Hes·so
Haute Ecole Spécialisée de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences and Arts Western Switzerland

Travail de Master of Science HES-SO en Engineering

BUILDING A RELIABLE CHATBOT USING GENERATIVE AI MODELS

Teo Ferrari

Professeur responsable : **Prof. Andrei Popescu-Belis**

HEIG-VD

En collaboration avec **le Swisscom Digital Lab**

DESCRIPTION

In their commitment to enhancing user experience, companies increasingly employ **chatbot services**. Generally, the existing chatbots rely on **traditional methods such as intent detection**, which involves identifying the user's intentions, and **dialogue models**, a structure that guides the conversation flow. While functional, this approach has **notable drawbacks**. Firstly, interactions often feel **rigid and unnatural**. Secondly, it requires **extensive and costly maintenance** due to the human creation of each dialogue state model and use case. Lastly, it struggles to adapt to the unpredictable nature of customer support.

Recognizing these challenges, many companies have started exploring **improvements to their chatbot technology** through various methods based on **large language models (LLMs).** These methods aim to provide **more flexible and dynamic interactions** compared to traditional chatbots.

One such example is **Swisscom**, which is currently testing a beta version of a chatbot that uses Retrieval Augmented Generation. This technique combines the ability to generate responses based on retrieved information from a database or a knowledge base, offering a more contextually relevant and dynamic conversation experience.

OBJECTIVES

The central concept of this thesis revolves around employing two techniques, **prompt engineering and model fine-tuning**, to develop a **system capable of addressing questions that are either related to facts about the company (static information) or to evolving details about the customer (dynamic information).** By leveraging these methodologies, **we create a versatile system** that can effectively respond to inquiries involving both static and dynamic information, and **we quantify its capabilities in both respects**.

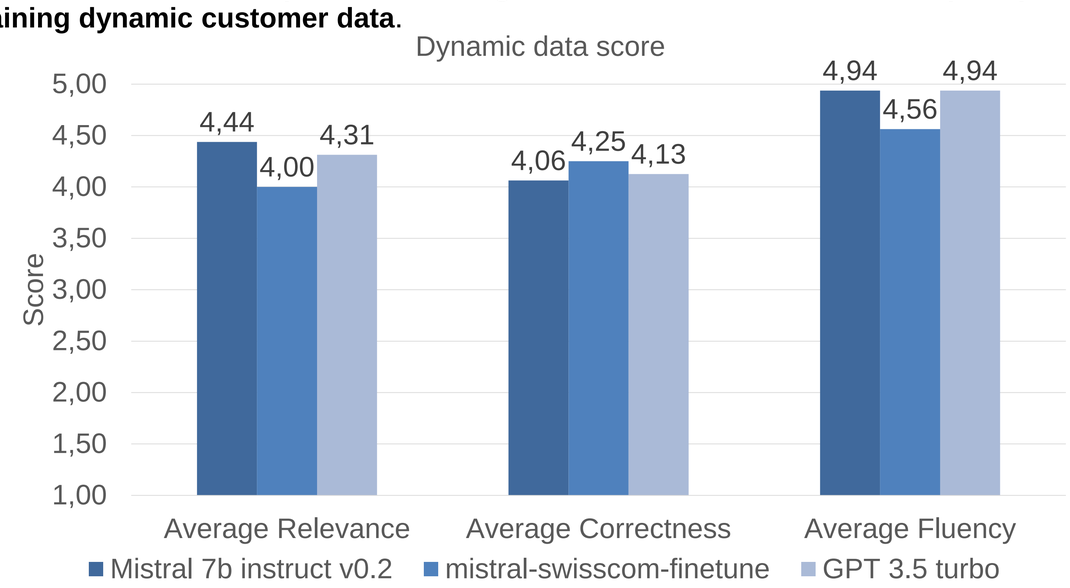
The resulting models are manually evaluated **on 24 questions, 16 pertaining to dynamic information present in the prompt and 8 pertaining to static information** fine-tuned in one of the models. The answers generated by the models are **evaluated along 3 criteria**:

- **Relevance with respect to the question**: the answer is on the same topic and appears to provide the type of information that is required by the question (be it correct or not). The system is penalized if it provides insufficient information, or more information than needed (superfluous), or the wrong type of information.
- **Correctness with respect to the knowledge base**: the answer contains information that is correct given the knowledge base, irrespective of its relevance or not to the question.
- **Fluency and style**: the answer is formulated in correct English, it avoids mathematical formulas, it adopts the right level of formality (or politeness), and does it include the appropriate greetings or addressing the customer.

The **Lickert scale** is used to score the answers, resulting in a score from 1 to 5.

RESULTS

The evaluation covers **3 models** and **24 questions**. Two of these models are foundational pre-trained models namely, *Mistral 7B instruct v0.2* and *GPT-3.5 Turbo* while the last model is fine-tuned on static **Swisscom-specific knowledge**. All the models have access to **a prompt containing dynamic customer data**.



The scores pertaining to the 16 questions about **dynamic customer data** are **close for all models**. This means that **all models reacted well to the prompt** containing said data. An exception comes when the questions involve mathematical operations, which is quite a well-known issue for LLMs.

When it comes to the remaining **8 questions about static Swisscom-related data** the **fine-tuned model is clearly superior in terms of correctness with respect to the knowledge base**. This superiority is **because only the fine-tuned model has access to the specific knowledge** which **demonstrates the effectiveness of the fine-tune**. Additionally, **only *Mistral 7B instruct v0.2* exhibits low relevance concerning the questions**. This lower relevance is due to the model's tendency to **avoid answering the question directly**, while **GPT-3.5 Turbo provides direct answers but often fabricates the necessary knowledge**. The **fluency and style are always very good**, as is expected with LLMs.

CONCLUSION

This work explored the possibility of **creating a hybrid customer support system** that was able to answer two types of questions using **a fine-tuned LLM and appropriate prompt engineering**.
Firstly, to answer questions regarding **static information about Swisscom**, we proposed to **fine-tune the LLM on documents** (or question-answer pairs) containing this information, **in order to inject this knowledge into the system**. This has been **implemented using a Parameter Efficient Fine-Tuning (PEFT) technique called Low Rank Adaptation (LoRA)** that provides the additional advantage of lowering the computational cost of fine-tuning and the risk of overfitting to the fine-tuning data. The **results of the fine-tuning part are good**: the answers generated for the questions in our test set make **it clear that the static Swisscom knowledge has been successfully injected into the model**.
Secondly, to answer questions **about dynamic billing related customer data**, we proposed a computationally light technique, specifically **prompt engineering based on this data**. By carefully crafting an adequate prompt and processing the customer data to keep only what is relevant to customer support**, good results have been attained**. **All the three models that we tested were able to answer questions pertaining to dynamic billing data correctly**.
The final results show that the **fine-tuned model has better capabilities to answer questions related to static data than the out-of-the-box model, while all models reach close results when it comes to using and understanding the prompt.**
In conclusion**, we have proven that creating a versatile system using different techniques for different types of data is possible**. The system **successfully integrated static Swisscom data using fine-tuning and dynamic billing-related customer data using prompt engineering**.

SWISS UNIVERSITIES APPLIED SCIENCES AND ARTS

MSE | MASTER OF SCIENCE IN ENGINEERING