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 pretrained 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.