An LLM-based Chatbot for Joint Access to Customer and Corporate Data

Teo Ferrari and **Vincent Coriou** and **Rastislav Kovac**
Vuk Vuković and **Daniel Dobos** and **Fatemeh Borran** and **Andrei Popescu-Belis**
teo.ferrari@heig-vd.ch

Abstract

Large Language Models (LLMs) give access to huge amounts of public-domain knowledge and have robust interactive capabilities. When it comes to corporate or personal data, several techniques enable their integration with LLMs: in-context learning, fine-tuning, or retrievalaugmented generation (RAG). In the present work, we demonstrate that two of these techniques can be combined, and design a system that can answer questions about company-level and userlevel data at the same time. Specifically, we present the design and evaluation of a customer support system which combines a fine-tuned version of an open-source LLM, for answering questions related to static company-level data, with in-context learning for answering questions about dynamic customer-specific data. This system has been designed with a Swisscom customer support use-case in mind. To inject static knowledge into the LLM, we employed a Parameter Efficient Fine-Tuning (PEFT) technique, specifically Low-Rank Adaptation (LoRA) (Hu, et al. 2021). This is cost-effective and at the same time has low overfitting risks. The static data used for fine-tuning contains companyspecific knowledge formulated as question-answer pairs. To enable the LLM to access dynamic, customer-specific data, based on previous studies (White, et al. 2023) and our own experiments, we engineered a prompt which combines instructions concerning the desired behaviour of the chatbot with a structured representation of customer-related information (here, mostly about billing). The evaluation data includes several dozen questions about static and dynamic knowledge, with the system's answers being assessed along three criteria used to evaluate free-form question answering (Sai et al. 2022): relevance, correctness, and fluency. In other words, we assess if the answer is on the same topic and provides the required type of information; then, if the answer is factually correct given the knowledge base (irrespective of its relevance to the question); and finally, if the answer is formulated in correct English and is appropriate in terms of politeness and greetings. Following preliminary assessments of several LLMs, we selected Mistral 7b for our implementation. The evaluation results revealed that fine-tuning a Mistral (Jiang, Sablayrolles, Mensch, et al. 2023) model enabled it to handle static data queries satisfactorily, while prompt engineering ensured effective access to dynamic data. The results confirm the feasibility of a versatile, efficient customer support system through the combination of fine-tuning and prompt engineering, tailored to the specific data sources encountered in customer service scenarios.