





Towards a Visual Perception-Based Analysis of Clustering Quality Metrics

Graziano Blasilli , Daniel Kerrigan , Enrico Bertini , and Giuseppe Santucci 

Abstract—Clustering is an essential technique across various domains, such as data science, machine learning, and explainable artificial intelligence. Information visualization and visual analytics techniques have been proven to effectively support human involvement in the visual exploration of clustered data to enhance the understanding and refinement of cluster assignments. To support the human involvement, several perceptual studies and visual quality metrics have already been proposed. However, the visual perception of clustering quality metrics, also known as Cluster Validity Indexes (CVIs), still remains to be further explored. This paper presents the first attempt of a deep and exhaustive evaluation of the perceptive aspects of clustering quality metrics, focusing on the Davies-Bouldin Index, Dunn Index, Calinski-Harabasz Index, and Silhouette Score. Our research is centered around two main objectives: a) assessing the human perception of common CVIs in 2D scatterplots and b) exploring the potential of Large Multimodal Models, in particular GPT-4o, to emulate the assessed human perception. To this end, we conducted two systematic data studies and a user study covering a broad collection of datasets. By discussing the obtained results, highlighting limitations, and areas for further exploration, this paper aims to propose a foundation for future research activities.

Index Terms—Clustering, Clustering Quality Metrics, Clustering Validity Indexes, Visual Perception, Large Multimodal Models, LMM.

1 INTRODUCTION

Clustering is an important aspect across various fields, including data science, machine learning, and explainable artificial intelligence, which is often supported by human involvement using information visualization and visual analytics techniques [3, 11, 15, 25, 44]. Scatterplots and dimensionality reduction techniques have become a standard approach for exploring high-dimensional data and their clusters [15]. When dealing with clusters, human perception plays a crucial role, and several perceptual studies and visual quality metrics have been proposed [4, 17, 36, 42, 45]. However, clustering quality metrics, also known as Cluster Validity Indexes – CVIs – (e.g., Davies-Bouldin Index, Silhouette Score), remain the standard approach to measure the quality of clustering because they provide quantitative and objective evaluations of the clustering structure, assessing aspects such as compactness, separation, and consistency [23]. Despite their widespread use, it is to be further explored how these metrics align with human perception of clustering quality in visual representations, especially in bidimensional scatterplots.

This paper presents the first attempt of a deep and more exhaustive evaluation of the perceptive aspects of clustering quality metrics in 2D scatterplots. By focusing on common metrics such as Davies-Bouldin Index, Dunn Index, Calinski-Harabasz Index, and Silhouette Score, our research is centered around two main objectives (Sec. 4): assessing the human perception of common CVIs in 2D scatterplots and exploring the potential of Large Multimodal Models (LMMs), to emulate this human perception.

To assess human perception, we designed and conducted a systematic data study and a user study (Sec. 6) covering a broad collection of synthetic datasets. The results demonstrated clear trends in human perception of CVI variations: the higher the variation, the easier the perception. However, the ease and consistency vary, with the Davies-Bouldin Index and Silhouette Score that resulted to be more easily perceived than the Dunn Index and Calinski-Harabasz Index.

LMMs have already been used in visualization research to replicate human behavior for crowdsourcing experiments [13, 18, 20, 29, 39, 48]. Investigating how accurately such models can emulate the human visual perception of cluster validity indexes in 2D scatterplots,

the possibility of simulating crowdsourcing experiments opens up, reducing extensive human participation and making the whole process more efficient. To this end, we assessed (Sec. 7) the potential of OpenAI's GPT-4o model [27] to accurately emulate the assessed human perception of common clustering quality metrics on 2D scatterplots. Results demonstrated the ability of GPT-4o to perceive CVI variation, but a considerable difference between the model visual perception and human visual perception emerged, making the model unsuitable to simulate crowdsourcing experiments.

This paper presents an initial step toward the assessment of the visual perception of clustering quality metrics. By presenting findings, discussing limitations and emerged concerns, this paper aims to propose a foundation for future research activities.

2 RELATED WORK

Past research has analyzed human perception of clusters in scatterplots from several angles. Tatu et al. [42] proposed several guidelines for validating visual quality metrics for data visualization. Following these guidelines, they also proposed a preliminary evaluation of two visual cluster separation metrics, *Distance Consistency* and *Distribution Consistency* [37], to estimate the degree to which they reflect user perception. Their study demonstrated that the metrics require further validation regarding their alignment with human perceptual judgments.

Albuquerque et al. [2] proposed a visual quality measure that can be used to rank scatterplots by class separation based on human perception. Sedlmair et al. [36] extended the work Tatu et al. [42] and extensively evaluated the two visual cluster separation metrics, *Distance Consistency* and *Distribution Consistency* [37], which had been identified as the most effective state-of-the-art measures. Initially, the authors aimed to use these measures to guide dimensionality reduction techniques and visual encoding choices. An extensive user study involving over 800 plots found the two metrics ineffective in more than half of the cases, particularly with real datasets. Furthermore, the authors concluded that other existing perceptive metrics could behave in a similar way since their design and validation are similar to those of the chosen ones.

Sedlmair and Aupetit [34] proposed a new framework for evaluating visual quality metrics that is based on bootstrapping. They demonstrated it by evaluating various visual class separation metrics for labeled scatterplots and found *Distance Consistency* to be the best. In later work [6], they proposed numerous visual class separation metrics and evaluated them using this approach. The majority of the metrics outperformed *Distance Consistency*.

Abbas et al. [1] proposed a novel visual quality measure, *ClustMe*, designed to rank monochrome scatterplots based on perceived cluster

- Graziano Blasilli and Giuseppe Santucci are with: Sapienza Università di Roma, Italy. E-mail: {graziano.blasilli, giuseppe.santucci}@uniroma1.it.
- Daniel Kerrigan and Enrico Bertini are with: Northeastern University, USA. E-mail: {kerrigan.d, e.bertini}@northeastern.edu.

patterns. Their work involved conducting a study where participants labeled 1000 synthetic scatterplots as having either one or more than one cluster. Their measure excelled in predicting human judgments, showing a strong agreement with human rankings. Unlike our study, this work is based only on unlabeled data. Aupetit et al. [7] used the dataset of scatterplots and human labels from Abbas et al. to evaluate a range of clustering techniques based on how the techniques' chosen number of clusters agreed with the humans' labels. Jeon et al. [21] also used the dataset from Abbas et al. to create a method for estimating the amount of ambiguity in the clusters of a scatterplot.

Etemadpour et al. [16, 17] conducted user studies to analyze the performance of different projection methods according to human perception. They concluded that the performance of projection methods is task-dependent and varies depending on the nature of the data. Some tested tasks were about clustered data, such as estimating the number of clusters, identifying the closest clusters, and ranking clusters by density.

Quadri and Rosen [31] modeled the effects that factors such as the size and opacity of dots in scatterplots have on users' perception of the number of clusters. Quadri et al. [30] then expanded on this work to create a tool that automatically optimizes the design of scatterplots for cluster identification by adjusting the previously studied factors. Xia et al. [46] generated a synthetic dataset of scatterplots containing two clusters where they varied the shapes, densities, relative positions, and other aspects of the clusters. They studied the effect that these aspects had on whether or not humans labeled the scatterplots as having two clusters. In later work, Xia et al. [47] evaluated various dimensionality reduction techniques through experiments where participants completed visual cluster analysis tasks using scatterplots of projected datasets created by the techniques. Hartwig et al. [19] constructed a dataset of scatterplots with human-labeled clusters and used it to train a neural network to cluster two-dimensional data. They evaluated their approach against popular clustering algorithms and found it to have better agreement with human labels.

Lewis et al. [26] conducted a user study where they presented participants with multiple versions of a scatterplot that differed in class labels. They had participants select the best two clusterings and the worst clustering. They then compared several clustering quality metrics to the participants' choices. They found that some metrics, such as Silhouette Score [32] and Calinski-Harabasz Index [10], correlated well with human judgment, whereas others, such as Gamma [8] and weighted inter-intra [40], did not.

