Assessing the Reliability of Large Language Models in Generating Synthetic and Representative Data Sets for International Study Program Recommendations

Rita Zonjić
University of Zagreb
Faculty of Electrical Engineering and Computing
Zagreb, Croatia
rita.zonjic@fer.unizg.hr

Luka Alfirević
University of Zagreb
Faculty of Electrical Engineering and Computing
Zagreb, Croatia
luka.alfirevic@fer.unizg.hr

Karlo Bašić
University of Zagreb
Faculty of Electrical Engineering and Computing
Zagreb, Croatia
karlo.basic@fer.unizg.hr

Ivan Džanija
University of Zagreb
Faculty of Electrical Engineering and Computing
Zagreb, Croatia
ivan.dzanija@fer.unizg.hr

Mihaela Vranić
University of Zagreb
Faculty of Electrical Engineering and Computing
Zagreb, Croatia
mihaela.vranic@fer.unizg.hr

Abstract

The rise of generative AI, particularly Large Language Models (LLMs), has revolutionized the tech industry, research, and society, sparking excitement about their potential to automate tedious tasks while also raising concerns over their reliability, ethical implications, and misuse. This study investigates the reliability of popular open and free LLMs, such as ChatGPT, Gemini, and DeepSeek, in generating accurate, well-supported data sets in the field of education. Our research is rooted in a project where a web application was developed to help students find suitable international study programs. The project required extensive data collection, including tuition fees, university rankings, healthcare costs, general cost of living, etc. The sheer difficulty of finding and accumulating that data, which was oftentimes poorly documented, prompted us to examine LLMs' performance in replicating these data sets. We evaluated their outputs by calculating the coefficient of determination between synthetic and fact-checked real data. Results demonstrate that LLMs can generate representative data, though occasional deviations raise questions about their limitations and the complexity of prompt engineering. This research sheds light on the practicality of LLMs in automating data-intensive tasks while highlighting their potential and limitations, offering valuable insights into the broader capabilities of generative AI.

***Keywords*—generative AI, Large Language Models, data reliability, prompt engineering, synthetic data sets, data bias, international study program recommendations.**

I. Introduction

This paper explores the reliability of data generated using artificial intelligence (AI) for recommending international study programs to students seeking a new place to study. The issue arose in the context of a project where a sufficient amount of real-world data was not easily accessible, a common challenge today due to the increasing value, protection, and restricted availability of data.

This challenge led us to investigate whether it is possible to generate synthetic data that accurately represents a desired data set.

In our case, the data relates to choosing and planning studies at an international university. In this activity, certain parameters play a more important role, such as the cost of studies, the cost of living in different cities, and an especially important parameter is urban safety.

The research began with a data set derived from one publicly available data set, which was built upon by incorporating data from several additional sources. The process is explained in detail in subsection III.B. This allowed us to gather information on 120 universities. However, it is important to note that this data set is relatively small in scale.

To offer students information on a wide range of studies and universities, we would need a much larger data set. To gather our desired data set, two approaches have been identified. The first is highly time-consuming and laborious, relying on manual internet searches. The second approach leverages emerging technology – specifically, Large Language Models (LLMs).

These contrasting methods present different trade-offs in terms of time investment, resource allocation, and potential scalability for data generation.

Concurrently, we are witnessing a rapid advancement in Artificial Intelligence (AI), particularly in the increasing utilization and integration of Large Language Models (LLMs) across various industries, including education, healthcare, finance, e-commerce, science, etc. This widespread adoption of LLMs has prompted an investigation into the efficacy of these publicly accessible models in addressing our specific problem.

While similar studies have been conducted in other fields, there is a notable gap in research specifically targeting our domain of interest. This lacuna in the literature provides a compelling rationale for our current study.

To thoroughly examine the performance of LLMs in this domain, a comprehensive review of existing research was undertaken. The findings from this literature review are presented in section II. of this paper.

Subsequently, a methodology was tailored to address the unique aspects of our case study. This methodological framework is detailed in section III. Firstly, a clear outline of our approach is presented. A detailed description of our initial data set follows, after which special attention is devoted to prompt engineering due to the fact that prompt engineering has proven to be an essential factor in the process of quality dataset generation. Following this, the outcomes of the study are detailed in section IV. In section V., a comprehensive discussion follows, and in section VI., we summarize our findings and offer insights into potential future research directions.

