

# Understanding Grey Sheep Users in Recommender Systems: A Systematic and Critical Review of Detection, Modeling, and Evaluation

Yong Zheng<sup>a,\*</sup>

<sup>a</sup>Center for Decision Making and Optimization, School of Computing, Illinois Institute of Technology, Chicago, 60616, Illinois, USA

## ARTICLE INFO

### Keywords:

Grey sheep users  
recommender systems  
collaborative filtering  
outlier detection  
unusual preferences

## ABSTRACT

Grey sheep users (GSUs), characterized by weak or unstable alignment with the broader user population, remain a challenge in collaborative filtering-based recommender systems. Despite extensive efforts devoted to their detection and treatment, a clear and consistent understanding of GSUs is still lacking. Existing studies typically adopt model-dependent definitions and often assume that GSUs necessarily reduce recommendation quality. This paper presents a systematic and critical review of GSUs, covering their definition, detection strategies, modeling paradigms, and evaluation practices. We revisit the classical similarity-based interpretation and show that the effect of GSUs depends on model design rather than being inherently harmful. Based on this observation, we introduce a refined conceptual perspective that describes GSUs as users with consistently misaligned preference patterns, and further propose a dynamic view that distinguishes between short-term irregular behavior and stable preference differences. We then provide a comprehensive taxonomy of existing approaches, including detection-dependent methods and detection-free strategies, and analyze their underlying assumptions and limitations. We further show that the form in which a detection result is expressed constrains which handling strategies are available, and that this correspondence is rarely exploited in practice. A key finding of this study is that current research is limited by weak evaluation practices. In particular, the absence of ground truth, inconsistent definitions, and dependency on aggregate performance metrics lead to circular reasoning and unreliable conclusions. Finally, we outline future research directions toward more robust modeling frameworks and user-centered evaluation protocols, with the aim of establishing a clearer conceptual foundation and guiding the development of more reliable and inclusive recommender systems.

## 1. Introduction

Recommender systems have become an essential component in a wide range of online platforms, where they support users in navigating large volumes of information by providing personalized suggestions. Among various recommendation paradigms, collaborative filtering (CF) remains one of the most widely adopted approaches due to its simplicity and effectiveness [27, 52, 20, 60]. CF methods typically rely on the assumption that users with similar historical preferences will exhibit similar future consumption behaviors, and therefore leverage user–user or item–item similarities to generate recommendations [48, 43]. Despite their success, CF continues to face several well-known challenges, including data sparsity, the cold-start problem, and the presence of grey sheep users (GSUs) [15, 65]. Current literature includes several reviews addressing rating sparsity [21, 8, 49, 63] and cold-start problems [28, 38, 58, 29] in recommender systems, whereas dedicated reviews focusing specifically on grey sheep user problems remain absent.

The concept of “grey sheep users” (also spelled “gray sheep users” in the community) was introduced to describe users whose preferences are unusual or weakly correlated with those of the broader user population [9]. GSUs actively engage with the system but do not consistently align with any major user groups [15]. In contrast, “white sheep” users demonstrate strong correlations with many others, and “black sheep” users exhibit minimal interaction history and unusual tastes. Consequently, neighborhood-based CF often fails to identify reliable neighbors for these users. Prior

© 2026. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <https://creativecommons.org/licenses/by-nc-nd/4.0/>. This is the accepted manuscript of an article published in *Computer Science Review*; it is not the Version of Record. The Version of Record is available at <https://doi.org/10.1016/j.cosrev.2026.101078>

\*Corresponding author

 [yzheng66@illinoistech.edu](mailto:yzheng66@illinoistech.edu) (Y. Zheng)  
ORCID(s): 0000-0003-4990-4580 (Y. Zheng)

studies commonly argue that the presence of GSUs leads to reduced recommendation accuracy for these individuals and may introduce noise that degrades the overall system performance [65, 2]. Although early discussions of GSUs are often framed around neighborhood-based CF, the problem is not limited to this setting. Latent-factor based models, such as matrix factorization (MF) [26], also exhibit degraded performance for GSUs, because global optimization objectives tend to capture dominant preference structures and may underrepresent minority or unusual users. Gras et al. show that standard MF still provides inaccurate recommendations for GSUs [18]. Their results indicate that MF can improve GSU recommendation accuracy when training is adapted to these users, but standard MF alone does not fully resolve the problem. More recent work further shows that latent-factor vectors provide a more effective representation for detecting GSUs, suggesting that latent-factor models can help characterize GSUs while still requiring explicit mechanisms for identification or adaptation [62]. Therefore, the GSU issue is a general challenge in the family of collaborative filtering methods, including both neighborhood-based approaches and latent-factor based models.

The most common approach to address this involves detecting GSUs within the user population, excluding them from CF, and applying alternative recommendation strategies, such as content-based methods [32, 1], specifically for these users. Despite more than two decades of research on the detection and handling of GSUs, spanning clustering techniques [15, 45], outlier detection [17, 62], classifications [2, 24], a unified and rigorous understanding of this user group remains limited. Existing work reveals several fundamental issues. First, definitions of GSUs are typically model-dependent and fragmented, relying on similarity thresholds, clustering structures, or anomaly detection outcomes, without a consistent conceptual basis. Second, it is often implicitly assumed that GSUs exhibit consistently low similarity with the rest of the user population, resulting in the absence of highly correlated neighbors. Under this condition, GSUs are expected to degrade recommendation quality. However, this assumption does not fully consider the internal mechanisms of modern CF models. For example, in the user-based  $K$ -nearest neighbor CF (UBCF) method [43], the influence of low-similarity users can be reduced by applying similarity thresholds during neighbor selection [18]. This observation raises a critical question regarding whether GSUs are inherently problematic or whether their impact depends on specific modeling choices. In addition, current research is further limited by the absence of ground truth labels and standardized evaluation protocols for GSUs. Recently, Boualaoui et al. [6] also recognized the limitations of existing evaluation practices and proposed a dedicated evaluation protocol for grey sheep user detection, further highlighting the growing need for more rigorous and standardized assessment frameworks. Existing studies often rely on model-dependent definitions and indirect performance comparisons, which do not provide definitive evidence that the identified users are genuinely grey sheep users.