3 CLUSTER VALIDITY INDEXES

Clustering quality metrics, also known as *cluster validity indexes* (CVIs) [23], are essential metrics in cluster analysis. They offer objective measures of the quality of clustering results, such as assessing how well the clusters are formed, how well they match the ground truth (if available), and comparing different clustering results to select the best method and parameters. Two main families of metrics exist: external and internal. The external validation metrics evaluate the clustering quality by comparing the clustering results to an external ground truth or reference. These metrics can be only used when labeled data are available. Conversely, internal validation metrics evaluate the clustering quality based on the data itself without using external information. These metrics assess the compactness, separation, and overall structure of the clusters. By evaluating the relationship between intra-cluster cohesion (within-cluster) and inter-cluster separation (between-cluster), they can determine the quality of a clustering solution [24].

Among the several internal metrics that exist [23], the most common ones are Davies-Bouldin Index, Dunn Index, Calinski-Harabasz Index, and Silhouette Score [23, 28, 38]. Our work concentrates on these four indexes. A summary of these CVIs can be found in Tab. 1, with further details provided in the following. All the metrics depend on the concept of distance. We employed the standard approach based on using Euclidean distance. In the rest of the paper, we will use the terms *cluster quality metrics*, *validity indexes*, and *CVIs* interchangeably to refer to the same concept.

Table 1: List of the CVIs evaluated in this study.

CVI	Abbr.	Range	Best	Ref.
■ Davies-Bouldin Index	DB	$[0, +\infty]$	⬇️	[12]
■ Dunn Index	DN	$[0, +\infty]$	⬆️	[14]
■ Calinski-Harabasz Index	CH	$[0, +\infty]$	⬆️	[10]
■ Silhouette Score	SL	$[-1, +1]$	⬆️	[32]

■ **Davies-Bouldin Index** The Davies-Bouldin Index [12] is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. The common approach is to use Euclidean distance. The Davies-Bouldin Index ranges $[0, +\infty]$ with lower values indicating a better clustering quality.

■ **Dunn Index** The Dunn Index [14] is defined as the ratio of the minimum between-cluster distance to the maximum within-cluster distance. The metric ranges $[0, +\infty]$, where higher values indicate better clustering quality with well-separated and compact clusters. This metric is based on geometrical considerations that have the same basic rationale of the Davies-Bouldin Index; they are both designed to identify clusters that are compact and well separated [9]. Many variations of the metric exist since they are based on the different interpretations of inter-cluster distance and intra-cluster distance measures. Thus, as the Dunn Index is not available in the *scikit-learn* library [28], which is at the base of our experiments, we implemented our own version to facilitate this analysis. We considered the inter-cluster distance as the distance between the centroids and the intra-cluster distance as the mean distance of cluster points to its centroid.

■ **Calinski-Harabasz Index** The Calinski-Harabasz Index [10], also known as the Variance Ratio Criterion [28], assesses how well the clusters are formed by considering the ratio of the sum of between-cluster distance to the sum of within-cluster dispersion. The metric ranges $[0, +\infty]$ with higher values indicate a better clustering quality, meaning the clusters are well-separated and internally cohesive.

■ **Silhouette Score** The Silhouette Score [32] assesses how similar each point in a cluster is to its own cluster compared to other clusters. The metric ranges in $[-1, +1]$. A score close to 1 indicates that the point is well-matched to its own cluster and poorly matched to neighboring clusters. A score close to 0 suggests that the point is on or very close to the boundary between clusters. A negative score indicates that the point might have been assigned to the wrong cluster. The overall Silhouette Score is the average score of all points. Higher values indicate better-defined and well-separated clusters.

4 RESEARCH OBJECTIVES

This paper aims to bridge the algorithmic/mathematical validation of clustering quality and human visual perception. By focusing on common CVIs, this research investigates how well these metrics align with human visual interpretations of cluster quality in 2D scatterplots. The research is structured around two research questions (RQs). First, we examined the minimal variation in CVI values that humans can perceive. Second, we explored the potential of Large Multimodal Models (LMMs) to emulate human perception in evaluating clustering results on 2D scatterplots.

RQ1 What is the minimal variation of a given cluster quality metric in a bidimensional scatterplot that a human can perceive?

RQ2 Can Large Multimodal Models emulate the human visual perception of cluster quality metrics in a bidimensional scatterplot to accurately simulate crowdsourcing experiments?

Research Question #1 Bidimensional scatterplots are commonly used to represent clustering results because they allow the datasets to be visualized through projections into a two-dimensional plane [15, 36].

Cluster validity indexes (CVIs) are objective measures that detect clustering quality under various geometrical configurations [23]. These metrics can vary widely, with some having potentially infinite ranges, such as the Davies-Bouldin Index or the Dunn Index. When a clustered dataset is projected in different ways, using various projection techniques or different parameterizations of the same method, it can result in different CVI values. Determining the best plot for presenting the dataset to users is challenging because a better CVI value does not necessarily imply that the difference between two CVI values is visually perceptible to humans. Thus, despite having different CVI values, two scatterplots might appear similar or even indistinguishable to the human eyes. Therefore, it is important to understand the minimal variation in a given CVI that humans can visually perceive.

Research Question #2 User studies involving participants analyzing thousands of plots present a significant challenge due to the required time and effort. These tasks demand extensive attention and can be highly time-consuming for the users. Large Language Models have already been used in visualization research to replicate human behaviors [29,39,48]. They have also been used for crowdsourcing [13,18,20]. If these models can emulate the human visual perception of cluster validity indexes in 2D scatterplots, the possibility of accurately simulating crowdsourcing experiments opens up. This capability would significantly reduce the need for extensive human participation by allowing AI to perform deeper evaluations and make the process more efficient.

5 DESIGN CONSIDERATIONS

To answer our research questions, we designed two experiments. The first experiment (Sec. 6) aims to answer to **RQ1** using a two-phase approach, integrating a *data study* and a *user study*. The second experiment (Sec. 7) aims to answer to **RQ2** using a *data study* on which is employed the Large Multimodal Model GPT-4o as a user.

In this section, we introduce the preliminary choices and the rationale that guided the design of our experiments, while detailed descriptions of each experiment are provided in the following section. Our approach started by defining a reliable method to measure CVI variations. After that, we designed a specific evaluation task to assess the human perception of such variations. This task involved comparing pairs of scatterplots to evaluate their clustering representations. Then, we needed to select appropriate datasets for our experiments, and we opted to generate synthetic datasets to ensure a comprehensive study.

CVI Variation Cluster Validity Indexes play an essential role in evaluating the quality of clusterings. However, one of the challenges faced when working with CVIs is the variation in their ranges, which, for some metrics, can be infinite. Additionally, the value of a CVI is influenced by various factors, such as the number of instances in the dataset, the number of dimensions, and the number of clusters. This intrinsic variability makes it difficult to directly compare CVI values obtained from different clusterings or two projections plotting the same dataset. The four CVIs chosen for our analysis exemplify this issue, as each has distinct ranges and is sensitive to the aforementioned factors. To overcome this limitation and to ensure a fair and consistent comparison, it is essential to use a metric that normalizes these variations.

To achieve this, we have chosen the *Relative Percent Difference* (RPD) as a metric for measuring the variation between two CVI values and providing a quantitative measure of its variation. RPD (Eq. (1)) is a metric that quantifies the change between two values relative to their average, offering a standardized way to express the difference.

$$RPD(x,y) = \frac{|x-y|}{\frac{1}{2}(|x|+|y|)} \quad (1)$$

By using RPD, we can effectively compare two CVI values without being affected by their ranges or the specific characteristics of the dataset, such as the number of instances, dimensions, or clusters. The use of RPD ensures that the comparison is independent of these variables, making it a robust method to assess the variation of CVI metrics between two different clusterings.

Task We designed a simple and fast perceptual task to assess the human perception of CVI variations. The user is provided with two scatterplots representing a clustered dataset. These two scatterplots represent the same dataset projected into 2D in different ways. This ensures that it is possible to compare the difference in terms of CVIs between the two plots. The task for the user is to determine whether these scatterplots look similar in representing the clusters or if one of them provides a better representation of the clustering. To accomplish this task, the user is suggested to consider several key factors such as the shapes of the clusters, the purity of the clusters with respect to their labels (i.e., whether points of the same color are grouped together correctly), the separation between clusters, and how tightly the points within each cluster are grouped. An example of such pairs is shown in Fig. 6 and Fig. 7. If the user finds that the two scatterplots look similar, meaning both represent the clustering in a similar manner, this indicates that the variations in CVIs between the two plots have not been perceived. In other words, the visual differences are not significant enough to impact the user's perception of clustering quality. Conversely, if the user chose one scatterplot as a better representation of the clustering, this suggests that the variations in the CVIs have been noticed.

