chang et al., 2008). In recent years, successful *zero-shot intent classification* approaches (Liu et al., 2019; Yan et al., 2020; Yin et al., 2019) have received greater attention, whereby learning conducted using labelled examples of a subset of *seen* intent labels is transferred to *unseen* intents. However, these methods still require human-labelled data, and tend to bias towards seen intents, with the number of unseen intents also generally much lower that seen intents (Liu et al., 2022; Zhang et al., 2022). With the significant recent advancements in the quality of text embedding models (Muennighoff et al., 2023), we explore the potential for dataless intent classification methods using a number of recent state-of-the-art text embedding models. We introduce several approaches for generating intermediate textual representations for intents, most notably using intent label descriptions, and formalise our methodology. We perform extensive evaluation of our methods, including scenarios with large numbers of intents from different domains, using three commonly used intent classification datasets. We summarise our contributions as follows:

- We introduce a new scheme for generating intent descriptions with an aim to minimise reliance on human expert input.
- We show that our intent descriptions yield significant improvements over label tokenization and synthetic utterances through extensive evaluation.
- We aggregate and explore the potential of a multitude of current SOTA text embedding models for dataless classification.
- We implement and evaluate a method for generating and utilising synthetic examples for dataless classification.
- We extensively evaluate our methodology on three commonly used intent classification

datasets and report on the results.

- We provide qualitative error analysis aimed at guiding future work.

2 Related Works

2.1 Generalized Zero-Shot Learning

Zero-shot learning (ZSL) (Yin et al., 2019) aims to leverage learning previously performed on labeled examples from seen tasks to unseen tasks, of which there are no labeled examples available for supervised training. ZSL has seen increasing popularity in the domain of intent classification (Liu et al., 2019; Yan et al., 2020) in recent years, whereby models are trained on a subset of intent labels and evaluated on another disjoint subset of intent labels. In more recent years, the concept of generalized zero-shot learning (GZSL) has seen an increase in prominence in the domain, in which the performance on both seen and unseen classes are considered in tandem (Zhang et al., 2022; Lamanov et al., 2022). Several GZSL approaches learn a label prototype space during training, which is transferred to unseen classes through methods such as interclass relationship modelling (Zhang et al., 2021) and prototype adaptation (Zhang et al., 2022). Approaches such as (Lamanov et al., 2022) encode the utterance and labels in a sentence-pair setup, with template-based lexicalisation of labels used as class prototypes. Other approaches exist that use label prototypes as centroids in Gaussian mixture models trained on seen class utterances (Yan et al., 2020; Liu et al., 2022). An issue that can occur with GZSL is bias towards seen classes (Zhang et al., 2022), which can lead to significantly lower performance on unseen classes. It is also difficult to see the efficacy of transfer to a large number of diverse unseen classes, as the number of unseen classes in evaluation are also typically much smaller than the number of seen classes.

2.2 Dataless Classification

Dataless text classification (Chang et al., 2008) is defined as tackling text classification without prior training on any labelled data. Generally regarded as a precursor to zero-shot text classification, this approach typically leverages sentence representations without any training on labelled data, by comparing the semantic representations between a sentence and that of the intent classes (Song and Roth, 2014). (Zha and Li, 2019) utilises “seed" words associated with each intent class to further contextualise the intent class representation, as a single word may often be insufficient to encapsulate the meaning of the class (Chen et al., 2015). Some approaches further leverages class hierarchy to augment classification performance (Li et al., 2016; Popov et al., 2019).

3 Methodology

3.1 Problem Definition

Let $\mathcal{C}$ be a set of intents supported by a task-oriented dialogue system, $\mathcal{U} = \bigcup\{\mathcal{U}_c\}_{c\in\mathcal{C}}$ defines the set of all user utterances, $\mathcal{U}_c = \{u_i\}_{1\le i\le n_c}$ is the set of utterances belonging to intent class c. The model undergoes no task-specific training and is tasked with making an intent prediction $\hat{y}_i$ for a previously unseen utterance u_i at inference time. We follow the paradigm set by previous works in dataless text classification (Chang et al., 2008; Song and Roth, 2014) to conduct nearest-neighbour classification over the sentence embedding space. For a given utterance u_i, an encoder $\mathbf{h}(\cdot)$ and a set of class label representations $\{l_c\}_{c\in\mathcal{C}}$, we make a prediction $\hat{y}_i$ as follows:

$$\hat{y}_i = \arg\max_c s(\mathbf{h}(u_i), \mathbf{h}(l_c))$$

where $s(\mathbf{u}, \mathbf{v}) = \mathbf{u}\cdot\mathbf{v}/||\mathbf{u}||_2||\mathbf{v}||_2$ is the cosine similarity between two vectors.

In order to conduct nearest-neighbour classification using intent labels, we require an intermediate representation, or prototype, which encapsulates to some degree the meaning of a class (Zha and Li, 2019), from which we can obtain a suitable embedding. A commonly used approach in dataless classification is to use the labels (Chang et al., 2008).

3.2 Label Tokenization

A class prototype is obtained by tokenizing intent labels directly, inserting spaces and replacing character separators, i.e.

```
AddToPlaylist →Add To Playlist
oil_change_how→ Oil Change How
```

However, this approach depends on the descriptiveness of the original intent labels, which can vary significantly between datasets and tasks. As such, we propose an additional step to produce intent label *descriptions* which we hypothesise can (1) better align the semantic representation with the characteristics of the class and (2) provide more consistent performance across datasets or approaches without requiring in-task data, which previous works (Lamanov et al., 2022) have shown could improve performance over purely using tokenized labels.

3.3 Our Approach

3.3.1 Intent Description

Our objective for generating intent label descriptions is to produce a brief description of the intent expressed by the user in a given utterance, while ensuring the process requires minimal expert human effort as to remain scalable for large numbers of intent classes. We formalise our process for writing intent descriptions as follows:

Label Preservation The resulting intent description must contain tokens from the original intent label i.e. `car_rental` → `User wants to rent a car`, or replace with an appropriate word (lexical cognates, synonyms etc.).

Format Consistency Descriptions should be written in the declarative form, beginning with either `"User is [asking|saying]"`, or `"User wants [to]"`, and aim to introduce minimal extraneous tokens. Our approach differs from the template-based approach in (Lamanov et al., 2022) in that we use exclusively the declarative form in writing our descriptions to maintain consistency across intent classes and datasets. Example descriptions can be seen in Table 1, more examples can be found in Appendix A.1.

Label	Description
`abbreviation`	“user is asking what an abbreviation stands for or means"
`flight_no`	“user is asking about a flight number"
`AddToPlaylist`	“user wants to add a song to a playlist"
`food_last`	“user wants to know how long a food lasts"
`maybe`	“user is expressing uncertainty"

Table 1: Example descriptions for intent labels from each of the datasets used in our experimentation (Section 4.1).