To address these limitations, this paper presents a systematic review and critical analysis of grey sheep users, with the objective of re-examining their definition, modeling, and evaluation. Relevant studies were collected by searching major academic platforms, including Google Scholar, ACM Digital Library, IEEE Xplore, and ScienceDirect, using the keywords “grey sheep”, “gray sheep” and “recommender system”, covering publications up to early 2026. A total of 35 papers were identified, among which 5 focus on conceptual discussions without experimental validation, while the remaining 30 propose concrete methodologies with experimental evaluation.

To the best of our knowledge, our work presents the first comprehensive review of grey sheep users in recommender systems. The contributions of this work are threefold:

- **A refined conceptual and dynamic perspective:** This study revisits the conventional static definition of GSUs and argues that grey sheep behavior is contingent on model design. A dynamic perspective is further introduced to capture the evolution of user behavior over time and its interaction with cold-start and sparsity effects.
- **A comprehensive taxonomy of existing approaches:** Existing methodologies are systematically categorized into detection-dependent strategies (e.g., clustering-based approaches, outlier detection, classification methods) and detection-free strategies (e.g., neighborhood augmentation, representation learning, semantic decoupling), with an analysis of their underlying assumptions and limitations, and of how the form of a detection output constrains the handling strategies available.
- **A critical analysis of evaluation challenges:** This work examines fundamental issues in current evaluation practices, including the absence of ground truth, inconsistent and model-dependent definitions, and the dependency on aggregate prediction metrics. It highlights how these factors lead to circular reasoning and unreliable conclusions, and discusses directions for developing more rigorous, user-centric, and standardized evaluation frameworks.

The remainder of this paper is organized as follows. Section 2 revisits the definition of grey sheep users and introduces a refined conceptual framework. Section 3 reviews strategies for handling grey sheep users in recommendation tasks. Section 4 presents current evaluation protocols and provides a critical analysis of evaluation challenges. Section 5 outlines future research directions. Finally, Section 6 concludes the paper.

## 2. Understanding Grey Sheep Users

A clear and consistent understanding of GSUs is essential for both theoretical analysis and practical modeling in recommender systems. This section revisits the classical definition of GSUs and examines its underlying assumptions, particularly the dependence on user-user similarities. It further analyzes the relationship between GSUs and other commonly discussed user types, including black sheep and cold-start users, and highlights the limitations of existing categorizations. Finally, this section introduces a dynamic perspective and a refined conceptual definition to provide a more comprehensive foundation for subsequent analysis.

### 2.1. Classical Definition of Grey Sheep Users

The concept of GSUs was originally introduced to describe a subset of users whose preferences do not consistently align with those of the broader user population [9]. Within the framework of CF, user preferences are typically inferred from similarity relationships derived from historical interactions, based on the assumption that users with similar past behaviors will exhibit similar future interests [43]. Under this assumption, prior studies classify users into three groups: white sheep, grey sheep, and black sheep [35]. White sheep users exhibit strong and stable correlations with many others, which enables reliable neighbor identification, whereas black sheep users are characterized by extremely limited interaction data and are commonly associated with cold-start scenarios. In contrast, grey sheep users actively interact with the system but neither consistently agree nor disagree with any particular user groups, resulting in weak or low similarity relationships with other users [15, 65]. Consequently, neighborhood-based CF often fails to identify reliable neighbors for these users, which may reduce recommendation accuracy. Latent-factor based models, such as matrix factorization (MF) [26], also exhibit degraded performance for GSUs, as global optimization objectives tend to capture dominant preference patterns and underrepresent users with unusual tastes. Empirical studies show that while MF can be adapted to improve GСУ recommendations, standard MF alone does not fully resolve the issue, indicating that the GСУ problem persists across both neighborhood-based and latent-factor based CF methods [18, 62].

Existing literature further characterizes GSUs from two complementary perspectives. The first perspective emphasizes behavioral inconsistency, where grey sheep users exhibit opinions that do not systematically align with those of any identifiable user group and are often located near the boundaries of user clusters [15]. The second perspective focuses on similarity structure, where GSUs tend to have low correlations with many users and only a few highly correlated neighbors [65]. These characteristics are closely associated with the distribution of user-user similarities, where grey sheep users typically exhibit predominantly low similarity values and highly skewed distributions compared to typical users. These observations indicate that the classical definition of GSUs is primarily descriptive and remains dependent on similarity structures, motivating further examination of their practical impact and underlying assumptions.

The formation of grey sheep behavior can be related to multiple factors. First, some users may have genuinely unique or niche preferences that are not shared by the majority, resulting in inherently low similarity with others. Second, inconsistency or noise in user behavior may lead to unstable similarity relationships across different items. Third, data-related factors such as sparsity or limited interaction history may cause users to appear as grey sheep, even when their underlying preferences are not fundamentally unusual. These observations suggest that the classical definition of GSUs is influenced not only by intrinsic user characteristics but also by data properties and modeling assumptions, which may lead to inconsistencies across different systems and studies.

### 2.2. Are Grey Sheep Users Inherently Problematic?

A widely held assumption in the recommender systems literature is that GSUs are inherently problematic. Prior studies commonly argue that users with unusual preferences and weak correlations with the majority not only receive inferior personalized recommendations but also introduce noise that degrades overall system accuracy [9, 15, 65].

A closer examination of CF mechanisms provides a more nuanced perspective. Consider the standard UBCF, which is frequently adopted to analyze the grey sheep problem [17, 62]. In this setting, rating prediction is derived from pairwise similarity computations, such as Pearson correlation or cosine similarity, followed by aggregation over the

top- $K$  most similar users. In practice, neighbor selection can be further constrained by similarity thresholds to exclude weakly related or negatively correlated users [65]. Under such conditions, if grey sheep users consistently exhibit low or negative similarity with others, they are unlikely to be selected during neighbor formation. Consequently, these users behave as isolated nodes rather than sources of interference, and their impact on the broader community is structurally limited.