Perception Probability Perception probability in our study refers to the likelihood that participants can visually detect CVI variations. Given a pair of scatterplots (e.g., Fig. 6), for each CVI we have an RPD value of the pair. For each CVI under analysis, we evaluate the perception probability separately. When participants (humans, LMMs) consider both plots similar or equal in representing the clustering, it implies that the visual difference has not been noticed, and therefore, the corresponding RPD value has not been perceived. Conversely, if participants identify one plot as better, it implies that the visual difference has been detected. In the participant's responses, we assign a perception score of 1 when the visual difference is perceived and a score of 0 when it is not. By averaging the perception scores across different RPD values (for a given CVI), we can compute an average perception curve (e.g., Fig. 5), where the x-axis represents the RPD values and the y-axis represents the Perception Probability. The precision of the average perception probability improves with the number of perception scores associated with each RPD value; this can be obtained with many pairs sharing a similar RPD value or with many users. To account for variability in user responses, we fit a polynomial regression model (degree 5) to the user responses. This modeling approach helps to estimate the underlying perception probability distribution more accurately. As a result, we obtain an average perception probability curve and its confidence interval, representing how perceptible the CVI variations are to human (or LMM) observers. In the following analyses, we adopted the statistical concept of probability $p \geq 0.95$, a commonly used threshold to define a high probability value.

Datasets To answer our research questions, a primary objective was to ensure that the dataset instances (scatterplot pairs) used to evaluate the CVI variations encompassed a broad spectrum of RPD values across all the CVIs under consideration. To achieve that, we opted for the generation of synthetic data, which allowed us to control the properties of the datasets and ensure the desired variability in the CVI metrics. Although several datasets have been proposed for benchmarking clustering metrics (e.g., [33]), they do not exhaustively cover all the RPD values we need, and also, they do not provide little variations of the same dataset that are suited for our study.

We used the *repliclust* [49] Python package, a versatile tool designed for creating synthetic datasets with defined cluster structures. By leveraging this tool, we were able to produce datasets that not only varied in the number of clusters but also exhibited a range of spatial arrangements, thereby influencing the CVI values. The core of our data generation process involved creating pairs of two-dimensional datasets. Each pair consisted of two datasets that shared the same cluster points but differed in spatial configuration. By using *repliclust*, we generated 1732 bi-dimensional datasets arranged in 879 pairs. Datasets have 1000 data entries each and a number of clusters ranging

from 2 to 5. This limited range has been chosen to cover a variety of clustering complexities, from simpler structures with fewer clusters to more complex ones with more clusters, while remaining within users’ cognitive limits [41]. Each dataset pair was designed to have identical points per cluster, ensuring that any differences in the CVI values were due to changes in spatial arrangement rather than variations in the data points themselves. By varying several parameters during the generation, we obtained a dataset collection that captured a wide range of RPDs for each considered CVI, providing a robust basis for evaluating the CVI variations.

We called this collection of datasets “SDPC”, which stands for *Synthetic Dataset Pairs Collection*, available for replicability or future research at <https://github.com/blasilli/ClusteringMetricsPerception>.

Large Multimodal Model Large Multimodal Models (LMMs) are advanced artificial intelligence systems designed to understand, generate, and manipulate multiple types of data, including images and human language. To conduct our study, we needed to choose an appropriate LMM capable of working with both text and image input. After considering various options, we selected GPT-4o [27] for its advanced capabilities in reasoning across multiple modalities, including audio, vision, and text. GPT-4o is an advanced model developed by OpenAI, designed to understand and generate human-like text based on the input it receives. Building on the capabilities of its predecessors, GPT-4o has enhanced abilities in natural language processing, enabling it to engage in complex conversations, provide detailed explanations, and perform a variety of language-related and vision-related tasks with high accuracy.

6 EXPERIMENT 1 – ASSESSING THE HUMAN PERCEPTION

To answer to **RQ1**, we designed the experiment *E1* that adopted a two-phase approach, integrating a *data study* in the first phase (*E1.1*) and a *user study* in the second one (*E1.2*). This dual-phase approach has been designed to understand human perception of variation in CVIs while managing the practical constraints of user participation. Furthermore, this approach has been designed to ensure that our findings are robust and well-supported by empirical evidence.

A conceptual scheme of the experiment is shown in Fig. 1. The first phase, by using the almost thousand pairs in SDPC, aimed at analyzing a wide spectrum of CVI variations. The obtained preliminary results allowed us to define the subset SDPC-72 to be used in the second phase. A user study allowed us to better understand and confirm the perception patterns that emerged in the previous phase.

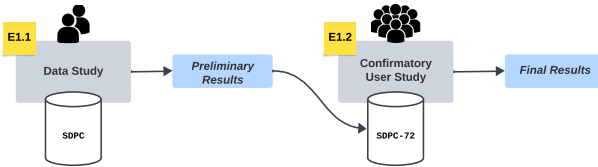


Fig. 1: Conceptual scheme of Experiment 1, designed to answer to **RQ1**. We conducted a data study (*E1.1*) with two participants analyzing 879 scatterplot pairs. Then, we conducted a confirmatory user study (*E1.2*) with 13 participants that analyzed 72 scatterplot pairs, accurately selected from the original set, by considering the perceptive patterns that emerged in (*E1.1*).

6.1 E1.1 – Data Study

6.1.1 Method

In the first phase, our goal was to explore a wide spectrum of CVI variations by considering all the pairs in the SDPC collection. However, presenting such an extensive and time-consuming study to many users was believed to be infeasible. Therefore, we adopted the *data study* approach [35]. This approach is complementary to a user study: rather

than a few datasets observed by many people, a few people observe many datasets. This wide analysis enabled us to identify preliminary patterns in the perception of the CVIs and to determine the perceptive threshold of such variation to be further investigated.

6.1.2 Procedure

Two authors of this paper analyzed the thousand scatterplot pairs in SDPC collection. Each author individually performed the evaluation task (Sec. 5) on all the pairs. This task involved determining whether the two scatterplots in each pair appeared similar in representing the clusters or if one provided a better representation. As a result of the analysis conducted in this data study, we defined RPD threshold values for the perception of the four CVIs. These threshold values represent the RPD values that, below them, the variations in the CVIs become perceptible to human observers. By identifying these thresholds, we gained a preliminary understanding of the human perception of these metrics.

6.1.3 Results

From the responses of the two participants, we computed their perception probability of CVI variations by employing a polynomial regression model as explained in Sec. 5. This approach allowed us to estimate the underlying perception probability distributions shown in Fig. 2.

The estimation of the perception probability of the four CVIs demonstrates clear trends in human perception of CVI variations. Both the Davies-Bouldin Index and Silhouette Score show easier perception, with convergence ($p \geq 0.95$) that starts after RDP values 38% and a fast-growing trend, indicating a strong perception of differences above these thresholds. The Dunn Index also shows an increasing trend with higher variability, converging after RDP values exceed 61%. The Calinski-Harabasz Index, however, has more variability and converges after RDP values exceed 82%, suggesting it is harder for participants to visually perceive differences captured by this CVI.

Overall, while participants can reliably perceive variations in clustering quality across all CVIs, the ease and consistency vary, with the Davies-Bouldin Index and Silhouette Score being more easily perceived than the Dunn Index and Calinski-Harabasz Index. In the following, we provide a detailed analysis of each considered CVI.

E1.1 – Davies-Bouldin Index The perception probability chart for the Davies-Bouldin Index (Fig. 2a) shows a clear increasing trend, where higher RDP values correspond to a higher probability of being visually perceived. There is some variability in the participants’ responses, but this variability is limited and does not significantly affect the overall trend. The curve starts to converge ($p \geq 0.95$) after the RDP value exceeds 38%, indicating that participants consistently note differences in the CVI above this threshold. The confidence interval suggests that the two participants of the study are reasonably consistent in perceiving visual differences in clustering quality as quantified by the Davies-Bouldin Index. Overall, the perception of Davies-Bouldin Index variations seems to be easily perceived.

