Hi, I’m Martí Mas, and in collaboration with DXC Technology and supervised by Josep Suy, we are presenting the project “Tuning a Generative Language Model for the Creation of Customized Chatbots for Public Administrations”, supervised by Josep Suy.

We built a system that gives a Large Language Model access to databases/websites and is customizable to client needs through BPMNs using Flowable. The databases/websites, provided by the clients, contain up-to-date information. The system can answer user queries truthfully by augmenting the model’s ability to generate text with the ability to retrieve information from these databases/websites. With this system we avoid hallucinations while allowing for a wide range of possible configurations different clients might require.

The system for retrieving from the database is called Retrieval Augmented Generation or RAG. We implement a custom RAG solution called Small-To-Big-Retrieval, which is able to query the documents from multiple semantic levels at the same time. On the frontend, for accessibility reasons, we deploy on-edge Machine Learning to perform hotword recognition on the user’s machine. This is done on-edge for privacy reasons, as we don’t want to constantly send microphone data to the cloud. We also implement measures to restrict the behavior of the Language Models, to stick to the stakeholder’s needs and guidelines.

Retrieval is done intelligently by finding the most semantically relevant pieces of information in the database.

Using Machine Learning, we first classify user queries into two categories: hard ones and easy ones. We then choose to use a larger or a smaller Language Model to answer the queries based on this classification, reducing costs. This is called RouteLLM.

Here's a quick demo of how it all comes together: Here a stakeholder, the Departament de Drets Socials de la Generalitat, wants to help citizens know how social rights work and what social benefits they might have access to. The system asks a few questions to understand the user’s situation and recommends possible benefits that might be available to them. The system refuses to answer questions outside the scope defined by the stakeholders or questions that might abuse the system.