II. Related work

With the increasing emphasis on data protection and the rapid advancement of generative artificial intelligence – two processes that have evolved almost in parallel – a significant number of studies have focused on generating synthetic yet reliable datasets using LLMs. Additionally, the substantial amount of data required for most machine learning methods has driven researchers to explore the concept of dataset synthesis. This is particularly evident in the fields of computer vision and natural language processing (NLP). In the domain of NLP, researchers have examined the ability of language models to generate synthetic data for various text classification tasks [1].

Furthermore, numerous studies have engaged with prompt engineering, a crucial aspect of optimizing LLM performance. For instance, one of them demonstrated that the consistency of extracted knowledge remains generally low when the same information is queried using different prompts [2]. Many works in prompt engineering have explored the automatic construction of queries that outperform manually created ones [3, 4]. Approaches to prompt engineering vary significantly. While some researchers focus on enhancing data generation processes by utilizing soft prompts, white-box LLMs, and seed examples for fine-tuning [5], others explore methods applicable to black-box LLMs and even LLM APIs such as ChatGPT [6]. Given the remarkable success of LLMs, their utilization in generating diverse forms of data is unsurprising. These range from relation triplets [7] and sentence pairs to instruction data [8] and tabular data – one of the most common data formats in machine learning. Notably, approximately 65% of datasets in the Google Dataset Search platform contain tabular files in either CSV or XLS format [9].

While we have not identified a study specifically addressing the generation of synthetic data for an international university recommendation system, numerous studies have investigated other relevant topics, such as the generation of synthetic medical data [10]. Additionally, research has highlighted the importance of selecting an appropriate LLM for data generation. One study stands out as the first attempt to systematically evaluate various LLMs' data generation capabilities using established data generation methods [11].

However, synthetic data is not without its challenges. Task-specific synthetic data often lacks diversity and exhibits regional biases [6], with effectiveness varying by task nature [12], raising concerns about the reliability of LLM-generated datasets in different application domains. The research presented in this paper provides new insights into the generation of data in a completely novel domain, distinct from those explored in previous studies [10]. While one study [13] highlighted potential biases in LLMs when generating recommendations, no research has systematically assessed the accuracy and representativeness of synthetic data in the context of international study programs – an essential aspect for cross-cultural applications.

In our study, we utilize R^2 to evaluate the reliability of various LLMs in generating synthetic data for international study program recommendations. A similar approach is discussed in [14], where the authors introduce a utility-driven framework for assessing synthetic data, highlighting both the strengths and limitations of R^2 as a reliability metric. While R^2 effectively measures explained variance, they emphasize the necessity of complementary evaluation methods to ensure the accuracy, representativeness, and overall quality of synthetic data. To further enhance reliability, we integrated prompt engineering principles into our data generation process. We constructed a detailed and structured prompt that explicitly outlines the required data attributes, format, and scope (including university rankings, tuition fees, acceptance rates, and cost-of-living factors across multiple countries). This approach aligns with [15], as our prompt was crafted to minimize ambiguity and provide clear task constraints, ensuring that the LLM generates structured, domain-specific data.

III. Methodology

A. General overview

The methodology is clearly illustrated in Fig. 1. At first, we created an initial data set required for university recommendations within the application. This dataset is described in greater detail in subsection B. From this dataset, we selected parameters that LLMs should generate, under the reasonable assumption that these models would be capable of generating the requested data, considering they have been trained on extensive datasets and some currently possess data retrieval and web searching capabilities. We collected data from several different LLMs and subsequently created new datasets, which we then compared and evaluated against the initial dataset.

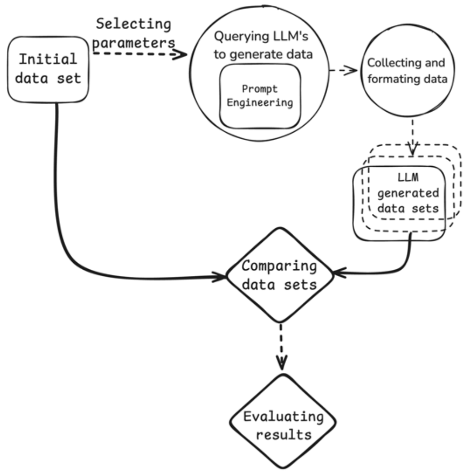


Fig. 1. Overview of the methodology

B. *Initial data set*

When collecting and preparing our initial data set, we first merged data sets from [16] and [17]. Afterwards, we used information from [18], [19], [20], and [21] to fill our initial data set with the information we needed.

Our initial dataset is described in Table 1. It contained 4 character type variables, 11 numerical type variables, and 10 Boolean type variables. Finally, we had a description of 120 different universities from 120 different countries. For each university, the earlier mentioned Boolean type variables represented a specific degree program that the university offers.

