Project Paper

Retail Car Price Prediction using Regression methods

Christobal Rupp
Paul Seiter-Neininger
Kevin Rieger
Emil Schallwig
Adrian Tanko

December 3, 2024

Submitted to

Data and Web Science Group
Dr. Sven Hertling
Franz Krause
Andreea Iana
University of Mannheim

1 Introduction

1.1 Motivation and Idea

The field of data science has increasingly focused on synthetic data generation as a method for augmenting datasets to enhance the performance of machine learning models (Neves et al. 2022). This study investigates the impact of synthetic data integration on the predictive accuracy of machine learning models, specifically in the context of forecasting retail prices for automobiles. By utilizing both authentic and synthetic datasets, the research provides a detailed evaluation of the advantages and limitations associated with synthetic data usage.

Moreover, although algorithms for predicting car prices are prevalent, their methodologies are frequently proprietary, creating barriers to broader applicability and understanding. To address this, the study aims to develop a predictive model capable of performing effectively under constrained data availability. Furthermore, this work seeks to identify and rigorously quantify the critical factors that influence vehicle retail pricing, thereby promoting greater transparency and reproducibility within this domain.

1.2 Dataset

The dataset utilized in this study comprises real-world car retail data obtained from an online platform, featuring 4,009 entries and 12 attributes. It offers a diverse representation of car brands, models, and specifications, thereby serving as a robust foundation for analyzing the relationship between a vehicle's features and its retail price. Key attributes include **brand**, **model**, and **model year**, which provide essential identification and chronological context. Additionally, features such as **mileage**, **fuel type**, **engine specifications**, and **transmission type** capture technical details that are likely to influence pricing. Attributes such as **exterior** and **interior color**, along with historical data like **accident history** and **title condition**, contribute to a comprehensive assessment of the car's overall condition.

The target variable, **price** (in USD), reflects the car's retail value but necessitates preprocessing to address inconsistencies in currency symbols and formatting to enable numerical analysis. Similarly, attributes such as **mileage** and **engine specifications** are stored as textual data and require conversion to numeric formats. Data challenges include missing values, notably in **fuel type** ($\sim$4%) and **clean title** ($\sim$15%), alongside inconsistencies in textual fields, particularly in **engine specifications**. Addressing these issues through preprocessing is critical for preparing the data for effective modeling.

Complementing this real-world dataset, the study incorporates a synthetic dataset containing **320,000 entries**, generated during a Kaggle challenge based on the original dataset. While the neural network architecture used for synthetic data generation is not specified, the synthetic dataset exhibits inconsistencies, such as mixed-up car features resulting in unrealistic vehicle offers. These anomalies raise concerns about the synthetic data's quality and its alignment with real-world scenarios. Furthermore, the synthetic data's sampling, derived entirely from the distribution of the original dataset, poses challenges for creating effective train-test splits.

Despite these limitations, the synthetic dataset's significantly larger size provides an opportunity to explore the potential benefits of data augmentation, particularly in enhancing model performance when faced with constraints due to the limited size of the original dataset. Still, the question remains whether the inclusion of synthetic data can improve predictive accuracy or if it introduces biases that undermine model reliability.

2 Preprocessing

2.1 Feature Extraction

The original dataset included diverse features, such as categorical, numerical, and textual attributes: `brand`, `model`, `model_year`, `milage`, `fuel_type`, `engine`, `transmission`, `ext_col`, `int_col`, `accident`, `clean_title`, and `price`. Consistent data formatting was applied to standardize these features. For instance, `clean_title` was converted into a binary feature, with "Yes" represented as `1` and "None" as `0`, while `price` values were transformed into numeric values by removing currency symbols. Similarly, `milage` was cleaned to retain only numeric values, and `accident` was recoded into a binary format, where reported damage was `1` and "None Reported" was `0`.

Additional predictive features were engineered from textual fields using regular expressions. For example, `cylinders`, `horsepower`, and `cubic_capacity` were extracted from `engine`, and a binary `turbo` flag was created based on the presence of the keyword "Turbo." From `transmission`, features such as `is_automatic`, `gears`, and `dual_shift` were derived. Fuel type data enabled the creation of binary indicators for `is_hybrid`, `is_diesel`, and `is_gasoline`, while electric vehicles were identified using "Electric Motor" in `engine`. For colors, `ext_col_mon` differentiated between monochromatic colors (non-standard black/white) and standard ones, while `color_match` flagged matching interior and exterior colors. Although initially thought to be significant, `color_match` showed no predictive value and was excluded from the final model.