E1.1 – Dunn Index The perception probability chart for the Dunn Index (Fig. 2b) shows a clear increasing trend, where higher RDP values correspond to a higher probability of being visually perceived. The curve starts to converge ($p \geq 0.95$) after the RDP value exceeds 61%, indicating that the participants consistently note differences in the CVI above this threshold. The confidence interval and the provided responses indicate that the two participants of the study are less consistent in perceiving visual differences in clustering quality as quantified by the Dunn Index. Overall, the perception of Dunn Index variations seems to be less easily perceived than the one of Davies-Bouldin Index.

E1.1 – Calinski-Harabasz Index The perception probability chart for the Calinski-Harabasz Index (Fig. 2c) shows an increasing trend, where higher RDP values correspond to a higher probability of being visually perceived. There is considerable variability in the participants’ responses. This suggests that the characteristics of the CVI could be difficult to visually perceive. The curve starts to converge ($p \geq$

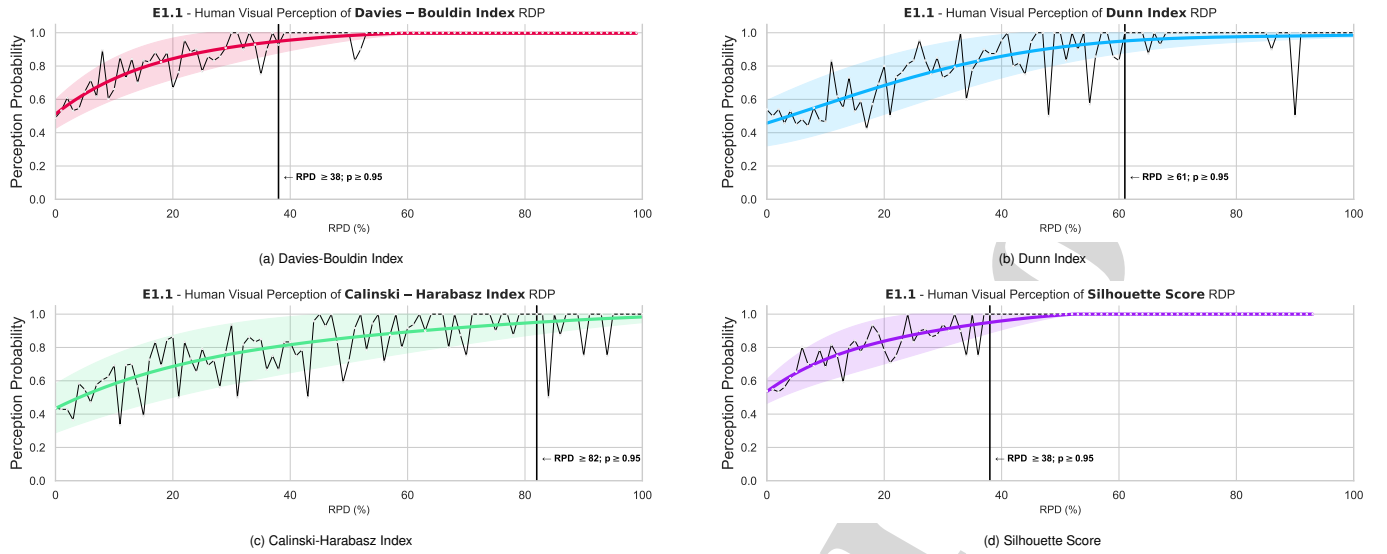


Fig. 2: **Experiment 1.1 – data study.** Preliminary results about human perception of variations in the four Clustering Validity Indexes (CVIs) under analysis. The charts illustrate the human perception probability for different Relative Percent Difference (RPD) values across four CVIs: (a) Davies-Bouldin Index, (b) Dunn Index, (c) Calinski-Harabasz Index, and (d) Silhouette Score. The black dots represent the actual average responses of the users, while the thin black curve shows the trend. The thick colored line shows the estimated probability distribution obtained using a polynomial regression model. The colored area represents the confidence interval. Additionally, the vertical line highlights when the perception probability starts to be greater than 0.95, meaning that higher RPD values seem to always be noticed.

0.95) after the RDP value exceeds 82%, indicating that the participants consistently note differences in the CVI above this threshold. Overall, the perception of Calinski-Harabasz Index variations seems to be the most difficult to perceive with respect to the other CVIs.

■ **E1.1 – Silhouette Score** The perception probability chart for the Silhouette Score (Fig. 2d) shows a clear increasing trend, where higher RDP values correspond to a higher probability of being visually perceived. There is some variability in the participants' responses, but this variability is limited and does not significantly affect the overall trend. The curve starts to converge ($p \geq 0.95$) after the RDP value exceeds 38%, indicating that the participants consistently note differences in the CVI above this threshold. The global trend indicates that the two participants of the study are reasonably consistent in perceiving visual differences in clustering quality as quantified by the Silhouette Score. Overall, the perception of Silhouette Score variations seems to be easily perceived, more than the other CVIs.

6.2 E1.2 – User Study

6.2.1 Method

This user study aims to confirm or adjust the preliminary results obtained in E1.1. Proposing a user study with almost the thousand plot pairs in SDPC was considered unfeasible due to the practical constraints of user participation. The insights gained from the previous phase allowed us to generate SDPC-72, a smaller (72 pairs) and carefully selected subset from the original collection, but being at the same time representative of the distribution of SDPC.

By using the convergence threshold assessed in E1.1, we first started to discard, from SDPC, all the pairs of plots with RPD values over these thresholds. Then, we selected a subset of pairs that is representative of the distribution of the bigger collection. Additionally, this subset has been chosen to ensure that all four CVIs under consideration in this study agree on the same scatterplot being the better representation for each pair. That is, for each pair of scatterplots, one plot is considered superior by all the CVIs under analysis. This consensus is significant because it is not guaranteed that all CVIs will agree on the same plot. Each CVI metric considers different aspects of clustering, such as cluster separation, cohesion, and purity, which can lead to different outputs. An example of such disagreement is shown in Fig. 7.

Using this subset, we designed a *user study* in which participants have to perform the comparison task presented in Sec. 5.



Fig. 3: Example of a question from the user study in Experiment 1. Participants are presented with two scatterplots, each displaying the same dataset but with different spatial arrangements, and are asked to determine whether the scatterplots look similar in representing the clustering or if one of them provides a better representation. The task involves considering cluster shapes, purity, separation, and compactness.

6.2.2 Procedure

From the empirical evidence obtained in E1.1, we defined the subset SDPC-72. Using this data collection, we developed a user study implemented as a web application, using an adapted version of the STEIN tool [5]. An example of a question from the user study is visible in Fig. 3. Participants were first provided with a short presentation explaining the objectives of our study and the task they will accomplish. After some profiling questions, participants proceeded to complete the online study independently.

