Unveiling Statistical Relationships Among Popular LLM Benchmarks: A Quantitative Framework

Haonan Wang[1] **Ziang Xiao**[1]
Johns Hopkins University[1]
hwang298@jh.edu,ziang.xiao@jhu.edu

Abstract

The rapid advancement of Large Language Models (LLMs) has led to a proliferation of evaluation benchmarks, raising concerns about redundancy and overlapping assessment dimensions. In this study, we propose a quantitative framework to analyze the relationships among 16 popular benchmarks using 12 widely recognized LLMs.Correlation analysis revealed clusters of related benchmarks, including strong correlations in mathematical reasoning (e.g., GSM8K, MATH, DROP), moderate correlations in general knowledge (e.g., MMLU, TriviaQA), and weak correlations in code generation (e.g., HumanEval, MBPP EvalPlus) and reasoning tasks (e.g., ARC Challenge, HellaSwag). Next, factor analysis identified three latent dimensions in strong correlations benchmarks, mathematical and reasoning, and general knowledge and code explaining 95.13% of the variance. Our results underscore that simply adding more benchmarks does not improve evaluation. Instead, meta-benchmarks, which integrate diverse but complementary dimensions, are critical for holistic and efficient LLM evaluation. This study provides insights for designing scalable and interpretable evaluation systems.

1 Introduction

In recent years, the development of Large Language models(LLMs) has advanced the field of Natural Language Processing(NLP)tasks. LLMs learn mainly through the pre-training task of predicting the next word to have a strong ability to solve various language tasks and diverse domain tasks in all fields. Accompanying the rapid development of Large Language Models (LLMs) and their diverse real-world downstream applications, accurately evaluating these models has become crucial, and an upsurge of efforts in evaluating LLMs introduces a series of benchmarks(Bender et al., 2021). Benchmarks play a critical role in assessing the performance of LLMs, providing a standardized way to measure the diverse capabilities of these models(Hendrycks and et al., 2020). Evaluation in LLMs from some technology report(Achiam et al., 2023; Dubey et al., 2024) has shifted from using datasets tailored to single tasks to more comprehensive benchmarks such as MMLU(Hendrycks and et al., 2020), BIG-Bench(Srivastava et al., 2022), HELM(Liang et al., 2022), AGIeval(Zhong et al., 2023) and et. c. These benchmarks aims to create a comprehensive evaluation system for the overall evaluation of the performance and capability of LLMs from different angles and levels.

However, as these benchmarks grow in complexity, their effectiveness and results in practical applications do not always reflect the capability to meet human needs in real-world scenarios. The issue has been identified as "*social-technical gap*" in (Liao and Xiao, 2023). Although existing benchmarks attempt to evaluate models by covering a wide range of tasks and datasets from all fields, we still need to explore the "*social-technical gap*" reflected between the capacity of LLMs measured by current benchmarks and real-world downstream applications. Specifically, most LLMs evaluate primarily on upstream benchmarks, but there is also a significant gap between these benchmarks and actual downstream applications. (Ethayarajh and Jurafsky, 2021) noted that performance on standardized benchmarks does not always correlate with success in practical applications, indicating a discrepancy between benchmark scores and actual utility.Some existing methods, like GLUE(Wang et al., 2018) and SuperGLUE(Wang et al., 2019), attempt to capture the broad capabilities of models through multi-task evaluation and cross-domain testing. However, these methods still focus mainly on the performance of standardized tests and do not fully consider the effectiveness and ability of the models to meet user needs in complex real-world applications. Although (Liang et al., 2022) introduced the HELM framework to provide a more holistic evaluation, it still lacks a systematic quantification of the power of upstream benchmarks for practical utility.

Our research focuses on understanding the statistical relationships among mainstream benchmarks through a quantitative framework. The key contributions of this work are as follows:

- We developed a framework that combines correlation analysis and factor analysis to quantify the relationships among popular benchmarks (e.g. MMLU, GSM8K) in 12 popular large language models (LLMs) to explore the interdependencies and independence of benchmarks.

- We also provide a publicly available dataset containing the scores of 12 popular LLMs on 16 popular benchmarks for researchers to further analyze benchmark relationships, optimize assessment methods, and develop new LLM capability assessment tools.

2 Related work

Evaluation and Benchmarks of LLMs. Evaluation of LLM has become a central focus in current research. In order to comprehensively assess the performance of LLMs, a widely adopted approach involves selecting various capability dimensions and designing corresponding evaluation tasks. These dimensions serve as criteria for testing and comparing the performance of the model in different aspects(Zhao et al., 2024). Based on the evaluation methodologies, the assessment of these capability dimensions can be categorized into three primary approaches: benchmark-based(Hendrycks and et al., 2020), human-based evaluation(Zheng et al., 2023), and model-based(Li et al., 2023). Benchmarking has become a standard way for evaluating the performance of LLMs, the advantages of benchmarking are its high degree of automation and reusability and can reduce the need for manual intervention.GLUE (Wang et al., 2018)was one of the first benchmarks designed to test a model's general language understanding abilities across a diverse set of tasks. As models like GPT-3 began to surpass human-level performance on GLUE, SuperGLUE(Wang et al., 2019)was introduced, offering more challenging tasks that require deeper language understanding. Knowledge-faced evaluation benchmarks like MMLU(Hendrycks and et al., 2020), and C-eval(Huang et al., 2024)focus on evaluating the understanding and application. Reasoning-faced evaluation benchmarks like GSM8K(Gao et al., 2023), and BBH(Suzgun et al., 2022) focus more on LLM's performance in solving reasoning problems. In addition, some comprehensive evaluation benchmarks like OpenCompess(Contributors, 2023) also attempt to combine these two types of evaluation tasks to fully evaluate the comprehensive capabilities.