In our experimentation (Section 4), our intent descriptions added on average 6.6 tokens to the tokenized intent labels (1.9 → 8.5), with 98.3% of descriptions containing at least one of the label tokens in exact form, and 82.7% of all label tokens preserved.

3.3.2 Synthetic Examples

We compare additionally against synthetic utterance generated for each intent class. We leverage `gpt-3.5-turbo` (OpenAI, 2023) for this purpose, by including the tokenized intent labels and label description within the prompt to generate a set $\mathcal{S}$ of questions or commands fitting said intent i.e. “*Given a category* `tokenized_intent` *and the description* `description`*, Please generate* `n` *different example sentences of users asking questions or making commands that fit the given category.*". At inference time, we sample k synthetic examples for c classes and make prediction $\hat{y}_i$ as follows:

$$\hat{y}_i = \arg\max_c \frac{\sum_m^k s(\mathbf{h}(u_i), \mathbf{h}(s_m^c))}{k}$$

where s_m^c denotes the m^{th} example utterance belonging to intent class $c \in \mathcal{C}$. Examples of synthetic utterances can be found in Appendix A.1. We report on the results separately in Section 5.5 and the full results can be seen in Appendix A.2. We also consider synthetic examples generated using `gpt-4` but found the average performance to be lower on our task (Appendix A.3).

4 Experiments

4.1 Datasets

We evaluate our methods on three commonly used English task-oriented dialogue (TOD) system intent classification datasets. (1) **ATIS** (Hemphill et al., 1990) is an English air-travel information system dataset containing 18 intent classes. For comparison, we follow previous works (Zhang et al., 2022) in filtering out intent classes containing fewer than 5 examples. (2) **SNIPS-NLU** (Coucke et al., 2018) contains 7 intent classes, totalling 14,484 utterances. (3) **CLINIC** (Larson et al., 2019) is a dataset for out-of-scope intent classification, with 150 intents and 22,500 utterances spanning 10 domains. As our method does not involve fine-tuning on task-specific data, we consider *entire* datasets to consist of unseen data for evaluation.

4.2 Models

We select 11 models from the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al.,

2023) that are in the top 20 at the time of writing[1]. Our selections are based of the following criteria: (1) the model weights must be released (2) documentation of training methods and experimentation details must be readily available. Additionally, owing to computational limits[2], we only consider models up to 3GB in size. Basic model specifications are shown in Table 2.

Model	s	d_h	l	μ_{MTEB}
InstructOR$_{Large}$	1.34	768	512	61.59
E5-v2$_{Base}$	0.44	768	512	61.50
E5-v2$_{Large}$	1.34	1024	512	62.25
Multilingual-E5$_{Large}$	2.24	1024	514	61.50
E5$_{Large}$	1.34	1024	512	61.42
GTE$_{Small}$	0.07	384	512	61.36
GTE$_{Base}$	0.22	768	512	62.39
GTE$_{Large}$	0.67	1024	512	63.13
BGE$_{Small}$	0.13	384	512	62.17
BGE$_{Base}$	0.44	768	512	63.55
BGE$_{Large}$	1.34	1024	512	64.23
OpenAI-Ada-002	-	1536	8191	60.99

Table 2: Specifications of selected models grouped by training method. Column s shows model size (GB), d_h embedding dimensions, l maximum sequence length and μ_{MTEB} averaged performance on MTEB benchmark.

InstructOR (Su et al., 2023) embeds the utterance with a task description, allowing for task-specific conditioning at inference time, with good performance on unseen domains. Trained on 330 datasets using a contrastive learning objective (Ni et al., 2022). This family of models is initialised from GTR (Ni et al., 2022) models, which are in-turn initialised from T5 (Raffel et al., 2020) models.

E5 (Wang et al., 2022) performs unsupervised pretraining on the model on ~270M text pairs using an InfoNCE (van den Oord et al., 2019) objective with other utterances within the batch acting as negative examples, followed by supervised fine-tuning on 3 datasets. We select the *Base* and *Large* variants, initialised from *bert-base-uncased* and *bert-large-uncased-whole-word-masking* respectively.

GTE (Li et al., 2023) pretrains the model on ~800M text pairs and fine-tunes using 33 datasets. The contrastive learning objective used in this work considers, for each query-document pair (q_i, d_i) in a batch, the pairwise relation to the remaining examples $\{(q_j, d_j)\}_{j\neq i}$. The embedding similarities $s(q_i, d_j)$, $s(q_i, q_j)$, $s(d_i, d_j)$ are added to the partition function, where $s(q, d)$ is the cosine similarity between two embeddings.

BGE The work (Xiao et al., 2023) initialised from BERT (Devlin et al., 2019) models and trained using RetroMAE (Xiao et al., 2022) whereby both the input sentence and sentence embeddings in an autoencoder setup are randomly masked during MLM training. The authors use `[CLS]` token embeddings as the sentence representation. Our experimentation showed a slight improvement when using averaged token embeddings (Mean performance +0.82% *Tokenized-labels*, +1.06% *Class-description*).

We report results in Section 5 for all E5, GTE and BGE models using averaged token embeddings as sentence representations. We additionally compare model performances against a commonly used embedding model in OpenAI's `text-embedding-ada-002` (Neelakantan et al., 2022) which we refer to in our tables as 'OpenAI-Ada-002'.

[1] November-December 2023

[2] All experiments conducted using a single 9GB GPU

5 Results

5.1 Baselines and Terminology

We compare the performance of our methods against several unknown intent classification methods previously detailed in Section 2. Here we clarify the terminology used henceforth to refer to these methods in our results. We refer to scores on unseen intent labels reported by (Zhang et al., 2021) as **ICR**, (Yan et al., 2020) as **SEG**, (Liu et al., 2022) as **ML-SEG**, dataless approach trained using original data from (Lamanov et al., 2022) as **TIR**$_{Orig}$ and likewise **TIR**$_{Syn}$ for training on synthetic data. We refer to the results of the adapted method of (Gidaris and Komodakis, 2018) reported in (Zhang et al., 2022) as **CosT** and the reported main results as **LTA**.