A high correlation between `model_year` and `milage` was observed, reflecting the natural relationship between a car's age and usage. However, due to the absence of exact sale dates, these features were retained separately to avoid introducing inaccuracies.

Missing values were handled using a k-Nearest Neighbors (kNN) imputer for features with less than 25% missing data, while those with more missing values were dropped (Anil Jadhav and Ramanathan 2019). An exception was `horsepower`, which was imputed using `cubic_capacity` and `cylinders` before the latter two were removed, given their high multicollinearity with `horsepower`.

Outlier detection focused on reducing the distortion caused by extreme values. A simpler z-score method was employed to remove rows where the z-score of `price` exceeded 3. This approach preserved statistical integrity while mitigating risks associated with anomalies, such as niche luxury vehicles that might misrepresent general trends.

Car brands were grouped into three categories to reduce complexity: *luxury*, *upper class*, and *normal class*. Brands in the top 25% of prices across most bins (80-100%) were classified as *luxury* (e.g., Ferrari, Bentley), while those appearing in 50-80% of bins were labeled *upper class* (e.g., Tesla, Porsche). Remaining brands were considered *normal class*. This grouping created binary indicators (`is_luxury`, `is_upper_class`) while removing original `brand` and `model` attributes.

Feature selection refined the dataset by eliminating features with Spearman correlations below 15% with `price`, addressing multicollinearity, and retaining the most predictive variables. For example, `cubic_capacity` and `cylinders` were removed in favor of `horsepower`. The final dataset included `model_year`, `milage`, `horsepower`, `price`, `clean_title`, `accident`, `is_automatic`, `turbo`, `dual_shift`, `is_hybrid`, `is_diesel`, `is_gasoline`, `is_luxury`, and `is_upper_class`, representing a balance between relevance and simplicity.

2.2 Feature Engineering

A significant preprocessing challenge was the highly right-skewed distributions of the `mileage` and `price` features, which can hinder the performance of many machine learning algorithms. To address this, we applied log-transformations to both variables to normalize their distributions and reduce the impact of extreme values.

The original `mileage` distribution showed a sharp concentration of lower values and a long tail of very high mileage readings. This variability reflects realistic vehicle usage, with diesel cars often accumulating extremely high mileage. Since these values are plausible, we retained them as part of the dataset. After applying the log-transformation, the distribution became more symmetric and concentrated around the mean, effectively mitigating skewness. Although the distribution is not perfectly normal, the transformation improved its suitability for modeling.

A similar transformation was applied to `price`, which also exhibited right skewness, dominated by lower-priced vehicles and a smaller number of high-value entries. The log-transformation normalized this distribution, producing a bell-like shape that reduced the influence of outliers and improved the model's ability to generalize.

To assess the impact of this transformation, we created a secondary dataset with log-transformed features and compared its performance against the original. The results showed marked improvements in model stability and accuracy, particularly for algorithms sensitive to feature scaling, such as Linear Regression and Polynomial Regression.

By addressing skewness in `mileage` and `price`, we improved the alignment of these features with algorithmic assumptions, enhancing the dataset's predictive power while preserving the integrity of plausible extreme values. This process highlights the importance of feature engineering in optimizing machine learning models.

2.3 Synthetic Dataset

To improve the quality and reliability of the synthetic dataset, we rigorously identified and removed erroneous entries. The dataset, initially containing 320,000 entries, included numerous cases of poorly generated data with mismatched features. These flaws stemmed from a flawed generation process, resulting in unrealistic combinations such as engines associated with one brand appearing in vehicles of another (e.g., a Ferrari engine in a Toyota). To address this, we focused on three features: `brand`, `model`, and `engine`, systematically removing entries with mismatches. This process eliminated unrealistic configurations, reducing noise and ensuring the remaining dataset comprised coherent and plausible vehicle data. By resolving these inconsistencies, we enhanced the dataset's suitability for training meaningful patterns.

