Applied Data Science Project

Designer AI 1

Mentors:
Silvia Chiusano
Alessandro Fiori
Andrea Avignone
Giuseppe Rizzo

Group Members:
Mehrbod Noworuz S315925
Masoud KhalilianS308102
Bahrak QaderiS296876

Table of contents

01
Introduction

Describes the project's Objecitve and stakeholders

02
Research Question

The questions that are to be answered by this presentation

03
Data

Explanations on data features and how they were prepared.

04
Implementation

Describes the structure of our model and methods and LLMs

05
Evaluation

Demonstrates the results of our project

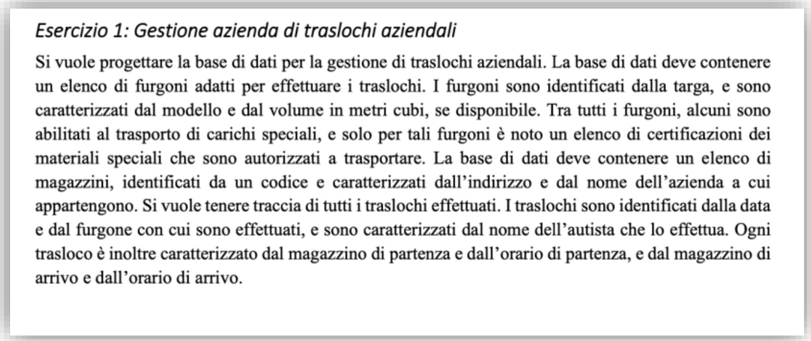
06
Conclusion

Explains the answer to the research questions.

Introduction

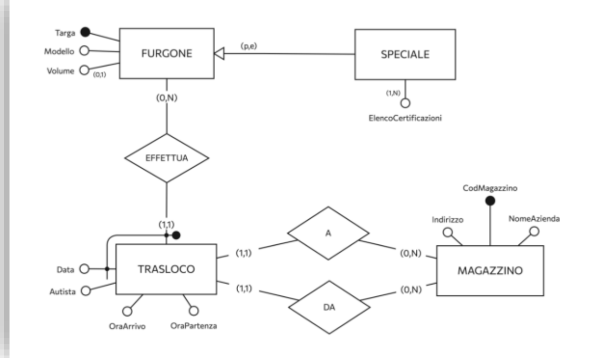
As Artificial Intelligence is moving forward, it is making its way into educational systems. With examples like ChatGPT, we have seen the role of these AI models in helping students with better understanding of various topics. Therefore, implementing AI into education serves as the future of education and brings higher quality education which is the 4th goal of UNSD.

Our project is also aiming to develop a module that would help students with creating and understanding ER models through turning their text instruction of the logic into an ER model.



Esercizio 1: Gestione azienda di traslochi aziendali

Si vuole progettare la base di dati per la gestione di traslochi aziendali. La base di dati deve contenere un elenco di furgoni adatti per effettuare i traslochi. I furgoni sono identificati dalla targa, e sono caratterizzati dal modello e dal volume in metri cubi, se disponibile. Tra tutti i furgoni, alcuni sono abilitati al trasporto di carichi speciali, e solo per tali furgoni è noto un elenco di certificazioni dei materiali speciali che sono autorizzati a trasportare. La base di dati deve contenere un elenco di magazzini, identificati da un codice e caratterizzati dall'indirizzo e dal nome dell'azienda a cui appartengono. Si vuole tenere traccia di tutti i traslochi effettuati. I traslochi sono identificati dalla data e dal furgone con cui sono effettuati, e sono caratterizzati dal nome dell'autista che lo effettua. Ogni trasloco è inoltre caratterizzato dal magazzino di partenza e dall'orario di partenza, e dal magazzino di arrivo e dall'orario di arrivo.