Participants A total of 13 participants were involved in the user study. The group includes individuals across various academic positions: 2 Full Professors, 1 Associate Professor, 2 Assistant Professors, 1 Postdoc researcher, and 7 PhD students. Participants had previous experience interpreting the results of clustering analyses using scatterplots, and

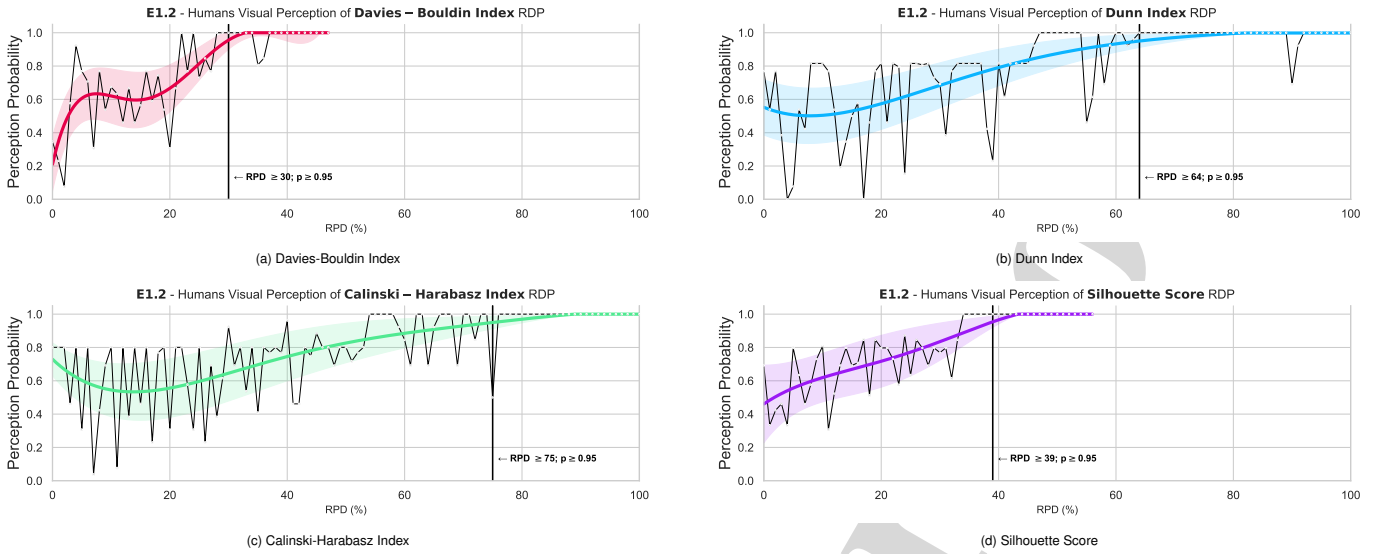


Fig. 4: **Experiment 1.2 – user study.** User study results about human perception of variations in the four Clustering Validity Indexes (CVIs) under analysis. The charts illustrate the human perception probability for different Relative Percent Difference (RPD) values across four CVIs: (a) Davies-Bouldin Index, (b) Dunn Index, (c) Calinski-Harabasz Index, and (d) Silhouette Score. The black dots represent the actual average responses of the users, while the thin black curve shows the trend. The thick colored line shows the estimated probability distribution obtained using a polynomial regression model. The colored area represents the confidence interval. Additionally, the vertical line highlights when the perception probability starts to be greater than 0.95, meaning that higher RPD values seem to always be noticed.

most of them felt moderately to very comfortable performing such tasks. Only two participants represented an exception; they felt slightly uncomfortable in interpreting the results of clustering analyses using scatterplots. The group is comprised of individuals actively engaged in different research fields. The fields most familiar to them are artificial intelligence, data science, machine learning, deep learning, data mining, data visualization, and visual analytics. This diverse expertise enriched the study by providing a broad perspective on the visual interpretation of clustering quality metrics. This paper, being a foundation for future research activities, involved a limited number of participants. Future research activities should involve a higher number to obtain more accurate results.

6.2.3 Results

From the responses of the participants, we computed their perception probability of CVI variations by employing a polynomial regression model as explained in Sec. 5. This approach allowed us to estimate the underlying perception probability distributions as shown in Fig. 4.

The estimation of the perception probability of the four CVIs demonstrates clear trends in human perception of CVI variations; the higher the RPD, the higher the perception probability. Both the Davies-Bouldin Index and Silhouette Score show easier perception, i.e., faster-growing trend and limited variability and fluctuations. The Dunn Index shows a slower increasing trend with higher variability. The Calinski-Harabasz Index, as in the *E1.1*, shows more variability, suggesting it is harder for participants to visually perceive differences captured by this CVI. Overall, these results follow the trends that emerged in *E1.1*, with the values of probability that are, in general, lower than the one assessed in the experiment *E1.1*.

While the averaged perception probability, computed by the polynomial model, is similar to the one obtained in the qualitative data study, the actual users' responses show considerable fluctuations, especially the Dunn Index and Calinski-Harabasz Index. It could be interesting, as future work, to involve more users in a future user study to limit the variability potentially generated by a few groups of users.

6.3 Overall Results

To obtain the overall resulting human perception of CVI variations, we merged the findings from both the data study (*E1.1*) and the user

study (*E1.2*). This approach allowed us to integrate the strengths of both methods to form a more comprehensive understanding of how variations in clustering quality metrics (CVIs) are perceived.

We employed polynomial regression models to synthesize the two preliminary results (*E1.1* and *E1.2*) by averaging them. This averaged approach has been chosen to emphasize the importance of the user study while still incorporating the extensive results from the quantitative analysis. The polynomial regression models provided a curve that captured the overall trend of human perception across the different CVIs, combining the insights from both studies. The resulting perception probability curves, which illustrate the overall assessed human perception of CVI variations, are presented in Fig. 5.

The perception probability of CVI variations can be categorized into two distinct groups based on the trends observed in the perception probability charts. The first group includes the Davies-Bouldin Index and the Silhouette Score, which show a similar trend based on a rapid increase in perception probability. This suggests that humans can easily perceive variations in these indexes, with $p \geq 0.95$ occurring after both RPD values exceed 39% and 40%, respectively. Their confidence intervals show a considerable variance in the RPD range of 0-25%, rapidly decreasing after this range. The second group includes the Dunn Index and the Calinski-Harabasz Index, which show smoother increasing trends, with higher thresholds for $p \geq 0.95$ at 63% and 80% RPD, respectively. Their confidence intervals show an almost constant variance over the trend, which smoothly decreases. The overall trends of the two groups highlight the distinct perceptual characteristics between the two groups of CVIs.

■ **Davies-Bouldin Index** The perception probability chart for the Davies-Bouldin Index (Fig. 5a) shows a clear increasing trend, where higher RPD values correspond to a greater likelihood of being visually perceived. The confidence interval suggests higher variability in perception at lower RPD values, which decreases as RPD values increase. The curve starts to converge ($p \geq 0.95$) after the RPD value exceeds 43%, indicating that humans consistently recognize differences in this CVI above this threshold. Overall, the rapid increase in the perception probability suggests that variations in the Davies-Bouldin Index are well perceived.

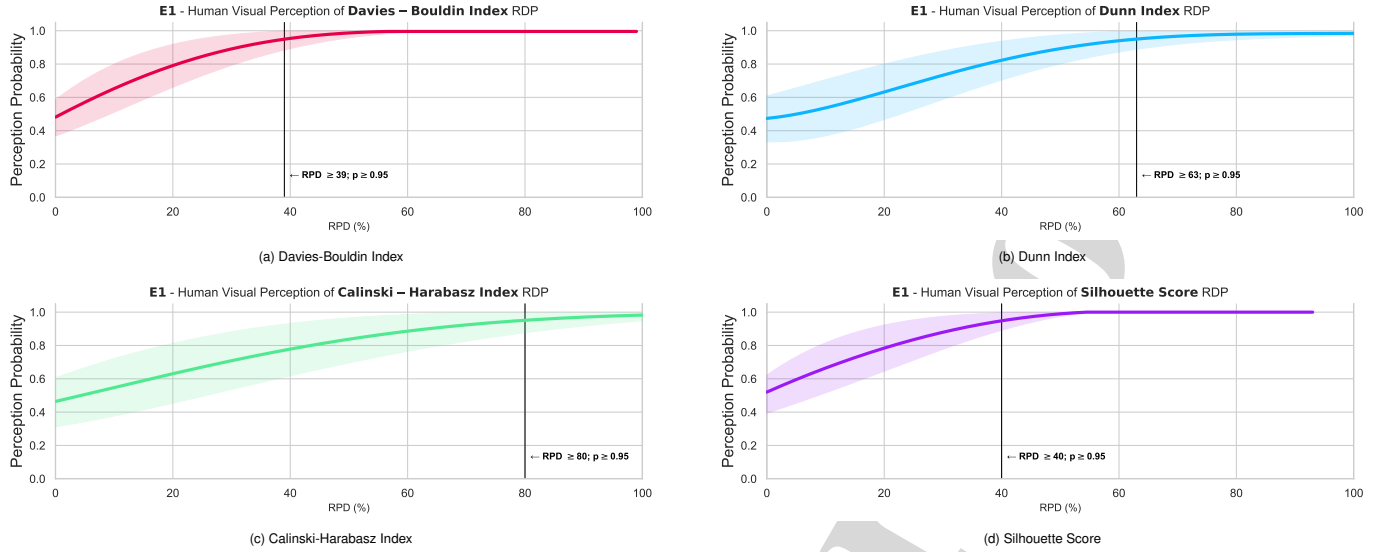


Fig. 5: **Experiment 1 – Overall Results.** Overall results about human perception of variations in the four Clustering Validity Indexes (CVIs) under analysis. The charts illustrate the human perception probability for different Relative Percent Difference (RPD) values across four CVIs: (a) Davies-Bouldin Index, (b) Dunn Index, (c) Calinski-Harabasz Index, and (d) Silhouette Score. The thick colored line shows the estimated probability distribution obtained using a polynomial regression model that considered both the results of *E1.1* and *E1.2*. The colored area represents the confidence interval. Additionally, the vertical line highlights when the perception probability starts to be greater than 0.95, meaning that higher differences seem to always be noticed.