Another key step involved removing entries in the synthetic dataset that closely mirrored test data vehicles. This ensured the synthetic data did not inadvertently replicate or mimic test data, which could compromise evaluation integrity. We examined `brand`, `price`, and `mileage`, excluding synthetic entries that matched test data exactly or varied by 10% or less in `price` or `mileage`. This careful extraction maintained the independence of the synthetic dataset while retaining its utility for model training. By addressing these critical issues, we refined the synthetic dataset to better support modeling and ensure fair evaluations.

3 Explorative Data Analysis

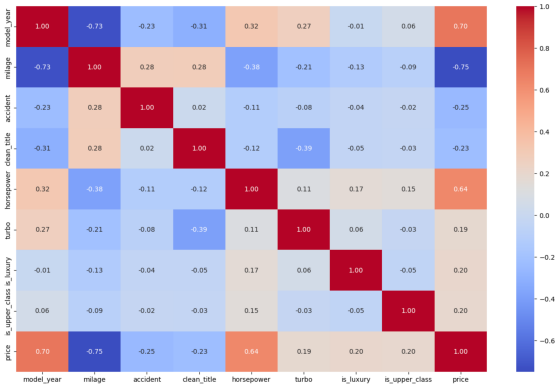
3.1 Spearman Correlation Matrix

To understand the relationships between features in our dataset, we computed a Spearman correlation matrix, which measures monotonic relationships between variables. The heatmap revealed key associations with the target variable, `price`, highlighting three features with the strongest

correlations: `model_year` (0.70), `mileage` (-0.75), and `horsepower` (0.64). These features were prioritized for further exploration due to their strong predictive value for model training.

In addition, we analyzed **depreciation per mileage per class**, integrating the impact of mileage across car classes like `is_luxury` and `is_upper_class`. While not directly shown in the correlation matrix, exploratory analysis revealed a combined correlation of 40% with `price`, underscoring the importance of class-related segmentation in price determination.

The high negative correlation (-0.75) between `mileage` and `price` aligns with the intuitive understanding that higher mileage reduces a car's value, while the strong positive correlation (0.70) between `model_year` and `price` reflects the higher value of newer cars. Notably, `mileage` and `model_year` are themselves highly correlated (-0.73), as older cars typically have higher mileage. However, due to missing sale date information, these features were retained separately to avoid inaccuracies in modeling.



The Spearman correlation matrix provided essential insights for feature selection, emphasizing the predictive importance of mileage, model year, horsepower, and class-related depreciation. These findings laid the groundwork for further analyses, detailed in the following sections.

3.2 Model Year and Price

The relationship between `model_year` and `price` reveals a clear trend: older vehicles generally command lower prices, reflecting the depreciation associated with age. Newer vehicles retain higher market values due to technological advancements, improved efficiency, and perceived reliability. However, this trend is not uniform and is accompanied by notable exceptions and insights.

One prominent outlier is a **Ford Bronco** from an earlier model year, which retains a high value despite its age. This highlights the unique behavior of rare or collectible vehicles, whose values can appreciate due to scarcity, historical importance, and collector demand. Such anomalies underscore the need to account for market segmentation when analyzing price trends, as vintage cars do not follow typical depreciation models.

Price variance also differs significantly across model years. For newer vehicles (2020–2024), variance is high, driven by a mix of luxury cars, high-performance models, and electric vehicles. In contrast, older cars, especially those from before 2000, exhibit much lower price variance, reflecting their reduced functional value and the diminishing

impact of brand and model differences. The convergence of prices for older cars suggests a more uniform valuation as vehicles age.

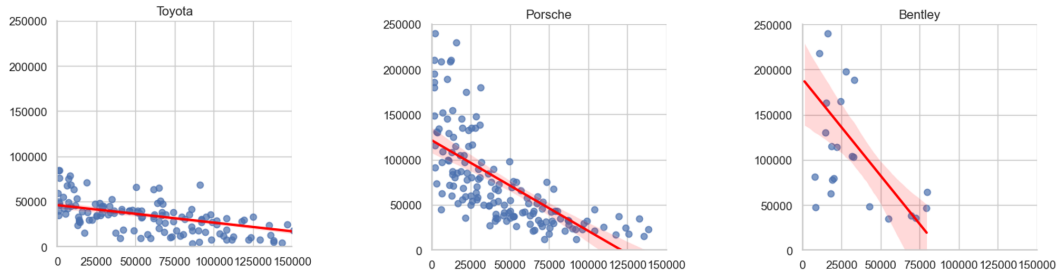
A box plot of `price` grouped by `model_year` illustrates these patterns. Newer cars show a broader interquartile range, indicating greater diversity in pricing, while older vehicles display tighter price distributions. Median prices steadily decline with older model years, affirming the expected depreciation trend. The plot also highlights outliers, such as high-value vintage cars, which stand out above the general price range for older model years.