Evaluation and Meta-benchmark of LLMs

For more and more traditional benchmarks design in the LLMs evaluation field, based on the traditional benchmarks the model evaluation is always limited to performance in a single dimension. However, as the capabilities of large language models (LLMs) continue to grow, researchers have increasingly recognized that single-dimensional evaluations cannot fully capture the practical utility of these models. There are so many researchers creating more complex and multi-dimensional model evaluation systems named Meta-benchmark design. These meta-benchmark design aims to solve the limitations of traditional benchmark design and provide more practical application requirements. These meta-benchmark design evaluations are not only limited to evaluating the performance score of the model but also comprehensively consider the efficiency, robustness, resource, and other dimensions of the model in real-world applications. For example, based on traditional benchmarks in real criteria can not fully capture the complexity of real queries, and the LLM as a benchmark for reviewers has the problem of scoring bias and the limited number of queries. The Mixeval(Ni et al., 2024)proposes a new paradigm for building efficient LLM assessments by strategically mixing existing benchmarks for evaluating LLMs. MixEval combines real user queries with efficient, reality-based benchmarks to provide a fast, cost-effective, and repeatable evaluation method, with the capability for dynamic assessment. To evaluate the validity of the benchmarks themselves, the BenchBench(Perlitz et al., 2024) authors team introduces Benchmark Agreement Testing (BAT) as a method for validating new benchmarks against established ones but highlights the lack of standardized procedures for such testing. By analyzing over 40 prominent benchmarks, the team demonstrates that overlooked methodological choices can significantly impact BAT results, potentially undermining the validity of conclusions drawn from them. Finally, the authors introduce the BenchBench, a Python package for conducting BAT, and release the BenchBench leaderboard, a meta-benchmark designed to evaluate benchmarks against their peers.

3 Method

3.1 Correlation analysis

Correlation analysis is a widely-used statistical method to measure the linear relationship between two variables. Our primary objective is to determine whether significant statistical associations. To achieve this, we utilized correlation matrix methods, specifically Pearson correlation coefficients(PCC) and Spearman rank correlation coefficients(SRCC), to present the relationships between different benchmarks comprehensively(Taylor, 1990).

3.1.1 Pearson Correlation Coefficient

The Pearson correlation coefficient r quantifies the linear relationship between two variables X and Y. Its value ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear relationship.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Steps to Compute Pearson Correlation:

1. Calculate the means of X and Y:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

2. Compute deviations:

$$(X_i - \bar{X}), \quad (Y_i - \bar{Y})$$

3. Compute the numerator (covariance):

$$\text{Cov}(X, Y) = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

4. Compute the denominator (product of standard deviations):

$$\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

5. Compute r:

$$r = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

3.2 Factor analysis

Factor Analysis (FA) is a widely used statistical method for identifying latent variables or factors underlying a set of observed variables, that reduces the dimensionality of observed variables by representing them with a smaller number of common factors. It is then used to analyze the relationships between the variables. The underlying principle is to decompose the original variables into two components:

- 1.A component explained by common factors, which accounts for most of the information in the original variables.

- 2.A unique component, independent of the common factors, reflects the deviation of the original variables from the common factor linear combination.

Steps in Factor Analysis

1. **Variable Selection:**
   - Variables with **strong correlations** (e.g., Pearson r > 0.7) are selected to ensure shared variance among the chosen variables.
   - Variables with **weak correlations** (e.g., r < 0.3) are excluded, as they may represent independent constructs unsuitable for shared factor models.
   - All variables are **standardized** (e.g., z-scores) to eliminate scale effects, ensuring each variable contributes equally to the analysis.

2. **Suitability Tests:**
   - **1. Kaiser-Meyer-Olkin (KMO) Test** The KMO test evaluates whether the partial correlations among variables are small. A higher KMO value (> 0.7) indicates suitability for Factor Analysis (FA). The formula for the KMO test is:

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^2}$$

where r_{ij} is the correlation coefficient, and a_{ij} is the partial correlation coefficient.

- **2. Bartlett's Test of Sphericity** This test evaluates whether the correlation matrix significantly differs from an identity matrix (where all off-diagonal values are 0). A significant $p < 0.05$ confirms that variables are sufficiently correlated for FA. The formula for Bartlett's test is:

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right) \ln |R|$$

where R is the correlation matrix, n is the sample size, and p is the number of variables.

3. It can be expressed as:

$$x = Af + \varepsilon$$

or in expanded form:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

Here:

- $f = [f_1, f_2, \ldots, f_m]^T$ represents the common factors extracted, explaining the shared variance among p variables.
- $A = (a_{ik})$ is the factor loading matrix, with a_{ik} representing the correlation between variable x_i and factor f_k.