STAKEHOLDERS MAP

Core

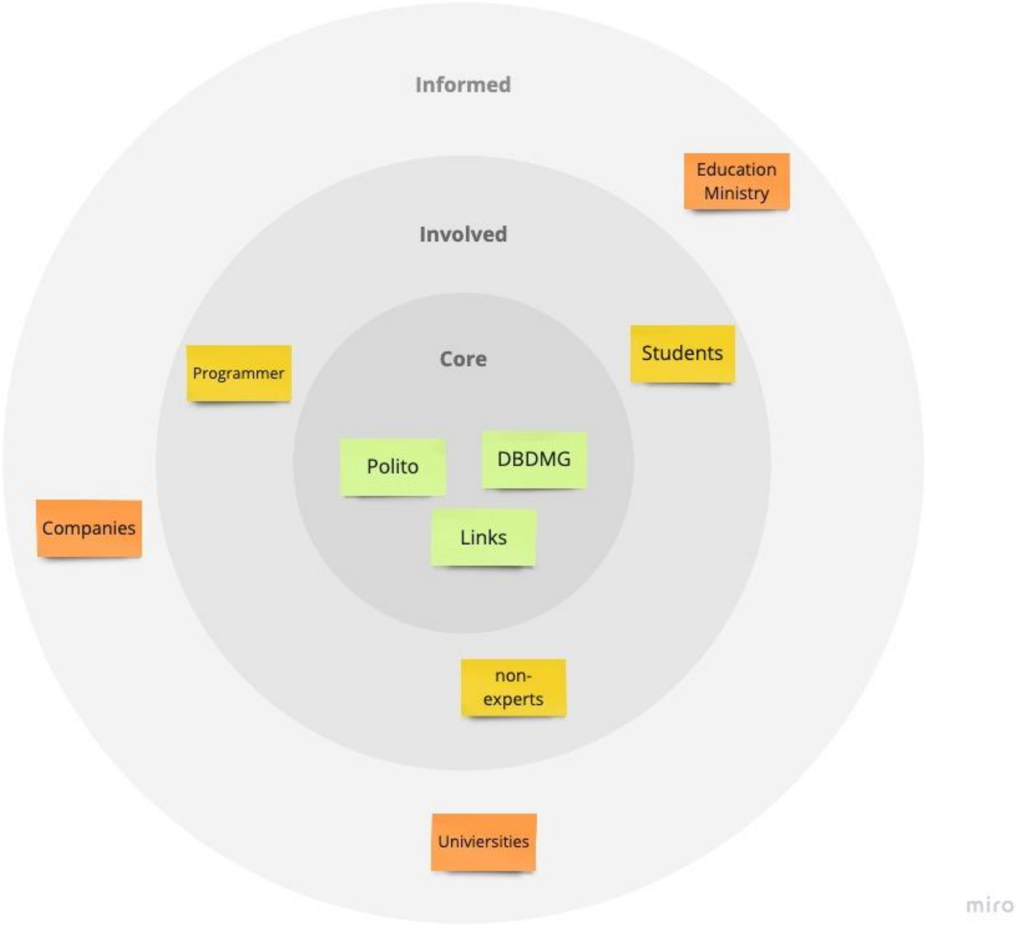
End users, Direct beneficiaries

Involved

Facilitators that can help to promote and encourage the adoption

Informed

Supervisors or experts to keep informed



Research Questions

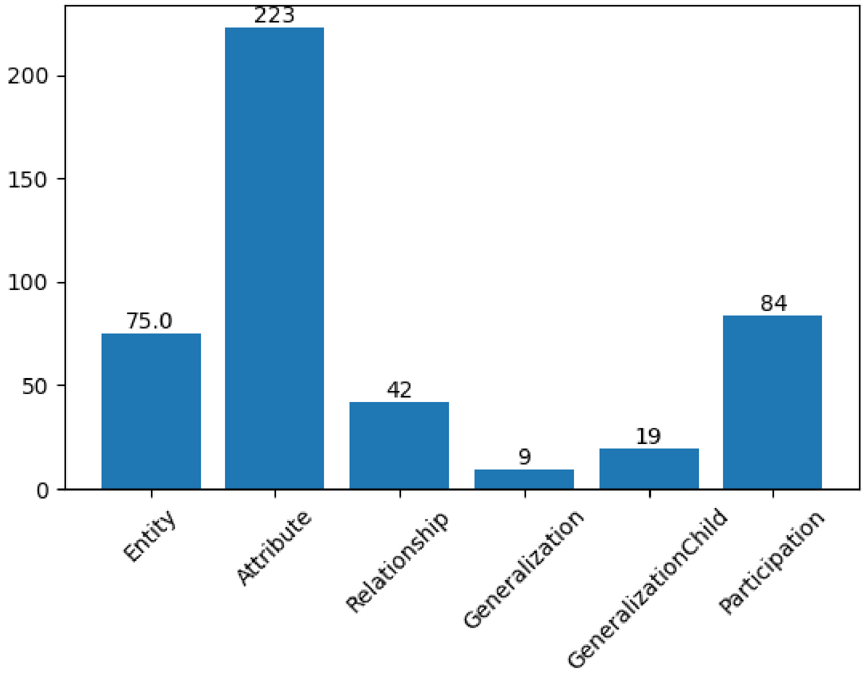
1. Are Large Language models developed enough to be a reference in database designing, particularly as an educator or an assistant tool?

