User Simulation in Task-Oriented Dialog Systems based on Large Language Models via In-Context Learning

Ronny Horst

Supervisors: Prof. Dr. Bettina Harriehausen-Mühlbauer, Prof. Dr. Jutta Groos
Darmstadt University of Applied Sciences

Introduction

Task-oriented dialog systems (TODS) are designed to assist users with specific tasks within defined domains. This approach can help overcome the challenges faced in deploying new TODS, as their quality often needs to be ensured through resource-intensive human evaluation. To ensure the quality of TODS pre-deployment, interactive evaluation through a user simulator can be used instead of relying on costly human resources. By using a user simulator, the quality of the TODS can be evaluated interactively, which can help to reduce the cost and time required for human evaluation. Developing a user simulator can be a labor-intensive task due to the need for hand-crafted rules, heuristics, or large amounts of annotated data to train a model. However, recent advancements in Large Language Models (LLMs) and their In-Context Learning (ICL) abilities can simplify this process by utilizing combinations of instructions and/or demonstrations as context. This thesis investigates the utilization of recent developments by addressing the following questions:

- Can LLMs, with their emerging ICL capabilities, effectively serve as user simulators in facilitating interactive conversations with TODS?
- How would an LLM-based user simulator be set up architecturally, and what in-context learning strategies could be used to increase the humaneness of the simulator?
- How can various prompting strategies be formally described to ensure reproducibility and enable meaningful comparisons to understand their effectiveness?

ICL-Based User Simulator

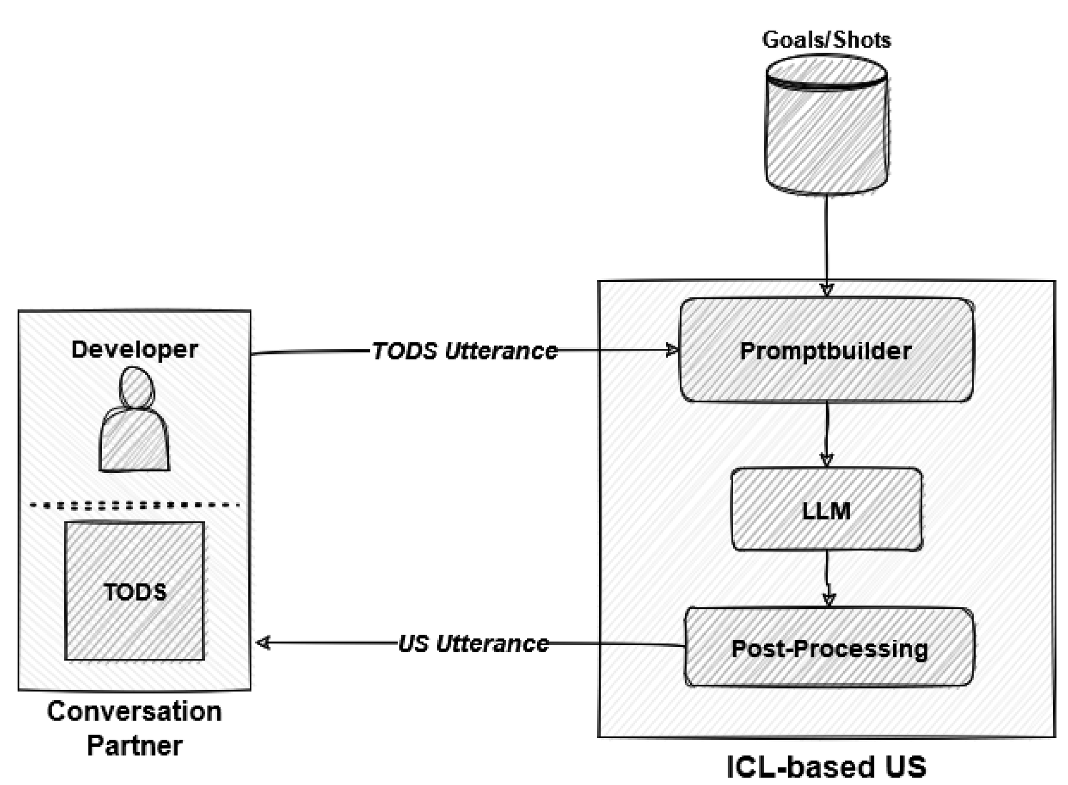

Figure 1. Sketch of the basic ICL-based User Simulator setup.