5.2 Metrics

Following from previous works (Zhang et al., 2022; Lamanov et al., 2022), we report Accuracy and Macro-F1 scores for intent classification on each of the datasets. In addition, we also compute the average of Accuracy and F1 score for direct model

	Model	ATIS			SNIPS			CLINIC		
		Acc	F1	Mean	Acc	F1	Mean	Acc	F1	Mean
Baselines	ICR (Zhang et al., 2021)	35.54	34.54	35.04	-	-	-	-	-	-
	SEG (Yan et al., 2020)	-	-	-	69.61	69.31	69.46	-	-	-
	ML-SEG (Liu et al., 2022)	-	-	-	77.08	75.97	76.53	-	-	-
	TIR$_{Orig}$ (Lamanov et al., 2022)	-	-	-	-	-	-	63.90	73.10	68.50
	TIR$_{Syn}$ (Lamanov et al., 2022)	-	-	-	-	-	-	58.00	61.30	59.65
	CosT (Zhang et al., 2022)	46.04	45.21	45.62	47.73	62.84	55.28	62.73	70.28	66.50
	LTA (Zhang et al., 2022)	66.09	**55.02**	**60.55**	**90.09**	84.22	87.16	73.18	75.74	74.46
Tokenized Intent Labels	InstructOR$_{Large}$	12.41	25.03	18.72	82.71	82.07	82.39	64.50	61.02	62.76
	E5-v2$_{Base}$	13.20	27.58	20.39	77.30	76.96	77.13	65.33	62.40	63.87
	E5-v2$_{Large}$	14.67	38.61	26.64	70.83	69.15	69.99	61.56	59.24	60.40
	Multilingual-E5$_{Large}$	16.41	28.53	22.47	59.90	58.80	59.35	59.13	55.56	57.34
	E5$_{Large}$	44.71	36.43	40.57	75.68	73.21	74.44	70.27	67.96	69.11
	OpenAI-Ada-002	21.88	30.09	25.98	83.32	82.19	82.75	68.25	65.70	66.97
	GTE$_{Small}$	14.28	27.21	20.75	74.94	73.04	73.99	69.38	67.55	68.47
	GTE$_{Base}$	68.99	42.34	55.66	82.37	81.14	81.75	71.56	69.74	70.65
	GTE$_{Large}$	45.14	34.42	39.78	80.13	78.60	79.36	70.44	68.64	69.54
	BGE$_{Small}$	11.40	27.60	19.50	79.20	76.81	78.00	71.67	69.89	70.78
	BGE$_{Base}$	52.15	39.34	45.74	77.73	75.88	76.81	73.85	72.24	73.05
	BGE$_{Large}$	48.24	40.11	44.17	80.60	78.74	79.67	74.05	72.45	73.25
Intent Label Descriptions	InstructOR$_{Large}$	42.44	42.97	42.70	85.85	85.35	85.60	78.35	76.98	77.67
	E5-v2$_{Base}$	64.73	40.20	52.47	87.75	87.23	87.49	72.38	69.87	71.12
	E5-v2$_{Large}$	60.48	41.80	51.14	87.84	86.77	87.31	72.34	70.50	71.42
	Multilingual-E5$_{Large}$	**73.23**	38.69	55.96	84.64	83.11	83.88	73.17	71.48	72.33
	E5$_{Large}$	60.22	41.33	50.77	89.00	88.83	88.92	75.45	74.20	74.83
	OpenAI-Ada-002	58.97	43.71	51.34	89.71	**89.28**	**89.50**	78.75	76.86	77.81
	GTE$_{Small}$	68.87	42.92	55.90	84.62	84.22	84.42	71.35	69.41	70.38
	GTE$_{Base}$	67.05	42.27	54.66	86.60	86.22	86.41	75.60	73.91	74.75
	GTE$_{Large}$	66.52	44.71	55.62	86.65	86.01	86.33	76.71	75.12	75.92
	BGE$_{Small}$	57.22	40.19	48.70	86.01	85.01	85.51	73.05	70.96	72.00
	BGE$_{Base}$	55.88	44.21	50.05	88.66	87.98	88.32	78.10	76.52	77.31
	BGE$_{Large}$	59.26	47.50	53.38	89.58	89.01	89.30	**79.63**	**78.38**	**79.00**

Table 3: Performance of baseline and selected models on 3 intent classification tasks. We report accuracy, macro-f1 score and the mean of both for each dataset. For each metric, **bold** denotes highest score, underline denotes second-highest

comparison similar to (Gritta et al., 2022). Results are shown in full in Table 3.

5.3 Methods using Tokenized Labels

Despite a lack of task-specific fine-tuning, models using tokenized intent labels generally performed comparably to most of the baselines on unseen intents. The average performance across all models for each dataset is shown in Table 4. The best-performing model (GTE$_{Base}$) outperforms ICR (+20.63 Mean) on the ATIS dataset, SEG (+12.30 Mean) and ML-SEG (+5.23 Mean) on the SNIPS-NLU dataset and both TIR approaches (+2.15 Mean vs TIR$_{Orig}$, +11.00 Mean vs TIR$_{Syn}$) on the CLINIC dataset. GTE$_{Base}$ outperforms CosT on all 3 datasets (+10.04 Mean ATIS, +26.47 Mean SNIPS-NLU, +4.15 Mean CLINIC); however, it also significantly underperforms LTA on all 3 datasets (-4.89 Mean ATIS, -7.79 Mean SNIPS-NLU, -4.92 Mean CLINIC). We note the average performance across 12 models remains competitive with baselines other than LTA, though this approach appears quite sensitive to model as indicated by the comparatively high standard deviation (Table 4).

Method		ATIS	SNIPS	CLINIC
Tokenized Intent Labels	μ	31.70	76.30	67.18
	σ	12.72	6.53	5.05
Intent Label **Desc**ription	μ	51.89	86.91	74.54
	σ	3.77	3.02	3.02
Desc-Tok	μ	20.19	10.61	7.36
	σ	-8.95	-3.51	-2.03

Table 4: Performance mean μ and standard deviation σ across all 12 selected models for each of the 3 evaluation datasets. **Desc-Tok** denotes the individual differences in performance between using tokenized labels and intent descriptions.

5.4 Methods using Intent Descriptions

Our method using intent label descriptions yields a significant improvement over using tokenized labels (Table 4), with an average increase per model of $+20.19\%$ on the ATIS dataset, $+10.61\%$ on the SNIPS dataset and $+7.36\%$ on the CLINIC dataset. This appears to support our hypthesis (1) (Section 3.2) in that the additional contextualisation added through describing the label via a declarative sentence better encapsulates the semantic information represented by a label. We also note from Table 4 that the standard deviation in performance across models is significantly lower when using descriptions, supporting our hypothesis (2) that descriptions can improve consistency across models and approaches. Our overall best performing model (BGE$_{Large}$) also considerably outperforms the strongest baseline (LTA) in both SNIPS (89.30 vs 87.16) and CLINIC (79.00 vs 74.46). We do note that all of our approaches underperform on the ATIS dataset compared to the baseline, with our overall best-performing approach yielding 53.38 vs 60.55, we provide further insight into possible reasons in Section 6 to help guide future research.