Beyond their potential impact on other users, GSUs face inherent difficulties in obtaining accurate recommendations for themselves. User-based CF relies on the identification of highly similar neighbors, and the presence of unusual or inconsistent preferences makes it challenging to construct reliable neighborhoods for these users. Even when neighbors are identified, their relevance may be limited, which leads to unstable or low-quality predictions. As a result, the performance of UBCF for GSUs is often substantially lower than that for typical users, consistent with empirical observations reported in prior work [65].

The interpretations above, however, rely on the assumption that grey sheep users exhibit uniformly low similarity with others, which may not hold in practice. GSUs may display distinct patterns in terms of the local similarity (i.e., similarities based on a subset or a cluster of the items [44]), where they align with other users on a subset of items, particularly popular ones, but remaining globally inconsistent. In such cases, they may still be selected as neighbors in specific contexts, which can introduce instability into the prediction process. Therefore, the extent to which grey sheep users influence others depends on the interaction between similarity distributions and neighbor selection mechanisms, rather than solely on their overall similarity levels. Moreover, prior work on the magic barrier suggests that a portion of prediction error cannot be eliminated due to noise and inconsistency in user ratings, implying that some behaviors attributed to GSUs may reflect inherent uncertainty rather than a distinct user category [4]. Consistent with this view, Gras et al. [17] report a correlation of only  $-0.68$  between their likelihood-based greyness measure and per-user prediction error, which indicates that a substantial portion of the error observed for these users is not accounted for by their position in the similarity structure.

The discussion above concerns neighborhood-based CF, but the question of whether GSUs are inherently problematic can also be examined empirically across model families. Gras et al. [18] compare several matrix factorization variants that differ only in how the ratings of GSUs are used during training. A model trained exclusively on GSN ratings raises the median RMSE by approximately 8% relative to a model trained on the full population, and a model built for an individual GSN still requires a subset of normal users in order to estimate the latent factors. Retaining GSUs in the shared user-item matrix and increasing the weight assigned to their prediction errors performs better than either alternative. These results indicate that the difficulty faced by GSUs in matrix factorization arises from the estimation of global latent factors, which requires interaction data beyond that available within the group itself, and they align with the neighborhood-based finding that inconsistent users cannot be served using information drawn only from users similar to them [4]. The same conclusion therefore holds in two model families that rely on different mechanisms, which suggests that it reflects a property of how preferences are estimated rather than a property of the users.

More recent neural models, including neural collaborative filtering [20] and graph-based architectures [55], share the same optimization structure as matrix factorization in one important respect, since they minimize a global objective in which a minority group contributes few training signals. Samir et al. [45] report that NCF [20], LightGCN [19] and an autoencoder variant all perform worse on an identified GSN subset than on the population as a whole, and attribute this to the absence of any mechanism that treats high-uncertainty users differently during optimization. Direct evidence remains limited, however, because studies that adopt neural architectures typically evaluate at the aggregate level and do not report results for this subgroup separately (Section 3.2.2). It therefore remains unclear whether greater model expressiveness reduces the problem or leaves it undetected in aggregate metrics, and this uncertainty follows directly from the evaluation practices discussed in Section 4.

This conclusion is nevertheless conditional on the information available to the model. Sánchez-Moreno et al. [46] combine rating-based similarity with the structural position of users in a social network, and report the opposite pattern: when their models are trained on the full population and evaluated on the subset of users with unusual tastes, performance is worse than when training is restricted to that subset. A plausible explanation is that their similarity function draws on an additional source of information that is independent of the rating matrix, so that neighbors can still be identified within a small group of users whose ratings alone would not support reliable similarity estimates. Under these conditions the mainstream population contributes less, and may introduce neighbors whose apparent similarity derives from popular items rather than from shared preferences. We note that the two studies also differ in domain, feedback type and in how unusual users are defined, so the comparison cannot be attributed to model family alone. Taken together, however, these results support the view that isolation is neither uniformly harmful nor uniformly

beneficial, and that its effect depends on whether the model in question can estimate preferences from the information contained within the isolated group.

These observations suggest that the grey sheep problem should not be interpreted solely in terms of its impact on the broader user community. Instead, it reflects two distinct challenges: the conditional influence of GSUs on others, which depends on model design and similarity structures, and the inherent difficulty of modeling these users within similarity-based frameworks. This perspective indicates that the grey sheep problem is fundamentally model-dependent. Whether GSUs affect others, and whether isolating them helps or harms, both follow from the way a given model estimates preferences rather than from any intrinsic property of the users themselves. Section 3.1.4 examines how this dependence plays out in the handling strategies adopted in practice.

### 2.3. Grey Sheep vs Cold-Start vs Black Sheep

The recommender systems literature commonly distinguishes among grey sheep, cold-start, and black sheep users to explain variations in recommendation performance. Cold-start users are defined as users with no interaction history (i.e., strict cold-start) or with very limited interactions (i.e., practical cold-start), which hinders reliable estimation of their preferences [28, 38, 58]. Black sheep users are typically described as individuals whose preferences are highly inconsistent with others, often leading to very low similarity with the broader population; however, such low similarity may also arise from sparse or unreliable interaction data, which complicates its interpretation in practice. Namely, black sheep users are usually those with very low or even zero user-user similarities with others due to noise in limited interactions [65, 64]. GSUs are generally positioned between these two extremes, exhibiting partial or unstable similarity relationships with other users [15, 65].

Although these distinctions are widely adopted, they are primarily derived from user-user similarity patterns, which depend on model assumptions and data representations. As discussed in Section 2.2, similarity-based measures may not fully capture the structure of user preferences, and their interpretation can vary across algorithms. Consequently, the boundaries among grey sheep, cold-start, and black sheep users are often ambiguous in practice. For instance, users with sparse interaction histories may display low or unstable user-user similarities due to insufficient user-item interactions, which can lead to their misclassification as GSUs. Likewise, the distinction between grey sheep and black sheep users is not always clear, as both groups may exhibit weak correlations with others, differing mainly in magnitude rather than in nature.

