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

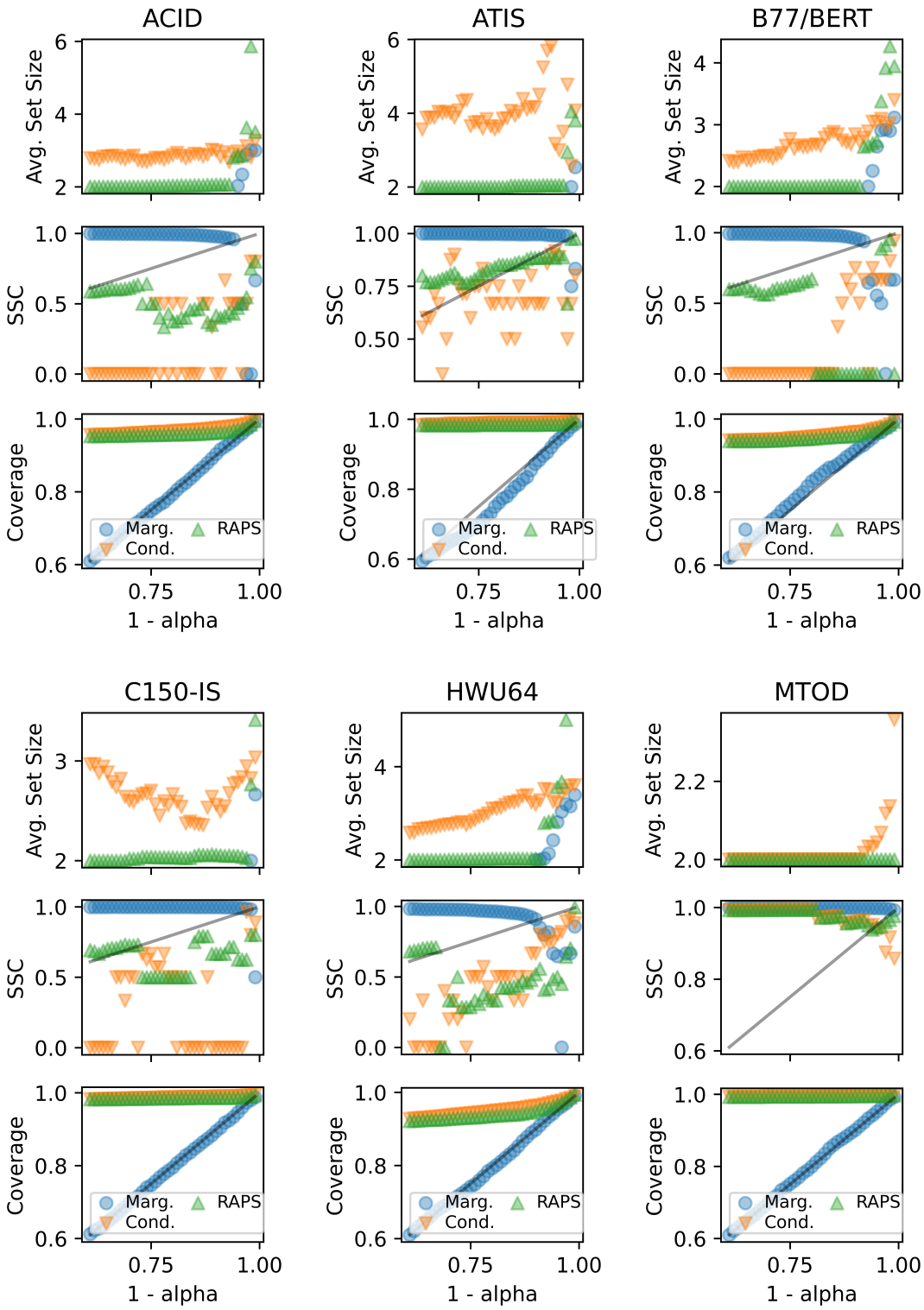

Figure 2: Test set results for varying error rate α.

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.