Conformal Intent Classification and Clarification for Fast and Accurate Intent Recognition

Floris den Hengst
Vrije Universiteit Amsterdam
f.den.hengst@vu.nl

Ralf Wolter
ING Group NV
Ralf.Wolter@ing.com

Patrick Altmeyer
TU Delft
P.Altmeyer@tudelft.nl

Arda Kaygan
ING Group NV
Arda.Kaygan@ing.com

Abstract

We present Conformal Intent Classification and Clarification (CICC), a framework for fast and accurate intent classification for task-oriented dialogue systems. The framework turns heuristic uncertainty scores of any intent classifier into a clarification question that is guaranteed to contain the true intent at a pre-defined confidence level. By disambiguating between a small number of likely intents, the user query can be resolved quickly and accurately. Additionally, we propose to augment the framework for out-of-scope detection. In a comparative evaluation using seven intent recognition datasets we find that CICC generates small clarification questions and is capable of out-of-scope detection. CICC can help practitioners and researchers substantially in improving the user experience of dialogue agents with specific clarification questions.

1 Introduction

Intent classification (IC) is a crucial step in the selection of actions and responses in task-oriented dialogue systems. To offer the best possible experience with such systems, IC should accurately map user inputs to a predefined set of intents. A widely known challenge of language in general, and IC specifically, is that user utterances may be incomplete, erroneous, and contain linguistic ambiguities.

Although IC is inherently challenging, a key strength of the conversational setting is that disambiguation or *clarification* questions (CQs) can be posed (Purver et al., 2003; Alfieri et al., 2022). Posing the right CQ at the right time results in a faster resolution of the user query, a more natural conversation, and higher user satisfaction (van Zeelt et al., 2020; Keyvan and Huang, 2022; Siro et al., 2022). CQs have been considered in the context of information retrieval (Zamani et al., 2020) but have received little attention in the context of task-oriented dialogue.

Deciding when to ask a CQ and how to pose it are challenging tasks (DeVault and Stone, 2007; Keyvan and Huang, 2022). First, it is not clear when the system can safely proceed under the assumption that the true intent was correctly identified. Second, it is not clear when the model is too uncertain to formulate a CQ (Cavalin et al., 2020). Finally, it is unclear what the exact information content of the clarification question should be.

We present Conformal Intent Classification and Clarification (CICC), a framework for deciding when to ask a CQ, what its information content should be, and how to formulate it. The framework uses conformal prediction to turn a models' predictive uncertainty into prediction sets that contain the true intent at a predefined confidence level (Shafer and Vovk, 2008; Angelopoulos et al., 2023). The approach is agnostic to the intent classifier, does not require re-training of this model, guarantees that the true intent is in the CQ, allows for rejecting the input as too ambiguous if the model is too uncertain, has interpretable hyperparameters, generates clarification questions that are small and is amenable to the problem of detecting out-of-scope inputs.

In a comparative evaluations with seven data sets and three IC models, we find that CICC outperforms heuristic approaches to predictive uncertainty quantification in all cases. The benefits of CICC are most prominent for ambiguous inputs, which arise naturally in real-world dialogue settings (Zamani et al., 2020; Larson et al., 2019).

2 Related Work

We discuss related work on ambiguity and uncertainty detection within IC and CP with language models.

Clarification Questions Various works acknowledge the problem of handling uncertainty in intent classification and to address it with CQs. Dhole

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2412–2432

(2020) proposes a rule-based approach for asking discriminative CQs. The approach is limited to CQs with two intents, lacks a theoretical foundation, and provides no intuitive way of balancing coverage with CQ size. Keyvan and Huang (2022) survey ambiguous queries in the context of conversational search and list sources of ambiguity. They mention that clarification questions should be short, specific, and based on system uncertainty. We propose a principled approach to asking short and specific questions based on uncertainty of any underlying intent classifier for the purposes of task-oriented dialogue.

Alfieri et al. (2022) propose an approach for asking a CQ containing a fixed top-k most likely intents with intent-specific uncertainty thresholds. This approach does not come with any theoretical guarantees and its hyperparameters need to be tuned on an additional data set whereas our approach comes with guarantees on coverage of the true intent and with intuitively interpretable hyperparameters that can be tuned on the same calibration set. We do not compare directly to this method but include top-k selection in our benchmark.

CQs have been studied in other domains, including information retrieval (Zamani et al., 2020), product description improvement (Zhang and Zhu, 2021), and open question-answering (Kuhn et al., 2023). In contrast to the task-specific domain investigated in this work, these domains leave more room for asking generic questions for clarification and do not easily allow for incorporating model uncertainty. Furthermore, the proposed methods require ad hoc tuning of scores based on heuristic metrics of model uncertainty, and do not provide ways to directly balance model uncertainty with CQ size.

Uncertainty and out-of-scope detection The out-of-scope detection task introduced by Larson et al. (2019) is a different task from the task of handling model uncertainty and ambiguous inputs (Cavalin et al., 2020; Yilmaz and Toraman, 2020; Zhan et al., 2021; Zhou et al., 2021). However, predictive uncertainty is often used in addressing the out-of-scope detection task. Although the tasks of handling ambiguous input and detecting out-of-scope input are different, we briefly discuss approaches that leverage model uncertainty for out-of-scope detection here.

Various out-of-scope detection approaches train an intent classifier and tune a decision boundary based on a measure of the classifier's confidence (Shu et al., 2017; Lin and Xu, 2019; Yan et al., 2020; Hendrycks et al., 2020). Samples for which the predictive uncertainty of the model lies on one side of the boundary are classified as out-of-scope. These approaches use the models' heuristic uncertainty to decide whether an input is out-of-sample whereas we first turn the models' heuristic uncertainty into a prediction with statistical guarantees and then use this prediction to decide when and how to formulate a clarification question. We additionally propose an adaptation of the CICC framework for out-of-scope detection.

Conformal Prediction on NLP tasks Conformal Prediction has been used in several NLP tasks, including sentiment classification by Maltoudoglou et al. (2020), named-entity recognition by Fisch et al. (2022) and paraphrase detection by Giovannotti and Gammerman (2021). However, the application to intent classification, task-oriented dialogue and the combination with CQs presented here is novel to our knowledge.

3 Methodology

We address the problem of asking CQs in task oriented dialogue systems in the following way. We take a user utterance and a pre-trained intent classifier, and then return an appropriate response based on the predictive uncertainty of the model. Algorithm 1 lists these steps, and an example input is presented in Figure 1. In this section we describe and detail the components of CICC. We start by providing a background on conformal prediction.

3.1 Conformal Prediction

Conformal Prediction is a framework for creating statistically rigorous prediction sets from a heuristic measure of predictive uncertainty of a classifier (Shafer and Vovk, 2008; Angelopoulos et al., 2023). We here focus on split conformal prediction as it does not require any retraining of the underlying model, and refer to it simply as conformal prediction from here on out.