4. **Standardization**: Standardize variables to ensure each has a mean of 0 and variance of 1:

$$x_{ij} = \frac{x_{ij} - \frac{1}{n}\sum_{j=1}^{n} x_{ij}}{\sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \frac{1}{n}\sum_{j=1}^{n} x_{ij})^2}}$$

5. **Covariance Matrix**: Compute the covariance matrix S with elements:

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n} x_{ik}x_{jk}$$

6. **Eigenvalue Decomposition**: Perform eigenvalue decomposition on S and retain the first m eigenvectors:

$$\hat{A} = [\sqrt{\lambda_1}v_1, \sqrt{\lambda_2}v_2, \ldots, \sqrt{\lambda_m}v_m]$$

7. **Determine Number of Factors**:

$$m = \arg\min \left(\sum_{i=1}^{m} \lambda_i \middle/ \sum_{i=1}^{p} \lambda_i \geq r \right)$$

8. **Estimate Factor Scores**:

$$f_j = \hat{A}^T S^{-1} x_j$$

Through these steps, the factor loading matrix $\hat{A}$ and common factors f are obtained, representing the shared structure underlying the observed variables.

$$\hat{\varepsilon}_j = x_j - \hat{A}f_j$$

4 Results and Analysis

4.1 Dataset and Benchmark Scores

In our study, we present the dataset and benchmark scores collected from multiple authoritative sources Table 1, including the **Mix-Eval paper**(Ni et al., 2024), **Open LLM Leaderboard**(HuggingFace, 2024), **OpenCompass Leaderboard**(OpenCompass, 2024), and **detailed technical evaluation reports** for various large language models (LLMs). The benchmarks and scores are meticulously curated and aligned with our evaluation goals, focusing on five key capabilities: **General**, **Knowledge**, **Code**, **Math**, and **Reasoning**. Our data collection process adheres to the following principles:

1. **Uniform Scoring Standards:** To ensure consistency, all scores are standardized based on either the **0-shot** or **few-shot** settings, depending on the common evaluation practice for each benchmark.

2. **Authoritativeness of Sources:** Benchmark scores are sourced from either published technical reports or widely recognized evaluation leaderboards. This ensures fairness and reliability by excluding any self-reported or unpublished scores from model developers.

3. **Selection of Popular Models and Benchmarks:** We prioritize LLMs and benchmarks that are widely recognized and extensively cited by the community. This selection ensures the inclusion of prominent models (e.g., GPT-4, Claude) and benchmarks (e.g., MMLU, GSM8K).

4. **Categorization by Key Capabilities:** Each benchmark is mapped to one of the five core capabilities (General, Knowledge, Code, Math, and Reasoning), as highlighted in technical reports, to facilitate targeted analysis.

5. **Platform Consistency:** For each benchmark, scores are sourced from a single platform to eliminate variations caused by different evaluation settings or configurations.

The resulting dataset Table 1 provides a comprehensive and well-structured collection of benchmark scores, which enables rigorous and reproducible analysis of the statistical relationships between benchmarks and their corresponding model capabilities.

4.2 Study1:Correlation Matrix Analysis

In the first experiment, we aim to investigate two main objectives:

- **Uncover Linear Correlations Between Benchmarks**: We explore to analyze the relationships between different benchmarks, can better determine whether benchmarks evaluate similar abilities, and assess their redundancy and the overlap in the capabilities they measure.

- **Identify Task Independence**: We explore to identify benchmarks that assess distinct capabilities, which can provide a more comprehensive evaluation of model abilities across various dimensions.

To achieve this, we will use the evaluation results of several mainstream benchmark tasks, categorized as follows:

- **General**:MMLU,MixEval,MMLU-hard,Mixeval-hard,ComonsenseQA

- **Knowledge**: TriviaQA, TrivialQA-Hard,

- **Math**: GSM8K, DROP,DROP-hard,MATH

- **Code**: MBPP EvalPlus, HumanEval

- **Reasoning**:ARCChallenge,HellaSwag,GPQA

4.2.1 Pearson Correlation Coefficient

The coefficient will be used to measure the linear relationship between benchmarks, which ranges from [-1, 1], where a value close to 1 indicates a strong positive correlation, close to 0 indicates no correlation, and close to -1 indicates a negative correlation. Based on the Pearson correlation results(Figure 1,Table 2) ,we can see:

- **High-Correlation Variables** exhibit strong pairwise relationships (e.g., $r > 0.7$), indicating that they assess similar or overlapping aspects of the observed data. These variables can be grouped into potential dimensions based on their shared characteristics. Below, we categorize and analyze these variables into distinct groups:

  - Mathematics Factor:**GSM8K, DROP, DROP-Hard, MATH:** GSM8K is highly correlated with DROP ($r = 0.944$), DROP-Hard ($r = 0.928$), and MATH ($r = 0.902$).These strong correlations suggest that these variables collectively measure mathematical reasoning and problem-solving abilities. However, due to the redundancy between DROP and DROP-Hard, retaining only one (e.g., DROP) is recommended to reduce overlap and ensure analytical efficiency.
  - General Knowledge Factor:**MMLU, MixEval, CommonsenseQA, TriviaQA, TriviaQA-Hard:**MMLU shows a moderate correlation with MixEval ($r = 0.76$) and strong correlations with CommonsenseQA ($r = 0.919$) and TriviaQA ($r = 0.937$).TriviaQA and TriviaQA-Hard have an exceptionally high correlation ($r = 0.9$), suggesting significant overlap. It is therefore suggested to retain TriviaQA as a representative variable for general knowledge evaluation tasks.
  - Code Generation Factor: **HumanEval, MBPP EvalPlus:**The correlation between HumanEval and MBPP EvalPlus is strong ($r = 0.885$), indicating that both benchmarks evaluate programming and code generation abilities. This high correlation reflects their shared assessment focus, suggesting that either benchmark could serve as a representative for evaluating this task dimension.