5.5 Methods using Synthetic Data

We evaluate the efficacy of methods using synthetic examples by generating a set of $n = 20$ synthetic examples, from which we sample k to act as class prototypes, we repeat this procedure 20 times and compute the average performance across all samples. Table 5 shows averaged model performance across all 12 selected models and samples for $k = [1, 3, 5, 10, 15]$. For full results see Table 11 in Appendix A.2. We conducted additional experimentation with $k > 15$ but found further increasing k did not yield significant improvements

k	Metric	ATIS		SNIPS		CLINIC	
		μ	σ	μ	σ	μ	σ
$k=1$	Mean	23.59	8.42	71.37	5.51	53.87	5.42
	Δ_{Label}	-6.15	-4.23	-4.94	-1.02	-13.31	0.37
	Δ_{Desc}	-24.08	4.38	-15.54	2.57	-20.60	2.48
$k=3$	Mean	28.63	7.41	77.27	4.16	64.65	3.21
	Δ_{Label}	-1.10	-5.23	0.96	-2.37	-2.53	-1.84
	Δ_{Desc}	-19.03	3.37	-9.64	1.22	-9.82	0.27
$k=5$	Mean	30.05	6.74	78.54	3.98	67.29	2.81
	Δ_{Label}	0.31	-5.90	2.24	-2.55	0.11	-2.23
	Δ_{Desc}	-17.62	2.70	-8.36	1.04	-7.18	-0.13
$k=10$	Mean	30.80	5.33	79.63	3.57	69.24	2.48
	Δ_{Label}	1.06	-7.31	3.32	-2.96	2.06	-2.57
	Δ_{Desc}	-16.87	1.29	-7.28	0.63	-5.23	-0.46
$k=15$	Mean	31.12	5.15	80.06	3.46	69.99	2.50
	Δ_{Label}	1.38	-7.49	3.75	-3.07	2.80	-2.55
	Δ_{Desc}	-16.55	1.12	-6.85	0.52	-4.49	-0.44

Table 5: Averaged mean of accuracy and macro-f1 scores experiments conducted across 20 samples and 12 models using k number of synthetic examples per intent class. Δ_{Label} and Δ_{Desc} are differences to the averaged performance of methods using tokenized labels and intent descriptions respectively.

in performance. We note our method using $k = 15$ synthetic examples outperforms tokenized labels on SNIPS (80.06 vs 76.30) and CLINIC (69.99 vs 67.18) datasets, but underperforms slightly on the ATIS dataset (31.12 vs 31.70). Synthetic examples underperforms description-based methods by a considerable margin on all datasets, suggesting single intent label descriptions can be more powerful as class prototypes than synthetic instances. We note also the higher standard deviation σ in performance compared to the description-augmented method but lower compared to methods using tokenized labels.

6 Analysis

Figure 1 shows the embeddings generated by our best-performing model (BGE$_{Large}$) on the 3 evaluation datasets visualised using t-SNE (van der Maaten and Hinton, 2008), along with the embedding for the intent label description. Due to the challenge to readability posed by the large number of intents in the CLINIC dataset, instead sample the 15 top-performing (100% accuracy) and lowest-performing (24.47% accuracy) intent classes for illustration, with the results shown in Figures 1c and 1d respectively.

In-Domain Saturation We observe a poor alignment on the ATIS dataset between the intent label descriptions (Figure 1a) and utterance embeddings corresponding to each class, possibly explaining

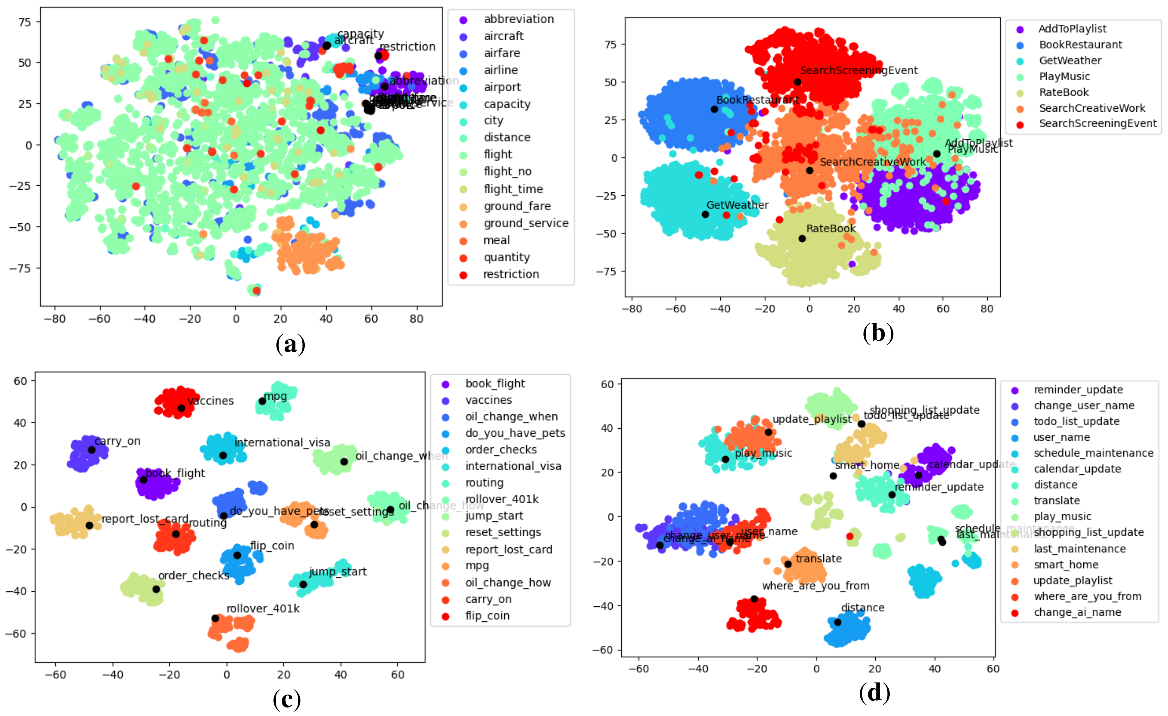

Figure 1: t-SNE (van der Maaten and Hinton, 2008) visualisation of embeddings computed using BGE$_{Large}$, class label description embeddings are shown in black and labelled. (**a**) Embeddings of ATIS (**b**) Embeddings of SNIPS (**c**) Embeddings of top 15 classes from CLINIC (**d**) Embeddings of bottom 15 classes from CLINIC.