For a classification task with classes $\mathcal{Y} : \{1, \ldots, K\}$, a test input $X_t \in \mathcal{X}$ with label $Y_t \in \mathcal{Y}$, and a user-defined error level $\alpha \in [0, 1)$, CP returns a set $\mathcal{C}(X_t) \subseteq \mathcal{Y}$ for which the following holds (Vovk et al., 1999) even when using a finite amount of samples:

$$\mathbb{P}\left(Y_t \in \mathcal{C}(X_t)\right) \geq 1 - \alpha \tag{1}$$

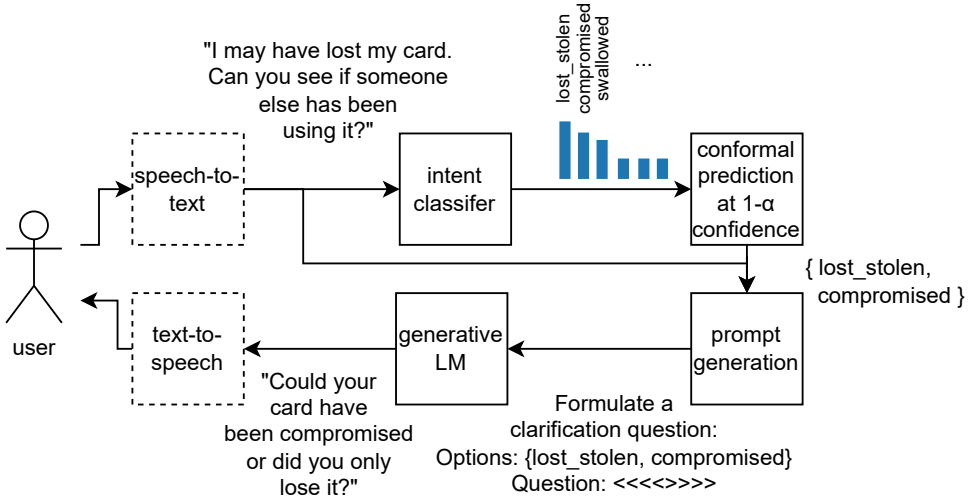

Figure 1: The conformal intent classification and clarification interaction loop.

If e.g. $\alpha = 0.01$ the set $\mathcal{C}(X_t)$ is therefore *guaranteed* to contain the true Y_t in 99% of test inputs.

Conformal prediction uses a heuristic measure of uncertainty of a pretrained model and a modestly sized calibration set to generate prediction sets. Formally, we assume a held-out calibration set $D : \{(X_i, Y_i)\}$ of size n, a pre-trained classifier $\hat{f} : \mathcal{X} \rightarrow \mathbb{R}^K$, and a nonconformity function $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ that returns heuristic uncertainty scores where larger values express higher uncertainty. An example of a nonconformity function for a neural network classifier is one minus the softmax outputs of the true class:

$$s(X_i) := 1 - \hat{f}(X_i)_{Y_i}. \tag{2}$$

This score is high when the softmax score of the true class is low, i.e., when the model is badly wrong.

The nonconformity function s is evaluated on D to generate a set of nonconformity scores $\mathcal{S} = \{s(X_i, Y_i)\}$. Next, the quantile $\hat{q}$ of the empirical distribution of $\mathcal{S}$ is determined so that the desired coverage ratio $(1 - \alpha)$ is achieved. This can be done by choosing $\hat{q} = \lceil (n+1)(1-\alpha) \rceil / n$[1] where $\lceil \cdot \rceil$ denotes the ceiling function. Then, for a given test input X_t, all classes $y \in \mathcal{Y}$ with high enough confidence are included in a prediction set $\mathcal{C}(X_t)$:

$$\mathcal{C}(X_t) := \{y : s(X_t, y) \leq \hat{q}\}. \tag{3}$$

This simple procedure guarantees that (1) holds i.e. that the true Y_t is in the set at the specified confidence $1 - \alpha$. Note the lack of retraining or ensembling of classifiers, that the procedure requires little compute and that D can be relatively small as long as it contains a fair number of examples for all classes and is exchangeable[2] with the test data (Papadopoulos et al., 2002).

There are various implementations of conformal prediction with different nonconformity functions and performance characteristics. The most simple approach is known as *marginal* conformal prediction and it uses the nonconformity function in (2). Marginal conformal prediction owes its names from adhering to the guarantee (1) marginalized over $\mathcal{X}$ and $\mathcal{Y}$, i.e. it satisfies the coverage requirement (1) on average, rather than e.g. for a particular input X_t. Marginal CP can be implemented following the steps described previously: (i) compute nonconformity scores S using (2), (ii) obtain $\hat{q}$ as described previously, and (iii) construct a prediction set using (3) at test time. A benefit of this approach is that it generates prediction sets with the smallest possible prediction set size on average. A limitation is that its prediction set sizes may not reflect hardness of the input (Sadinle et al., 2019).

Alternatively, one can ensure conditional adherence to (1) with so-called conditional or adaptive conformal predictors. A benefit of conditional approaches is that higher model uncertainty results in larger prediction sets. However, a downside is that these sets are expected to be larger on average than those obtained with a marginal approach.Romano et al. (2020) introduce a conditional CP approach that consists of broadly the same steps as marginal CP but with a different nonconformity function s and a different prediction set construction. First we define a permutation $\pi(X)$ of $\{1 \ldots K\}$ that sorts

[1] this is essentially the $\hat{q}$ quantile with a minor adjustment

[2] distributed identically but not necessarily independently

$\hat{f}(X)$ in descending order. Conditional CP can defined as: (i) sum all predictor outputs $\hat{f}(X_i)_k$ for all $\{k \in K | \hat{f}(X_i)_k \geq \hat{f}(X_i)_{Y_i}\}$, (ii) obtain $\hat{q}$ as before, and (iii) include all for a test input X_t:

$$\mathcal{C}(X_t) := \{\pi_1(X_t), \ldots, \pi_k(x)\}, \quad (4)$$

where

$$k = \sup \left\{ k' : \sum_{j=1}^{k'} \hat{f}(X_t)_{\pi_j(X_t)} < \hat{q} \right\} + 1. \quad (5)$$

Angelopoulos et al. (2021) introduce an approach with a term to regularize the prediction set size: their approach is therefore known as Regularized Adaptive Prediction Sets (RAPS). It effectively adds an increasing penalty to the ranked model outputs in the first step of conditional CP in order to promote smaller prediction sets where possible. Since the second and third step are similar to conditional CP, its prediction sets still adhere to the coverage guarantee (1).

In general, a suitable conformal prediction technique strikes the right balance between three desiderata: (i) adhering to the coverage requirement in (1), (ii) producing small prediction sets and (iii) adaptivity. Whereas the former two can be measured easily, metrics for adaptivity require some more care. Angelopoulos et al. (2021) introduce a general-purpose metric for adaptivity. It is based on the coverage and referred to as the size-stratified classification (SSC) score:

$$\text{SSC} = \min_{b \in \{1,\ldots,K\}} \frac{1}{|\mathcal{I}_b|} \sum_{i \in \mathcal{I}_b} \mathbb{1}\left\{Y_i \in \mathcal{C}\left(X_i\right)\right\} \quad (6)$$

for a classification task defined as above and $\mathcal{I}_b \subset \{1, \ldots, n\}$ the set of inputs with prediction set size b, i.e. $\mathcal{I}_b := \{X_i, |\mathcal{C}(X_i)| = b\}$.

Within CICC, conformal prediction is applied to a pre-trained intent classifier to create a set of intents that contains the true user intent at a predefined confidence for any user utterance. The sets are then used in making a decision on when to ask a clarification question and how to formulate it. We continue to discuss when and how such questions are asked based on Algorithm 1 in the following section.

3.2 When to Ask a Clarification Question

For a user utterance X, a pre-trained intent classifier $\hat{f}$ and a nonconformity function s, we generate a prediction set that covers the true user intent with confidence $1 - \alpha$ (Algorithm 1, ln 1). If the set contains a single intent, the model is confident that the true intent has been detected and the dialogue can be handled as usual (ln 2-3).