- **Low-Correlation Variables** exhibit weak pairwise relationships with most other variables ($r < 0.4$). This indicates that these variables assess distinct and independent aspects of the dataset. Despite not contributing to shared dimensions, they are valuable for identifying unique capabilities. Below is a detailed analysis:

- **ARC Challenge (Reasoning):** The ARC Challenge has a maximum correlation of $r = 0.573$ with other variables, which is relatively low compared to other benchmarks. This suggests that it evaluates unique reasoning skills that are not captured by other benchmarks, making it an independent and meaningful measure of reasoning ability.
- **HellaSwag (Reasoning):** exhibits a maximum correlation of $r = 0.445$, reflecting its focus on distinct reasoning tasks. Its weak correlations indicate that it evaluates capabilities not heavily overlapping with other reasoning benchmarks.
- **GPQA (Reasoning):** demonstrates extremely low correlations ($r < 0.3$) across all variables.This highlights its role in assessing a highly independent and specialized reasoning capability that does not align with other shared dimensions.
- Low-correlation variables are essential for capturing independent abilities, as they represent unique evaluation dimensions that shared factors might overlook. While these variables may not contribute significantly to shared dimensions, retaining them ensures that the analysis encompasses the full spectrum of distinct capabilities.

- **Variables with High Redundancy** Certain pairs of variables demonstrate excessively high correlations ($r > 0.85$), leading to redundancy in the data. Retaining all of these variables could result in unnecessary duplication of information, so careful selection is necessary to maintain efficiency while preserving diversity. Below is an analysis of these redundant variables:

  - **DROP and DROP-Hard:**The correlation between DROP and DROP-Hard is exceptionally high ($r = 0.928$).Both variables assess mathematical reasoning abilities; however, their redundancy suggests that retaining only one (e.g., DROP) is sufficient for analysis without losing critical information.
  - **TriviaQA and TriviaQA-Hard** exhibit a very high correlation ($r = 0.9$). Since they both measure general knowledge and reading comprehension, it is advisable to retain TriviaQA as a representative variable for this category.
  - **MATH and GSM8K** have a high correlation ($r = 0.902$), indicating their shared focus on mathematical reasoning.
  - Despite their similarity, both benchmarks can be retained as they reflect different aspects of mathematical evaluation tasks, ensuring a more comprehensive representation of this domain. High-redundancy variables are a potential source of inefficiency in analysis. While some pairs can be reduced to a single representative variable, others, like MATH and GSM8K, should be retained to reflect task diversity and provide a broader evaluation scope.

4.2.2 Conclusion

The correlation analysis revealed significant relationships among several variables, highlighting clusters of tasks with shared dimensions. High-correlation variables were grouped into three categories: math reasoning, general, knowledge, and code, indicating overlapping evaluation focuses. Conversely, low-correlation variables such as ARC Challenge, HellaSwag, and GQA demonstrated independence, capturing unique task dimensions. Some highly redundant variables like DROP and DROP-Hard or TriviaQA and TriviaQA-Hard suggested potential simplification by retaining only representative variables. These findings established the foundation for further factor analysis by identifying shared and independent evaluation dimensions in the benchmarks.

4.3 Study2:Factor analysis

Based on the results of the correlation matrix Table 2, several variables demonstrate strong correlations ($r > 0.7$), suggesting the presence of shared dimensions among them. These high-correlation variables include **GSM8K, DROP, DROP-Hard, MATH**, **MMLU, MixEval, CommonsenseQA,**

TriviaQA, and **HumanEval, MBPP EvalPlus**, categorized into three potential groups: **mathematical reasoning**, **general knowledge**, and **code generation**.

Variables not included in the factor analysis (e.g., GQA, HellaSwag, ARC Challenge) showed low correlations with other variables, suggesting that they evaluate unique and independent dimensions of ability. As such, while these variables were deemed unsuitable for factor analysis, their distinct contributions merit separate discussion to highlight the specific capabilities they assess.

This finding provides a foundation for applying factor analysis to explore the latent dimensions underlying these variables. By reducing the dimensionality of the data, factor analysis enables us to group these variables into a smaller number of meaningful factors while retaining the core information shared among them. In the following analysis, we conducted a factor analysis on the identified high-correlation variables to determine the latent factors and interpret their contributions.

4.3.1 KMO Test and Bartlett's Test

Table 3 above presents the results of the Kaiser-Meyer-Olkin (KMO) test and Bartlett's Test of Sphericity, which evaluate the suitability of the dataset for factor analysis.

- The KMO value is 0.712, which is above the commonly accepted threshold of 0.6. This indicates that the variables have sufficient correlation to justify factor analysis.
- Bartlett's Test shows a chi-square value of 122.731 with 28 degrees of freedom and a significance level $P = 0.000001$.
- The P-value is highly significant ($P < 0.05$), rejecting the null hypothesis that the correlation matrix is an identity matrix.
- This result confirms that there is sufficient correlation among variables for factor analysis.

The results indicate that factor analysis is appropriate for this dataset. The KMO value suggests a moderate level of sampling adequacy, and Bartlett's Test supports the presence of correlations among variables.

4.3.2 Variance Explained Analysis

Table 4, Table 8 and Figure 2 above summarize the explained variance of the components before and after rotation, which is used to evaluate the contribution of each factor to the variance of the variables.

Explained Variance (Pre-Rotation):

- The first component explains 85.672% of the variance, which is significantly high, indicating its dominant role in capturing the data structure.
- The second component adds an additional 5.206%, bringing the cumulative explained variance to 90.878%.
- Subsequent components contribute minimally to the variance, with their eigenvalues falling below 1.0, suggesting they are less important for explaining the overall variance.

Explained Variance (Post-Rotation):

- After rotation, the explained variance is redistributed among the components to achieve a more balanced structure.
- The first three components collectively explain 95.129% of the variance, with contributions of 36.368%, 34.223%, and 24.538%, respectively.
- This suggests that three factors are sufficient to represent the data effectively while maintaining interpretability.