2. How reliable are they?

3. Which LLMs would best suit this task?

Data

Due to the nature of Entity Relationship Diagrams, Some elements tend to appear more often than the rest, leading to an imbalance in the dataset.



Data Preparations

01

Data Cleaned

Language, Extra Elements

02

Data Transformed

Into the format expected as output

03

Generate Knowledge

For Model to better understand the Task

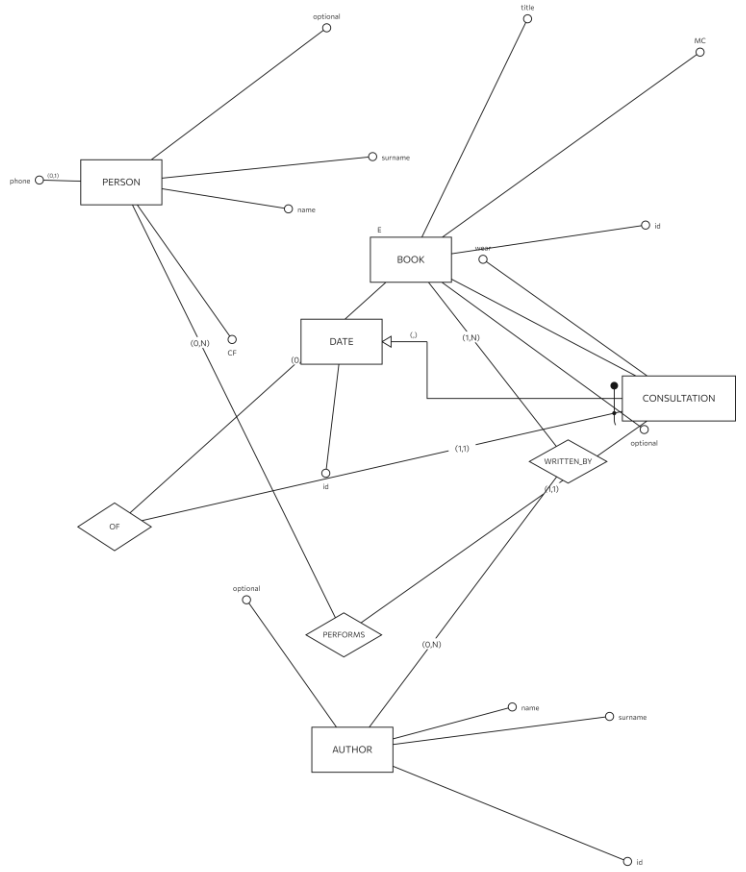
04

Fine Tune

Fine-tune of GPT and Lora/Qlora for other models.