If the set contains many intents, that is, more than a user-specified threshold $th \in \mathbb{N}_{>0}$, then there is no reasonable ground for formulating a clarification question. Instead, a generic request to rephrase the question can be asked (ln 4-5), or a hand-over to a human operator could be implemented here. In the remaining case, i.e. if the prediction set is of reasonable size, a CQ is asked (ln 6-7).

CICC comes with two parameters to control when a CQ should be asked. Both have clear semantics and can be interpreted intuitively. The first is the threshold th that controls when the input is too ambiguous to ask a CQ (Algorithm 1 ln 4-5). This parameter is set by the chatbot owner on the basis of best practices in, and knowledge of chat- and voicebot interaction patterns. In general, this number should remain small to reduce the cognitive load on users. We advise to set this value no higher than seven (Miller, 1956; Plass et al., 2010).

The second parameter is the error rate α. It controls the trade-off between the prediction set size and how certain we want to be that the prediction set covers the true intent. As $\alpha \to 0$, our confidence that the true intent is included in the set grows, but so does the size of the prediction set. Because conformal prediction is not compute intensive, α can be set empirically. Thus, CICC provides a means of selecting between *achievable* trade-offs between prediction set sizes and error rates. We continue to discuss how specific CQs are

Algorithm 1 CICC algorithm

Input: utterance X, classifier $\hat{f}$, chat/voice-bot c, calibration set D, generative LM g
Parameters: error rate α, threshold th, ambiguity response a
Output: response R

1: set $\leftarrow$ conformal prediction$\left(\hat{f}(X), D, \alpha\right)$
2: **if** |set| == 1 **then**
3: $\quad R \leftarrow c$(set.get()). {bot response}
4: **else if** |set| $> th$ **then**
5: $\quad R \leftarrow a$. {input too ambiguous}
6: **else**
7: $\quad R \leftarrow g$(set, X) {clarification question}
8: **end if**

formulated in CICC.

3.3 Generating a Clarification Question

When a CQ is in order (ln 6-7 in Alg. 1), it needs to be formulated. We propose to generate a CQ based on the original input X and the prediction set, as it is guaranteed to contain the true intent at a typically high level of confidence. Because the alternatives in the CQ are the most likely intents according to the model, and because the number of alternatives in the CQ corresponds to the models' uncertainty, asking a CQ provides a natural way of communicating model uncertainty to the user while quickly determining the true user intent.

CICC makes no assumptions about the approach for generating a CQ. Anything from hardcoded questions, templating, or a generative LM can be used. However, we recognize that the number of possible questions is large: it consists of the powerset of all n intents up to size th excluding sets of size one and zero. Therefore, we opt to use a generative LM in our solution.

We prompt the LM to formulate a clarification question by giving it some examples of clarification questions for a set of example intents to disambiguate between. We additionally provide the original utterance X to enable the formulation of CQ relative to the original utterance. See Appendix A for details.

3.4 Out-of-scope Detection

Ambiguity is a part of natural language which could lead to model uncertainty. Specific reasons for uncertainty in intent recognition are inputs that are very short and long, imprecise and incomplete inputs, etc. However, a particularly interesting type of uncertainty stems from inputs that represent intent classes that are not known at training time (Zhan et al., 2021). These inputs are referred to as out-of-scope (OOS) and detecting these inputs can be seen as a binary classification task for which data sets with known OOS samples have been developed.

CICC rejects inputs about which the model is too uncertain (Algorithm 1, ln 5) and this naturally fits with the OOS detection task as follows: we can view a rejection of an input as a classification of that input as OOS. Therefore, although handling ambiguity in the model gracefully and detection OOS inputs are separate challenges, vanilla CICC implements a form of OOS detection.

	samples	intents
ACID (Acharya and Fung, 2020)	22172	175
ATIS (Hemphill et al., 1990)	5871	26
B77 (Casanueva et al., 2020)	13083	77
B77-OOS	16337	78
C150-IS (Larson et al., 2019)	18025	150
C150-OOS (Larson et al., 2019)	19025	151
HWU64 (Liu et al., 2021)	25716	64
IND	∼20k	61
MTOD (eng) (Schuster et al., 2019)	43323	12

Table 1: Characteristics of datasets used

Additionally, the CICC framework can be leveraged for OOS detection if OOS samples are known at calibration time. Specifically, we can optimize parameters α and th to maximize predictive performance expressed by some suitable metric such as the F1-score on the calibration set. OOS samples can be obtained from other intent recognition data sets in other domains. This practice is described in detail by e.g. (Zhan et al., 2021) under the name of open-domain outliers. We refer to versions of CICC which have been optimized for F1-score in this way as CICC-OOS.

4 Experimental Setup

This section lists the experiments performed to comparatively evaluate CICC across seven data sets and on three IC models[3].

Data We evaluate CICC on six public intent recognition data sets in English and an additional real-life industry data set (IND) from the banking domain in the Dutch language. Table 1 shows the data sets and their main characteristics. All data sets were split into train-calibration-test splits of proportions 0.6-0.2-0.2 with stratified sampling, except for the ATIS data set in which stratified sampling is impossible due to the presence of intents with a single sample. Random sampling was used for this data set instead. We use an in-scope version (C150-IS) of the 'unbalanced' data set by Larson et al. (2019) in which all out-of-scope samples have been removed.

For evaluation on out-of-scope (OOS) detection, we use two datasets: a version of C150 with all OOS samples divided over the calibration and test splits, and no OOS samples in the train split (C150-OOS), and a version of B77 with so-called open-domain outliers in which samples from the ATIS dataset make up half of the samples in the calibra-

[3] https://github.com/florisdenhengst/cicc

tion and test splits to represent OOS inputs (B77-OOS) (Zhan et al., 2021).

Models We employ fine-tuned BERT by Devlin et al. (2019) for all public data sets and a custom model similar to BERT for the IND data set (Alfieri et al., 2022). We base the nonconformity scores on the softmax output in these settings. In order to test performance on a commercial offering, we additionally evaluate using DialogflowCX (DFCX) on the B77 data set.[4] This commercial offering outputs heuristic certainty scores in the range $[0, 100]$ for the top five most certain recognized intents. These outputs were normalized to sum to 1, all other scores were set to 0 to determine the nonconformity scores.

Baselines In practice CQs can be formulated using heuristics (Alfieri et al., 2022). We compare CICC to the following baselines using the models' heuristic uncertainty scores:

B1 select all intents with score $> 1 - \alpha$, select the top $k = 5$ if this selection is empty.

B2 select all intents with a score $> 1 - \alpha$.

B3 select the top $k = 5$ intents.

Metrics We evaluate the approaches on a set of metrics that together accurately convey the added benefit of asking a confirmation question. We use the *size* of the prediction set $\mathcal{C}(X_i)$ and how often the input is rejected as too ambiguous for the model (Algorithm 1, ln 5). For a test set of size n:

$$\text{Amb} := \frac{1}{n} \sum_{i=0}^{n} \begin{cases} 1 & \text{if } |\mathcal{C}(X_i)| \geq th \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

First, we report how often the true intent is detected for the $m \leq n$ inputs that are not rejected (Algorithm 1, lns 3 and 5). This metric is known as coverage (cov) and can be seen as a generalisation of accuracy for set-valued predictions:

$$\text{Cov} := \frac{1}{m} \sum_{i=0}^{m} \mathbb{1}_{\mathcal{C}(X_i)}(Y_i)\,. \tag{8}$$

Second, we report the average size of the clarification questions for accepted inputs (Algorithm 1, ln 7). This metric can be seen as an analogue to precision for set-valued predictions:

$$|\text{CQ}| = \frac{1}{m} \sum_{i=0}^{m} |\mathcal{C}(X_i)|. \tag{9}$$

[4] https://cloud.google.com/dialogflow/cx/docs

Finally, we report the relative number of times the prediction set is of size one

$$\text{Single} := \frac{1}{m} \sum_{i=0}^{m} \begin{cases} 1 & \text{if } |\mathcal{C}(X_i)| = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

in which case the dialogue can continue as usual (Algorithm 1, ln 3). We additionally report the SSC as defined above in (6).