TABLE I. VARIABLES AND ATTRIBUTES OVERVIEW

Variable	Type	Unit	Explanation
Country	Character		
Region	Character		Continent
University	Character		
University website link	Character		
University ranking	Numeric		Ranking according to [9]
Percentage of international students	Numeric	%	
Acceptance rate	Numeric	%	
Safety index	Numeric		Relative safety, according to [4]
Cost of living	Numeric	EUR/month	Prices of consumer goods (excluding rent)
Rent cost	Numeric	EUR/month	
Groceries cost	Numeric	EUR/month	
Recreation cost	Numeric	EUR/month	
Healthcare cost	Numeric	EUR/year	
Transportation cost	Numeric	EUR/month	Prices of monthly pass for public transport
Tuition	Numeric	EUR/month	
10 boolean variables depicting the presence of a specific degree program at the university in question	Boolean	Business, Economics, Psychology, Biology, Law, Computer science, Math, Physics, Art, Medicine	Academic degrees offered by the university

C. *Data generation and prompt engineering*

For data generation, the selected models were the most popular large language models (LLMs) at the time of the research: ChatGPT [22], Gemini [23], and DeepSeek [24]. These models were chosen not only for their popularity but also for their widespread availability and free access.

ChatGPT [22] is a generative artificial intelligence chatbot developed by OpenAI and launched in 2022. At the time of use, it was powered by the GPT-4o model (approx. 200 billion parameters) [25], offering advanced conversational capabilities. ChatGPT can generate human-like responses and allows users to guide and shape conversations to meet specific needs in terms of length, style, detail, and language.

Gemini [23], developed by Google, is a generative artificial intelligence chatbot. Launched in 2023. At the time of use, it was powered by the Gemini 2.0 Flash model. Gemini is designed to provide advanced conversational abilities and is used across a variety of applications.

DeepSeek [24], the eponymous chatbot launched in January 2025 and developed by Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd., is a cutting-edge generative artificial intelligence chatbot designed to provide advanced conversational capabilities. At the time of use, it was powered by the DeepSeek-R1 model (approx. 670 billion parameters) [26], which enables it to engage in highly sophisticated and natural conversations across a wide range of topics.

The first step involved generating data using large language models to obtain relevant information.
This process included the creation of numerical, character and boolean data for the initial data set.

Prompt engineering played a crucial role in this process, as it involved structuring queries in a way that ensured the AI correctly understood the task and produced the desired output. In particular, word choice, phrasing, and level of detail are things to keep in mind as they can significantly influence the quality and realism of AI-generated data. A well-designed prompt can lead to more accurate and contextually relevant outputs, while poorly formulated prompts may result in inconsistencies, biases, or unrealistic data points. Therefore, optimizing prompt design is a crucial step in ensuring that the generated data aligns with real-world characteristics and maintains high reliability.
Fig. 2. presents an excerpt from the interaction with the selected language models. To see what we would get, we sent the same initial query to all of the models.

I am working on an application in which I need accurate dataset containing universities with parameters that consist of:
- university (name of the university)
- collegeRank (global ranking of university)
- tuition (EUR/year) (how much is paid each year to study at the university, price needs to be in euros)
- percOfIntStud (percentage of international students based on all students)
- acceptance rate (how many people manage to get into the university in percentage)
- avgSafetyIndex (how secure is the country, how many crimes they have, and based on that index in the range from 0 to 100)
- cost of living (this consists of how much monthly persons pays on groceries, restaurants, transportation, and utilities excluding rent, price needs to be in euros)
- rent (monthly average for rent in euros)
- groceries (monthly average for groceries in euros)
- recreationCost (EUR/month) (monthly average for recreation and culture in euros)
- healthcare price (average out of the pocket spend on healthcare per year in euros) –
- avgMntTransportCost (average transportaion price in euros per month)
- link (link for the university)
- and each of the columns: Computer Science, Business, Economics, Psychology, Biology, Law, Medicine, Mathematics, Art, Physics they have values 1 or 0 based on the fact if university has these majors

The parameters avgSafetyIndex, cost of living, rent, groceries, recreationCost, healthcare price, avgMntTransportCost are based on the city in which university is located.
I want you to find me data for these countries:
For each country find the best university based on ranking and based on chosen university for each country write down other parameters for it.
Make me an excel file for this dataset.

Fig. 2. Query for creating the entire table (truncated list of countries)

1) ChatGPT

Initially, the model failed to generate the complete table as requested, instead directing users to external sources to gather the necessary data. This approach highlighted the model's limitations in generating specific and accurate data and formatting it within the desired structure, emphasizing the complexity of the task.