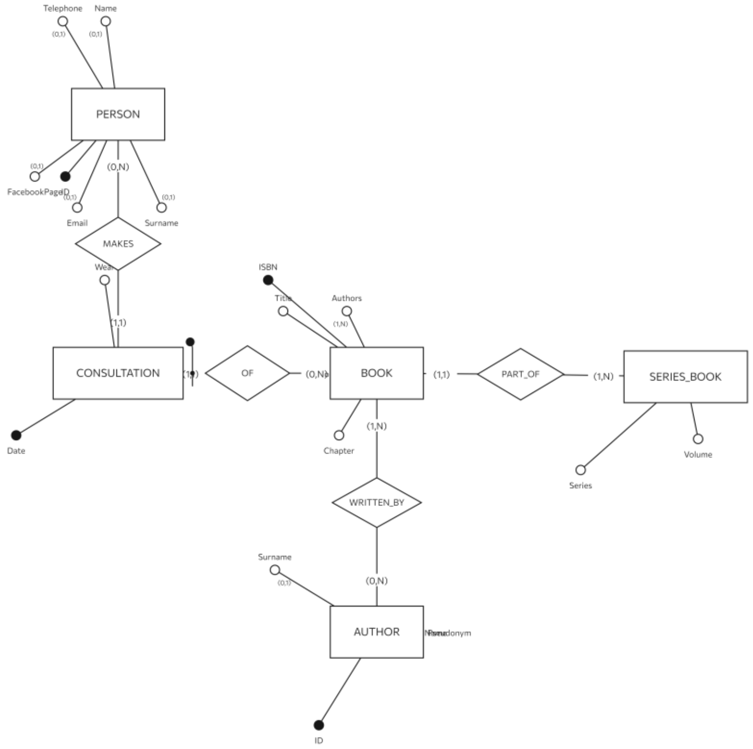
Positioning

In earlier versions of the outputs, positions where rather random and difficult to read and re-order.



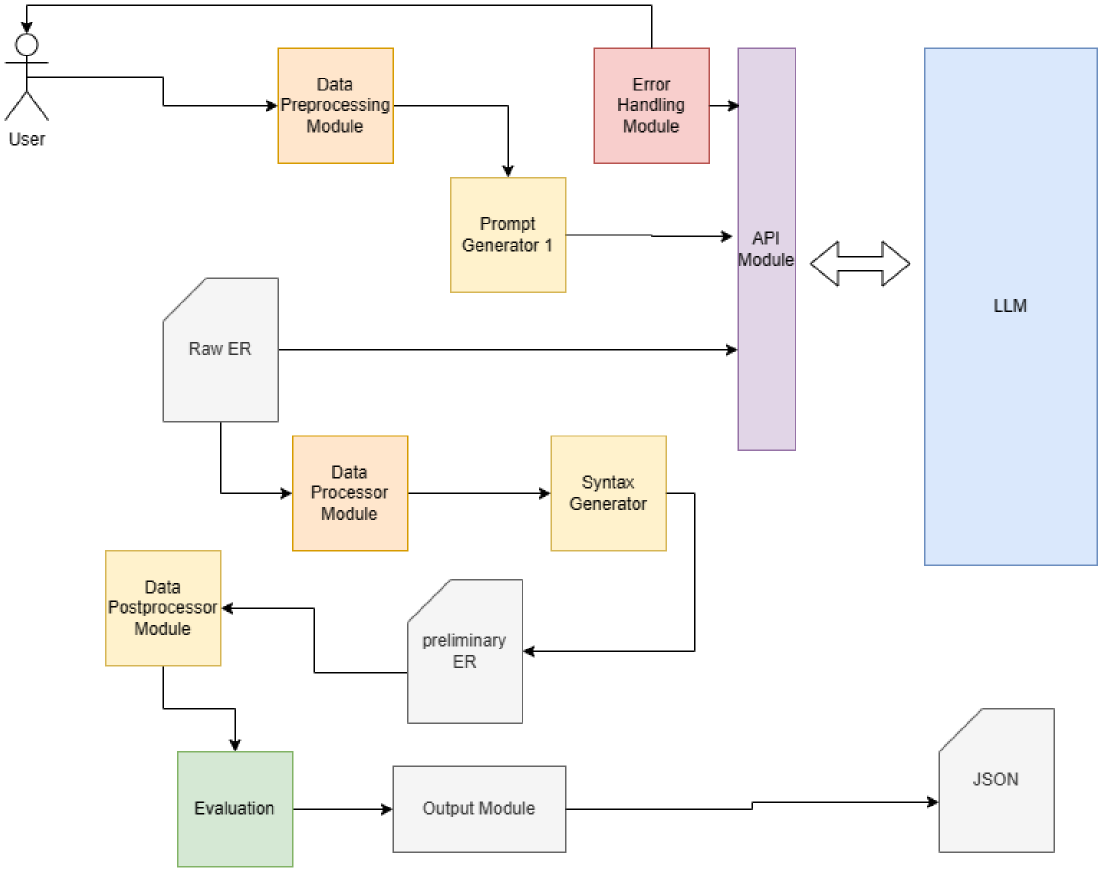
Improved Positioning

The algorithm starts from one point on the 2D plane and adds the positioning of the new elements below or to the right of the previous elements.