For out-of-scope detection we report the standard metrics F1-score and AUROC.

Parameters We varied α and found the best settings empirically on the calibration set. We report our key results for the best α and additionally investigate the effect of varying α.

We set the threshold th at seven to avoid excessive cognitive load for users for all experiments, except when using DFCX in which case we set th to four (Miller, 1956; Plass et al., 2010). The reason for this is that DFCX currently only outputs non-zero scores for the top five intents. Hence, the set contains all intents that have a non-zero confidence score with this setting.

We include the following conformal prediction approaches and select an approach that produces the best empirical results in terms of coverage and CQ size: marginal, conditional (also known as adaptive) (Romano et al., 2020) and RAPS (Angelopoulos et al., 2021). Marginal conformal prediction was selected in all experiments, details can be found in Figure 2.

5 Results

Table 2 lists the main results. The first column shows the coverage, i.e. the percentage of test samples in which the ground truth is captured in the prediction set. We see that only CICC and B3 adhere to the requirement of coverage $\geq 1 - \alpha$ in all settings. The second column shows the fraction of test samples for which a single intent is detected. We see that CICC outperforms the baselines that meet the coverage requirement in five out of seven data sets.

The third column lists the average size of the CQ. We see that CICC yields the smallest CQs and that the number of inputs that is deemed too ambiguous is relatively small for CICC. The last column denotes the relative number of inputs that is rejected as too ambiguous. CICC rejects a relatively low number of inputs. Upon inspection, many of these inputs could be classified as different intents based

Setting	$1-\alpha$	th		Cov↑	Single↑	|CQ| ↓	Amb
ACID	.98	7	CICC	.98	.87	**3.01**	.03
			B1	.98	**.88**	5	0
			B2	.95	1	–	0
			B3	.99	0	5	0
ATIS	.99	7	CICC	.99	**.98**	**2.54**	0
			B1	.99	.73	5	0
			B2	.98	1.00	-	0
			B3	1.00	0	5	0
B77/BERT	.97	7	CICC	.98	.73	**2.84**	.04
			B1	.97	**.84**	5	0
			B2	.93	1	–	0
			B3	.98	0	5	0
B77/DFCX	.90	4	CICC	.91	.66	**2.63**	.02
			B1	.95	**.71**	5	.27
			B2	.90	.98	2.26	0
			B3	.97	0	5	1
C150-ID	.99	7	CICC	.99	**.97**	**2.66**	0
			B1	.99	.82	5	0
			B2	.98	1	–	0
			B3	1	0	5	0
HWU64	.95	7	CICC	.95	**.82**	**2.81**	.01
			B1	.97	.70	5	0
			B2	.90	1	–	0
			B3	.98	0	5	0
IND	.90	7	CICC	.91	**.25**	**3.46**	.11
			B1	.88	.42	5	0
			B2	.70	1	–	0
			B3	.91	0	5	0
MTOD	.99	7	CICC	.99	**1**	–	0
			B1	1	.98	5	0
			B2	.99	**1**	–	0
			B3	1	0	5	0

Table 2: Test set results where underline indicates meeting coverage requirement. **Bold** denotes best when meeting this requirement, omitted for last column due to missing ground truth for ambiguous.

Dataset	Algorithm	1-α	th	F1↑	AUROC↑
C150-OOS	CICC	.990	7	.07	.88
	CICC-OOS	.995	6	**.91**	**.97**
B77-OOS	CICC	.970	7	.76	.92
	CICC-OOS	.994	6	**.90**	**.97**

Table 3: Results for the OOS detection task.

on the textual information alone (see Appendix B). For the B77/DFCX setting, we see that B1 predicts a single output frequently, at the cost of rejecting inputs as too ambiguous. This contrasts with CICC, which rejects inputs much less frequently and instead asks a small CQ.

We continue by looking at the results for OOS detection in Table 3. We find that vanilla CICC does not perform well on the OOS detection in comparison to the specialized CICC-OOS variant. The specialized CICC-OOS favours a relatively low α as this simultaneously forces the approach toward large prediction sets for OOS samples and small prediction sets for in-sample inputs. At the same time, using the CICC-OOS settings for parameters α and th in the regular CICC interaction loop would result in relatively many CQs of a relatively large size.

Next, we investigate how different conformal prediction approaches perform for varying levels of α in Figure 2. The top figures show that all conformal prediction approaches enable trading off set size with coverage, a desirable property in practice of intent classification. Looking at the adaptivity (center figures), we see mixed results. A possible explanation for this is in the general-purpose evaluation of adaptivity, which relies on the minimum coverage across classes (see Eq. 6). The data sets used in our experiments contain a relatively low number of examples for some classes and these rare classes may have an outsized effect on the SSC metric. Looking at the bottom figure for each data set, we see that all conformal prediction approaches lie at or above the x=y diagonal: conformal prediction always adheres to the coverage requirement with the marginal approach yielding the smallest average set sizes.

6 Conclusion

We have proposed a framework for detecting and addressing uncertainty in intent classification with conformal prediction. The framework empirically