Dataset	$\mu_{s_{in}}$	$\sigma_{s_{in}}$	$\mu_{s_{out}}$	$\sigma_{s_{out}}$	Δ_s	$\%\Delta_s$
ATIS	0.80	0.06	0.73	0.05	0.07	8.33
SNIPS	0.76	0.04	0.68	0.03	0.08	10.09
CLINIC	0.83	0.05	0.68	0.04	0.15	17.98

Table 6: Mean embedding similarity of sentences within the same class (*in*) and different classes (*out*). Δ_s denotes the average difference between *in*-class and *out*-class, $\%\Delta_s$ denotes the percentage average difference of similarity.

the poor performance in general on this dataset across models. We note the single-domain nature of the ATIS dataset, with all utterances relating to air-travel/flight, additionally, we note the significantly imbalanced nature of the ATIS dataset (Nan et al., 2021), with $\sim 74\%$ of utterances belonging to the `flight` class, which is a label that overlaps the domain of the dataset. We hypothesise this may lead to the intent label descriptions being much worse at capturing semantic information distinct to each class. This is supported by analysis on the pairwise embedding similarities of utterances belonging to the same class vs utterances belonging to difference classes (Table 6) where models' embeddings on the ATIS dataset consistently had lower percentage-difference in embedding similarity between *in*-class and *out*-class, implying more difficulty in distinguishing the utterances using solely embeddings. This issue does not appear as prominently in SNIPS or CLINIC likely due to domains being largely more distinct, though it is still visible in the lower-performing classes in CLINIC (Figure 1d).

Keyword/Lexical Overlap Another source of misclassifications may arise in situations whereby the class utterance embedding spaces overlap, whilst the intent label description embedding is aligned with the utterance embeddings. This can be seen for example with `SearchScreeningEvent` $\longleftrightarrow$ `SearchCreativeWork` in Figure 1b, `play_music` $\longleftrightarrow$ `update_playlist` and `user_name` $\longleftrightarrow$ `change_user_name` from Figure 1d. This appears to be due to the significant lexical overlap between utterances within the two classes, i.e. referring to common topics, keywords, irrespective of the domain of the classes.

Embedding Similarity Analysis We perform additional analysis on the mean embedding similarity of sentences within the same intent class (*in*-class) and of different intents (*out*-class). For a set of intent classes $\mathcal{C}$ and utterances $\mathcal{U}$, we calculate the mean *in*-class similarity $\mathbf{s}_{in}$ and *out*-class similarity $\mathbf{s}_{out}$ as

$$\mathbf{s}_{in} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \sum_{u_i \in \mathcal{U}_c} \sum_{u_j \in \mathcal{U}_c \setminus \{u_i\}} \frac{s(\mathbf{h}(u_i), \mathbf{h}(u_j))}{n_c(n_c - 1)}$$

Model	$\mathbf{s}_{in}$	$\mathbf{s}_{out}$	Δ_s	$\%\Delta_s$
InstructOR$_{Large}$	0.87	0.79	0.08	0.09
E5-v2$_{Base}$	0.82	0.74	0.08	0.09
E5-v2$_{Large}$	0.82	0.75	0.07	0.08
Multilingual-E5$_{Large}$	0.84	0.79	0.06	0.07
E5$_{Large}$	0.81	0.72	0.09	0.11
GTE$_{Small}$	0.84	0.76	0.07	0.09
GTE$_{Base}$	0.82	0.75	0.08	0.10
GTE$_{Large}$	0.83	0.75	0.08	0.09
BGE$_{Small}$	0.67	0.49	0.18	0.27
BGE$_{Base}$	0.71	0.56	0.15	0.21
BGE$_{Large}$	0.71	0.55	0.16	0.23
OpenAI-Ada-002	0.81	0.72	0.08	0.10

Table 7: Mean μ of pairwise embedding similarity between *in*-class ($\mathbf{s}_{in}$) and *out*-class ($\mathbf{s}_{out}$) utterances for each selected model. Δ_s denotes the difference between $\mathbf{s}_{in}$ and $\mathbf{s}_{out}$, $\%\Delta_s$

.

$$\mathbf{s}_{out} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \sum_{u_i \in \mathcal{U}_c} \sum_{u_j \in \mathcal{U}_{c'}} \frac{s(\mathbf{h}(u_i), \mathbf{h}(u_j))}{n_c n_{c'}}$$

where $\mathcal{U}_c$ and $\mathcal{U}_{c'}$ denotes the set of utterances belonging to class c and all classes other than c' respectively, n_c is the number of utterances in set $\mathcal{U}_c$. The mean *in*-class and *out*-class similarity scores are shown per dataset (Table 6), and per model (Table 7) . From a basic correlation analysis of the mean embedding similarity against a number of metrics, we note for model performance on the MTEB benchmark there exists a strong positive correlation to the difference Δ_s between *in*-class and *out*-class examples (Pearson $r = 0.72$, $p < 0.01$) as well as $\%\Delta_s$ (Pearson $r = 0.73$, $p < 0.01$), and there exists a strong negative correlation to the mean *out*-class similarity $\mu_{s_{out}}$ (Pearson $r = -0.71$, $p < 0.01$). Additionally we observe a strong correction between the aforementioned measures to model performance on the CLINIC dataset: mean difference (Pearson $r = 0.74$, $p < 0.01$), percentage-mean-difference (Pearson $r = 0.72$, $p < 0.01$) and mean *out*-class (Pearson $r = -0.71$, $p < 0.01$). We hypothesise that this indicates the quality of model embeddings as indicated by the mean difference between *in*-class and *out*-class to matter more with higher numbers of intent classes, and that this task in turn is a good indicator for text embedding model quality.

Analysis Summary Our proposed approach performs well overall against the strong baseline methods in unseen intent classification; however, it struggles in certain instances with overlaps in intents within the same domain, particularly if the class definition is non-distinct from other classes in domain i.e. `flight` from the ATIS dataset. To tackle such issues, future work may investigate the introduction of a hierarchical intent structure that is inferred in a dataless context to maintain scalability. The results of our experiments have shown intent label descriptions can perform well as intent prototypes in this problem setting, and that the naive addition of synthetic examples may yield worse performance; however, synthetic examples may be able to supplement dataless classification using intent label descriptions i.e. to tackle issues relating to lexical overlap between classes, hierarchical intent classes.

Limitations Our approach nonetheless contains a number of limitations: We have identified issues with the descriptiveness of individual labels earlier in this section, and textual labels may not be readily available for certain datasets, though summarisation methods may be effectively applied to few user utterances to produce such labels. Future work may also investigate the application of descriptions to tasks outside of intent classification, such as emotion recognition (Rashkin et al., 2019).

7 Conclusion

Dataless classification allows for scaling to a large number of unseen classes without requiring training on labelled, task-specific data. The benefits of such an approach can enhance development of task-oriented dialogue systems in application to data-poor or compute-limited scenarios where supported intents may also change as the system is developed. In this paper, we have explored the potential of current SOTA text embedding models in dataless intent classification settings using three different approaches for representing intent classes and compared our results against strong zero-shot learning baselines. We proposed a method for standardising the generation of intent label descriptions with an aim to minimise the amount of human annotations required to further support scaling to high numbers of intent classes. Our results have shown that description-augmented dataless classification methods can achieve comparable, and sometimes superior performance to zero-shot methods on the task of intent classifcation.

References

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, page 830–835. AAAI Press.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *ArXiv preprint arXiv:1902.10909*.