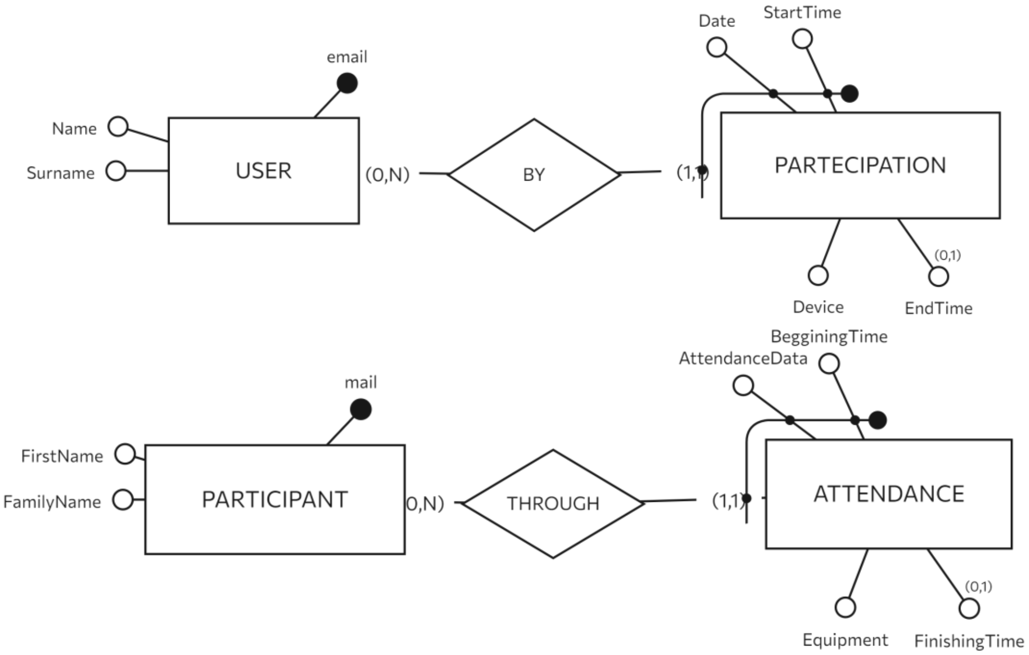


Final Implementation

Instead of expecting the LLM to also produce the syntax, we produce the logic using the LLM and this output is then read by a Syntax Generator module to guarantee a JSON readable by designer.io.



Evaluation Challenge



Evaluation Metric

$TP = PredictedCount$

$FP = GroundtruthCount\ - PredictedCount$

$FN = 0$

OR

$TP = GroundtruthCount$

$FP = 0$

$FN = GroundtruthCount\ - PredictedCount$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times Precision\ \times Recall}{Precision + Recall}$$

Implemented Models

GPT 3.5 & 4

Produced by OpenAI. A well known LLM widely used as a baseline of LLM performance.

Gemeni

Gemini is a family of multimodal large language models developed by Google DeepMind, serving as the successor to LaMDA and PaLM 2.

Code Llama V2

Produced by Meta Pretrained on 2 trillion tokens and has a context length of 4096. Code Llama is particularly designed for programming tasks.

Implemented Open Source Models

Toppy

is an open source model that has distilled knowledge from most famous LLMs with only 7B parameters.