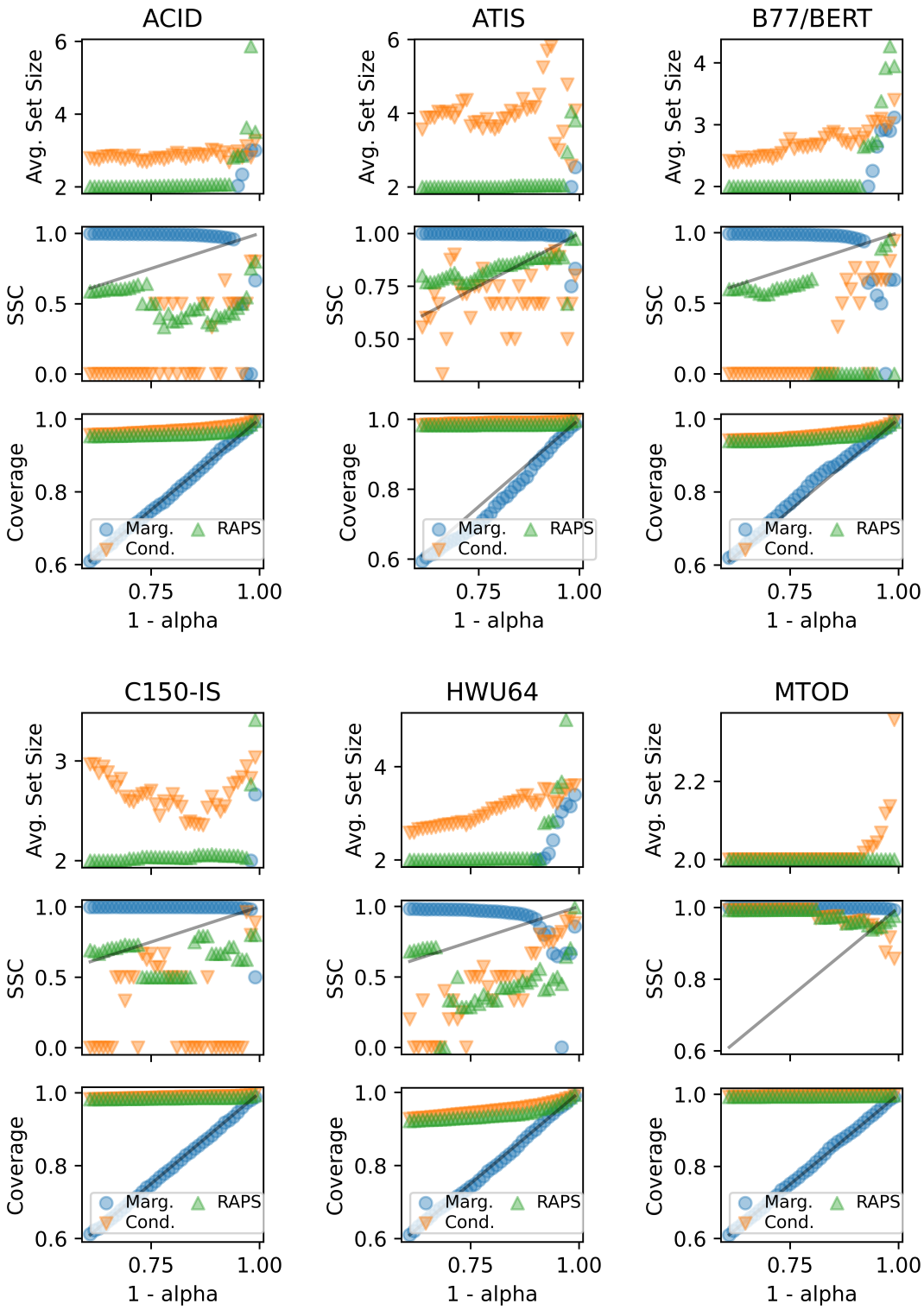

Figure 2: Test set results for varying error rate α.

determines when to ask a clarification question and how that question should be formulated. The framework uses a moderately sized calibration set and comes with intuitively interpretable parameters.

We have evaluated the framework in eight settings, and have found that the framework strictly outperforms baselines across all metrics in six out of eight cases and performs competitively in the other. The framework additionally handles inputs that are too ambiguous for intent classification naturally. We have additionally proposed and evaluated the usage of CICC for out-of-scope detection and found that it is suitable for this.

We finally believe that the framework opens promising avenues for future work, including the usage of intent groups for better adaptivity, an extension to Bayesian models to address data drift and unsupervised OOS with CICC (Fong and Holmes, 2021), to determine conversation stopping rules based on subsequent questions to rephrase or clarify and to combine it with reinforcement learning for, e.g., personalization (Den Hengst et al., 2019, 2020). We believe that CICC and/or conformal prediction may also prove useful in various other tasks, including entity recognition, detecting label errors (Ying and Thomas, 2022) and to empirically identify similar intents.

Limitations

A limitation of the framework is that it relies on a user determining values for the hyperparameters α and th. The former balances model certainty with CQ size. Arguably, this trade-off has to be made in any approach and CICC makes this an explicit choice between achievable trade-offs. The threshold th must be set not to reject too many inputs as too ambiguous while avoiding information overload in the user. We advise setting it to no more than seven based on established insights from cognitive science (Miller, 1956). However, more research on the impact of CQ size on user satisfaction in various context is in order. Another limitation is that the approach does not include a mechanism for stopping the dialogue. We leave the investigation of stopping criteria based on e.g. the number and size of CQs asked during the dialogue for future work. Furthermore, this work did not thoroughly investigate the quality of the CQs produced by the LLM. However, we view the CQ production component as a pluggable component and therefore believe a full-scale evaluation on this to be out-of-scope for this work. Additionally, using CICC for OOS detection requires the presence of OOS labels. While these can be obtained from other data sets using the practice of open-domain outliers (Zhan et al., 2021), fully unsupervised approaches based on e.g. hierarchical Bayesian modeling or with parameters that yield good performance across data sets as hinted at by Table 3. A final limitation is that we applied conformal prediction to the softmax of outputs of uncalibrated neural network outputs. This makes results consistent across settings (including DFCX), but smaller CQs may be achievable by applying Platt scaling prior to conformal prediction calibration (Platt et al., 1999).

Acknowledgements

We thank Mark Jayson Doma and Jhon Cedric Arcilla for their help in obtaining and understanding DialogflowCX model output. We kindly thank the reviewers for their time and their useful comments, without which this work would not have been possible in its current form.

References

Shailesh Acharya and Glenn Fung. 2020. Using optimal embeddings to learn new intents with few examples: An application in the insurance domain. In *KDD 2020 Workshop on Conversational Systems Towards Mainstream Adoption(KDD Converse 2020)*. CEUR-WS.org.

Andrea Alfieri, Ralf Wolter, and Seyyed Hadi Hashemi. 2022. Intent disambiguation for task-oriented dialogue systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 5079–5080.

Anastasios N Angelopoulos, Stephen Bates, et al. 2023. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. 2021. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*.

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

Paulo Cavalin, Victor Henrique Alves Ribeiro, Ana Appel, and Claudio Pinhanez. 2020. Improving out-of-scope detection in intent classification by using embeddings of the word graph space of the classes.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3952–3961.

Floris Den Hengst, Eoin Martino Grua, Ali el Hassouni, and Mark Hoogendoorn. 2020. Reinforcement learning for personalization: A systematic literature review. *Data Science*, 3(2):107–147.

Floris Den Hengst, Mark Hoogendoorn, Frank Van Harmelen, and Joost Bosman. 2019. Reinforcement learning for personalized dialogue management. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 59–67.

David DeVault and Matthew Stone. 2007. Managing ambiguities across utterances in dialogue. In *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue (Decalog 2007)*, pages 49–56.

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.

Kaustubh D Dhole. 2020. Resolving intent ambiguities by retrieving discriminative clarifying questions. *arXiv preprint arXiv:2008.07559*.

Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. 2022. Conformal prediction sets with limited false positives. In *International Conference on Machine Learning*, pages 6514–6532. PMLR.

Edwin Fong and Chris C Holmes. 2021. Conformal bayesian computation. *Advances in Neural Information Processing Systems*, 34:18268–18279.

Patrizio Giovannotti and Alex Gammerman. 2021. Transformer-based conformal predictors for paraphrase detection. In *Conformal and Probabilistic Prediction and Applications*, pages 243–265. PMLR.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751.

Kimiya Keyvan and Jimmy Xiangji Huang. 2022. How to approach ambiguous queries in conversational search: A survey of techniques, approaches, tools, and challenges. *ACM Computing Surveys*, 55(6):1–40.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. CLAM: Selective clarification for ambiguous questions with large language models. In *ICML Workshop Challenges of Deploying Generative AI*.

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.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5496.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, pages 165–183. Springer.

Lysimachos Maltoudoglou, Andreas Paisios, and Harris Papadopoulos. 2020. Bert-based conformal predictor for sentiment analysis. In *Conformal and Probabilistic Prediction and Applications*, pages 269–284. PMLR.

George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. 2002. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 345–356. Springer.