These findings align with the Spearman correlation analysis, where `model_year` demonstrated a strong positive correlation (0.70) with `price`. This emphasizes `model_year` as a critical predictor in the dataset while underscoring the importance of recognizing exceptions, such as vintage vehicles. The combination of descriptive statistics and visualizations captures both the general depreciation trend and unique anomalies, providing a comprehensive understanding of this feature's impact.

3.3 Milage and Price

The relationship between mileage and price follows an expected trend: as mileage increases, a car's value decreases. This negative correlation, identified in the Spearman correlation matrix (-0.75), reflects the perception that higher mileage indicates greater wear and tear, reducing a vehicle's market value and desirability. However, the impact of mileage on price varies significantly across vehicle classes.

Luxury brands like Bentley and Lamborghini exhibit steep depreciation per mile compared to upper-class brands such as Porsche and standard brands like Toyota and BMW. In the luxury market, buyers prioritize low mileage as a marker of exclusivity and pristine condition. Consequently, small increases in mileage can lead to substantial drops in value, with regression lines for Bentley and Lamborghini showing steep slopes. In contrast, upper-class brands like Porsche experience more moderate depreciation rates. While these cars retain elements of prestige, buyers are generally more accepting of moderate mileage, resulting in a less pronounced decline in value compared to luxury brands.



For standard brands like Toyota and BMW, the impact of mileage on price is the least severe. These vehicles are often valued for practicality and reliability, leading to slower

and steadier depreciation rates. Their regression lines exhibit gentler slopes, highlighting their resilience to value loss with increased mileage.

At lower mileage levels, luxury cars display greater price variability due to differences in model configurations, rarity, and demand. For instance, Lamborghinis and Bentleys at similar mileage levels may have vastly different valuations based on trim levels or limited-edition status. In contrast, standard brands like Toyota and BMW exhibit tighter price distributions at low mileage, reflecting a more uniform valuation process.

These findings underscore the importance of vehicle class in determining how mileage affects price. Recognizing the varying depreciation rates across classes enables predictive models to better capture real-world pricing dynamics. This nuanced approach enhances model accuracy by incorporating the interplay of mileage, vehicle class, and market expectations.

3.4 Horsepower and Price

The relationship between horsepower and price shows a clear positive correlation: vehicles with higher horsepower tend to command higher prices. This aligns with market expectations, as greater horsepower is associated with enhanced performance, advanced engineering, and premium vehicle categories, which justify higher price points. However, this relationship is not strictly linear and exhibits notable variability.

For vehicles with horsepower between 200 and 300, prices range widely from below $20,000 to over $60,000. This spread indicates that horsepower alone does not determine price; other factors, such as brand, model, mileage, and additional features, significantly influence market value. Additionally, most vehicles fall within the 100–300 horsepower range, reflecting the dominance of standard consumer vehicles in the dataset. In contrast, cars with over 400 horsepower are rarer and typically represent luxury or high-performance models with premium price tags.

Outliers in the upper-right portion of the scatter plot—vehicles with extreme horsepower and price—highlight niche categories such as sports cars or limited-edition models. These outliers emphasize the dataset's heterogeneity and the need to account for market segmentation in predictive modeling.

Horsepower values were not always available in the dataset and were imputed using related features such as cubic capacity and cylinder count. This imputation leveraged strong correlations between these attributes to estimate missing values. However, imputed values may introduce some uncertainty, particularly for vehicles where cubic capacity or cylinder count alone cannot fully capture performance characteristics.

In summary, while higher horsepower correlates with higher prices, significant variability exists within this trend, influenced by additional factors such as market demand and vehicle class. These insights highlight the importance of incorporating horsepower into predictive models while accounting for its interactions with other features.

