$\hat{f}(X)$ 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:

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

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

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**

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

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