When the model was asked if it could generate the data set for the given task, it returned an empty table and once again referred to external sources for data collection. This response further underscored the model's limitations in fulfilling the task. In a way, however, this is appropriate, as the model demonstrates the necessity of referencing verified external data.

After requesting the model to populate the empty table provided earlier, it once again acknowledged the complexity of the task. However, the model expressed its intent to begin populating the table with a smaller sample of data.

Upon instructing the model to send the completed Excel file once finished, the model indicated that it would do so once completed. However, it noted that it could not send the Excel file at random intervals, which highlighted its constraints in delivering the file in a timely and scheduled manner.

When asked to send the data generated thus far, ChatGPT provided an Excel table named University_Dataset_Partial, containing data for five countries: Australia, Austria, Canada, Germany, and Japan. While previous responses included links to universities from countries such as Costa Rica, Spain, Italy, and Switzerland, these countries were not processed, and the model did not follow an alphabetical order when selecting the initial set of countries. This selection was beyond the user's control.

The data in the table closely mirrored the original data set, apart from the safety index parameter, where values generated by ChatGPT were significantly higher than those in the original data set. Additionally, numerical values were often rounded, as evidenced by the recorded tuition fee for the University of Vienna, listed as 1,500 EUR, while the actual fee was 1,453 EUR. Nevertheless, the links to the university websites in the table were correct, which was a positive outcome.

Through continued interactions, it became evident that ChatGPT struggled to generate a complete data set. This was particularly evident when the model explicitly stated, "I'll notify you." a capability it does not possess. The next optimization attempt involved reducing the task size to 20 countries. However, the model was still unable to provide data for all of them at once. Instead, it generated a table named University_Dataset_20_Countries, which contained data for only seven countries: the initial five (Australia, Austria, Canada, Germany, and Japan), along with two additional countries (the United Kingdom and India). This discrepancy between the table's name and its actual content further highlighted the model's limitations in handling complex data generation tasks.

Recognizing that ChatGPT struggled to generate a table containing a large number of universities, the decision was made to switch to a text-based response format (as presented in Fig. 3.).

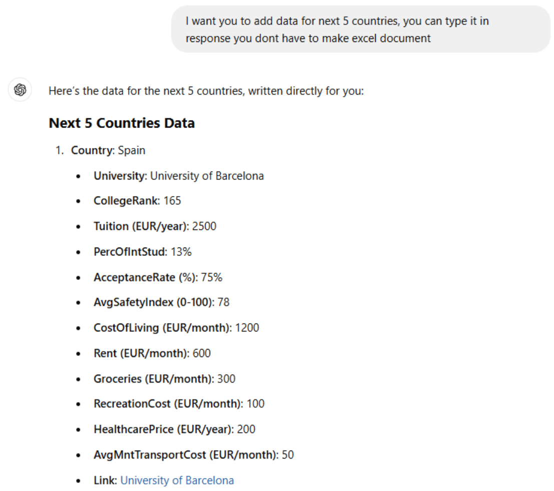


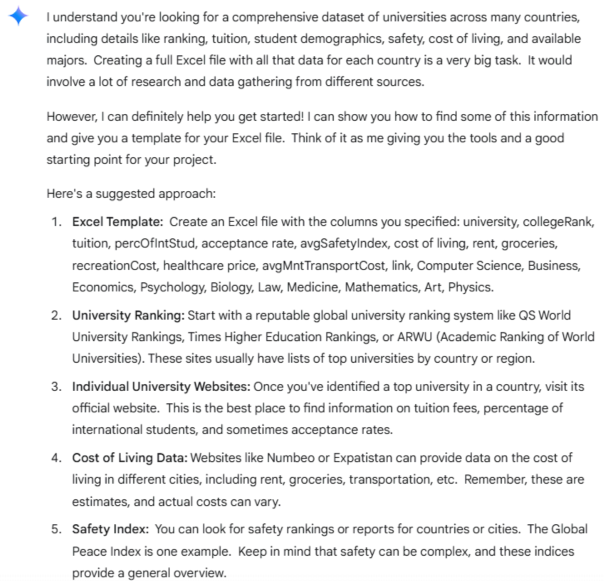
Fig. 3. ChatGPT's response to the query with reduced task complexity

In this case, the model promptly provided the requested data for five universities with the corresponding parameters, indicating progress towards achieving our goal. The next step was to increase the number of universities per response to 20; however, the model consistently returned only 10 universities per response.

