ABSTRACT

The growing importance of human-computer interaction and natural language processing is highlighted by significant advances such as the introduction of ChatGPT, a Large Language Model (LLM)-based dialogue system. While open-domain dialogue systems focus on user engagement, task-oriented dialogue systems (TODS) are designed to assist users with specific tasks within defined domains. However, the deployment of new TODS faces challenges, particularly in ensuring dialogue quality through resource-intensive human evaluation. Assessing the quality of TODS requires a nuanced understanding of user intent and the generation of contextually appropriate responses. Automated evaluation mechanisms play a crucial role in systematic testing prior to TODS deployment. Historically, two evaluation methods have been used: dataset-based and interactive. While dataset-based evaluation serves as a benchmark, it does not capture the dynamic nature of TODS, limiting its adaptability to real-world user responses. In contrast, the interactive evaluation involves a user simulator engaging in multi-turn dialogues with TODS, mimicking authentic conversational scenarios. Despite closely simulating real-world usage, creating effective simulators can be resource-intensive, with previous approaches relying on rules, heuristics, or large amounts of annotated data showing limitations in adapting to unexplored domains or resource-constrained environments.

This research presents an end-to-end ICL-based User Simulator (US) for TODS . Using LLMs, the proposed architecture can perform task-oriented user simulation in an interactive end-to-end manner with minimal data requirements. It addresses the need for user simulation for low-resource domains by evaluating zero and few-shot ICL strategies within the MultiWOZ domain against a pre-trained baseline US. In addition, an extended TELeR-RESPONDeR taxonomy for nuanced prompt descriptions is presented to increase the comparability of the proposed methods and to address the need for a standard ICL notation. The proposed In-Context Learning User Simulator (ICL-US) demonstrated proficient generation of lexically diverse user responses that closely matched real user baselines as quantified by MTLD. Although the ICL-US did not outperform the US baseline, it demonstrated promising conversational capabilities with task-oriented dialogue systems, highlighting the potential of ICL-based user simulation even in a zero-shot setting and emphasizing the potential of the ICL-based approach.