Jan L Plass, Roxana Moreno, and Roland Brünken, editors. 2010. *Cognitive load theory.* Cambridge University Press, New York, NY, US.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. *Current and new directions in discourse and dialogue*, pages 235–255.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. 2019. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234.

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

Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Lei Shu, Hu Xu, and Bing Liu. 2017. Doc: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2911–2916.

Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2022. Understanding user satisfaction with task-oriented dialogue systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2018–2023.

Mickey van Zeelt, Floris den Hengst, and Seyyed Hadi Hashemi. 2020. Collecting high-quality dialogue user satisfaction ratings with third-party annotators. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pages 363–367.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. 1999. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444–453.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert YS 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.

Eyup Halit Yilmaz and Cagri Toraman. 2020. Kloos: Kl divergence-based out-of-scope intent detection in human-to-machine conversations. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 2105–2108.

Cecilia Ying and Stephen Thomas. 2022. Label errors in banking77. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 139–143.

Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020*, pages 418–428.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert YS Lam. 2021. Out-of-scope intent detection with self-supervision and discriminative training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3521–3532.

Zhiling Zhang and Kenny Zhu. 2021. Diverse and specific clarification question generation with keywords. In *Proceedings of the Web Conference 2021*, pages 3501–3511.

Wenxuan Zhou, Fangyu Liu, and Muhao Chen. 2021. Contrastive out-of-distribution detection for pre-trained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

A Appendix: Implementation Details

We used `python v3.10.9` with packages `numpy` and `pandas` for data manipulation and basic calculations, `matplotlib` to generate illustrations, `mapie` for conformal prediction and reproduced these results in Julia and the package `conformalprediction.jl`. We used the `huggingface` API for fine tuning a version of `bert-base-uncased` using the hyperparameters below. For an anonymized version of the code and data see https://anonymous.4open.science/r/cicc-205A.

```
learning_rate = 4.00e-05
warmup_proportion = 0.1
train_batch_size = 32
eval_batch_size = 32
num_train_epochs = 5
```

A.1 Generative Language Model

We use the `eachadea/vicuna-7b-1.1` variant of the LLAMA model using the HuggingFace API for the experiments presented here. We here provide an example prompt:

```
Customers asked an ambiguous question. Complete each set with a disambiguation question.

Set 1: Customer Asked: 'The terminal I paid at wouldn't take my card. Is something wrong?'
Option 1: 'card not working'
Option 2: 'card swallowed'
Disambiguation Question: 'I understand this was about you card. Was is swallowed or not working?'
**END**

Set 2:
Customer Asked: 'I have a problem with a transfer. It didn't work. Can you tell me why?'
Option 1: 'declined transfer'
Option 2: 'failed transfer'
Disambiguation Question: 'I see you are having issues with your transfer. Was your transfer failed or
**END**

Set 3: Customer Asked: 'I transferred some money but it is not here yet'
Option 1: 'balance not updated after bank transfer'
Option 2: 'transfer not received by recipient'
Disambiguation Question:
```

More efforts can be spent on prompt engineering and more advanced generative LMs can be used, which we expect to improve the user satisfaction of CICC. Alternatively, simple text templates can be used. We consider the following alternatives and list some of their expected benefits and downsides:

Templates a simple template-based can be used in which the user is asked to differentiate between the identified intents. Benefits of templates include full control over the chatbot output but a downside is that the CQs will be less varied, possibly sounding less natural and will not refer back to the users' original utterance,

LM without user input when using a LM, it is possible to not incorporate the user input X in the prompt. This has the benefit of blocking any prompt injection but the downside of possibly unnatural CQs due to the inability to refer to the user query,

LM with user input by incorporating the user utterance into the LM prompt for CQ generation, the CQ can refer back to the user's phrasing and particular question, and therefore be formulated in a possibly more natural way.

We believe that more research is warranted to identify which of these approaches is most applicable in which cases, and how possible downsides of these alternatives can be mitigated in practice.

B Appendix: Sample ambiguous inputs

Tables 4- 5 list inputs that are considered ambiguous by CICC in the B77 and HWU64 data sets respectively. Some inputs could refer to multiple intents whereas some other inputs could be considered out-of-scope.

#	Utterance	Label	Prediction Set
1	what is the matter?	direct debit payment not recognised	activate my card, age limit, balance not updated after bank transfer, balance not updated after cheque or cash deposit, beneficiary not allowed, cancel transfer, card arrival, card delivery estimate, card not working, card swallowed, cash withdrawal not recognised, change pin, compromised card, contactless not working, country support, declined card payment, declined transfer, direct debit payment not recognised, exchange rate, failed transfer, get physical card, lost or stolen card, lost or stolen phone, pending card payment, pending cash withdrawal, pending transfer, pin blocked, Refund not showing up, reverted card payment?, terminate account, top up failed, top up reverted, transaction charged twice, transfer not received by recipient, transfer timing, unable to verify identity, why verify identity, wrong amount of cash received,
2	Can I choose when my card is delivered?	card delivery estimate	activate my card, card about to expire, card acceptance, card arrival, card delivery estimate, change pin, contactless not working, country support, get physical card, getting spare card, getting virtual card, lost or stolen card, order physical card, supported cards and currencies, top up by bank transfer charge, top up by card charge, visa or mastercard
3	My contanctless has stopped working	contactless not working	activate my card, apple pay or google pay, automatic top up, beneficiary not allowed, cancel transfer, card not working, card payment wrong exchange rate, contactless not working, declined card payment, disposable card limits, failed transfer, get disposable virtual card, get physical card, pending top up, pin blocked, top up failed, top up reverted, topping up by card, virtual card not working, visa or mastercard, wrong exchange rate for cash withdrawal
4	I misplaced my card and I dont know where the last place is where I used the card last. Can you look at my account and tell me the last place I used the card?	lost or stolen card	activate my card, atm support, card acceptance, card linking, card swallowed, cash withdrawal not recognised, compromised card, lost or stolen card, lost or stolen phone, order physical card, pin blocked
5	Is my card denied anywhere?	card acceptance	atm support, card acceptance, card not working, card payment fee charged, card swallowed, compromised card, contactless not working, declined card payment, lost or stolen card, lost or stolen phone, order physical card, unable to verify identity, visa or mastercard

Table 4: A sample of prediction sets on B77 of size $> th$ of seven with marginal conformal prediction on BERT outputs. Plausible labels have been highlighted with underscore.

C Appendix: LLM results

We here present a random sample of CQs on B77 and C150.