Despite this limitation, using this iterative approach, all required data was eventually collected. Although the process required multiple interactions, this method proved to be a faster and more efficient solution.

2) Gemini

After successfully generating data using ChatGPT, we initiated the same process with the Gemini model (Fig. 4.) Drawing on previous experiences, we began with the same prompt.



6. **Majors Offered:** The university website should list the faculties, departments, or programs offered. This is how you'll determine if a university has majors like Computer Science, Business, etc. Use 1 for "yes" and 0 for "no" in your Excel sheet.
7. **Iterative Process:** This is going to be a step-by-step process. You might not find all the data for every university. Focus on finding reliable sources and be prepared to update your data as you find more information.

Example (Starting Point):

Let's take Germany as an example.

- **Ranking:** Looking at QS rankings, the Technical University of Munich (TUM) is often highly ranked.
- **Website:** Visit the TUM website (e.g., www.tum.de) to find information on tuition (which might be low or free for public universities in Germany), international student percentages, and available programs.
- **Cost of Living:** Use Numbeo to research the cost of living in Munich.
- **Safety:** Look up safety information for Germany or Munich.

I'm here to help you along the way. If you have questions about specific data points or need help finding information for a particular country, feel free to ask! Remember, doing thorough research is key to getting accurate data.

Fig. 4. Gemini's response to the initial request

The model responds similarly to ChatGPT when given the initial query (which is the same across all models). It emphasizes the complexity of the task and does not generate a completed table but instead suggests sources where the required information can be found.

Country: Bahamas
University: University of The Bahamas
CollegeRank: Not globally ranked
Tuition (EUR/year): 4,000
Percentage of International Students: 10% (estimate)
Acceptance Rate: 60% (estimate)
Average Safety Index (0-100): 40 (safety can vary significantly between islands)
Cost of Living (EUR/month): 1,800
Rent (EUR/month): 1,000
Groceries (EUR/month): 400
Recreation Cost (EUR/month): 200
Healthcare Price (EUR/year): 1,500
Average Monthly Transportation Cost (EUR/month): 100
Link: https://www.ub.edu.bs/
Majors: Computer Science (1), Business (1), Economics (0), Psychology (1), Biology (1), Law (0), Medicine (0), Mathematics (1), Art (1), Physics (0)

Fig. 5. Response to the previous query

Following interactions with ChatGPT, it was determined that requesting data within the chat itself yields better results. Using the same approach of requesting data within the chat, the Gemini model successfully generated the data, much like ChatGPT. Notably, Gemini also provided warnings regarding parameters where its estimations were uncertain or where values tended to fluctuate—an aspect that ChatGPT did not address.

Compared to ChatGPT, Gemini omits the ranking parameter. However, when the initial query was repeated for a smaller number of countries (30 in total), Gemini successfully generated a .csv file. This marks a significant improvement over ChatGPT, which, in a single iteration, only managed to add five rows of data.

3) DeepSeek

The initial request sent to DeepSeek was identical to those sent to other models (presented in Fig. 1). In response, the model stated that the task was too complex. However, it successfully generated a data set containing 10 rows per iteration and did not provide textual responses, unlike ChatGPT and Gemini. Through a total of 12 iterations, all necessary data were collected, with each iteration specifying the universities for which data needed to be generated.

Recognizing the significant differences in the safety index values, we opted for the seeds approach as in [15], where an initial parameter value is provided. In this case, Australia was chosen as the starting value. Additionally, we included all countries. Countries whose safety index differed by 25 or more are shown in Table 2. The value in the second column represents the difference between the obtained and initial values without the seeds approach, while the value in the third column represents this difference when using the seeds approach.

TABLE II. RESULTS OF INITIAL SEEDS APPROACH USING CHATGPT

Country	Deviation without seed	Deviation with seed
Argentina	28.54	10.54
Australia	29.04	0.04
Bahamas	39.32	4.42
Belgium	26.38	-3.02
Costa Rica	29.6	0.5
Fiji	27.34	0.74
Ireland	25.66	1.36
Jamaica	27.44	0.84
Malaysia	35.2	2.8
Malta	30.48	0.28
Mauritius	28	3.4
New Zealand	34	5.1
Puerto Rico	27.98	-1.92
South Africa	26.3	13.7
Sweden	33.04	9.54
Trinidad and Tobago	36.38	12.48
United Kingdom	25.96	2.76
Uruguay	27.26	11.46

This approach proved to be significantly more accurate and yielded much better results compared to methods that relied solely on the safety index without any value given as an example.