The proposed ICL-based user simulator is set up as shown in Figure 1 using an end-to-end design. The simulator comprises a prompt builder that creates context for the LLM based on various ICL strategies and possible demonstration datasets, if available. The output of the prompt builder is then used to prompt the LLM. The resulting output from the LLM is post-processed to extract the user's utterance. This statement is then sent to the conversation partner, who may be a developer implementing new strategies or a TODS for testing purposes. To increase the comparability of the ICL strategies used, the TELeR taxonomy proposed by Santu et al. [6] was modified to allow for a finer grained description. The resulting taxonomy was named TELeR-RESPONDeR, with additional dimensions for **R**easoning, **E**nsemble, **S**elf-justification, **P**lanning, **O**utput, **N**otation, **D**emonstration and **R**etrieval. This extended taxonomy was applied to the existing work of Terragni et al. [7] and Davidson et al. [2], who published ICL-based user simulator approaches focusing on few-shot strategies during the preparation of this thesis.

Different ICL strategies have been investigated in this thesis, including zero-shot prompting and few-shot prompting with different characteristics regarding the level of instruction detail, role definitions, self-justification and output notation. For the few-shot approaches, various similarity-based retrieval techniques were used to extract similar goals and conversations given a seed goal. These retrieval methods allow sampling based on structured goals consisting of different slots via Jaccard sampling and textual goals via vector search. Furthermore, reasoning, planning and ensemble strategy concepts were applied to the task of generating the next user utterance.

Interactive Evaluation

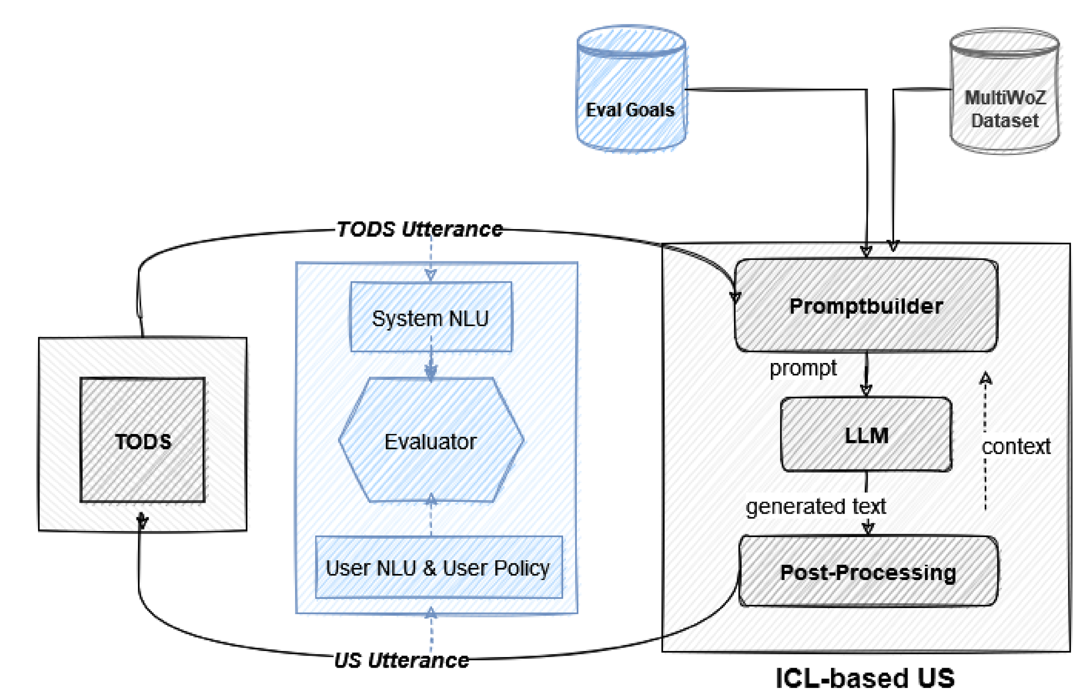

Figure 2. Sketch of the basic ICL-based User Simulator setup.