These observations suggest that conventional categorizations should be interpreted with caution. Rather than representing distinct and fixed user types, grey sheep, cold-start, and black sheep behaviors may correspond to different manifestations of underlying factors, including data sparsity, preference diversity, and behavioral inconsistency. From this perspective, user categories are better viewed as context-dependent constructs that arise from the interaction between user behavior and the modeling framework, rather than the static definitions above. To the best of our knowledge, existing research on GСУ detection has largely overlooked this temporal perspective on user classification.

### 2.4. A Dynamic View of Grey Sheep Users

The limitations of static user categorizations motivate a dynamic perspective on GSUs. Conventional definitions implicitly treat user types, including grey sheep, cold-start, and black sheep, as fixed attributes. In practice, user behavior evolves over time as additional interaction data is accumulated and preferences change [25, 40, 47]. As a result, users initially characterized by sparse or unstable interactions may later exhibit more consistent patterns, whereas previously stable users may develop more diverse or less predictable preferences.

From this perspective, grey sheep behavior should not be regarded as an inherent and permanent attribute, but rather as a transient state arising from the interaction among data availability, preference evolution, and model representation. As illustrated in Figure 1, users typically evolve from a cold-start phase with no interaction history to stages with increasing amounts of data, during which their observed preferences may appear unstable or misleading. In early stages with limited interactions, users may be tentatively classified as grey sheep or even black sheep due to apparent deviation from others; however, such classifications are highly uncertain and prone to misjudgment. As more interactions are observed, some users gradually align with common preference patterns, while others continue to exhibit persistent deviation, eventually forming more stable behavioral identities.

This dynamic view further indicates that the identification of GSUs depends on the temporal context in which user data is analyzed. Similarity patterns, which are commonly employed for classification, may vary across time as user interests evolve or as new items are introduced. Consequently, a user may transit across different behavioral states, such

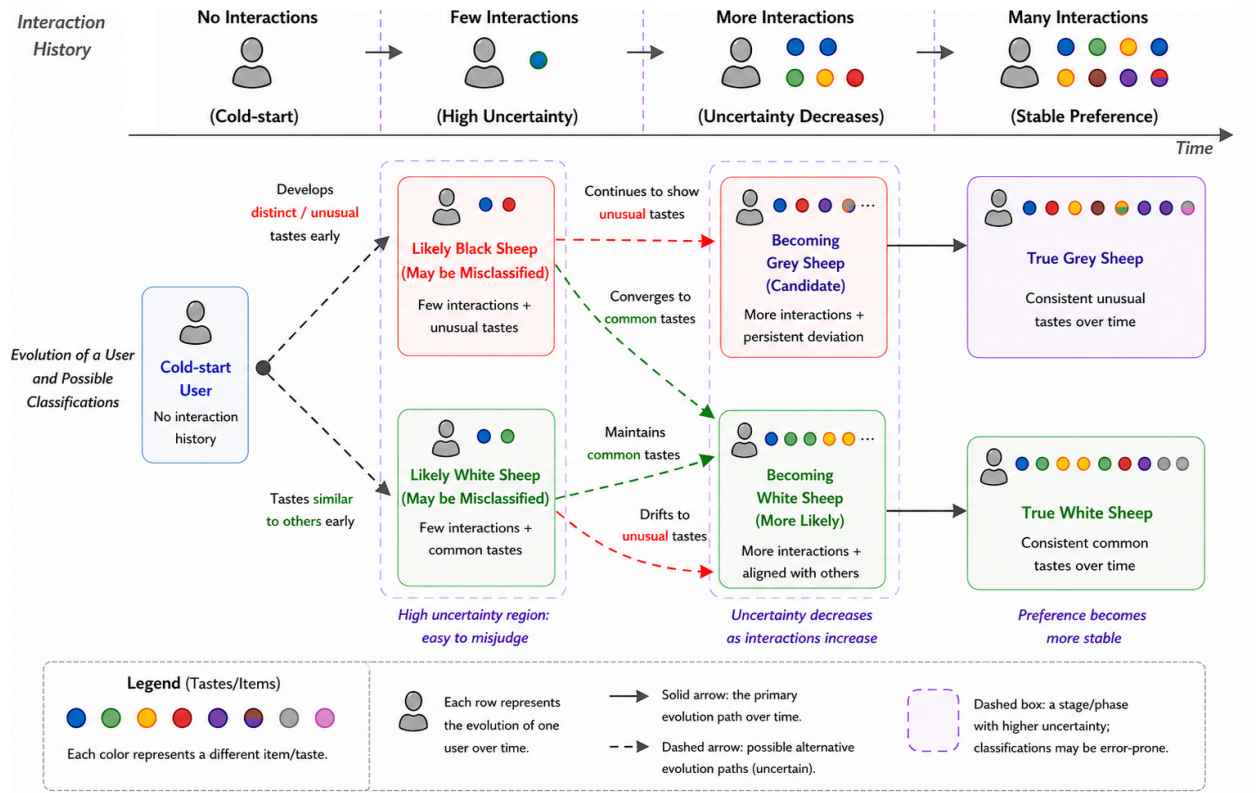


Figure 1: An illustrative example of dynamic user evolution in recommender systems

as from cold-start to an uncertain classification region, and subsequently toward either stable alignment with others or consistent deviation from the majority.

Adopting a dynamic perspective provides a more flexible and realistic foundation for understanding grey sheep behavior. It suggests that effective modeling should explicitly account for temporal evolution and uncertainty, rather than relying on static classifications. This perspective not only refines the interpretation of GSUs but also informs the design of recommendation models that can adapt to evolving user characteristics.

