PRINCIPAL ACADEMIC TUTOR
Prof. Stefano Ceri,
DEIB, Politecnico di Milano

ACADEMIC TUTORS
Prof.ssa Elena Baralis,
DAUIN, Politecnico di Torino

Prof. Luca Cagliero,
DAUIN, Politecnico di Torino

Prof.ssa Anna Bernasconi,
DEIB, Politecnico di Milano

Prof. Francesco Pierri,
DEIB, Politecnico di Torino

EXTERNAL INSTITUTIONS
Advanced Business Solutions Srl (A.B.S)

EXTERNAL TUTORS
Ing. Samuele Conti, A.B.S.

TEAM MEMBERS

Lorenzo Bertetto,
Politecnico di Torino,
Computer Science and Engineering

Francesca Bettinelli,
Politecnico di Milano,
Mathematical Engineering

Alessio Buda,
Politecnico di Milano,
Computer Science and Engineering

Marco Da Mommio,
Politecnico di Milano,
Mathematical Engineering

Simone Di Bari,
Politecnico di Torino,
Telecommunication Engineering

Claudio Savelli,
Politecnico di Torino,
Data Science and Engineering

ChatIMPACT

Executive summary

In recent years, Large Language Models (LLMs) such as ChatGPT (Brown, 2020) have become increasingly popular among the general public. Indeed, LLMs are powerful Artificial Intelligence (AI) tools that are able to learn patterns, meanings, and structures from huge amounts of textual data. This process, called training, makes them capable of solving complex tasks – including text generation, translation, or summarization – in multiple domains of application, from medicine to law. The huge potential of LLMs is hindered by the fact that, for non-expert users, it is hard to identify the model that is more suitable for a specific use-case because of the lack of a unified knowledge base.

Nowadays, the most relevant source of information on LLMs and LLM-related items, such as datasets or metrics, is the Hugging Face (HF) portal (Hugging Face – The AI community building the future., 2024). Despite its popularity, HF has several limitations: some details regarding models or datasets are not available on the platform, and some queries linking different entities cannot be performed.

The goal of the chatIMPACT project is to bridge this gap, building a platform that allows users to easily access the knowledge they need in the LLMs domain. The main outcome of our work is a proof of concept achieved in two steps:

1. Designing a comprehensive **conceptual map** of LLMs and LLMs related entities, namely datasets used for training, metrics to evaluate models and tasks that the models could solve.
2. Implementing the conceptual map in the form of an interactive, easy-to-navigate **web interface** accessible to different types of users.

The initial phase of the project was dedicated to literature review, mostly focused on the technical aspects of LLMs. This step was essential to acknowledge the lack of structured information in this domain.

The first outcome of the project was a conceptual scheme of LLMs and LLMs-related entities, such as datasets or evaluation metrics, which was presented in a paper (Bertetto, 2024) at the 36th International Conference on Advanced Information Systems Engineering (CAISE).

The subsequent step of the project was focused on listing the available sources of data to populate a concrete implementation of the conceptual graph. All the available information on HF was retrieved and data availability was estimated. This analysis supports our initial claim: HF is missing relevant information.

We decided to simplify the initial conceptual scheme for increased accessibility and prepared a set of manually curated data that has been loaded in a database on top of which a graphical interface has been implemented. A demo of the system can be found at http://geco.deib.polimi.it/chatIMPACT/.

In conclusion, the chatIMPACT project provided users with a tool to find the models that suit their needs and to understand their limits and potentiality. A current limit of our solution is represented by the fact that all the necessary data and related information was extracted manually, hence limiting scalability. Future directions of research include finding a way to automatically enrich HF's data, possibly with the help of LLMs themselves, an approach known as "Retrieval-Augmented Generation" (RAG), to extract structured information from different sources, such as other web portals or academic literature.

Key Words

Artificial Intelligence | Large Language Models | Knowledge base | Web Interface

CHATIMPACT

01. Literature Review

Literature review was essential to familiarize with technical aspects of LLMs and acknowledge the lack of structured information in this domain.