The cumulative explained variance suggests that three components adequately explain 95.129% of the dataset's variance, which is considered satisfactory. The redistribution of variance after rotation ensures that each factor captures unique aspects of the data.

4.3.3 Rotated Factor Loadings Analysis

The table 5 above displays the rotated factor loadings and communalities for each variable. Rotated factor loadings indicate the strength of the relationship between variables and the extracted factors, while communalities represent the proportion of variance in each variable explained by the factors.

Factor Loadings:

- Variables exhibit significant loadings on distinct factors, supporting the multidimensional structure of the data:

- **Factor 1:** Strongly associated with **HumanEval (0.742)** and **MBPP EvalPlus (0.833)**, indicating this factor captures programming and code generation abilities.
- **Factor 2:** Strongly associated with **CommonsenseQA (0.797)**, **GSM8K (0.792)**, and **DROP (0.567)**, highlighting a focus on reasoning and mathematical abilities.
- **Factor 3:** Dominantly associated with **MixEval (0.863)** and **TriviaQA (0.504)**, reflecting general knowledge and reading comprehension tasks.

Communalities:

- Communality values range from 0.913 to 0.978, indicating that the three extracted factors collectively explain a high proportion of variance for each variable.
- For instance, **MMLU (0.978)** and **MixEval (0.965)** are well-represented by the factors, validating the selection of these benchmarks for the analysis.

The rotated factor loadings confirm that three latent factors sufficiently explain the dataset's structure. Factor 1 captures programming tasks, Factor 2 focuses on reasoning and math, and Factor 3 highlights general knowledge. The high communalities suggest that the model effectively represents the variables.

4.3.4 Component Matrix Analysis

The table 6 above provides the component matrix, which displays the raw loadings of each variable on the three extracted components. These raw loadings represent the correlation between the variables and the components, helping to identify how each variable contributes to the factors.

Component 1:

- Strongly influenced by **MBPP EvalPlus (0.784)** and **HumanEval (0.506)**.
- This component reflects **code generation**, given the high contributions from benchmarks focused on this task.

Component 2:

- Strongly influenced by **CommonsenseQA (0.791)** and **GSM8K (0.732)**.
- This component appears to represent **mathematical and reasoning**, as benchmarks focused on reasoning tasks dominate the loadings.

Component 3:

- Dominated by **MixEval (1.165)**, followed by smaller contributions from **TriviaQA (0.203)**.
- This component highlights **general knowledge**, as reflected in the high loading of MixEval.

Conclusion: The analysis confirms that the extracted components meaningfully represent distinct dimensions of task performance. Benchmarks with high loadings on their respective components are critical for understanding the capabilities of the models and the component matrix reveals that each extracted factor aligns well with the task categories: programming, reasoning/mathematics, and general knowledge. Variables with negative loadings (e.g., **CommonsenseQA** on Component 1, -0.548) indicate minimal or opposite contributions to those specific components.

4.3.5 Factor Weight Analysis

Table 7 above presents the factor weights derived from the rotated variance explained ratios. These weights represent the relative contribution of each factor to the overall variance in the dataset. The weights are calculated using the formula:

$$\text{Weight } (\%) = \frac{\text{Rotated Variance Explained } (\%)}{\text{Cumulative Rotated Variance Explained } (\%)} \times 100$$

Factor 1:

- Explains 36.368% of the total variance, contributing a relative weight of 38.23%.
- This factor carries the largest proportion of information in the dataset, indicating its dominant role in explaining variability.

Factor 2:

- Explains 34.223% of the total variance, contributing a relative weight of 35.975%.
- This factor slightly trails behind Factor 1 in importance, providing additional explanatory power to the data structure.

Factor 3:

- Explains 24.538% of the total variance, contributing a relative weight of 25.794%.
- This factor provides complementary information, completing the representation of the dataset's variance.

The three factors together account for 95.129% of the total variance, confirming their ability to effectively summarize the dataset. Factor 1 has the highest weight, making it the most influential, while Factors 2 and 3 provide substantial but slightly lesser contributions. This distribution of weights supports the use of three factors for data analysis and interpretation.

4.3.6 Conclusion

The factor analysis identified three distinct latent factors: **code generation** (Factor 1), **mathematical reasoning** (Factor 2), and **general knowledge** (Factor 3). These factors collectively explain 95.129% of the dataset's variance, with Factor 1 contributing the most (38.23%). The adequacy of the model is supported by a moderate KMO value (0.712) and a highly significant Bartlett's Test, confirming the dataset's suitability for factor analysis. High communalities (0.913–0.978) indicate that the factors effectively represent the variables, ensuring a comprehensive and efficient summary of the dataset's structure.

5 Conclusion

In this study, we developed a quantitative framework to explore the statistical relationships between popular benchmarks for evaluating large language models (LLMs). Using correlation analysis, the factor analysis to analyze the interdependence and uniqueness of 16 benchmarks across different LLMs. Correlation analysis revealed clusters of strongly related tasks (e.g., mathematical reasoning, general knowledge, code generation) and independent benchmarks (e.g., ARC Challenge, HellaSwag, GQA), highlighting both overlapping and unique evaluation dimensions. Factor analysis identified three latent factors—code generation, mathematical reasoning, and general knowledge—explaining 95.13% of the variance, with strong model adequacy supported by a moderate KMO (0.712) and significant Bartlett's Test. These findings emphasize that increasing benchmarks alone does not improve evaluation. Instead, we advocate for meta-benchmarks that integrate diverse but complementary dimensions to create efficient, holistic, and interpretable evaluation frameworks for LLMs.