#	Utterance	Label	Prediction Set
1	olly	recommendation events	calendar set, general quirky, lists createoradd, music likeness, music query, play game, play music, play radio,
2	this song is too good	music likeness	audio volume mute, general affirm, general commandstop, general joke, general negate, lists remove, music dislikeness, music likeness
3	do i have to go to the gym	general quirky	calendar query, general quirky, lists query, recommendation events, recommendation locations, transport traffic, weather query
4	silently adjust	audio volume mute	audio volume down, audio volume other, audio volume up, iot hue lightchange, iot hue lightdim, iot hue lightup, music settings
5	how many times does it go	general quirky	datetime query, general quirky, lists query, qa factoid, qa maths, transport query, transport traffic
6	sports head lines please	news query	calendar set, general quirky, iot hue lightchange, music likeness, news query, qa factoid, social post, weather query
7	read that back	play audiobook	email addcontact, email query, email querycontact, email sendemail, general quirky, lists createoradd, music likeness, play audiobook, play music, social post,
8	i don't want to hear any more songs of that type	music dislikeness	audio volume mute, calendar remove, general commandstop, iot wemo off, lists remove, music dislikeness, music likeness
9	check celebrity wiki	general quirky	email query, general quirky, lists query, news query, qa factoid, social post, social query
10	Get all availables	lists query	email addcontact, email query, email querycontact, email sendemail, social post, social query, takeaway order,
11	rating	music likeness	cooking recipe, general quirky, lists createoradd, lists query, music likeness, music query, qa definition, qa factoid,
12	take me to mc donalds	transport query	play game, play podcasts, recommendation events, recommendation locations, recommendation movies, takeaway order, takeaway query
13	search	qa factoid	email querycontact, general quirky, lists createoradd, lists query, music query, qa definition, qa factoid,
14	unmute	audio volume up	audio volume down, audio volume mute, audio volume up, iot wemo off, music settings, play radio, transport query, transport traffic
15	please unmute yourself	audio volume mute	alarm remove, audio volume down, audio volume mute, audio volume up, iot cleaning, iot wemo on, music settings, play game
16	what's the best day next week to go out for pizza	datetime query	calendar query, cooking recipe, general quirky, qa factoid, recommendation events, recommendation locations, takeaway query
17	i need a manger	general quirky	calendar set, cooking recipe, general quirky, lists createoradd, music likeness, play game, qa definition, qa factoid, social post,
18	assistant shuffle entire library	play music	iot cleaning, iot hue lightchange, lists createoradd, music settings, play audiobook, play game, play music
19	put the disco lights on	iot hue lighton	alarm remove, iot cleaning, iot hue lightchange, iot hue lightoff, iot hue lighton, iot hue lightup, iot wemo on
20	hello how are you today	general greet	general greet, general praise, general quirky, play radio, recommendation events, recommendation locations, recommendation movies
21	where does tar work currently	email querycontact	cooking recipe, email querycontact, general quirky, lists query, qa definition, recommendation locations, takeaway query
22	can you pull up jeff	email querycontact	general quirky, iot cleaning, news query, play game, play music, play radio, social post
23	show me theatre where pirates of caribbean is playing	recommendation movies	general quirky, play game, play music, play podcasts, recommendation events, recommendation locations, recommendation movies
24	spot	play podcasts	calendar set, music likeness, music query, play music, play radio, qa factoid, weather query,
25	anything i need to know	general quirky	calendar query, calendar set, email query, email querycontact, general quirky, lists query, news query, social query
26	give me some good sounds	play music	audio volume down, audio volume up, music likeness, music query, music settings, play music, play radio
27	i don't like it	music dislikeness	calendar remove, general commandstop, general dontcare, general joke, general negate, lists remove, music dislikeness, music likeness, social post
28	what are you	general quirky	audio volume down, audio volume mute, audio volume up, music query, music settings, qa definition, recommendation locations, recommendation movies
29	i don't like this one	general quirky	audio volume mute, calendar remove, general joke, general negate, iot wemo off, lists remove, music dislikeness, social post

Table 5: Prediction sets on ACID of size $> th$ of seven with marginal conformal prediction on BERT outputs. Plausible labels have been highlighted with underscore.

Utterance	Prediction Set	LLM output
I hope you can help me. My account has been compromised from a different town that I haven't been to. Can you please place a hold on my account until I can make it to the bank?	cash withdrawal not recognised, compromised card	I understand you are having issues with your account. Was your account compromised or your card not recognised for cash withdrawal?
Is my card denied anywhere?	card not working, declined card payment	I understand you are asking about your card. Was it declined or not working?
Money that I have transferred hasn't arrived yet	balance not updated after bank transfer, transfer not received by recipient	I understand you are waiting for a transfer to be updated in your balance. Was the transfer not received by you or the recipient?
How long will the transfer take?	balance not updated after bank transfer,pending transfer, transfer timing	I understand you are asking about transfer timing. Was your transfer pending, not updated or not received?
What should I do with my stuck ATM?	card swallowed, declined cash withdrawal	I understand you have a stuck ATM. Was it your card that got stuck or was your cash withdrawal declined?

Table 6: Sample of LLM outputs on the B77 dataset where the user input has been made part of the prompt.

Utterance	Prediction Set	LLM output
how many more miles can i drive on my current tank	gas, mpg	I understand you are asking about your car's fuel. Are you asking about the number of miles you can drive on your current tank or the miles per gallon?
can i redeem my points for flights	redeem rewards, rewards balance	I understand you would like to redeem your points for flights. Are you looking to check your rewards balance or redeem your rewards?
are there any meetings between 2 and 5 pm	schedule meeting, meeting schedule	I understand you are looking for a schedule of meetings. Are you looking for a specific time range or a general schedule?
is the marathon on friday added to my calendar	calendar, calendar update	I understand you are referring to your calendar. Is it the marathon on friday that you are referring to or do you need an update on your calendar?
why didn't my card work	expiration date, card declined	I understand you are having issues with your card. Is it because of the expiration date or was it declined?

Table 7: Sample of LLM outputs on the C150 dataset.

D Appendix: Intent distributions

We here present the intent distributions for all public datasets across train, calibration and test splits.

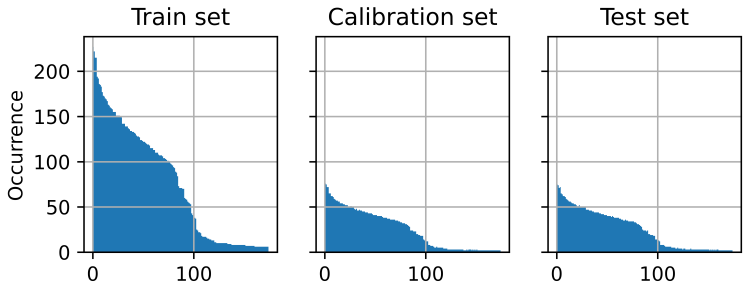

Figure 3: Intent distribution in ACID data set.

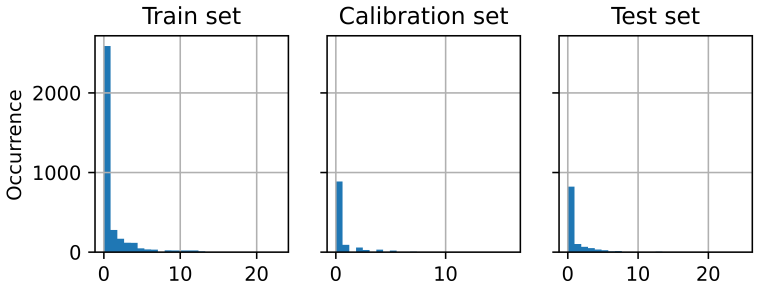

Figure 4: Intent distribution in ATIS data set.

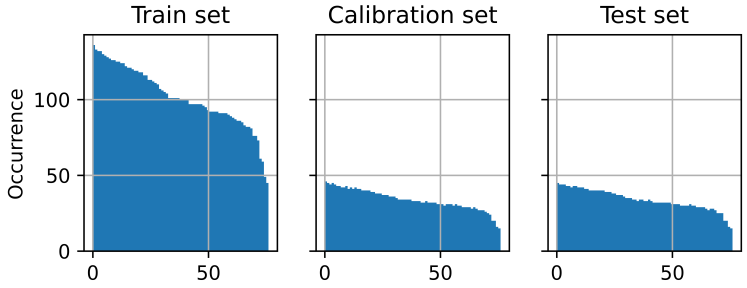

Figure 5: Intent distribution in B77 data set.