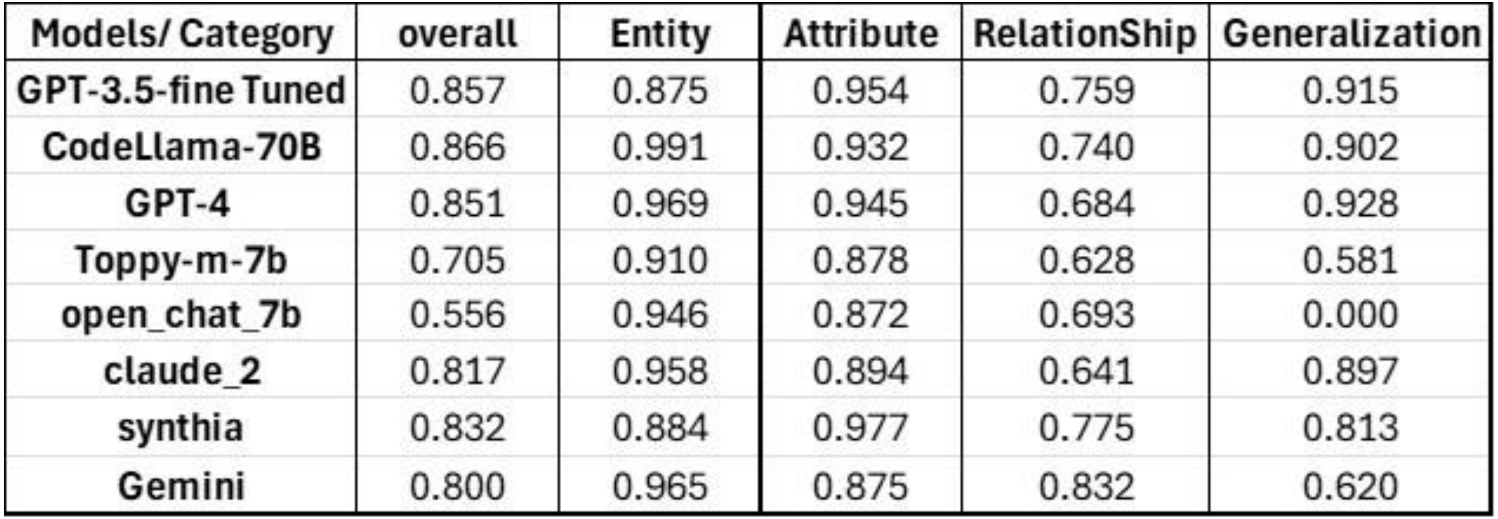
SynthiaIA

Synthetic Intelligent Agent is a Llama 2-70B model trained on Orca style Datasets.

Claude 2

is a leading LLM from Anthropic that enables a wide range of tasks from sophisticated dialogue and creative content generation to detailed instruction.

Experiments



Models/ Category	overall	Entity	Attribute	RelationShip	Generalization
GPT-3.5-fine Tuned	0.857	0.875	0.954	0.759	0.915
CodeLlama-70B	0.866	0.991	0.932	0.740	0.902
GPT-4	0.851	0.969	0.945	0.684	0.928
Toppy-m-7b	0.705	0.910	0.878	0.628	0.581
open_chat_7b	0.556	0.946	0.872	0.693	0.000
claude_2	0.817	0.958	0.894	0.641	0.897
synthia	0.832	0.884	0.977	0.775	0.813
Gemini	0.800	0.965	0.875	0.832	0.620

Future Work

Positioning

The positioning algorithm is capable of improvements.

Reverse System

LLMs to explain diagrams to students.

Positioning

The evaluation metric is rather naive.

Prompt

Prompt engineering is an under-development topic.

More models

Only GPT-3.5 has been fine-tuned.

More data

The model's performance is not at its peak

Conclusion

1 - Large language models have evolved to be valuable educational and assistant tools in database design, offering a wealth of knowledge and support for schema creation, query structuring, and performance optimization.

2 – The current results prove that even with minimum training. LLMs can guarantee nearly 85 % accuracy in this task.

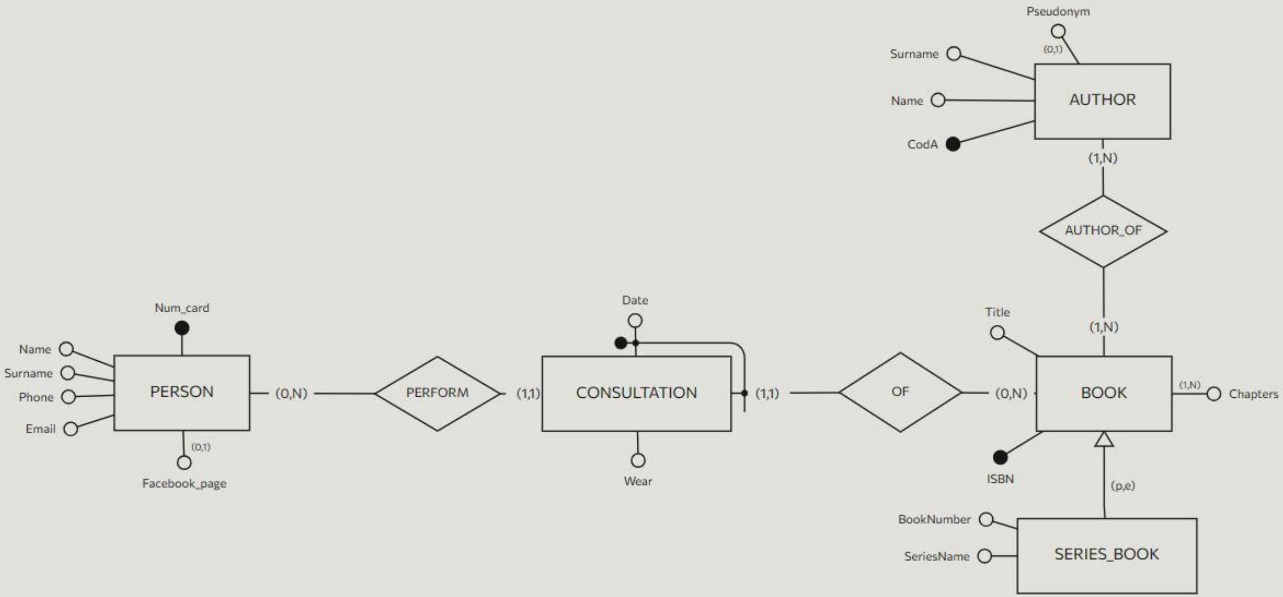
3 – Larger models, especially the ones that were fine-tuned on programming datasets are of the best performance and reliability among the LLMs.