## 2.5. A Refined Conceptual Definition

Building on the limitations of existing definitions and the dynamic perspective discussed above, we propose a refined conceptual definition of GSUs. Instead of defining GSUs solely by low or unstable similarity to others, we characterize them as *users whose preference patterns exhibit persistent structural misalignment with the broader population, even when sufficient interaction data is available*. This formulation shifts the emphasis from absolute similarity values to the consistency and temporal stability of deviation.

This refinement resolves several ambiguities in prior work. First, it differentiates GSUs from cold-start users by requiring sufficient interaction history, which reduces misclassification caused by data sparsity. Second, it separates GSUs from transient behaviors observed in early interaction stages, where apparent inconsistency may arise from limited evidence. Third, it sharpens the distinction between GSUs and black sheep users by focusing on persistence and structural characteristics, rather than on extreme or isolated similarity values.

Under this definition, GSUs are not identified from a single snapshot of similarity, but from temporal behavior and the stability of deviation from common preference patterns. A user is regarded as a GSU only when such deviation remains consistent across interactions and over time, indicating genuine preference divergence rather than noise or estimation uncertainty. This interpretation is consistent with the dynamic view illustrated in Figure 1, where only users who consistently exhibit unusual tastes in later stages are considered true grey sheep.

The proposed definition has direct implications for modeling and evaluation. It indicates that GSU detection should incorporate temporal signals, robustness to sparsity, and measures of preference stability, rather than relying exclusively on static similarity thresholds or clustering outcomes. By offering a more precise and principled characterization, this definition provides a foundation for more consistent analysis and comparison across studies.

### 3. Handling Grey Sheep Users: Paradigms

Traditionally, the handling of GSUs in collaborative filtering has followed a two-stage pipeline: systems first explicitly identify these users and subsequently apply customized interventions to mitigate their negative effects. Under this conventional view, GSUs are often perceived as anomalous entities or system burdens that require strict isolation, alternative algorithmic treatment, or even complete removal. However, recent advancements in recommender systems reflect a profound paradigm shift in how their preferences are managed. Rather than relying on explicit detection and isolation, recent approaches increasingly emphasize integrating such users directly into the recommendation process. By leveraging techniques such as global data augmentation and non-linear deep representation learning, modern end-to-end systems can implicitly accommodate unusual preferences without the need for isolation. Based on this evolution, this section categorizes the existing handling strategies into the following two fundamental paradigms: *Detection-Dependent Strategies*, which rely on prior discrete identification of GSUs, and *Detection-Free Strategies*, which integrate and address GSU issues within unified architectures without any form of explicit or implicit identification.

#### 3.1. Detection-Dependent Strategies

The defining characteristic of detection-dependent strategies is their dependence on a structured, typically two-stage pipeline. In the first stage, GSUs are explicitly identified and labeled using dedicated detection mechanisms. Subsequently, different handling strategies, such as isolating GSUs to use switched recommendation models [15, 3, 45] or adapting models to different types of users without isolation [18, 2], are applied to generate recommendations.

Table 1 summarizes these approaches, with the last three columns defined as follows. We use *greyness* to refer to the degree to which a user deviates from the broader population. The *greyness signal* refers to the quantity from which the degree of deviation is computed, such as the distance between a user and a cluster centroid, the likelihood of a rating profile under a population-level distribution, or a set of features describing a user's similarity relationships. The *detection output* refers to the form in which the result of detection is expressed, which may be a hard partition, a continuous score, a degree of membership, or a decision function defined over user features. The *handling strategy* refers to the treatment subsequently applied to the identified users. The first two dimensions are examined in the following subsections, and their relationship with the handling strategies is discussed in Section 3.1.4.

##### 3.1.1. Clustering-Based Detection

The earliest and most intuitive approach to managing GSUs is to isolate them through clustering algorithms, followed by a complete switch of the recommendation models. In this paradigm, GSUs are identified based on their distance from mainstream clusters. Specifically, these users either cannot be reliably assigned to any cluster or are located near cluster boundaries, and may therefore be separated as a distinct group. Ghazanfar et al. [14, 15] introduced this idea by adapting the *K-Means++* algorithm, where users with large distances to their nearest cluster centroid are isolated and assigned to a separate cluster. Malik et al. proposed an enhanced clustering-based approach, termed *PowKMeans++*, which extends earlier power-user weighting schemes to improve centroid initialization and clustering quality [33]. Based on the resulting clusters, GSUs are explicitly isolated and subsequently served using a hybrid recommendation framework. Ghorbani et al. also explored clustering-based separation strategies, where users are partitioned into groups based on similarity structures, and those not fitting well into any cluster are treated as grey sheep users [16], but they did not explore handling strategies after GSU detection. Recognizing that preference-based distance alone may be insufficient, Bejaoui and Ayachi [3] extended this framework by incorporating social trust information, such as TidalTrust, into the clustering process, grouping users who are distant from others in both preference and trust spaces into a GSU cluster. More recently, Samir et al. [45] introduced a neutrosophic *k*-means approach, in which each user is associated with degrees of truth, indeterminacy, and falsity with respect to cluster membership. In this framework, GSUs are characterized by higher levels of indeterminacy or unstable cluster assignments, reflecting the uncertainty in their preference patterns.

A fundamental theoretical issue arises regarding the interpretation of clustering-based isolation as outlier detection. Although early work [15] framed the problem in this manner, subsequent studies challenge this view [17]. As noted by