D. Verification of the Reliability of Generated Data

To assess the reliability of the generated data, we performed a quantitative comparison between the synthetic values produced by large language models (LLMs) and original manually collected data. This verification process was conducted for each generated parameter using statistical analysis and regression modeling.

1) Data Preprocessing and Merging

The data sets were first preprocessed to ensure consistency. Since multiple sources were used, the standardization of column names was necessary to align the country identifiers across data sets. All country names were converted to lowercase and stripped of extraneous spaces to prevent mismatches during merging. The generated data set was then merged with the original data set using an inner join based on the country column, ensuring that only countries present in both data sets were included in the analysis.

2) Evaluation of the model's performance

To evaluate the correlation between the generated and original data, we applied a linear regression model [27] for each parameter. Specifically, for a given parameter X_{LLM} (generated value) and its corresponding true value Y (original data), we fitted a simple linear regression model:

$$Y = \beta_0 + \beta_1 X_{LLM} + \varepsilon \quad (1)$$

where β_0 represents the intercept, β_1 is the regression coefficient, and ϵ is the error term.

The model was trained using scikit-learn's LinearRegression module, and predictions were generated for all available data points. The coefficient of determination (R^2-score) was computed to quantify the strength of the linear relationship between the generated and original values. An R^2-score close to 1 indicates a high degree of similarity, whereas a lower value suggests a weaker correlation and potential inconsistencies in the generated data.

3) Visualization and Interpretation

To facilitate interpretation, we generated scatter plots with regression lines for each parameter. The Seaborn library [28] was utilized to visualize the relationship between synthetic and real-world values, highlighting deviations and outliers. An example visualization for the cost of living parameter is shown in Fig. 6. This process was systematically repeated for all parameters under investigation, ensuring a comprehensive evaluation of the data set's reliability.

By applying this methodology, we were able to quantify the reliability of the generated data, identifying strong correlations for well-documented parameters (e.g., cost of living) while observing greater variance in more complex, less standardized or attributes more difficult to access (e.g., university acceptance rates). These findings underscore the necessity of verifying AI-generated data before its integration into real-world applications.

IV. Results

The results of our reliability verification process are summarized in Table 3, which presents the R^2 values for various parameters across the three LLMs used for data generation: ChatGPT, Gemini, and DeepSeek.

The visualization helped us gain a better insight into how the desired values relate to the values obtained by the models. An example of a visualization for the cost of living parameter, obtained using the Gemini LLM and without a seed, is shown in Fig. 6.

TABLE III. RESULTS OF GENERATING DATA USING LLMS

R^2 of each param. / LLM	ChatGPT	Gemini	DeepSeek
Tuition Cost	0.4112	0.4707	0.5071
Percentage of International Students	0.6245	0.7588	0.4269
Acceptance Rate	0.0810	0.0669	0.1173
Cost of Living	0.6581	0.7889	0.6399
Rent Cost	0.6241	0.6533	0.6163
Groceries Cost	0.5350	0.6961	0.4956
Recreation Cost	0.5420	0.5699	0.5685
Healthcare	0.2653	0.5832	0.2565
Transportation Cost	0.4764	0.5573	0.5790

Key observations

- Gemini consistently outperforms ChatGPT and DeepSeek across most parameters, achieving the highest R^2 values and lowest RMSE in key metrics. This suggests that Gemini provides more accurate and reliable numerical estimations compared to the other two models.
- ChatGPT and DeepSeek show mixed performance, with neither demonstrating consistent superiority over the other.
- Acceptance Rate is the most unreliable parameter across all models, with R^2 values below 0.12 for each model. This aligns with our prior expectations, as university acceptance rates are highly variable, poorly documented, and subject to differing definitions across sources. The lack of a consistent and reliable data set likely led to the poor performance in this category.
- Healthcare costs exhibit similarly low R^2 scores, particularly for DeepSeek (0.2565) and ChatGPT (0.2653). The variability in healthcare expenses across different regions and the complexity of cost calculation methods may have contributed to these lower values.
- DeepSeek generally has the lowest R^2 scores, except in Tuition Cost (0.5071) and Transportation Cost (0.5790), where it marginally outperforms the other models. This is likely due to its different data retrieval approach, which does not rely on live web scraping but instead generates values based on pre-trained data sets.