Acknowledgments

We sincerely thank Prof. Ziang Xiao for his invaluable guidance in shaping the research direction, refining the core ideas, identifying relevant literature, and developing the quantitative framework.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. *GitHub repository*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kawin Ethayarajh and Dan Jurafsky. 2021. Utility is in the eye of the user: A critique of nlp leaderboards. *Preprint*, arXiv:2009.13888.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Dan Hendrycks and et al. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

HuggingFace. 2024. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`. Accessed: November 29, 2024.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*.

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. 2024. Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. *Preprint*, arXiv:2406.06565.

OpenCompass. 2024. Opencompass: A universal evaluation platform for foundation models. `https://github.com/OpenCompassAI/OpenCompass`. GitHub repository.

Yotam Perlitz, Ariel Gera, Ofir Arviv, Asaf Yehudai, Elron Bandel, Eyal Shnarch, Michal Shmueli-Scheuer, and Leshem Choshen. 2024. Do these llm benchmarks agree? fixing benchmark evaluation with benchbench. *Preprint*, arXiv:2407.13696.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Richard Taylor. 1990. Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1):35–39.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

A Example Appendix

A.1 Dataset Benchmark

Table 1: Benchmark Dataset Scores for Different LLMs

LLMs	Mixeval	Mixeval-Hard	MMLU	MMLU-Hard	ComonsenseQA	TriviaQA	TriviaQA-Hard	HumanEval	MBPP EvalPlus	GSM8K	DROP	MATH	ARC Challenge	HellaSwag	GPQA
GPT-3.5-Turbo	79.7	43	74.5	35.1	81.6	85.2	46.4	48.1	82	57.1	84.8	54	93.7	93	30.8
GPT-4-Turbo	88.8	62.6	82.8	45.5	85.4	91.2	73.1	67	83.6	94.2	91	41	64.5	93.7	41.4
GPT-4.0	64.7	57.1	89	57.1	86.8	88	70.3	90.2	87.8	96.1	87.9	67.5	76.9	93.4	51.8
Llama-3-8B	75	45.6	71.9	40.7	78.3	71.7	40.2	72.6	72.8	72.6	86.4	67.6	51.9	81.8	32.2
Llama-3-70B	84	55.9	82.5	46.3	83.1	83.1	60.5	80.5	86	95.1	90.1	74.5	88	94.8	44.9
Owen-1.5-7B	71.4	35.5	68.7	29	82.1	64.1	29	37.2	35.8	75.3	76.4	51.2	52.6	81.2	28.9
Gemini 1.0 Pro	84.8	46.5	79.2	35.5	80.2	58.2	51.2	67.7	61.4	77.9	82.6	64.8	77.3	74.2	45.2
Gemini 1.5 Pro	84.8	58.3	78.2	54.5	84.4	85.3	67.8	84.1	74.6	90.8	82.5	87.3	79.5	91.2	26.2
Mistral-7B	51.9	28.4	51.5	33.5	66	73.7	33.5	42	49.5	53.2	72.8	27.3	11.8	71.4	28.9
Gemma1-1.2B	69.6	38.4	72.8	39	73.6	64.3	30.3	32.3	44.4	46.4	80.6	55.1	24.3	91.2	30.3
Gemma1-7B	69.9	51.6	81.9	40.3	73.6	73.3	39.3	92	90.5	96.4	93.7	80.4	71.1	95.7	56.4
Claude-3.5 Sonnet-0620	89.9	68.1	84.2	40.7	85.4	92.6	73.3	92	90.5	96.4	93.7	80.4	71.1	95.7	56.4