Dunn Index The perception probability chart for the Dunn Index (Fig. 5b) shows a smooth increasing trend, indicating that higher RPD values lead to a higher probability of being visually perceived. The confidence interval reveals variability at lower RPD values, which decreases as the RPD grows. The curve starts to converge ($p \geq 0.95$) after the RPD value exceeds 64%, demonstrating that humans consistently detect differences in the CVI above this threshold. Overall, the smooth growth trend implies that variations in the Dunn Index are less easily perceived than other indexes.

Calinski-Harabasz Index The perception probability chart for the Calinski-Harabasz Index (Fig. 5c) shows a smooth increasing trend, with higher RPD values leading to a higher probability of visual perception. The confidence interval indicates variability at lower RPD values, which reduces as the RPD values increase. The curve starts to converge ($p \geq 0.95$) after the RPD value exceeds 75%, indicating that participants consistently notice differences in the CVI above this threshold. The gradual increase suggests that variations in the Calinski-Harabasz Index are less easily perceived due to its smooth growing trend.

Silhouette Score The perception probability chart for the Silhouette Score (Fig. 5d) shows a clear increasing trend, where higher RPD values correspond to a higher probability of being visually perceived. The confidence interval shows higher variability at lower RPD values, which diminishes as RPD values rise. The curve starts to converge ($p \geq 0.95$) after the RPD value exceeds 45%, indicating that humans consistently recognize differences in this CVI above this threshold. Overall, the rapid increase in perception probability indicates that variations in the Silhouette Score are easily perceived.

7 EXPERIMENT 2 – ASSESSING THE LMM PERCEPTION

To address **RQ2**, we designed experiment *E2* using the *data study* approach. This experiment has been designed to understand the GPT-4o (gpt-4o-2024-05-13) model's perception of variations in CVIs.

7.1 Method

Our goal was to replicate the data study of Experiment 1, but instead of human participants, we employed the GPT-4o model as the evaluator. To achieve this, we utilized all the 1000 scatterplot pairs from the SDPC

collection. Each pair consisted of two scatterplots representing the same dataset but arranged differently. We passed these images to the GPT-4o model, asking it to determine whether the two scatterplots appeared similar in representing the clusters or if one of them provided a better representation. In addition, we also requested the model to provide its reasoning and motivations for each choice. This information was important for understanding the model's decision-making process and for comparing its reasoning to human participants' reasoning. By analyzing the choices and the explanations provided by GPT-4o, we aimed to assess how the model perceives variations in CVIs and whether its behavior aligns with human perceptions.

7.2 Procedure

One of the primary challenges we faced was specifying the task in a way that aligned the model capabilities with human visual perception. After several iterations and extensive testing, we defined a prompt that communicates the desired task and the expected approach to GPT-4o. The finalized prompt, reported in the following, was designed to guide the model in performing the evaluation as a human by considering different clustering factors and explaining the choices.

You are a human participating in a user study on visual perception of clustering quality metrics. You will see an image with two scatterplots: A on the left and B on the right, plotting the same dataset differently. Assume that the labeling is correct, so the points of the same colors should stay closer to each other. Your task is to determine whether the scatterplots look similar in representing the clusters or whether one of the scatterplots looks better. You should consider factors such as cluster shapes, how pure the clusters are concerning the labels, how separated the clusters are, and how tight the clusters are. Remember that humans cannot perceive these factors if their difference is small, pushing to conclude that the scatterplots look similar. Answer either with A, B, or BOTH. Think step by step and respond in JSON format with 'plot' and 'motivation' tags, providing detailed analysis. Do not include any text other than the JSON object.

7.3 Results

From the responses of GPT-4o, we computed its perception probability of CVI variations by employing a polynomial regression model as explained in Sec. 5. This approach allowed us to estimate the underlying perception probability distributions shown in Fig. 8 on the appendix.

The estimation of the perception probability of the four CVIs demonstrates clear trends in the perception of CVI variations: the higher the RPD, the higher the perception probability. Similarly to the assessed human perception (Sec. 6), both the Davies-Bouldin Index and Silhouette Score show easier perception, while Dunn Index and Calinski-Harabasz Index suffer from higher variability in the responses. Conversely, all the CVIs show faster-growing trends, with convergence ($p \geq 0.95$) that starts when RPD exceeds about 20% for all the CVIs. This suggests a higher visual perception of details by GPT-4o than humans.

The perception probability charts of the four CVIs demonstrate the ability of GPT-4o to perceive CVI variations. Differently from the human perception (Sec. 6), the model was able to perceive minimal visual differences that, instead, humans missed. As a result, we found a considerable difference between the GPT-4o visual perception and human visual perception that makes the model not suitable to simulate crowdsourcing experiments.

8 DISCUSSION AND LIMITATIONS

In this paper, we focused on evaluating the human and GPT-4o perception of differences in clustering quality metrics without assessing the correctness of these perceptions. Specifically, when a user looks at a pair of scatterplots and selects one because it is better at representing the clustering, it could be important to also determine whether this choice aligns with the CVIs' choices. Even if a participant notices a difference, they might select the worse plot, as illustrated in Fig. 6. Interestingly, the user study (E1.2) revealed a lack of consensus among participants: 8 users indicated that both plots were similar, 3 users chose plot A, and 2 users chose plot B. In contrast, both participants in experiment E1.1 chose plot B. Also GPT-4o selected plot B, motivating its choice: "Plot B shows better separation between the clusters and tighter grouping of points within each cluster. The blue cluster is more distinct and less mixed with the pink cluster compared to Plot A, where there is more overlap between the clusters". Additionally, CVIs may disagree on which plot in a pair represents the better clustering; see Fig. 7. This discrepancy highlights another important area for further investigation: understanding why these CVIs sometimes provide conflicting assessments and which aspects influence human perception more. Another aspect to discuss is whether the Relative Percent Difference is the appropriate metric for measuring variations in CVIs. While RPD is a standard statistical approach, it could be important to consider that the suitability of RPD may vary depending on the specific CVI being analyzed. Different CVIs measure different aspects of clustering quality, such as compactness, separation, and overall structure. Therefore, while RPD might be effective for one CVI, it may not be suitable for another. Future research should explore if the RPD accurately reflects the variations in CVIs. Furthermore, it is also important to consider how the representation of datasets in scatterplots might influence user responses. We used a standard scatterplot, but various techniques (e.g., [30]) for optimizing scatterplots could be explored to minimize such possible biases.

