This thesis focuses on developing and improving task-oriented dialogue systems design in the rapidly growing landscape of artificial intelligence and natural language processing. We propose techniques that can substantially decrease development and deployment costs, motivated by the desire to make these systems more adaptable and scalable. We introduce multiple novel approaches to achieving these goals.

Firstly, we present a weakly supervised automatic data annotation pipeline that can transform raw dialogue transcript into a refined set of semantically coherent concepts, bypassing the need for exhaustive manual annotations in natural language understanding for a given domain and significantly streamlining the development process.

We also explore the largely uninvestigated field of latent variable models in task-oriented dialogue system modeling. These models offer excellent capabilities with the potential to uncover the structure of behavioral patterns seen in the dialogue through inspection of the latent space and comparison with actions taken by the model. Furthermore, we explore the potential of these models to form hierarchical representations using our proposed architecture.

Following recent progress in the field, we harness the power of pre-trained large language models using in-context learning. We explore how easily language models can transfer the learned knowledge to previously unseen domains. We also propose a method based on retrieval-augmented LLM prompting that performs well with merely a few training examples. It shows great promise in our human evaluation trial, implying a substantial leap in efficiently using computational resources to train conversational AI. This brings us closer to more flexible and general-purpose systems.