A.2 Study1 Result

A.2.1 Pearson Correlation Coefficient between Benchmarks

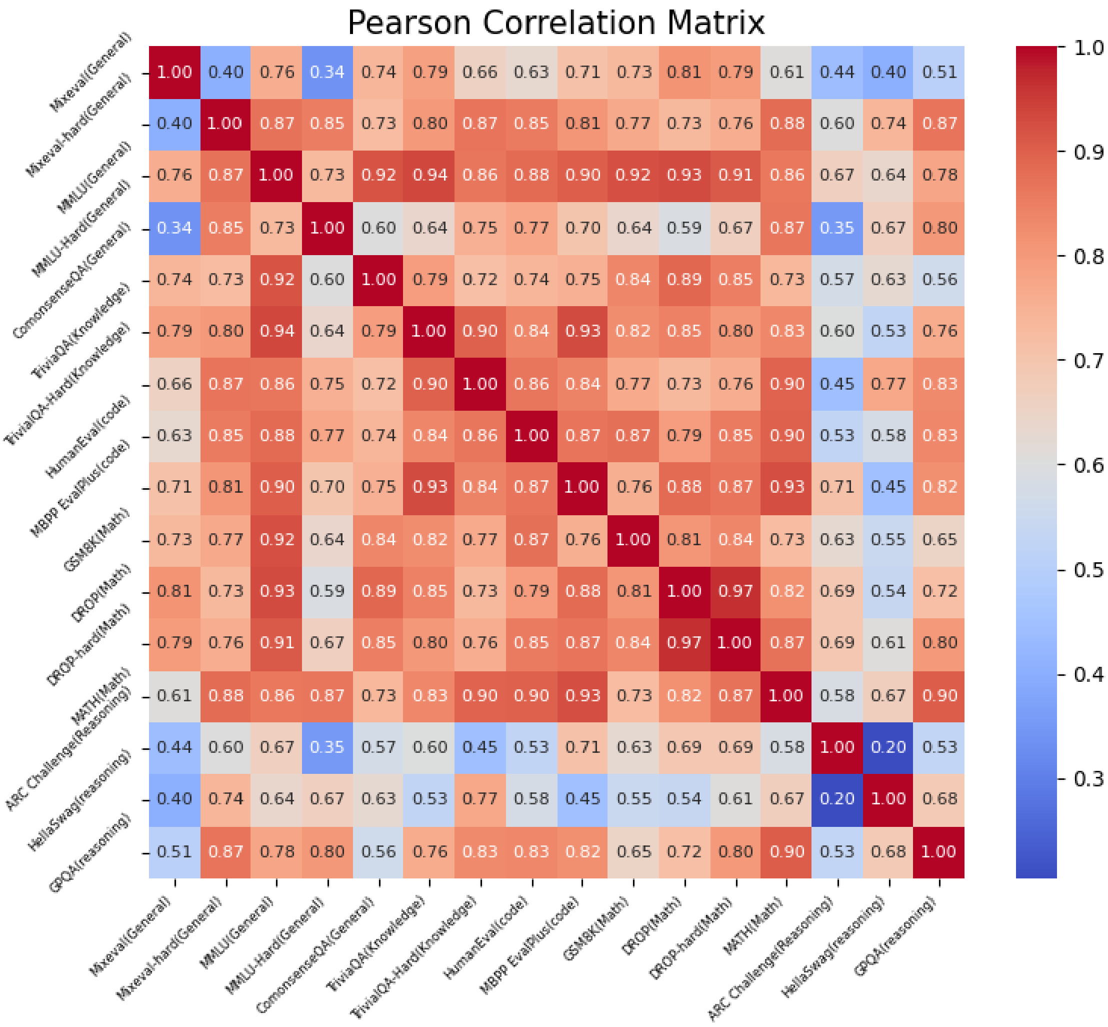

Figure 1: Pearson Correlation Coefficient between Benchmarks

Table 2: Pearson Correlation Coefficient between Benchmarks

	Mixeval (General)	**Mixeval-hard (General)**	**MMLU (General)**	**MMLU-Hard (General)**	**CommonsenseQA (General)**	**TriviaQA (Knowledge)**	**TriviaQA-Hard (Knowledge)**	**HumanEval (Code)**	**MBPP EvalPlus (Code)**	**GSM8K (Math)**	**DROP (Math)**	**DROP-hard (Math)**	**MATH (Math)**	**ARC Challenge (Reasoning)**	**HellaSwag (Reasoning)**	**GPQA (Reasoning)**
Mixeval (General)	1	0.401	0.76	0.339	0.74	0.788	0.655	0.634	0.709	0.679	0.812	0.795	0.606	0.438	0.401	-0.19
Mixeval-hard (General)	0.401	1	0.872	0.852	0.726	0.788	0.715	0.867	0.861	0.81	0.788	0.73	0.795	0.66	0.74	-0.282
MMLU (General)	0.76	0.872	1	0.882	0.919	0.937	0.9	0.887	0.902	0.904	0.928	0.91	0.881	0.665	0.637	-0.386
MMLU-Hard (General)	0.339	0.852	0.882	1	0.603	0.641	0.752	0.782	0.705	0.679	0.591	0.666	0.665	0.351	0.666	-0.268
CommonsenseQA (General)	0.74	0.726	0.919	0.603	1	0.794	0.715	0.828	0.801	0.788	0.889	0.849	0.606	0.573	0.621	-0.467
TriviaQA (Knowledge)	0.788	0.788	0.937	0.641	0.794	1	0.91	0.877	0.848	0.813	0.889	0.846	0.758	0.573	0.629	-0.377
TriviaQA-Hard (Knowledge)	0.655	0.715	0.9	0.752	0.715	0.91	1	0.801	0.848	0.826	0.884	0.808	0.753	0.601	0.526	-0.241
HumanEval (Code)	0.634	0.867	0.887	0.782	0.828	0.877	0.801	1	0.873	0.739	0.809	0.758	0.693	0.514	0.629	-0.223
MBPP EvalPlus (Code)	0.709	0.861	0.902	0.705	0.801	0.848	0.848	0.873	1	0.887	0.889	0.885	0.817	0.514	0.62	-0.461
GSM8K (Math)	0.679	0.81	0.904	0.679	0.788	0.813	0.826	0.739	0.887	1	0.888	0.889	0.875	0.69	0.607	-0.375
DROP (Math)	0.812	0.788	0.928	0.591	0.889	0.889	0.884	0.809	0.889	0.888	1	0.965	0.894	0.693	0.609	-0.248
DROP-hard (Math)	0.795	0.73	0.91	0.666	0.849	0.846	0.808	0.758	0.885	0.889	0.965	1	0.935	0.692	0.607	-0.331
MATH (Math)	0.606	0.795	0.881	0.665	0.606	0.758	0.753	0.693	0.817	0.875	0.894	0.935	1	0.585	0.607	-0.489
ARC Challenge (Reasoning)	0.438	0.66	0.665	0.351	0.573	0.573	0.601	0.514	0.514	0.69	0.693	0.692	0.585	1	0.204	0.051
HellaSwag (Reasoning)	0.401	0.74	0.637	0.666	0.621	0.629	0.526	0.629	0.62	0.607	0.609	0.607	0.607	0.204	1	0.051
GPQA (Reasoning)	-0.19	-0.282	-0.386	-0.268	-0.467	-0.377	-0.241	-0.223	-0.461	-0.375	-0.248	-0.331	-0.489	0.051	0.051	1

A.3 Study2 result:

Table 3: KMO Test and Bartlett's Test of Sphericity

Measure	**Value**	**Details**
KMO Value	**0.712**	-
Bartlett's Test of Sphericity	**Chi-square Approximation**	**122.731**
	df	**28**
	P-value	**0.000001**

Note: *** indicates significance at the 1% level, ** at the 5% level, and * at the 10% level.