Xingyuan Chen, Yunqing Xia, Peng Jin, and John Carroll. 2015. Dataless text classification with descriptive lda. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2224–2231. AAAI Press.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Spyros Gidaris and Nikos Komodakis. 2018. Dynamic few-shot visual learning without forgetting. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4367–4375.

Milan Gritta, Ruoyu Hu, and Ignacio Iacobacci. 2022. CrossAligner & co: Zero-shot transfer methods for task-oriented cross-lingual natural language understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4048–4061, Dublin, Ireland. Association for Computational Linguistics.

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

Dmitry Lamanov, Pavel Burnyshev, Katya Artemova, Valentin Malykh, Andrey Bout, and Irina Piontkovskaya. 2022. Template-based approach to zero-shot intent recognition. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 15–28, Waterville, Maine, USA and virtual meeting. 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.

Yuezhang Li, Ronghuo Zheng, Tian Tian, Zhiting Hu, Rahul Iyer, and Katia Sycara. 2016. Joint embedding of hierarchical categories and entities for concept categorization and dataless classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2678–2688, Osaka, Japan. The COLING 2016 Organizing Committee.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

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.

Han Liu, Siyang Zhao, Xiaotong Zhang, Feng Zhang, Junjie Sun, Hong Yu, and Xianchao Zhang. 2022. A simple meta-learning paradigm for zero-shot intent classification with mixture attention mechanism. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2047–2052, New York, NY, USA. Association for Computing Machinery.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 480–496, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering main causalities for long-tailed information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural*

Language Processing, pages 9683–9695, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Artem Popov, Victor Bulatov, Darya Polyudova, and Eugenia Veselova. 2019. Unsupervised dialogue intent detection via hierarchical topic model. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 932–938, Varna, Bulgaria. INCOMA Ltd.

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(1).

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, page 1579–1585. AAAI Press.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip Yu. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3090–3099, Brussels, Belgium. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert Y.S. Lam. 2020. Unknown intent detection using Gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1050–1060, Online. Association for Computational Linguistics.

Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4618–4625. AAAI Press.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Daochen Zha and Chenliang Li. 2019. Multi-label dataless text classification with topic modeling. *Knowl. Inf. Syst.*, 61(1):137–160.

Yiwen Zhang, Caixia Yuan, and Xiaojie Wang. 2021. Generalized zero-shot text classification via inter-class relationship. In *2021 IEEE 7th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, pages 413–417.

Yiwen Zhang, Caixia Yuan, Xiaojie Wang, Ziwei Bai, and Yongbin Liu. 2022. Learn to adapt for generalized zero-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 517–527, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Table of intents, descriptions and sampled synthetic examples generated using gpt-3.5-turbo

See Table 8 (ATIS), Table 9 (SNIPS) and Table 10 (CLINIC).

A.2 Full table of results for approach using synthetic examples generated using gpt-3.5-turbo

See Table 11.

A.3 Table of averaged mean and standard deviation statistics for examples generated using gpt-4

See Table 12.

Intent	Description	Synthetic Examples
abbreviation	user is asking what an abbreviation stands for or mean	“what does eta stand for?" “can you tell me the meaning of atc?" “what is the abbreviation vfr referring to?"
aircraft	user is asking about an aircraft	“what is the maximum speed of this aircraft?" “can you provide me with the dimensions of the aircraft?" “how many passengers can this aircraft accommodate?"
airfare	user is asking about fares, costs or airfares	“what are the airfare options for a round-trip flight from new york to los angeles?" “can you provide me with the cost of a first-class airfare from london to paris?" “how much does it usually cost for a one-way airfare from tokyo to sydney?"
airline	user is asking about an airline/airlines	“which airline offers the most affordable tickets from los angeles to new york?" “can you recommend any airlines that provide extra legroom for tall passengers?" “what are the baggage restrictions for this airline?"
airport	user is asking about an airport/airports	“which airports in new york have direct flights to los angeles?" “can you provide me with information about the nearest airport to my current location?" “how long does it take to get from the city center to heathrow airport?"
capacity	user is asking about capacity (of an aircraft)	“what is the seating capacity of a boeing 747 aircraft?" “can you tell me the maximum passenger capacity of a airbus a380?" “what is the cargo capacity of a cessna 172 aircraft?"
cheapest	user is asking about the cheapest (fare)	“can you find me the cheapest flight from new york to los angeles?" “i need the cheapest airfare available for a one-way trip from london to barcelona." “what is the cheapest flight i can get from chicago to miami during the christmas holidays?"
city	user is asking about a city or place	“can you provide me with flight options to new york city?" “what are the popular attractions in san francisco?" “which airlines operate flights to tokyo?"
day_name	user is asking about a day (of the week)	“which day of the week is the best to book a flight?" “can you tell me the day of the week for my flight to new york?" “what is the departure day for the flight to london?"
distance	user is asking for the distance between places/locations	“what is the distance between new york and los angeles?" “calculate the distance from london to paris." “how far is it from sydney to melbourne?"
flight	user is asking about available flights	“what flights are available from new york city to los angeles tomorrow?" “can you please check if there are any direct flights from london to tokyo?" “i need to book a one-way flight from chicago to miami on the 15th of june."
flight_no	user is asking about a flight number	“what is the flight number for the flight from new york to london?" “can you provide me with the flight number for the 6:00 am departure to los angeles?" “i need to know the flight number for the red-eye flight to chicago."
flight_time	user is asking about departue time or schedule for a flight	“what is the flight time for the next available flight to new york?" “can you tell me the departure time for flight 123 to london?" “i need to know the schedule for flights leaving tomorrow morning."
ground_fare	user is asking about the ground fare at a destination	“what is the average ground fare in los angeles?" “can you provide information about ground fares in paris?" “how much should i expect to pay for ground transportation in london?"
ground_service	user is asking about ground service at a location	“what are the available ground services at this airport?" “can you provide me with information about ground services at the destination airport?" “is there wheelchair assistance available as part of the ground services?"
meal	user is asking about meals/catering	“what meal options are available for the flight?" “can i request a vegetarian meal for my flight?" “do you have any special meals for passengers with dietary restrictions?"
quantity	user is asking about the quantity/amount of something	“how many flight attendants are there on this flight?" “could you tell me the total weight of the luggage allowed per passenger?" “how many passengers are currently on board the plane?"
restriction	user is asking about restrictions	“can you please provide me with the baggage restrictions for my upcoming flight?" “what are the restrictions on carrying liquids in my hand luggage?" “are there any age restrictions for children traveling alone on your flights?"

Table 8: Intents, descriptions and synthetic examples for the ATIS dataset.