**Table 1**

Detection-dependent approaches: detection mechanisms, output forms, and handling strategies

| Detection Approach                      | Greyness Signal                                                                                   | Detection Output                                           | Handling Strategy                                                                                                                                                                                |
|-----------------------------------------|---------------------------------------------------------------------------------------------------|------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Clustering: centroid distance           | Distance to cluster centroid in preference (or preference–trust) space                            | Hard partition ( $K+1$ cluster)                            | Isolate + switch to content-based or hybrid models [14, 15, 22, 33]; isolate + within-cluster prediction [3]; isolate to improve accuracy for mainstream users [41, 42, 57]; detection only [16] |
| Clustering: soft membership             | Indeterminacy degree derived from user–centroid proximity                                         | Soft membership; highest-indeterminacy cluster             | Isolate + item-based CF [45]                                                                                                                                                                     |
| Clustering: local community             | Centrality, reciprocity, neighbor frequency                                                       | Continuous greyness score + threshold                      | Detection only [7]                                                                                                                                                                               |
| Clustering: cluster weighting           | Agreement with collective signals; cluster power                                                  | Cluster-level label                                        | Detection only [5]                                                                                                                                                                               |
| Outlier detection: distribution-based   | Likelihood that a user's ratings were generated by the population-level rating distribution       | Continuous score + cutoff; decision intrinsic to the score | Detection only [17]; model adaptation without isolation [18]                                                                                                                                     |
| Outlier detection: representation-based | Per-user feature vectors: statistics of the user's own similarity distribution, or latent factors | Anomaly score from an external detector (LOF, ABOD, etc.)  | Detection only; isolation with alternative models assumed [65, 64, 61, 62]; isolate + content-based switching [53]                                                                               |
| Classification: constructed labels      | Pseudo-labels from clustering + IQR, personality traits, or heuristic rules                       | Inductive decision boundary                                | Remove GSUs [51, 24]; unified modeling [2]; detection only [37]                                                                                                                                  |

Gras et al. [17], classical clustering-based outlier detection identifies instances in sparse or marginal clusters. However, when users are grouped into a single GSU cluster, they are, by construction, similar to each other, which contradicts the assumption that GSUs exhibit uncorrelated preferences. This observation indicates that clustering-based methods impose an artificial structure that may not reflect the true nature of grey sheep behavior. Accordingly, these clustering-based approaches can be treated as spatial partitioning methods guided by heuristic criteria rather than formal outlier detection techniques.

Regardless of this theoretical concern, most clustering-based methods continue to locate greyness by measuring the distance between a user and a cluster centroid, and much of the subsequent work refines this measurement rather than replacing it. Comparative analyses of clustering techniques for GSU identification have also been reported [23], although they lack empirical validation. Jindal et al. [22] introduced power-user weighting into centroid initialization, later extended by [33], while Rashidi et al. [41, 42] examined which features of the rating matrix support separation and combined enhanced K-Means variants with meta-heuristic search, an approach further developed by Yu and Rashidi [57]. More recent work questions the centroid assumption itself. Boualaoui et al. [7] argue that a single distance to a centroid, which is often not a real user, ignores how a candidate interacts with its local community, and instead combine centrality, reciprocity and neighbor frequency into a continuous greyness score. The same authors subsequently shifted the decision from the user level to the cluster level, weighting each cluster by how strongly it exhibits grey sheep characteristics and labeling the least influential clusters accordingly [5].

Once GSUs are detected or isolated by these clustering-based approaches, the handling stage diverges considerably across studies, ranging from switching to content-based or demographic models [14, 15, 22, 33, 50], to restricting collaborative filtering to the isolated cluster itself [45, 3], to providing no recommendation at all when isolation serves only to protect the accuracy of mainstream users [41, 42, 57]. These strategies are systematically discussed in Section 3.1.4.

### 3.1.2. Outlier Detection Approaches

Early work on GSUs focused on isolating them from mainstream users and applying alternative recommendation models. Building on this line of work, subsequent research has increasingly focused on improving the detection stage

itself. One popular category of the work formulates GSU identification as an outlier detection problem. Existing approaches differ in the location of the normality model. Some evaluate users directly against a population-level distribution, whereas others first construct a user-specific representation and subsequently use a separate anomaly detection algorithm for identification.

Early distribution-based approaches formulate the identification of GSUs by modeling the global rating patterns rather than extracting user-specific statistical features. Specifically, Gras et al. [17] estimate a global, non-parametric probability density distribution for each item across the entire user population. Instead of representing users as low-dimensional feature vectors, their approach directly evaluates a user's raw rating profile against these global item distributions to compute a single *Likelihood* score (e.g., DILikelihood). This score inherently quantifies the probability that a user's unusual tastes were generated by the mainstream preference model, while also accounting for individual rating bias and rating imprecision through dynamic intervals. In this framework, the likelihood calculation itself serves as the outlier detection mechanism; users with the lowest likelihood scores are directly flagged as GSUs via a simple cutoff threshold, entirely eliminating the dependence on external or complex machine learning-based outlier detection algorithms.

In contrast, subsequent studies place greater emphasis on constructing informative user representations before applying anomaly detection techniques. Zheng et al. [65] proposed to represent each user by the distribution of similarities between that user and all other users. Instead of relying on raw similarity values, this approach extracts statistical features from the similarity distribution, such as the descriptive statistics of the similarity distributions including min, max, mean,  $q_1$ ,  $q_2$ ,  $q_3$ , standard deviation and skewness, to characterize the overall structure of user-user relationships. They observed that mainstream users typically exhibit left-skewed similarity distributions, indicating the presence of several relatively similar neighbors, whereas GSUs tend to have distributions dominated by low similarity values, reflecting weak connections to the rest of the population. Based on this representation, standard outlier detection methods, such as Local Outlier Factor (LOF), can be applied to identify users whose similarity distributions deviate significantly from the dominant pattern. In addition, Zheng et al. further extended this framework by employing histogram intersection [64] to measure the similarity between users' similarity distributions, enabling a more robust comparison of distributional shapes and improving the discrimination between GSUs and mainstream users. Tennakoon et al. [53] adopted the same representation in the context of news recommendation, using the skewness and mean of each user's similarity distribution to pre-select candidate GSUs and mainstream users, followed by the application of standard outlier detection algorithms to the combined sample. Their results indicated that ABOD and KNN performed best among the evaluated detectors, supporting the observation that performance is primarily driven by the representation rather than the detector.

More recent work highlights the limitations of using raw rating data, which is often sparse and noisy. The reliability of similarity calculations depends on the number of co-rated items between two users. When the number of co-rated items is limited, the resulting similarity values and their distributions may not accurately reflect underlying user characteristics. To address this issue, latent-factor-based representations have been introduced, where users are embedded into a dense vector space learned through matrix factorization or related models. Zheng [62] demonstrated that such latent representations yield more stable and discriminative features for identifying GSUs compared to sparse rating-based statistics. This effect was attributed to latent factors capturing underlying preference structures while mitigating sparsity and noise.