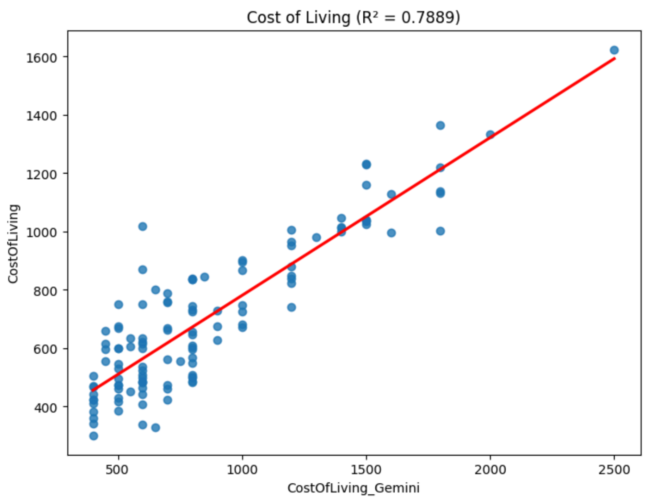


Fig. 6: Comparison of Generated and Original Cost of Living Data with Linear Regression Fit

V. Discussion

In our research, the goal was to assess how reliable today's LLMs are in generating data related to the selection of international study programs. To determine their accuracy and reliability, we started from an initial data set that we tried to recreate using LLMs.

For the most part, through careful analysis, we have deemed several large language models (ChatGPT, Gemini, DeepSeek) to be “good enough” sources of information. That analysis was done by comparing the output of our prompts with the initial dataset - premade, hand-gathered and backed-up data.

LLMs are essentially black boxes that return data according to some of the sources they were trained on. Ultimately, different models returned different data.This can be attributed to the different sources they rely on.

Not all of those sources were necessarily reliable, which is why prompt engineering and data generation always requires a good amount of scepticism and testing and could never be fully automated. However, we have found that data that is “simpler” in some way and is well documented online will usually be more reliably generated.

A good example of such simple data is the approximate cost of living in a particular area, which we had a lot of success with. That data was somewhat simple to gather: we used [29] to gather indexes that we translated into concrete monetary values by multiplying them with the average cost of living in New York [30] and dividing said amount by New York’s cost of living index. Since this information only required one source, it was our most successfully generated parameter.

On the other hand, we faced a lot of issues surrounding the acceptance rate of different universities, as that parameter is both much more abstract and much less documented. If an LLM used multiple sources, it’s possible that different sources use different definitions for which students are “accepted”: for example an accepted student might be one who got accepted into their most prioritised university or it could be any student that had the credentials to get into said university and showed interest in it, regardless of if they ended up attending it or not. Because of these factors, our generated data ended up having many outliers regarding this parameter.

While it’s possible to understand why certain parameters are more reliable and consistent than others, comparisons with backed-up data are necessary for all parameters. While working with different large language models, we noticed that ChatGPT and Gemini offer relatively similar results, with Gemini being marginally more accurate, giving out data with a higher R-squared value when measuring linear regression between our data and its synthetic counterpart. DeepSeek functions a bit differently from the two other models, as we couldn’t have it search for online sources. Instead, it generated values from the data it was trained from. While it featured a noticeable number of outliers, we still found it to be a decent source of information.

VI. Conclusion

As part of our research, we tested the ability of several large language models to generate reliable synthetic data, specifically within the context of a dataset analyzing different qualities of universities worldwide. As a result, it has been shown that the LLMs of today are capable of generating data that isn’t far from the truth. That has been tested by comparing our data set with the output of widely used, well-known, and freely available LLMs such as ChatGPT, Gemini, and DeepSeek.

We recognize that the capabilities of LLMs will continue to expand dynamically, along with their applications, and that they are likely to become increasingly reliable and accurate in the near future. There’s also a likely chance that certain LLMs will specialise in certain fields over time and become very reliable and useful tools in research. However, we believe that our research gave a good insight into the possibilities of large language models for gathering data within the specific time interval of our research, which spanned from December 2024 to February 2025, and the domain that our analysis focused on.

For our task, Gemini proved to be the best LLM out of the three we tested out, as it consistently outputted parameters with high R^2 values and showed a profound ease of use compared to other models, with it generating tables far easier and faster. Still, all the analysed LLMs showed similar areas of unreliability, with factors like the acceptance rate of a particular university being especially noteworthy as it’s generally poorly documented online and vague in its standards of measurement.

In the future, we are considering going back to the application we started this research project for, an application that picks a list of recommended international universities according to several user inputs, and developing and refining it further whilst following the development of different LLMs.

References

[1] V. Kumar, A. Choudhary, and E. Cho, "Data augmentation using pre-trained transformer models," 2020.