4 Models

To evaluate the predictive capabilities of machine learning models, we tested four algorithms—Linear Regression, Polynomial Regression, Random Forest, and XGBoost—on our dataset. Additionally, we assessed a model trained on synthetic data using XGBoost to explore the impact of synthetic data augmentation. Linear Regression served as the baseline model due to its simplicity and interpretability, providing a reference for comparing more complex approaches. This progression from simpler to advanced models enabled an incremental analysis of improvements from added complexity and feature interactions.

Our modeling process followed a standardized approach inspired by coursework methodology. For each model, we defined a parameter grid (`param_grid`) of various parameter combinations. A five-fold cross-validation was performed for each, normalizing train and test data independently. Polynomial Regression included polynomial transformations before normalization to capture non-linear patterns. After evaluating all parameter combinations, the best-performing configuration was used to train the final model on the entire training set, ensuring optimal performance.

Linear Regression achieved an MAE of 13,632.41, a MAPE of 50.05%, and an R^2 score of 0.6253, serving as a benchmark for more complex models. Its simplicity made it easy to interpret, but its inability to capture non-linear relationships limited its predictive performance. Polynomial Regression improved on these results, achieving an MAE of 11,004.47, a MAPE of 33.48%, and an R^2 score of 0.7372, demonstrating its ability to model complex relationships. However, the model required careful tuning to prevent overfitting, particularly with higher-degree polynomials.

The Random Forest Regressor further enhanced performance, with an MAE of 9,599.32, a MAPE of 29.47%, and an R^2 score of 0.7960. Its ensemble approach, averaging predictions from multiple decision trees, provided robustness against noise and outliers. Key hyperparameters, including 50 estimators (`n_estimators`), a maximum depth of 10 (`max_depth`), and a minimum of 5 samples to split a node (`min_samples_split`), balanced complexity and generalization. Random Forest excelled in handling mixed data types and offered insights into feature importance, making it a valuable addition to the analysis.

XGBoost emerged as the best-performing model on the original dataset, with an MAE of 9,391.27, a MAPE of 27.94%, and an R^2 score of 0.7985. The gradient boosting framework allowed it to iteratively refine predictions, capturing subtle data patterns overlooked by simpler models. The optimal configuration included 100 estimators (`n_estimators`), a maximum depth of 5 (`max_depth`), and a learning rate of 0.1 (`learning_rate`), among other parameters. These settings balanced precision and generalization while avoiding overfitting. XGBoost's ability to handle missing data and complex feature interactions contributed to its superior accuracy, making it the most effective model in our analysis.

To test synthetic data augmentation, we trained XGBoost on a dataset with 100,000 synthetic entries. While XGBoost was chosen for consistency, its performance degraded

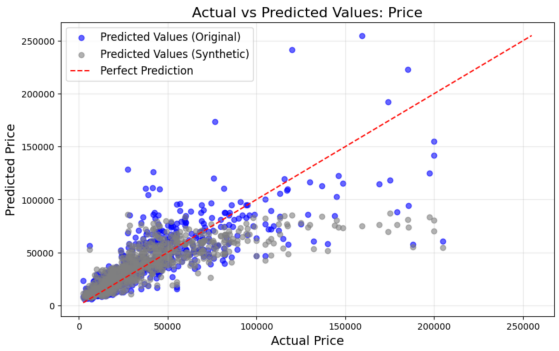
significantly, with an MAE of 9,777.54, a MAPE of 41.53%, and an R^2 score of 0.6235. This decline highlights challenges with synthetic data, such as inconsistencies and noise introduced during generation, which obscured meaningful patterns and hindered generalization.

This analysis underscores the progressive improvements achieved by increasingly sophisticated models, starting with Linear Regression and culminating in XGBoost. Parameter tuning and cross-validation maximized each model's performance. The inclusion of synthetic data revealed the importance of data quality over quantity, emphasizing that effective preprocessing and careful model selection are essential for achieving optimal predictive accuracy.

5 Results and Evaluation

The evaluation began with a baseline comparison using Linear Regression, which achieved an MAE of 13,632.41, a MAPE of 50.05%, and an R^2 score of 0.6253. Although modest in performance, it provided a valuable benchmark for assessing the improvements offered by more complex models.

