Background

**LLM Selection and Testing**

As part of our sample creation process, we decided to focus on the potential generation of synthesizing data using the power of different Large Language Models. There are several reason for this decision, mainly:

* **Model’s Capabilities** - LLMs exceptional ability to generate coherent and contextually accurate text, which is crucial for creating realistic and diverse synthetic samples.
* **Accessibility and User Friendly** - Using any LLM for our assignment does not require any prior knowledge or expertise, which makes this process easy to replicate and adapt if needed.
* **Scalability**  - These models can quickly produce large volumes of high-quality data, significantly reducing the time and effort required.

Therefore, LLMs became an integral part for our project - serving as powerful tools for generating synthetic samples that mirror the complexities of real data. To ensure robust and diverse sample creation, we evaluated several LLMs, including ChatGPT versions 4 and o1-mini, Ginimi, LLaMA, and Cloude. Each of these models offers unique capabilities in understanding and generating text, which is crucial for producing high-quality, representative data. By testing multiple LLMs, we aimed to assess their effectiveness and reliability in generating valid samples, ensuring that the augmentation process enhances model performance without introducing biases. This approach allows us to select the most suitable LLMs for our specific needs and then evaluate its performances.

To ensure the reliability and quality of the synthetic samples generated by the various Large Language Models (LLMs), we implemented a simple but effective validation process. This process began with basic checks to assess each model's convenience and ease of use, confirming that they could be operated with ease (so the process could be replicated easily). We then evaluated the model's knowledge of the dataset and its ability to generate new samples without memorizing, ensuring that the synthetic data was genuinely novel and not merely copied from existing sources. Additionally, we tested each LLM's capacity to produce cohesive and contextually accurate samples that made logical sense within the dataset's framework and in real life (like proper age). These validation steps were crucial in verifying that the generated data was both diverse and representative, and therefore a valid candidate for generating synthesized samples using causal inference.

Attached are some of the conversations made with each Model during this process. We aimed to use the same prompts for each model in order to truly test the model’s capabilities (without introducing any bias or advantages specific to a particular model).

Across all models tested, several common capabilities emerged. All models were able to successfully generate a valid new sample based on the Adult Income dataset. In Addition, each model demonstrated some understanding of the dataset and maintained logical consistency with real-world features. However, there were still significant differences in the models' overall performance and capabilities.

At the beginning, we tested **Chat-GPT** models (both the 4o version and the newer mini-o1 version) and found them to be highly effective for generating synthetic data, showcasing several key strengths that were crucial for our project. It should be noted that the o1-version is still in preview mode, so certain features are not yet available or fully developed, resulting in occasional performance limitations. Firstly, Chat-GPT demonstrated familiarity with the Adult Income dataset, showing a clear understanding of its features and their possible values. Additionally, it was able to generate large volumes of samples, from single entries to tens of thousands, and was the only model that succeeded in creating over 10,000 samples. The model provided a comprehensive explanation about the sample generation process and its ability to adjust parameters and generate realistic samples with meaningful correlations between features, ensuring that the synthetic data closely mirrored real-world relationships.

Importantly, ChatGPT produced genuinely novel data by randomly selecting values for each feature.

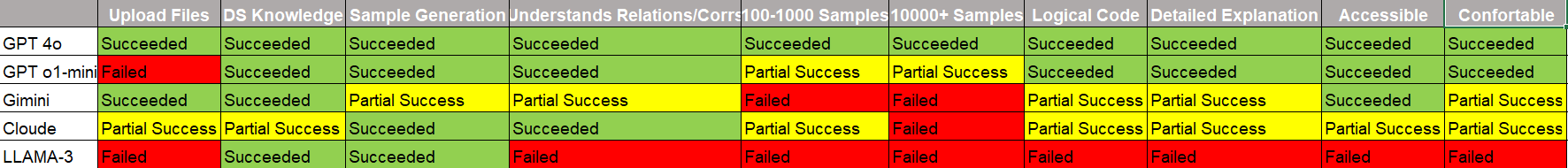
However, while ChatGPT generated logically consistent data, it lacked inter-feature correlations, meaning it did not fully capture the relationships present in the original dataset. Despite this limitation, ChatGPT’s ability to consistently create diverse, realistic, and customizable samples without copying existing data made it an invaluable tool for our data augmentation process.

We found **Gemini** to be a suitable model for general data creation, though it had a few limitations compared to the GPT models. On the positive side, Gemini easily identify the columns and target features, and it could generate up to 10 samples with logical accuracy, where each individual feature made sense. However, some of the generated samples were exact replicas of existing ones, with only minor changes, which hindered its ability to create genuinely new data. Additionally, when tasked with generating 100 or more samples, Gemini struggled to create the data himself and instead provided Python code to generate random feature values. Furthermore, even after tweaking the prompts and requesting new samples, Gemini failed to significantly improve, often changing only a single feature or producing repetitive output. Despite these challenges, Gemini’s solid understanding of individual features resulted in logically consistent samples, making it effective for smaller-scale data generation.

We found **Cloude** to be effective for generating smaller batches of synthetic data, successfully creating up to 100 new and original samples of the Adult Income Dataset with ease. Additionally, Cloude provided a detailed explanation of the correlations and logical inferences used in the data generation process, demonstrating a solid understanding of the dataset’s complexities and constraints. However, Cloude struggled significantly when tasked with producing larger datasets, such as 1,000 samples, often leading to incomplete or failed requests. In fact, after providing Python code to generate samples, Cloude had difficulty executing the task itself, failing to scale beyond smaller datasets. Even when Cloude did manage to generate samples, its performance was inconsistent, often requiring multiple attempts before success, making it unsuitable for extensive data generation tasks.It also should be noted that Cloude has limitations on the number of requests it can handle, making it difficult to use for repetitive or more complex requirements. This makes Cloude appropriate for targeted data augmentation but limits its utility for more comprehensive data synthesis needs. ​

We found **LLaMA** to be capable of generating individual samples effectively when provided with specific prompts, such as asking for a single row from the Adult Income Dataset. However, its performance significantly declined with larger requests, as it was notably slow and ultimately crashed when tasked with generating 100 rows, limiting its practicality for extensive data augmentation.

Here is a summarized comparison between all Models:



After conducting the tests and analyzing the performance of various LLMs, we have decided to primarily use ChatGPT, as it consistently proved to be the most effective model for our needs. While we may also consider using Cloude or Gemini for smaller-scale tasks, ChatGPT remains the best overall choice for generating high-quality, diverse synthetic data.