The results of experiment E.2 show a considerable difference between the model's perception and human visual perception. This difference raises concerns about the ability of the GPT-4o model to accurately emulate human perception. Determining whether the prompt "successfully" communicates the task and the expected behavior is challenging. Therefore, one potential explanation for this mismatch could be related to the prompt we used to instruct the model. An aspect that we did not explore is the model's behavior when presented with identical scatterplots in a pair. If GPT-4o still picks one over the other, it indicates a level of randomness or bias in its responses that needs to be considered. So, future research should focus on refining prompts to better guide the model in emulating human perception. Another possible explanation for the observed differences is the model's advanced visual capabilities compared to humans. The GPT-4o model abilities might detect slight variations that are, instead, imperceptible to humans, leading to the high probability values assessed in the results. Addressing this issue could require fine-tuning the model to align its perception to the human visual perception. In summary, to improve the alignment between the model perception and human perception, we highlight several future

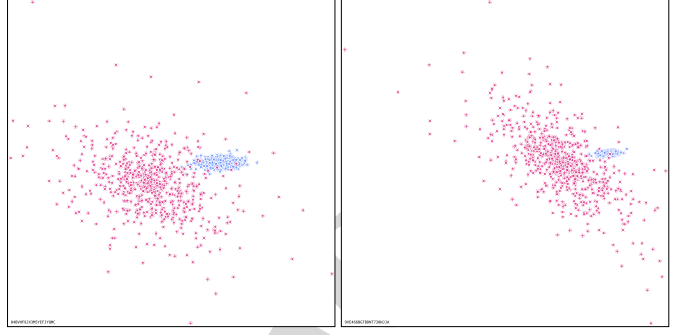


Fig. 6: The four CVIs consider **Plot A** (left) to be better (RPDs: DB=25%, DN=38%, CH=75%, SL= 28%). Both participants in experiment E1.1 and GPT-4o (E2) chose **Plot B** (right). The experiment E1.2 shows a lack of consensus among participants: 8 users indicated that both plots were similar, 3 users chose plot A, and 2 users chose plot B.

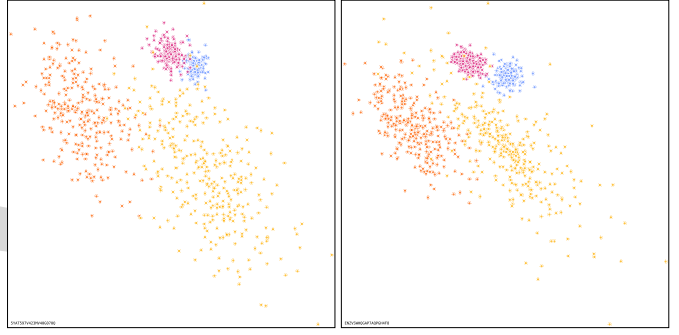


Fig. 7: The four CVIs disagree on the better plot. Davies-Bouldin Index and Calinski-Harabasz Index prefer **Plot A** (left), while Dunn Index and Silhouette Score prefer **Plot B** (right). RPDs: DB=17%, DN=52%, CH=34%, SL= 1%. Both participants in E1.1 and GPT-4o (E2) chose plot B.

research directions: • *Prompt Engineering*: Develop better prompts that clearly instruct the LMM on how to replicate the human perception. • *Fine-Tuning*: Perform fine-tuning of an LMM to better align its perception with human visual perception. • *Assess Correctness in Terms of CVI*: Investigate not only the model's perception but also its correctness in choosing the plot with the better CVI, considering potential disagreements between different CVIs. • *Explore Different LMMs*: Test different LMMs, such as LLaMa [43], Mistral [22], etc., to compare their performance in emulating human perception.

9 CONCLUSION

This paper presented an initial step toward assessing the visual perception of clustering validity indexes (CVIs), specifically focusing on the Davies-Bouldin Index, Dunn Index, Calinski-Harabasz Index, and Silhouette Score. Our research aimed to evaluate how humans perceive these CVIs on 2D scatterplots and to explore the potential of Large Multimodal Models, particularly GPT-4o, in emulating this perception. The obtained results show clear trends in the human perception of these CVIs. Conversely, GPT-4o showed a different perception than humans, making it ineffective to emulate humans in crowdsourcing experiments. By presenting findings, discussing limitations and emerged concerns, this paper aimed to propose a foundation for future research.

All the materials used in this paper have been made publicly accessible for replicability and future research at the following URL <https://github.com/blasilli/ClusteringMetricsPerception>.

REFERENCES

- [1] M. M. Abbas, M. Aupetit, M. Sedlmair, and H. Bensmail. Clustme: A visual quality measure for ranking monochrome scatterplots based on