Intent	**Description**	**Synthetic Examples**
AddToPlaylist	user wants to add a song to a playlist	“hey, can you please add this new release to my workout playlist?" “add the latest hit by taylor swift to my party playlist, please." “can you include this classic rock track in my road trip playlist?"
BookRestaurant	user wants to book/make a reservation at a restaurant	“can you help me book a table at a fancy restaurant for this saturday?" “i would like to make a reservation for two at the most popular restaurant in town." “what is the best way to book a restaurant online?"
GetWeather	user wants to know about the weather	“what will be the weather like tomorrow?" “can you provide me with a detailed weather forecast for the next week?" “is it going to rain today?"
PlayMusic	user wants to play a song	“hey, playmusic! can you play 'shape of you' by ed sheeran?" “playmusic, please play some soothing music to help me relax." “i'm in the mood for some throwback tunes. playmusic, can you play 'don't stop believin' by journey?"
RateBook	user wants the rating of/to rate a book	“can anyone recommend a ratebook website where I can find reviews and ratings for the latest bestsellers?" “what's the highest-rated ratebook on the market right now? i want to make sure i'm picking something worthwhile." “i'd like some suggestions for popular ratebooks in the fantasy genre. any recommendations?"
SearchCreativeWork	user wants to find a creative work (book, song etc.)	“can you help me search for a creative work that is similar to harry potter?" “i'm looking for a book recommendation, search for a creative work with a thrilling mystery plot." “find me a song that has won multiple awards and has a catchy melody."
SearchScreeningEvent	user wants to know when a movie is on/screening time of a movie	“when is the next screening event for the movie avengers: endgame?" “what are the screening times for the romantic comedy crazy, stupid, love?" “can you tell me the showtimes for the movie joker in theaters nearby?"

Table 9: Intents, descriptions and synthetic examples for the SNIPS dataset.

Intent	Description	Synthetic Examples
timezone	user is asking about timezone	“can you please tell me the current timezone in new york city?" “what is the timezone difference between san francisco and tokyo?" “i need to know the exact timezone utc offset for london."
fun_fact	user wants to know a fun fact	“tell me a fun fact about the eiffel tower!" “i need a fun fact to impress my friends at dinner tonight." “give me a fun fact about dolphins, please."
date	user wants to know the current date	“what is today’s date?" “can you tell me the current date?" “i need to know the date, please."
what_can_i_ask_you	user wants to know what can be asked	“what kind of questions are appropriate to ask you?" “can you provide me with a list of topics that i can ask you about?" “i’m curious, what type of questions am i allowed to ask in this category?"
todo_list_update	user wants to update or change todo list	“how can i edit my current todo list?" “can you show me how to modify my existing tasks in the todo list?" “is there a way to change the priority of items on the todo list?"
bill_balance	user wants to know their bill balance	“what is my current bill balance?" “can you please provide the details of my bill balance?" “i need to know how much is due on my bill."
schedule_meeting	user wants to schedule meeting	“can you help me schedule a meeting for next week?" “i need assistance in setting up a meeting with our new client." “how do i go about scheduling a team meeting for tomorrow?"
routing	user wants to know about routing number	“what is a routing number and why is it important for banking?" “how can i find the routing number for my bank account?" “can you explain the specific purpose of a routing number in online transactions?"
food_last	user wants to know how long a food lasts	“how long can i safely keep cooked chicken in the refrigerator?" “what is the shelf life of fresh milk at room temperature?" “can you give me some tips on how to extend the life of avocados?"
bill_due	user wants to know when a bill is due	“hey, can you remind me when my electricity bill is due?" “what’s the due date for my credit card bill this month?" “i need to know when my phone bill is due. can you help me with that?"
time	user is asking for the time	“what is the current time?" “could you please tell me what time it is?" “do you have the time?"
freeze_account	user wants to freeze their account	“how can i freeze my account temporarily?" “i need to put a hold on my account, can you assist me?" “please freeze my account until further notice."
rollover_401k	user wants to know about 401k rollover	“how can i rollover my 401k into a new retirement account?" “can you explain the process of a 401k rollover to me?" “what are the benefits of doing a rollover with my 401k?"
travel_alert	user wants to know about travel alerts	“are there any current travel alerts that i should be aware of?" “notify me if there are any travel alerts for my upcoming destination." “can you provide me with the latest travel alerts for international travel?"
translate	user wants to translate	“can you translate this document from english to french?" “excuse me, i need assistance translating this menu into spanish." “how can i translate this phrase into italian?"

Table 10: Intents, descriptions and synthetic examples for 15 intents from the CLINIC dataset.