To evaluate the proposed ICL-based user simulator, an interactive evaluation approach using the ConvLab-3 [9] and the MultiWoZ 2.1 [3] dataset as target domain was chosen. In order to use the evaluator module of ConvLab-3 [9], the ICL-US architecture was extended by user- and system-level natural language understanding modules and a policy module, as shown in figure 1. For the evaluation, a dataset of 100 user goals was generated. Zero-shot role prompting and instruction prompting as well as few-shot with structured and unstructured goal representations and the described sampling methods were then evaluated using the evaluation dataset. Three LLMs, namely GPT-3.5 Instruct [5], Llama2[8] and FLAN-T5 [1], were used to generate the user utterances based on the designed ICL strategies. Since GPT-3.5 gave superior results in the instruction tuning evaluations, further experiments were conducted using only this LLM. The dialogues were evaluated with a set of success and diversity metrics, addressing task success and naturalness of the conversations. To enable a meaningful comparison, a pre-trained user simulator baseline was also evaluated on the evaluation dataset. Furthermore, the diversity metrics were applied to a sampled dataset of the MultiWoZ user utterances, as they are based on human dialogues and thus form the baseline for the diversity metrics.

Evaluation Results

	Completion Rate	Success Rate	MTLD
Baseline Diversity MultiWoZ Data Set	-	-	**60.88**
Baseline User Simulator ConvLab-3	**0.61**	**0.40**	37.98
Zero-Shot Instruction	0.33	**0.28**	42.79
Zero-Shot Instruction & Role Definition	**0.35**	0.25	41.24
Few Shot FAISS Sampling Json Goal Format	0.24	0.21	**58.15**

Table 1. Goal Success Metrics and Diversity Metrics for baselines and best performing ICL-Based Strategies on GPT-3.5-Instruct.

An extract of the quantitative evaluation results is shown in Table 1.The ICL-based user simulator could not compete with the pre-trained user simulator on the success metrics, but the proposed approach produced more natural user utterances as measured by MTLD and human analysis. In the few-shot setting and demonstrations of real user conversations, the naturalness of the generated utterances was close to the baseline diversity dataset, in contrast to the zero-shot setting, which resulted in lower task success metrics.

Contributions, Limitations and Future Work

This thesis answered the research questions by making the following contributions:

- An end-to-end ICL-based user simulator architecture has been designed and possible extensions are presented.
- The basic end-to-end architecture has been revised for use in the ConvLab-3 evaluation framework.
- A selection of ICL strategies have been applied to the task of user simulation for task-oriented dialogue.
- The existing TELeR taxonomy has been extended and applied to all strategies to allow meaningful comparisons for future work.
- A selected set of strategies (Instruction, Role and Few-Shot Prompting) were evaluated in detail through interactive conversations between a TODS from the Convlab-3 framework and the proposed ICL-US, using automatic measures as well as human error analysis.

The infinite solution space of in-context learning approaches leaves room for endless experimentation. Due to resource limitations, only a selection of strategies could be evaluated. In the future, the conceptualised advanced prompting techniques, including chain-of-thought reasoning, least-to-most prompting and ensemble prompting methods, should also be evaluated. Furthermore, larger LLMs and newer open source LLMs should be tested in comparison to the closed source GPT, as manual tests showed promising results but were incompatible with the chosen evaluation framework. As revealed by the human error analysis, the interactive evaluation process and metrics of the proposed ICL-US are highly dependent on the interlocutor TODS. As discussed by Davidson et al. [2], the task of a user simulator should be to mimic a real human as closely as possible, rather than maximising goal success metrics by communicating as effectively as possible, which may result in artificial communication. Therefore, more appropriate ways of evaluating user simulators remain a gap in research and should be further explored.

References

[1] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[2] Sam Davidson, Salvatore Romeo, Raphael Shu, James Gung, Arshit Gupta, Saab Mansour, and Yi Zhang. User simulation with large language models for evaluating task-oriented dialogue. *arXiv preprint arXiv:2309.13233*, 2023.

[3] Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*, 2019.

[4] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[5] OpenAI. Openai gpt-3 api [gpt-3.5-turbo-instruct], 2023.

[6] Shubhra Kanti Karmaker Santu and Dongji Feng. Teler: A general taxonomy of llm prompts for benchmarking complex tasks. *arXiv preprint arXiv:2305.11430*, 2023.

[7] Silvia Terragni, Modestas Filipavicius, Nghia Khau, Bruna Guedes, André Manso, and Roland Mathis. In-context learning user simulators for task-oriented dialog systems. *arXiv preprint arXiv:2306.00774*, 2023.

[8] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[9] Qi Zhu, Christian Geishauser, Hsien chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gašić, and Minlie Huang. Convlab-3: A flexible dialogue system toolkit based on a unified data format. *arXiv preprint arXiv:2211.17148*, 2022.