XGBoost emerged as the best-performing model on the original dataset, achieving an MAE of 9,391.27, a MAPE of 27.94%, and an R^2 score of 0.7985. Its gradient boosting framework captured subtle patterns and feature interactions, outperforming Random Forest (MAE: 9,599.32, R^2: 0.7960) and highlighting the advantages of XGBoost's iterative optimization and regularization techniques.



Models trained on the original dataset consistently outperformed those using synthetic data. For example, XGBoost trained on synthetic data achieved an MAE of 9,777.54, a MAPE of 41.53%, and an R^2 score of 0.6235, significantly lower than its performance on the original dataset. Despite the synthetic dataset's larger size (100,000 entries vs. 3,200), poor quality, including mixed-up car configurations, outliers, and unrealistic combinations, likely contributed to this disparity. The differing distributions between synthetic and original data further amplified the performance gap, especially since evaluation was conducted on a subset of the original dataset.

Finally, our XGBoost model trained on the original dataset achieved a Root Mean Squared Error (RMSE) of 16,457, far surpassing the competition's median RMSE of 71,900 on Kaggle. This significant performance gap prompted thorough checks for

potential information leakage, but no evidence of such issues was found. The results reflect robust preprocessing and effective modeling strategies. Further exploration of this disparity is provided in the critical reflection section.

6 Critical Reflection

This study offers valuable insights into the potential impact of synthetic data on predictive modeling for car retail prices but is subject to several limitations affecting the validity and generalizability of its findings.

The dataset lacks critical features, such as interior specifications, reasons for sale, ownership history, or overall condition, which are essential for capturing nuanced determinants of pricing. Without these variables, the model's predictive power is inherently constrained. Additionally, inconsistent model name representations introduced significant noise. For instance, variations in labeling a "VW Golf" hindered the model's ability to extract meaningful distinctions. Given the dataset's small size, this issue further limited learning, forcing the omission of the model name feature, despite its relevance in determining price.

The ambiguity surrounding price determination further complicates the analysis. It is unclear whether prices were set by owners, dealerships, or other entities. We assumed owner-set prices, which could explain anomalies like a Bentley priced below $50,000 with fewer than 8,000 miles, likely reflecting a quick-sale priority. However, without clarity on price-setting methods, this assumption introduces uncertainty, potentially impacting the model's generalizability.

While the dataset includes the **model year**, the absence of **sale dates** prevents accurate calculation of vehicle age at the time of sale. This missing data forced us to retain **model year** and **mileage** as separate features, despite their high correlation, avoiding speculative inferences but potentially missing more informative insights.

The dataset's origin from a Kaggle competition allowed for performance benchmarking. Our model achieved an RMSE of 16,457, significantly outperforming the competition's median of approximately 71,900. While this highlights the robustness of our approach, the discrepancy invites scrutiny. Despite thorough checks, no evidence of information leakage was found. Our decision to exclude sales above $1 million likely reduced noise, potentially explaining the performance gap if other groups included these outliers. Alternatively, the discrepancy could result from random factors, such as favorable imputations or test set splits.

These findings emphasize the importance of high-quality, well-documented data. Improved datasets with richer attributes, consistent formatting, and clear price-setting mechanisms would enable more robust modeling and enhance the applicability of machine learning techniques in car price prediction.

Bibliography

Anil Jadhav, D. P. and K. Ramanathan (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence 33*(10), 913–933.

Neves, D. T., J. Alves, M. G. Naik, A. J. Proença, and F. Prasser (2022). From missing data imputation to data generation. *Journal of Computational Science 61*, 101640.

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Bachelor-, Master-, Seminar-, oder Projektarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und in der untenstehenden Tabelle angegebenen Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Declaration of Used AI Tools

Tool	Purpose	Where?	Useful?
ChatGPT	Rephrasing	Throughout	+
DeepL	Translation	Throughout	+
ResearchGPT	Summarization of related work	Sec. **??**	-
Dall-E	Image generation	Figs. 2, 3	++
GPT-4	Code generation	functions.py	+
ChatGPT	Related work hallucination	Most of bibliography	++

Unterschrift
Mannheim, den XX. XXXX 2024