02. HF Schema Definition

The conceptual schema underlying the HF platform was derived. This was crucial to understand HF limits

03. Extended Schema Design

The HF conceptual schema was enriched leading to an extended conceptual schema presented at the CAiSE Forum Conference 2024

04. Simplified Schema Design

The extended schema was simplified, retaining essential information. This step was fundamental for the implementation of the final tool

05. HF Data Extraction

Available data on HF has been extracted, setting the foundation of the next steps

06. HF Availability Analysis

Data extracted from HF has been analyzed to assess information availability on the platform

07. Manual Data Enrichment

Data regarding models produced by the most well-known companies (e.g., Meta, Google) have been manually extracted, filling all the information required by our schema

08. DB Implementation

A database instance has been populated with the manually extracted data. This was essential in order to build the final interface

09. Explorable Interface

A complete and easy-to-navigate interface was made available to the public

Project description written by the Principal Academic Tutor

The project developed along four main directions: (a) an in-depth analysis of the features of Large Language Models, including foundations, training, biases, privacy / legislative concerns and hallucinations; (b) definition of a conceptual map for Large Language Models, with the objective of understanding the main properties (entities, relationships, attributes) and then identifying Hugging Face (HF) as the main source of information to generate the map; (c) definition of a strategy for transforming the map into a linked database and generate its content, by a mixed strategy including access to HF's APIs and scraping; (d) definition of a user-friendly interface supporting interesting queries over the database.

In this work, students have addressed and then partially solved many critical problems due to the novel and uncoordinated nature of the information at hand, as well as typical inconsistencies and incompleteness of the HF source. Results are relevant, as this conceptual map allows understanding of the large language model's properties, exposed through a variety of interesting queries.

Team description by skill

The chatIMPACT team is vertical and specializes in Information Engineering.

The team's composition was beneficial to tackle the technical nature of the problem. However, each member contributed by applying his skills to different aspects of the project: the following will clarify each student's role.

First, let us highlight that each member was directly involved in studying and preparing the group meetings to present the LLM phenomenon to the tutors; these meetings were essential to developing the conceptual maps and the paper published at the 2024 CAiSE Forum in Cyprus.

Everyone contributed equally to the development of the paper, from both the content and the writing perspectives.

Here we report a brief description of the skills of each team member:

- **Lorenzo Bertetto**: Lorenzo is a Computer Science student from Politecnico di Torino. Lorenzo went with Claudio to present the team's work in Cyprus at the CAiSE conference. He applied his mastery of a wide range of software tools to build the backend part of the graphical interface; more specifically, he curated the MongoDB database containing the data and developed the application with Docker.
- **Francesca Bettinelli**: Francesca is a Mathematical Engineering student from Politecnico di Milano who is also currently studying Computational Science in Lausanne. She applied her knowledge in computational methods to analyze the HF Data that Alessio and Claudio gathered, extracting sharp insights that highlight the shortcomings of the HF platform.
- **Alessio Buda**: Alessio is a Computer Science student from Politecnico di Milano, where he is specializing in Artificial Intelligence. His deep understanding of Entity-Relationship (ER) models has been essential for the team, and he worked with Claudio on the extraction of the Hugging Face Data. In the past year he coordinated the efforts of the team with dedication, and he was in charge of organizing the recurrent meetings with the tutors.
- **Marco Da Mommio**: Marco is a Mathematical Engineering student from Politecnico di Milano, specializing in Statistical Learning. He worked closely with Lorenzo to build the graphical interface of the Conceptual Map, curating the frontend part of the application using Streamlit and HTML.
- **Simone Di Bari**: Simone is a Telecommunication Engineering student from Politecnico di Torino, and he obtained a double degree with the University of Chicago. In the past year he curated all the presentations and deliverables with assiduity, providing support to all his teammates despite being located in Chicago with a completely different timezone.
- **Claudio Savelli**: Claudio is studying Data Science and Engineering in the Politecnico di Torino. His enthusiasm for the world of Large Language Models has been a driver during the whole project, and he presented the work of the team in Cyprus at the 2024 CAiSE Forum conference. He also curated the extraction of the HF Data with Alessio, applying his data science skills to automatize the scraping.

Goal

The goal of the chatIMPACT project is to allow users to explore the Large Language Model domain. While this field is gaining increasing importance for companies, universities and individuals, there is currently a lack of a unified source of information about available models, training datasets, evaluation metrics and solvable tasks. This gap hinders the application of models to new fields, limiting the potential of this technology.

Our industrial partner, the tech consulting firm Advanced Business Solutions (A.B.S), has expressed interest in a tool that would allow them to gather all the information they need to choose and deploy a model for their customers simply. Such information is currently available, only partially, on the HuggingFace platform.