Once an appropriate user representation is obtained, the detection of GSUs can be formulated as a standard anomaly detection task. A broad scope of existing outlier detection methods can then be applied to identify users whose representations deviate from the dominant population. These include density-based approaches such as LOF, angle-based methods such as Angle-Based Outlier Detection (ABOD), as well as proximity-based methods that rely on distance or neighborhood structures. In addition, Zheng [61, 62] also explore neural-network-based approaches that learn decision boundaries for distinguishing GSUs in the representation space. Empirical results indicate that, across these diverse detection techniques, the choice of user representation has a more significant impact on performance than the specific anomaly detection algorithm employed.

These studies [65, 64, 61, 62] do not address the handling stage directly and instead assume that established solutions from earlier work, such as the isolate-and-switch strategy [14, 15], can be applied after detection. A notable exception adapts matrix factorization to retain GSUs in the shared user-item matrix rather than isolate them [18], an alternative discussed in Section 3.1.4.

### 3.1.3. *Classification Approaches with Constructed Labels*

A major challenge in applying supervised learning to GSU identification is the absence of ground truth labels, as datasets do not inherently distinguish between different user types. To address this issue, recent studies adopt offline label construction, where heuristic rules or auxiliary information are used to generate labels prior to training classification models for GSU identification and handling.

In the label construction phase, researchers partition the training data using derived criteria. Alabdulrahman et al. [2] generate labels through a combination of unsupervised learning and statistical thresholds. They first apply Expectation-Maximization clustering to group users based on rating and demographic features, and then use the interquartile range (IQR) method within each cluster to identify extreme cases. Users falling outside predefined statistical bounds are labeled as GSUs, while the remaining users are treated as normal. In contrast, Srivastava et al. [51] and Kaur et al. [24] construct labels using psychographic features rather than rating data alone. Based on the Big Five personality model, they derive user similarity matrices from behavioral profiles and assign the GSU label to users exhibiting consistently low similarity or correlation in these personality-driven representations.

After pseudo-labels are constructed, the resulting dataset is used to train classification models. User features paired with their corresponding labels are partitioned using standard protocols, such as hold-out splits or cross-validation, to train supervised models including neural networks, boosted decision trees, and decision forests. These models approximate the decision boundaries defined by the offline labeling process, enabling efficient identification of GSUs for new users based solely on their features during online deployment, without requiring recomputation of global clustering or similarity structures.

Despite relying on the same form of detection output, these studies adopt divergent handling strategies. More specifically, some remove identified GSUs from the CF process altogether [51, 24], others report classification accuracy without addressing recommendation at all [37], while the GSOR framework [2] retains GSUs and generates preference estimates through its internal probabilistic mechanisms.

Although classification-based approaches provide an automated pipeline, the dependence on constructed labels introduces a fundamental limitation. The labels are derived from heuristic criteria, such as clustering boundaries or statistical thresholds, rather than independently validated annotations. As a result, the classification models effectively learn to reproduce the assumptions embedded in the label construction process. This dependency may propagate biases or errors present in the initial labeling stage, which can affect both identification and evaluation of GSUs.

### 3.1.4. *From Detection Outputs to Handling Strategies*

The preceding subsections describe detection methods that use different types of signals, but the handling strategies used after detection are relatively limited. Table 1 highlights this asymmetry. This subsection first clarifies the distinctions among the three detection families beyond their choice of heuristic, then consolidates the handling strategies reported in the literature, and finally examines the extent to which particular detection outputs support corresponding strategies.

A recurring ambiguity concerns the boundary between detection and label construction. Classification approaches derive training labels from clustering structures or statistical thresholds, which are also used by clustering-based detectors. The distinction therefore lies not in the heuristic itself, but in the output of the procedure. Clustering and outlier detection assign a status to each user in the population under analysis, and this assignment does not directly extend to users outside that population. Classification transforms the same heuristic into a decision function defined over user features, which can subsequently be applied to a new user without recomputing global cluster structures or similarity matrices. The label construction stage is thus better understood as a means of transferring a transductive criterion into an inductive form rather than as an independent detection mechanism. This interpretation also explains why classification approaches inherit the assumptions embedded in the labeling stage, an issue further discussed in Section 4.2.1.

There are four handling strategies that recur across the three detection families (i.e., clustering, outlier detection, classification-based approaches). The first isolates the identified users and serves them with a different recommendation paradigm, typically content-based or demographic filtering [14, 15, 22, 33, 50]. The second isolates the users but continues to apply collaborative filtering within the separated group, either through item-based prediction [45] or by averaging over the cluster [3]. The third isolates the users without serving them, treating separation as a means of improving accuracy for the remaining population [41, 42, 57, 51, 24]. The fourth retains the users within a single model and adjusts the learning procedure to accommodate them [18, 2]. The first three require multiple recommendation models to be maintained and coordinated, while the third leaves the identified users without any recommendation.