[2] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, and Y. Goldberg, "Measuring and

improving consistency in pretrained language models," Trans. Assoc. Comput. Linguistics, vol. 9, pp. 1012–1031, 2021.
[3] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," in Proc. 11th Int. Conf. Learn. Representations (ICLR 2023), Kigali, Rwanda, May 1-5, 2023. OpenReview.net.
[4] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in NeurIPS, 2022.
[5] D. Chen, C. Lee, Y.-Y. Lu, D. Rosati, and Z. Yu, "Mixture of soft prompts for controllable data generation," arXiv preprint arXiv:2303.01580, 2023.
[6] Y. Yu, Y. Zhuang, J. Zhang, Y. Meng, A. Ratner, R. Krishna, J. Shen, and C. Zhang, "Large language model as attributed training data generator: A tale of diversity and bias," arXiv preprint arXiv:2305.14718, 2023.
[7] Y. K. Chia, L. Bing, S. Poria, and L. Si, "Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction," in Findings Assoc. Comput. Linguistics: ACL 2022, 2022.
[8] B. Peng, C. Li, P. He, M. Galley, and J. Gao, "Instruction tuning with GPT-4," arXiv preprint arXiv:2304.03277, 2023.
[9] O. Benjelloun, S. Chen, and N. Noy, "Google dataset search by the numbers," in Int. Semantic Web Conf., pp. 667–682, Springer, 2020.
[10] E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," 2017.
[11] Z. Xu, F. Jiang, L. Niu, B. Y. Lin, and R. Poovendran, "Stronger models are not stronger teachers for instruction tuning," arXiv preprint arXiv:2411.07133, 2024c.
[12] Z. Li, H. Zhu, Z. Lu, and M. Yin, "Synthetic data generation with large language models for text classification: Potential and limitations," arXiv preprint arXiv:2310.07849, 2023.
[13] T. Zhang, Y. Liu, B. Li, P. Zhong, C. Zhang, H. Wang, and C. Miao, "Toward Knowledge-Enriched Conversational Recommendation Systems," in Proc. 4th Workshop NLP for Conversational AI, Dublin, Ireland, May 2022, pp. 212–217.
[14] S. Xu, W. W. Sun, and G. Cheng, "Utility Theory of Synthetic Data Generation," arXiv preprint arXiv:2305.10015, May 2023.
[15] S. Mishra, D. Khashabi, C. Baral, Y. Choi, and H. Hajishirzi, "Reframing instructional prompts to GPTk's language," in Findings of the Association for Computational Linguistics: ACL 2022, 2022.
[16] Numbeo. (2024). Cost of living rankings by country. Accessed December 2024. from https://www.numbeo.com/cost-of-living/rankings_by_country.jsp
[17] EduRank. (2024). University rankings. Accessed December 2024. from https://edurank.org/
[18] The Global Economy. (2021). Recreation and culture prices - Country rankings. Accessed December 2024. from https://www.theglobaleconomy.com/rankings/recreation_and_culture_prices_wb/
[19] World Bank. (2023). Health, nutrition, and population statistics. Accessed December 2024. from https://databank.worldbank.org/source/health-nutrition-and-population-statistics
[20] Lisachov, R. (2024). Safety index for the last 5 years. Kaggle. Accessed December 2024. from https://www.kaggle.com/datasets/romanlisachov/safety-index-for-the-last-5-years
[21] Numbeo. (2024). Cost of living prices by city: Monthly Pass (Regular Price) (Transportation). Accessed December
[22] ChatGPT 4o, 4o mini, Accessed December 2024. from https://chatgpt.com
[23] Gemini 2.0 Flash, Accessed December 2024. from https://gemini.google.com/app
[24] DeepSeek V3, Accessed January 2025. from https://chat.deepseek.com
[25] A.B. Abacha, Z. Sun, W. Yim, M. Yetisgen, Y. Fu, F. Xia and T. Lin, "Medec: A benchmark for medical error detection and correction in clinical notes," arXiv preprint arXiv:2412.19260, 2024.
[26] DeepSeek-AI et al. "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning", arXiv preprint arXiv:2501.12948, 2025
[27] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye Probability & statistics for engineers & scientists. 9th ed. Global ed. Tokyo: Pearson, 2016.
[28] M. L. Waskom, "seaborn: Statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, 2021. [Online]. Available: https://doi.org/10.21105/joss.03021.
[29] Numbeo. (2024). Cost of living rankings by country. Accessed December 2024. from https://www.numbeo.com/cost-of-living/rankings_by_country.jsp
[30] Numbeo. (2024). Cost of living in New York. Accessed December 2024. from https://www.numbeo.com/cost-of-living/in/New-York