Table 4: Variance Explanation Table (Before and After Rotation)

Component	**Pre-Rotation Variance Explanation**		**Post-Rotation Variance Explanation**	
	Eigenvalue	Variance (%)	Variance (%)	Cumulative Variance (%)
1	6.854	85.672	36.368	36.368
2	0.416	5.206	34.223	70.591
3	0.340	4.251	24.538	95.129
4	0.183	2.292	-	-
5	0.132	1.648	-	-
6	0.050	0.621	-	-
7	0.022	0.272	-	-
8	0.003	0.039	-	-

Table 5: Rotated Factor Loading Coefficients and Communality

Benchmark	**Factor 1**	**Factor 2**	**Factor 3**	**Communality**
MMLU (General)	0.625	0.647	0.410	0.978
Mixeval (General)	0.316	0.348	0.863	0.965
CommonsenseQA (General)	0.353	0.797	0.438	0.951
GSM8K (Math)	0.462	0.792	0.349	0.963
DROP (Math)	0.540	0.567	0.547	0.913
HumanEval (Code)	0.742	0.589	0.171	0.927
MBPP EvalPlus (Code)	0.833	0.351	0.392	0.970
TriviaQA (Knowledge)	0.739	0.378	0.504	0.943

Table 6: Component Scores for Benchmarks

Name	**Component 1**	**Component 2**	**Component 3**
MMLU (General)	0.096	0.228	-0.112
Mixeval (General)	-0.380	-0.351	1.165
CommonsenseQA (General)	-0.548	0.791	-0.005
GSM8K (Math)	-0.303	0.732	-0.239
DROP (Math)	-0.045	0.064	0.261
HumanEval (Code)	0.506	0.174	-0.591
MBPP EvalPlus (Code)	0.784	-0.548	-0.046
TriviaQA (Knowledge)	0.527	-0.482	0.203

Table 7: Post-Rotation Variance Explanation and Weight Distribution

Name	**Post-Rotation Variance (%)**	**Cumulative Variance (%)**	**Weight (%)**
Factor 1	36.368	36.368	38.230
Factor 2	34.223	70.591	35.975
Factor 3	24.538	95.129	25.794

Table 8: Eigenvalues and Number of Factors

Number of Factors	**Eigenvalue**
1	6.8538
2	0.4165
3	0.3400
4	0.1834
5	0.1318
6	0.0495
7	0.0218
8	0.0031

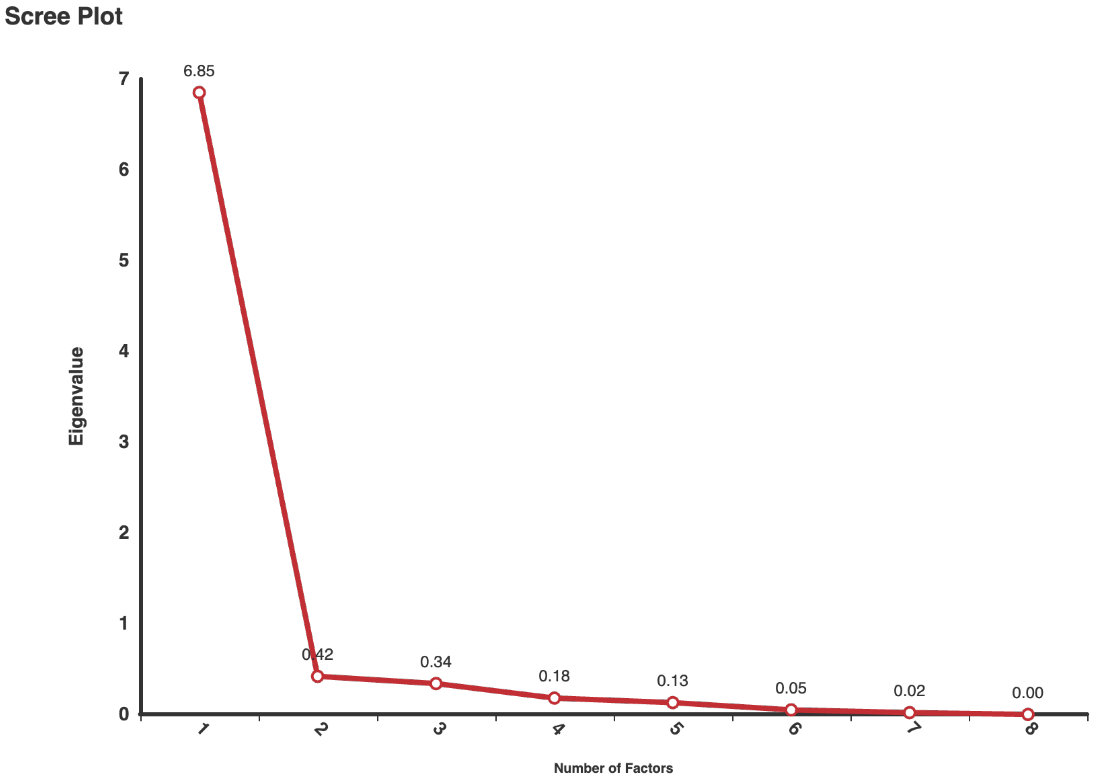

Figure 2: screen plot