- cluster patterns. *Computer Graphics Forum*, 38(3):225–236, 2019. doi: [10.1111/cgf.13684](https://doi.org/10.1111/cgf.13684) 1
- [2] G. Albuquerque, M. Eisemann, and M. Magnor. Perception-based visual quality measures. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 13–20, Oct. 2011. doi: [10.1109/VAST.2011.6102437](https://doi.org/10.1109/VAST.2011.6102437) 1
- [3] G. Andrienko and N. Andrienko. Visualization support to interactive cluster analysis. In A. Bifet, M. May, B. Zadrozny, R. Gavalda, D. Pedreschi, F. Bonchi, J. Cardoso, and M. Spiliopoulou, eds., *Machine Learning and Knowledge Discovery in Databases*, pp. 337–340. Springer International Publishing, Cham, 2015. 1
- [4] M. Angelini, G. Blasilli, S. Lenti, A. Palleschi, and G. Santucci. Effectiveness error: Measuring and improving radviz visual effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4770–4786, 2022. doi: [10.1109/TVCG.2021.3104879](https://doi.org/10.1109/TVCG.2021.3104879) 1
- [5] M. Angelini, G. Blasilli, S. Lenti, and G. Santucci. STEIN: Speeding up Evaluation Activities With a Seamless Testing Environment INtegrator. In J. Johansson, F. Sadlo, and T. Schreck, eds., *EuroVis 2018 - Short Papers*. The Eurographics Association, 2018. doi: [10.2312/eurovisshort.201810835](https://doi.org/10.2312/eurovisshort.201810835)
- [6] M. Aupetit and M. Sedlmair. SepMe: 2002 New visual separation measures. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 1–8, Apr. 2016. doi: [10.1109/PACIFICVIS.2016.7465244](https://doi.org/10.1109/PACIFICVIS.2016.7465244) 1
- [7] M. Aupetit, M. Sedlmair, M. M. Abbas, A. Baggage, and H. Bensmail. Toward Perception-Based Evaluation of Clustering Techniques for Visual Analytics. In *2019 IEEE Visualization Conference (VIS)*, pp. 141–145. IEEE, Vancouver, BC, Canada, Oct. 2019. doi: [10.1109/VISUAL.2019.8933620](https://doi.org/10.1109/VISUAL.2019.8933620) 2
- [8] F. B. Baker and L. J. Hubert. Measuring the Power of Hierarchical Cluster Analysis. *Journal of the American Statistical Association*, 70(349):31–38, Mar. 1975. doi: [10.1080/01621459.1975.10480256](https://doi.org/10.1080/01621459.1975.10480256) 2
- [9] J. Bezdek and N. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(3):301–315, 1998. doi: [10.1109/3477.678624](https://doi.org/10.1109/3477.678624) 2
- [10] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. doi: [10.1080/03610927408827101](https://doi.org/10.1080/03610927408827101) 2
- [11] M. Cavallo and c. Demiralp. A visual interaction framework for dimensionality reduction based data exploration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 13 pages, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2018. doi: [10.1145/3173574.3174209](https://doi.org/10.1145/3173574.3174209) 1
- [12] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979. doi: [10.1109/TPAMI.1979.4766909](https://doi.org/10.1109/TPAMI.1979.4766909) 2
- [13] P. Denny, S. Sarsa, A. Hellas, and J. Leinonen. Robosourcing educational resources – leveraging large language models for learnersourcing, 2022. 1, 3
- [14] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973. doi: [10.1080/01969727308546046](https://doi.org/10.1080/01969727308546046) 2
- [15] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea. Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173, 2021. doi: [10.1109/TVCG.2019.2944182](https://doi.org/10.1109/TVCG.2019.2944182) 1, 2
- [16] R. Etemadpour, R. C. da Motta, J. G. de Souza Paiva, R. Minghim, M. C. F. de Oliveira, and L. Linsen. Role of human perception in cluster-based visual analysis of multidimensional data projections. In *2014 International Conference on Information Visualization Theory and Applications (IVAPP)*, pp. 276–283, 2014. 2
- [17] R. Etemadpour, R. Motta, J. Paiva, R. Minghim, M. de Oliveira, and L. Linsen. Perception-based evaluation of projection methods for multi-dimensional data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 21(01):81–94, jan 2015. doi: [10.1109/TVCG.2014.2330617](https://doi.org/10.1109/TVCG.2014.2330617) 1, 2
- [18] P. Härmäläinen, M. Tavast, and A. Kunnari. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, article no. 433, 19 pages. Association for Computing Machinery, New York, NY, USA, 2023. doi: [10.1145/3544548.3580688](https://doi.org/10.1145/3544548.3580688) 1, 3
- [19] S. Hartwig, C. van Onzenoott, D. Engel, P. Hermosilla, and T. Ropinski. ClusterNet: A Perception-Based Clustering Model for Scattered Data, Mar. 2024. doi: [10.48550/arXiv.2304.14185](https://doi.org/10.48550/arXiv.2304.14185) 2
- [20] X. He, Z. Lin, Y. Gong, A.-L. Jin, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, and W. Chen. Annollm: Making large language models to be better crowdsourced annotators, 2024. 1, 3
- [21] H. Jeon, G. J. Quadri, H. Lee, P. Rosen, D. A. Szafrir, and J. Seo. CLAMS: A Cluster Ambiguity Measure for Estimating Perceptual Variability in Visual Clustering. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):770–780, Jan. 2024. doi: [10.1109/TVCG.2023.3327201](https://doi.org/10.1109/TVCG.2023.3327201) 2
- [22] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023. 8
- [23] A. José-García and W. Gómez-Flores. A survey of cluster validity indices for automatic data clustering using differential evolution. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '21, 9 pages, p. 314–322. Association for Computing Machinery, New York, NY, USA, 2021. doi: [10.1145/3449639.3459341](https://doi.org/10.1145/3449639.3459341) 1, 2, 3
- [24] M. Kim and R. Ramakrishna. New indices for cluster validity assessment. *Pattern Recognition Letters*, 26(15):2353–2363, 2005. doi: [10.1016/j.patrec.2005.04.007](https://doi.org/10.1016/j.patrec.2005.04.007) 2
- [25] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, R. Capobianco, E. Bertini, R. Giot, and M. Angelini. State of the art of visual analytics for explainable deep learning. *Computer Graphics Forum*, 42(1):319–355, 2023. doi: [10.1111/cgf.14733](https://doi.org/10.1111/cgf.14733) 1
- [26] J. M. Lewis, M. Ackerman, and V. R. de Sa. Human Cluster Evaluation and Formal Quality Measures: A Comparative Study. 2012. 2
- [27] OpenAI. Gpt-4o model. <https://openai.com/index/hello-gpt-4o/>, 2024. Accessed: 2024-06-28. 1, 4
- [28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Pasos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 2
- [29] L. Podo, M. Ishmal, and M. Angelini. Toward a Structured Theoretical Framework for the Evaluation of Generative AI-based Visualizations. In M. El-Assady and H.-J. Schulz, eds., *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2024. doi: [10.2312/eurova.20241118](https://doi.org/10.2312/eurova.20241118) 1, 3
- [30] G. J. Quadri, J. A. Nieves, B. M. Wiernik, and P. Rosen. Automatic Scatterplot Design Optimization for Clustering Identification. *IEEE Transactions on Visualization and Computer Graphics*, 29(10):4312–4327, Oct. 2023. doi: [10.1109/TVCG.2022.3189883](https://doi.org/10.1109/TVCG.2022.3189883) 2, 8
- [31] G. J. Quadri and P. Rosen. Modeling the Influence of Visual Density on Cluster Perception in Scatterplots Using Topology. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1829–1839, Feb. 2021. doi: [10.1109/TVCG.2020.3030365](https://doi.org/10.1109/TVCG.2020.3030365) 2
- [32] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: [10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7) 2
- [33] J. M. Santos and M. Embrechts. A family of two-dimensional benchmark data sets and its application to comparing different cluster validation indices. In J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, J. A. Olvera-Lopez, J. Salas-Rodríguez, and C. Y. Suen, eds., *Pattern Recognition*, pp. 41–50. Springer International Publishing, Cham, 2014. 3
- [34] M. Sedlmair and M. Aupetit. Data-driven Evaluation of Visual Quality Measures. *Computer Graphics Forum*, 34(3):201–210, 2015. doi: [10.1111/cgf.12632](https://doi.org/10.1111/cgf.12632) 1
- [35] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2634–2643, 2013. doi: [10.1109/TVCG.2013.153](https://doi.org/10.1109/TVCG.2013.153) 4
- [36] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Computer Graphics Forum*, 31(3pt4):1335–1344, 2012. doi: [10.1111/j.1467-8659.2012.03125.x](https://doi.org/10.1111/j.1467-8659.2012.03125.x) 1, 2
- [37] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009. doi: [10.1111/j.1467-8659.2009.01467.x](https://doi.org/10.1111/j.1467-8659.2009.01467.x) 1
- [38] S. Solimun and A. A. R. Fernandes. Cluster analysis on various cluster validity indexes with average linkage method and euclidean distance (study on compliant paying behavior of bank x customers in indonesia 2021). *Mathematics and Statistics*, 10(4):747–753, 2022. 2
- [39] M. Solomon, B. Genossar, and A. Gal. Copychats: Question sequencing

- with artificial agents. In *Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics*, HILDA 24, 7 pages, p. 1–7. Association for Computing Machinery, New York, NY, USA, 2024. doi: [10.1145/3665939.3665963](https://doi.org/10.1145/3665939.3665963) 1, 3
- [40] A. Strehl. Relationship-based clustering and cluster ensembles for high-dimensional data mining. 2002. 2
- [41] T. A. Sørensen and S. Kyllingsbæk. Short-term storage capacity for visual objects depends on expertise. *Acta Psychologica*, 140(2):158–163, 2012. doi: [10.1016/j.actpsy.2012.04.004](https://doi.org/10.1016/j.actpsy.2012.04.004) 4
- [42] A. Tatu, P. Bak, E. Bertini, D. Keim, and J. Schneidewind. Visual quality metrics and human perception: an initial study on 2d projections of large multidimensional data. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI ’10, 8 pages, p. 49–56. Association for Computing Machinery, New York, NY, USA, 2010. doi: [10.1145/1842993.1843002](https://doi.org/10.1145/1842993.1843002) 1
- [43] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023. 8
- [44] J. W. Tukey and M. B. Wilk. Data analysis and statistics: an expository overview. In *Proceedings of the November 7-10, 1966, Fall Joint Computer Conference*, AFIPS ’66 (Fall), 15 pages, p. 695–709. Association for Computing Machinery, New York, NY, USA, 1966. doi: [10.1145/1464291.1464366](https://doi.org/10.1145/1464291.1464366) 1
- [45] J. Xia, W. Lin, G. Jiang, Y. Wang, W. Chen, and T. Schreck. Visual clustering factors in scatterplots. *IEEE Computer Graphics and Applications*, 41(5):79–89, 2021. doi: [10.1109/MCG.2021.3098804](https://doi.org/10.1109/MCG.2021.3098804) 1
- [46] J. Xia, W. Lin, G. Jiang, Y. Wang, W. Chen, and T. Schreck. Visual Clustering Factors in Scatterplots. *IEEE Computer Graphics and Applications*, 41(5):79–89, Sept. 2021. doi: [10.1109/MCG.2021.3098804](https://doi.org/10.1109/MCG.2021.3098804) 2
- [47] J. Xia, Y. Zhang, J. Song, Y. Chen, Y. Wang, and S. Liu. Revisiting Dimensionality Reduction Techniques for Visual Cluster Analysis: An Empirical Study. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):529–539, Jan. 2022. doi: [10.1109/TVCG.2021.3114694](https://doi.org/10.1109/TVCG.2021.3114694) 2
- [48] F. Yanez and C. Nobre. User-Adaptive Visualizations: An Exploration with GPT-4. In D. Archambault, I. Nabney, and J. Peltonen, eds., *Machine Learning Methods in Visualisation for Big Data*. The Eurographics Association, 2024. doi: [10.2312/mlvis.20241126](https://doi.org/10.2312/mlvis.20241126) 1, 3
- [49] M. J. Zellinger and P. Bühlmann. repliclust: Synthetic data for cluster analysis, 2023. 3