Sample Task

We want to create the database for the consultation of historical books in a museum by some scholars. To consult the books each person must register at the museum and acquire a card. Persons are identified by the identification code of the card; you also know the name, surname, a telephone number, an e-mail address, and the address of the facebook page (if available). The books available at the museum are characterized by the ISBN, the title and the list of co-authors of the book. For each book the list of titles of the chapters of the book is also known. If the book is part of a series, you also know the name of the series and the number that characterizes the book within the series. Each author is identified by a unique code and characterized by the name, surname and any eventual pseudonym used to write. Each consultation of a book is characterized by the person who made it, the date on which it was made, the required book, and the state of wear of the book when withdrawn by the person for consultation. Consider that the same book can be consulted by at most one person on each day. Consider also that the same person can consult several different books in the same day and can consult the same book several times but on different dates.

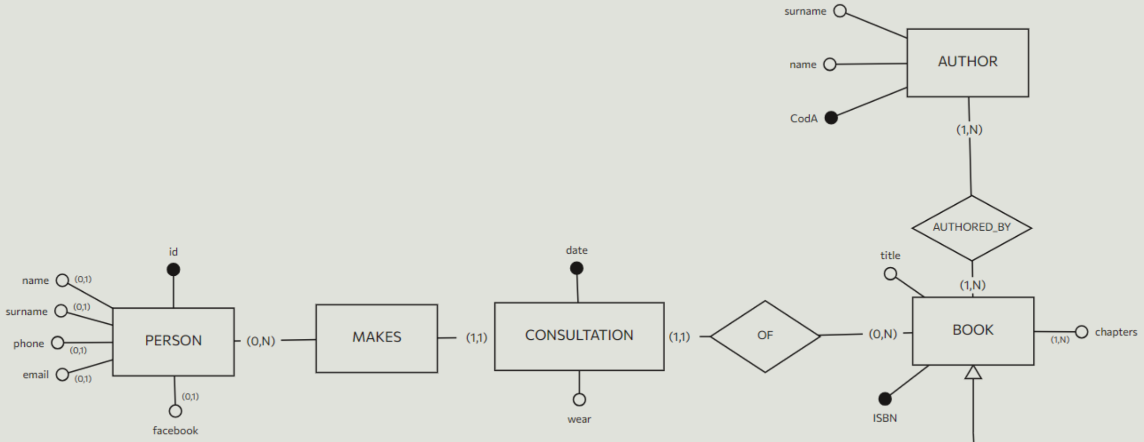
Challenge

Which one is the original ER diagram, and which is produced by our AI?



Challenge

Which one is the original ER diagram, and which is produced by our AI?


surname
name
CodA
AUTHOR
(1,N)
AUTHORED_BY
(1,N)
title
id
date
name (0,1)
surname (0,1)
phone (0,1)
email (0,1)
PERSON
(0,N)
MAKES
(1,1)
CONSULTATION
(1,1)
OF
(0,N)
BOOK
(1,N)
chapters
(0,1)
facebook
wear
ISBN
BookNumber
SeriesName
SERIES_BOOK