	Model	ATIS			SNIPS			CLINIC		
		Acc	F1	Mean	Acc	F1	Mean	Acc	F1	Mean
$n = 1$	InstructOR$_{Large}$	32.77	23.99	28.38	72.60	69.26	70.93	56.94	53.71	55.32
	E5-v2$_{Base}$	27.01	19.30	23.16	70.28	66.52	68.40	50.05	47.21	48.63
	E5-v2$_{Large}$	29.50	19.12	24.31	68.09	64.41	66.25	47.24	44.54	45.89
	Multilingual-E5$_{Large}$	23.85	18.37	21.11	64.02	60.24	62.13	45.68	43.54	44.61
	E5$_{Large}$	28.57	20.22	24.40	69.35	66.13	67.74	54.44	51.38	52.91
	OpenAI-Ada-002	30.86	19.40	25.13	75.35	72.78	74.07	57.70	54.42	56.06
	GTE$_{Small}$	25.87	20.15	23.01	65.42	62.17	63.80	51.37	48.41	49.89
	GTE$_{Base}$	25.34	20.33	22.83	69.09	65.89	67.49	53.10	50.04	51.57
	GTE$_{Large}$	29.94	21.83	25.88	70.02	66.56	68.29	54.95	51.72	53.34
	BGE$_{Small}$	27.44	21.32	24.38	66.60	62.76	64.68	52.69	49.56	51.13
	BGE$_{Base}$	24.57	20.62	22.59	70.39	66.52	68.46	55.24	52.21	53.72
	BGE$_{Large}$	33.97	23.83	28.90	71.31	67.29	69.30	58.17	54.73	56.45
$n = 3$	InstructOR$_{Large}$	39.20	29.25	34.22	76.71	72.39	74.55	67.88	64.84	66.36
	E5-v2$_{Base}$	35.75	26.97	31.36	76.25	71.56	73.90	63.52	60.63	62.08
	E5-v2$_{Large}$	40.41	27.85	34.13	75.68	70.98	73.33	62.35	59.47	60.91
	Multilingual-E5$_{Large}$	25.07	25.90	25.48	75.67	70.93	73.30	60.56	58.19	59.37
	E5$_{Large}$	37.33	29.64	33.48	74.57	70.24	72.40	67.18	64.25	65.72
	OpenAI-Ada-002	46.96	26.53	36.74	82.42	80.27	81.34	68.77	65.77	67.27
	GTE$_{Small}$	24.50	26.95	25.72	71.00	67.40	69.20	62.38	59.16	60.77
	GTE$_{Base}$	30.05	27.82	28.93	74.57	70.63	72.60	64.69	61.76	63.23
	GTE$_{Large}$	40.40	29.40	34.90	75.04	71.23	73.14	65.78	62.67	64.23
	BGE$_{Small}$	29.24	27.49	28.37	73.49	68.98	71.23	64.59	61.72	63.16
	BGE$_{Base}$	28.35	27.00	27.67	73.83	69.23	71.53	66.59	63.66	65.13
	BGE$_{Large}$	38.30	28.14	33.22	74.83	70.09	72.46	68.05	64.62	66.34
$n = 5$	InstructOR$_{Large}$	41.77	32.86	37.31	78.36	74.08	76.22	70.30	67.51	68.90
	E5-v2$_{Base}$	34.49	28.76	31.63	78.53	73.47	76.00	66.75	63.94	65.34
	E5-v2$_{Large}$	36.82	29.53	33.17	78.02	73.66	75.84	65.70	62.76	64.23
	Multilingual-E5$_{Large}$	31.29	29.28	30.29	76.21	72.18	74.19	64.36	61.78	63.07
	E5$_{Large}$	37.24	32.79	35.01	76.04	71.20	73.62	69.63	66.62	68.13
	OpenAI-Ada-002	45.01	28.38	36.70	84.56	82.60	83.58	70.81	68.03	69.42
	GTE$_{Small}$	32.92	30.05	31.48	73.21	69.16	71.18	65.63	62.58	64.10
	GTE$_{Base}$	29.90	30.02	29.96	76.54	72.13	74.33	67.11	63.95	65.53
	GTE$_{Large}$	41.92	32.41	37.17	75.73	71.18	73.45	68.48	65.38	66.93
	BGE$_{Small}$	35.33	32.64	33.99	72.85	68.06	70.46	67.15	64.35	65.75
	BGE$_{Base}$	27.94	29.49	28.72	76.61	71.90	74.25	69.42	66.52	67.97
	BGE$_{Large}$	35.79	32.38	34.08	76.26	71.00	73.63	70.68	67.64	69.16
$n = 10$	InstructOR$_{Large}$	47.38	33.77	40.58	80.58	76.50	78.54	72.37	69.68	71.03
	E5-v2$_{Base}$	37.04	32.17	34.60	80.31	74.92	77.61	69.59	66.86	68.23
	E5-v2$_{Large}$	46.80	32.53	39.66	79.11	74.31	76.71	68.65	65.70	67.17
	Multilingual-E5$_{Large}$	30.88	32.70	31.79	78.71	74.43	76.57	67.87	65.39	66.63
	E5$_{Large}$	41.44	34.74	38.09	77.83	73.35	75.59	72.42	69.62	71.02
	OpenAI-Ada-002	46.60	32.90	39.75	85.57	83.46	84.51	73.30	70.60	71.95
	GTE$_{Small}$	32.71	33.53	33.12	74.77	70.42	72.59	67.48	64.56	66.02
	GTE$_{Base}$	28.05	31.23	29.64	77.35	72.76	75.06	69.50	66.44	67.97
	GTE$_{Large}$	45.05	35.25	40.15	76.29	71.67	73.98	69.86	66.90	68.38
	BGE$_{Small}$	36.24	34.44	35.34	75.95	71.13	73.54	68.96	66.27	67.61
	BGE$_{Base}$	31.14	31.62	31.38	78.15	73.07	75.61	71.48	68.73	70.10
	BGE$_{Large}$	43.19	35.56	39.38	77.77	72.44	75.10	72.36	69.39	70.88
$n = 15$	InstructOR$_{Large}$	40.59	35.40	37.99	80.57	75.75	78.16	73.10	70.54	71.82
	E5-v2$_{Base}$	42.17	34.44	38.31	80.25	74.65	77.45	70.18	67.50	68.84
	E5-v2$_{Large}$	47.71	33.67	40.69	79.86	74.66	77.26	69.70	66.69	68.19
	Multilingual-E5$_{Large}$	28.31	33.48	30.89	79.91	75.32	77.61	69.31	66.76	68.03
	E5$_{Large}$	42.42	36.31	39.36	78.02	73.00	75.51	73.13	70.26	71.69
	OpenAI-Ada-002	48.13	34.26	41.20	87.04	85.03	86.03	73.97	71.36	72.66
	GTE$_{Small}$	38.54	34.38	36.46	75.03	70.32	72.68	68.63	65.60	67.12
	GTE$_{Base}$	33.68	32.35	33.02	78.27	73.56	75.92	69.86	66.73	68.29
	GTE$_{Large}$	37.98	34.38	36.18	77.78	72.93	75.36	70.51	67.62	69.07
	BGE$_{Small}$	28.06	34.30	31.18	75.43	70.54	72.98	70.20	67.56	68.88
	BGE$_{Base}$	27.20	31.08	29.14	78.92	73.65	76.29	71.93	69.15	70.54
	BGE$_{Large}$	42.22	37.06	39.64	78.76	73.43	76.10	73.17	70.24	71.71

Table 11: Results per model using k synthetic examples averaged across 20 samples.

k	Metric	ATIS		SNIPS		CLINIC	
		μ	σ	μ	σ	μ	σ
$k = 1$	Mean	24.51	10.15	67.63	5.48	51.63	5.13
	Δ_{Label}	-7.19	-2.58	-8.68	-1.05	-15.56	0.08
	Δ_{Desc}	-27.38	6.37	-19.29	2.46	-22.92	2.12
$k = 3$	Mean	31.19	8.61	73.25	4.49	63.71	2.76
	Δ_{Label}	-0.51	-4.11	-3.06	-2.04	-3.47	-2.29
	Δ_{Desc}	-20.70	4.84	-13.66	1.47	-10.83	-0.25
$k = 5$	Mean	33.29	7.90	74.73	4.16	66.54	2.35
	Δ_{Label}	1.59	-4.82	-1.57	-2.37	-0.64	-2.70
	Δ_{Desc}	-18.60	4.13	-12.18	1.14	-8.00	-0.67
$k = 10$	Mean	36.12	7.51	76.28	3.49	68.92	2.08
	Δ_{Label}	4.42	-5.21	-0.02	-3.04	1.73	-2.97
	Δ_{Desc}	-15.77	3.73	-10.63	0.48	-5.63	-0.94
$k = 15$	Mean	36.17	7.13	76.78	3.75	69.74	1.93
	Δ_{Label}	4.47	-5.59	0.48	-2.78	2.55	-3.12
	Δ_{Desc}	-15.72	3.36	-10.13	0.73	-4.81	-1.09

Table 12: Averaged mean of accuracy and macro-f1 scores experiments conducted across 20 samples and 12 models using k number of synthetic examples per intent class generated using `gpt-4-1106-preview`. Δ_{Label} and Δ_{Desc} are differences to the averaged performance of methods using tokenized labels and intent descriptions respectively.