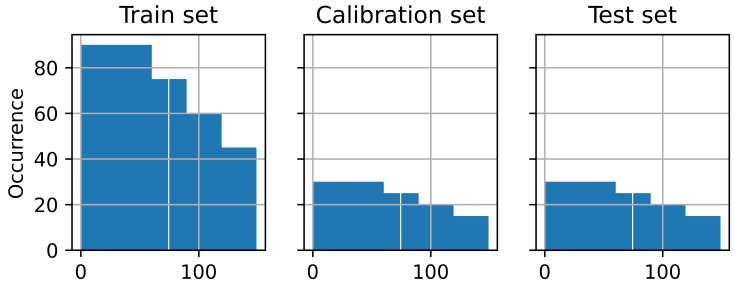

Figure 6: Intent distribution in C150-IS data set.

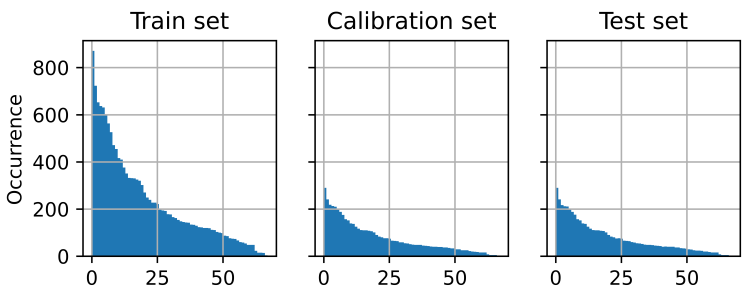

Figure 7: Intent distribution in HWU64 data set.

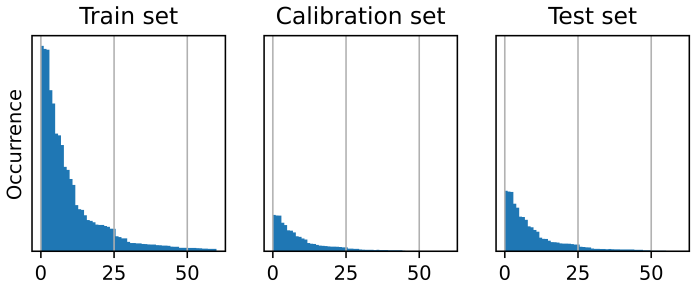

Figure 8: Intent distribution in IND data set.

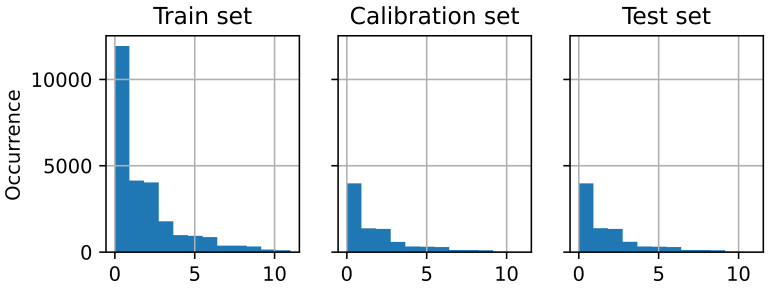

Figure 9: Intent distribution in MTOD data set.

E Appendix: Unoptimized α

This appendix contains results for an unoptimized α hyperparameter, arbitrarily set at .10 and .01. We see that for most data sets, there is no need to ask a clarification question as the model already achieves the desired coverage. Much higher coverages (as in Table 2) are achievable for these data sets. For some more challenging data sets such as C150, HWU64 and IND, CICC yields small clarification questions while retaining a reasonably large number of clarification questions of size 1.

Setting	$1-\alpha$	th		Cov↑	Single↑	|CQ| ↓	Amb
ACID	.90	7	CICC	.90	.92	–	0
			B1	.97	.93	5	0
			B2	.95	**1**	–	0
			B3	.99	0	5	0
ATIS	.90	7	CICC	.88	.89	–	0
			B1	.99	.93	5	0
			B2	.98	**1**	–	0
			B3	1	0	5	0
B77/BERT	.90	7	CICC	.98	.79	**2.90**	.04
			B1	.97	.90	5	0
			B2	.93	**1**	–	0
			B3	.99	0	5	0
B77/DFCX	.90	4	CICC	.91	.66	2.63	.02
			B1	.95	.71	4.79	.27
			B2	.90	**.98**	**2.26**	0
			B3	.97	0	5	1
C150	.90	7	CICC	.99	.97	**2.66**	0
			B1	.99	.82	5	0
			B2	.98	**1**	–	0
			B3	1	0	5	0
HWU64	.90	7	CICC	.90	.97	**2.00**	0
			B1	.96	.79	5	0
			B2	.90	**1**	–	0
			B3	.98	0	5	0
IND	.90	7	CICC	.91	**.25**	**3.46**	.11
			B1	.88	.42	5	0
			B2	.70	1	–	0
			B3	.91	0	5	0
MTOD	.90	7	CICC	.90	.90	–	0
			B1	.99	.99	5	0
			B2	.99	**1**	–	0
			B3	1	0	5	0

Table 8: Test set results for $1-\alpha = .90$ where underline indicates meeting coverage requirement. **Bold** denotes best when meeting this requirement, omitted for last column due to missing ground truth for ambiguous.

Setting	$1-\alpha$	th		|Cov↑	Single↑	|CQ| ↓	Amb
ACID	.99	7	CICC	1	**.77**	**3.00**	.10
			B1	.98	.85	5	0
			B2	.95	1	–	0
			B3	.99	0	5	0
ATIS	.99	7	CICC	.99	**.98**	**2.54**	0
			B1	.99	.73	5	0
			B2	.98	1	–	0
			B3	1	0	5	0
B77/BERT	.99	7	CICC	.98	**.79**	**2.90**	.04
			B1	.97	.90	5	0
			B2	.93	1	–	0
			B3	.99	0	5	0
B77/DFCX	.99	4	CICC	.97	0	5	1
			B1	.97	.05	5	.95
			B2	.90	1	–	0
			B3	.97	0	5	1
C150	.99	7	CICC	.99	**.97**	**2.66**	0
			B1	.99	.82	5	0
			B2	.98	1	–	0
			B3	1	0	5	0
HWU64	.99	7	CICC	.99	**.25**	**3.39**	.28
			B1	.98	.05	5	0
			B2	.90	1	–	0
			B3	.98	0	5	0
MTOD	.99	7	CICC	.99	**1**	–	0
			B1	1	.98	5	0
			B2	.99	**1**	–	0
			B3	1	0	5	0

Table 9: Test set results for $1-\alpha = .99$ where underline indicates meeting coverage requirement. **Bold** denotes best when meeting this requirement, omitted for last column due to missing ground truth for ambiguous.

F Appendix: Comparison results OOS detection

We here compare the results of OOS detection as reported by baselines. Note that these results were generated on different splits of the data and (where applicable), possibly using different open-domain samples, and that a direct comparison between results is invalid.

Dataset	Algorithm	F1↑	Accuracy↑
C150	CICC-OOS	.91	.68
	Zhan et al. (2021) 25%	.81	.88
	Zhan et al. (2021) 50%	.87	.88
	Zhan et al. (2021) 75%	.89	.88
	Cavalin et al. (2020)	.76	.73
B77	CICC-OOS	.90	.89
	Zhan et al. (2021) 25%	.74	.70
	Zhan et al. (2021) 50%	.80	.73
	Zhan et al. (2021) 75%	.87	.81

Table 10: Results for the OOS detection task.