To build the required tool, different sub-goals have been defined:

- **Characterize the state of the art**: To better understand the field, the literature on LLMs has been analyzed. Four main entities defining the domain have been identified: large language models, Datasets, Tasks, and Metrics.
- **Analyze information availability on HF**: HF is the most used platform to retrieve information on LLMs and related entities. While offering a large repository for models and datasets and providing lists of metrics and tasks, HF has some limitations. Information is loosely structured, allowing users to insert new entities even when some details about them are missing.
- **Data collection**: data to populate our platform was collected manually by selecting the most interesting items among models and datasets. Code has also been written to extract data from HF automatically. However, as of today, the missing information still needs to be filled out manually.
- **Interface Design**: An easy-to-navigate interface has been designed, and a database with manually curated data has been connected.

The outcome of this project is a proof of concept demonstrating the usefulness of our tool and the advantages it provides with respect to HF.

Understanding the problem

The popularity of Large Language Models has increased in recent years, particularly after the release of the popular “chatGPT” chatbot. LLMs are powerful tools that learn patterns from large amounts of textual data and exploit those patterns to generate answers to user queries.

However, the lack of a unified knowledge base about LLMs and related entities hinders their potential. It is difficult for a non-expert user, whether an employee or an academic researcher, to find information about available models, the data used during their training, the tasks they can solve, or the most suitable metrics to evaluate them.

The most relevant source of information on LLMs, the HuggingFace platform, presents several limitations:

- **Unstructured data**: HuggingFace data does not present a fixed structure. Many pieces of information describing the different entities are simply reported in a list of tags or described in free-text descriptions. While this approach allows for more flexibility, it limits the possibility for users to perform meaningful queries.
- **Missing data**: Only a small subset of fields of HuggingFace entities are mandatory for users to fill in when uploading them. This makes the upload process easier and faster but, on the other hand, once again restricts users' ability to conduct significant searches.
- **Complex querying**: Retrieving information on HuggingFace, especially linking different entities, is often complex, if not impossible. This is mainly due to the limits mentioned above but also to the general design of the HF platform, which is primarily structured as a simple catalogue of models and datasets.

These limitations demonstrate the need for a more structured data source for more complex querying.

The support of A.B.S. was paramount to better understanding the problem and the requirements for our platform. Thanks to their daily experience with LLMs and awareness of HF limitations, A.B.S. provided invaluable insights that helped us shape our platform based on end-users needs.

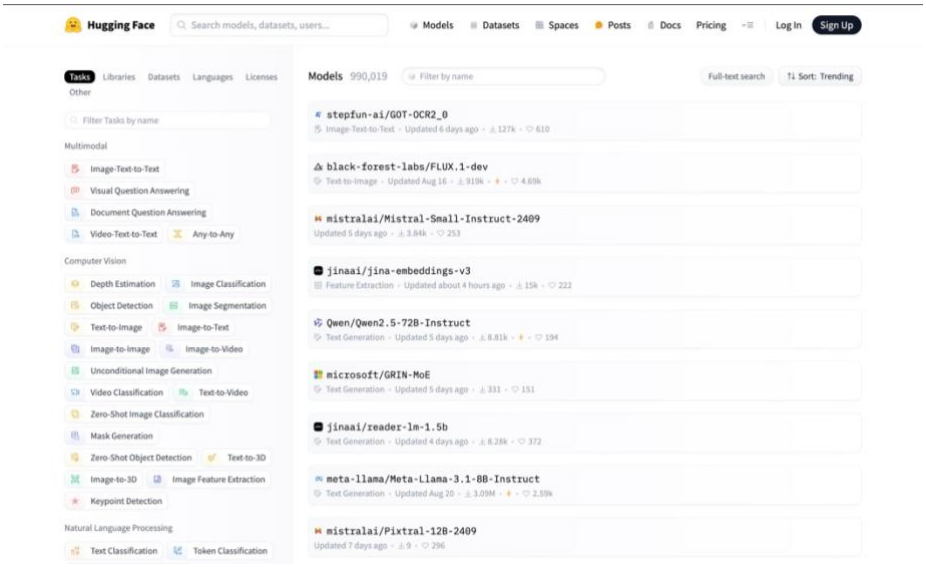

Figure 1: HuggingFace Models Search Page

Exploring the opportunities

The project offers numerous opportunities to meet various unaddressed needs in the fast-growing field of Large Language Models (LLMs). There is a rising need for an easy-to-use tool to assist users in navigating and selecting the most suitable model, considering the expanding usage of LLMs across different industries. Therefore, a system that simplifies the selection, analysis, comparison, and evaluation of various models is essential. This project utilizes its features to fulfill multiple requirements:

- **Improved LLM Discovery and Evaluation**: One of the leading market opportunities is the need for tools that provide in-depth information about LLMs and enable comparisons based on technical specifications, performance metrics, and application domains. Existing platforms like Hugging Face offer some features but fall short of providing structured and accessible data. We address this gap directly by offering an intuitive interface and a comprehensive conceptual model, making it easier for users to discover models that fit their needs. This opens opportunities in industries where model performance and domain specificity are critical.
- **Customization and Fine-Tuning for Specific Domains**: Companies increasingly use AI solutions customized to their needs. We offer a platform for identifying models that are or can be fine-tuned for specific tasks or domains, such as processing legal documents or summarizing medical reports. This customization capability creates opportunities for sectors like human resources, education, and content creation, where businesses need AI models optimized for their unique data and tasks.
- **Enterprise Integration and Workflow Optimization**: Businesses require AI solutions that seamlessly integrate with their existing workflows and data pipelines. We simplify the selection of LLMs and enable businesses to optimize their model deployment strategies. This reduces time-to-market and operational efficiency, creating opportunities for companies offering AI-driven automation in content generation, customer service, and business intelligence.

These opportunities make our solution versatile. It fills existing market gaps and facilitates new possibilities for AI adoption across industries.

Generating a solution

The ChatIMPACT solution offers a novel platform for exploring and comparing Large Language Models (LLMs), addressing the current lack of structured, easily accessible information in this domain. The project bridges gaps in existing resources like Hugging Face by implementing a comprehensive conceptual model that maps out LLMs, datasets, metrics, and downstream tasks in a detailed entity-relationship diagram. This model allows users to perform different queries, such as finding the most appropriate models for specific tasks, training datasets, or evaluation metrics, enhancing the usability and reliability of LLM selection.

The solution prototype consists of two main components. The first component is a MongoDB-based database that stores and manages extensive data about LLMs, including their

specifications, training data, performance metrics, and application domains. With its comprehensive coverage, this database embodies the conceptual model's flexibility, supporting a wide range of user queries and ensuring scalability for future data expansion. The second component is an interactive web interface developed using the Streamlit Python framework. This interface, designed with user-friendliness in mind, serves as the primary point of interaction between users and the database. It offers a seamless experience that simplifies LLMs' exploration, comparison, and evaluation. Users can filter models based on various attributes, making identifying models that meet their specific needs easier. Additionally, the interface provides visualization tools to help users understand the relationships between different LLMs, datasets, and metrics.

By integrating these two components, the ChatIMPACT solution enhances the accessibility and usability of LLMs for a diverse audience, from academic researchers to industry professionals. The database and interface work together to provide a comprehensive overview of the LLM landscape, offering users the tools to make informed decisions based on detailed model characteristics and structured data. Future enhancements may include automated data enrichment from additional sources, further expanding the database's coverage and utility.

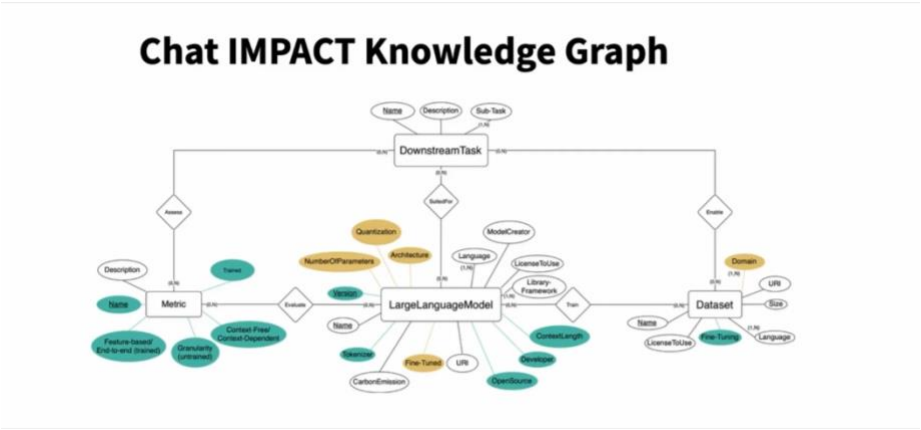

Figure 2: chatIMPACT Homepage

Main bibliographic references

Bertetto, L. B. (2024). Towards an Explorable Conceptual Map of Large Language Models. *Proceedings of the 36th International Conference on Advanced Information Systems Engineering.* Limassol: Springer.

Brown, T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165.*

Hugging Face – The AI community building the future. (2024, September 24). Tratto da Hugging Face: https://huggingface.co/