**Table 2**

Detection-free approaches: mechanisms, added signals, and evidence of their effect on GSUs

| Approach                                      | Mechanism                                         | Added signal                                      | Effect on GSUs                                           | Evaluation scope                |
|-----------------------------------------------|---------------------------------------------------|---------------------------------------------------|----------------------------------------------------------|---------------------------------|
| Neighborhood augmentation: synthetic [13, 12] | Virtual users from inverted ratings               | Synthetic profiles derived from the rating matrix | Larger pool of candidate neighbors                       | Aggregate only                  |
| Neighborhood augmentation: exogenous [46]     | Similarity combining ratings and network position | Social ties, independent of ratings               | Second similarity channel when ratings are uninformative | Aggregate and per user group    |
| Deep representation learning [11]             | Augmented matrix fed to an autoencoder            | Synthetic profiles                                | Neighborhood absorbed into latent space                  | Aggregate only                  |
| Semantic decoupling [31]                      | User and item attributes matched directly         | Demographic, health and item properties           | No neighbor construction required                        | Aggregate, application-specific |

These strategies are not equally compatible with the detection outputs listed in Table 1. A hard partition records only group membership and therefore directly supports the isolation-based strategies, but it provides no quantity that a downstream model can consume. Continuous scores retain more information, which Gras et al. exploit by assigning penalty weights during stochastic gradient descent in proportion to user status, allowing GSUs to remain in the shared user-item matrix while receiving greater emphasis during optimization [18]. Their comparison also indicates the cost of the alternative. *GSUOnly*, trained exclusively on the ratings of GSUs, raises the median RMSE by approximately 8% relative to a model trained on all users, while *SingleGSU*, constructed for an individual GSU, still requires a subset of normal users during training and deliberately excludes other GSUs [18]. Both results are consistent with earlier neighbor-based evidence that inconsistent users require information from outside their own cluster [4]. Soft membership degrees provide comparable information, although existing work reduces them to a hard assignment by selecting the cluster with the highest average indeterminacy [45]. Inductive decision boundaries yield a class posterior rather than a partition, which the GSOR framework uses to generate preference estimates for identified users within a unified probabilistic model [2].

In practice, this correspondence is rarely exploited. Most studies that compute a continuous score apply a cutoff and proceed as though a hard partition had been obtained, thereby discarding the graded information that distinguishes their detection stage from clustering. Several recent clustering-based methods stop at detection and report no handling procedure [16, 7, 5]. The apparent interchangeability of handling strategies across detection families is therefore a consequence of this reduction rather than evidence that the choice of detector is immaterial. Detection outputs differ in the information they preserve, and a handling strategy that ignores this information forfeits the advantage the detector provides.

### 3.2. Detection-Free Strategies

In contrast to detection-dependent strategies, detection-free approaches do not explicitly identify GSUs prior to recommendation. Instead, these methods acknowledge the presence of GSUs and aim to mitigate their impact by modifying the underlying recommendation process. The key idea is to design models that are inherently robust to unusual or weakly correlated user preferences, thereby reducing the dependence on explicit user categorization.

Rather than separating GSUs from the general population, detection-free strategies reformulate CF through mechanisms such as neighborhood augmentation, representation learning, and semantic decoupling. Neighborhood augmentation takes two distinct forms (i.e., the synthetic and exogenous approaches), depending on whether the additional signal is synthesized from the rating matrix itself or drawn from a source outside it. These approaches enhance the availability of informative signals or restructure the similarity space, enabling the system to generate reliable recommendations even when traditional user-user correlations are weak or inconsistent.

Table 2 summarizes these approaches. The column of *added signal* refers to the information that a method introduces into the recommendation process beyond the original rating matrix, such as synthetic user profiles, social

ties, or descriptive attributes of users and items. The *effect on GSUs* refers to the mechanism through which this additional information is expected to benefit users with unusual preferences, given that these users are never identified explicitly. The *evaluation scope* column records the population over which the reported improvements are measured, which in most cases is the full user base rather than the users the method is intended to help. While this paradigm avoids the complexity and potential errors associated with explicit detection, it also introduces challenges in directly quantifying its impact on GSUs. In the following, we review representative detection-free approaches and analyze how they address the grey sheep problem without relying on explicit identification.

### 3.2.1. Neighborhood Augmentation

A primary reason GSUs receive inaccurate recommendations in UBCF is the limited availability of highly correlated neighbors. Two lines of work (i.e., the synthetic and exogenous approaches) address this without explicit detection, differing in where the additional neighbors come from. The synthetic approach generates new profiles from the rating matrix itself. El Fazziki et al. [13] proposed a collaborative method based on opposite user inference. The main idea is that if a user  $X$  exhibits a strong negative correlation with a target user  $Y$ , then a virtual user  $\neg X$ , constructed by inverting the ratings of  $X$ , will exhibit a high positive correlation with  $Y$  [13]. By augmenting the rating matrix with such virtual neighbors, the method increases neighborhood density and enlarges the set of candidate neighbors for all users [13, 12].

The exogenous approach instead draws the additional signal from outside the rating matrix. Sánchez-Moreno et al. [46] combine rating-based similarity with the structural position of users in a social network, so that two users may be treated as neighbors on the basis of their network relations even when their rating overlap is insufficient to estimate similarity reliably. The mechanism differs from opposite user inference in an important respect. Synthetic profiles are constructed from the same data whose limitations they are intended to overcome, whereas exogenous signals constitute an independent source of evidence. This distinction also affects what can be verified, since the contribution of a synthetic neighbor cannot be traced back to any observed behavior.

Neither approach identifies GSUs explicitly. Instead, both restructure the neighborhood space so that users with unusual preferences can obtain sufficient informative neighbors. Consequently, users who would otherwise exhibit weak or unstable correlations can benefit from the enriched interaction structure. The effectiveness of these approaches is typically assessed through improvements in global evaluation metrics such as MAE or RMSE, compared with standard CF as baselines, so that the mitigation of the GСУ problem is inferred indirectly from overall performance gains. One exception is the work of Sánchez-Moreno et al. [46], which reports results separately for users with the lowest popularity scores at several thresholds, allowing the effect on this subgroup to be observed directly. As summarized in the last column of Table 2, the remaining methods report improvements only over the full user base, so it remains unclear whether the underlying datasets contain substantial GСУ effects or whether the reported gains arise primarily for GSUs or for the general user population.

### 3.2.2. Deep Representation Learning

Recent advances in neural collaborative filtering enable detection-free accommodation of GSUs by combining data augmentation with deep representation learning. El Youbi El Idrissi et al. [11] propose a hybrid framework that integrates opposite user inference with an Autoencoder-based prediction model. The overall pipeline consists of two sequential stages. In the first stage, the original rating matrix is enriched through opposite user inference, where virtual users are generated by inverting the ratings of existing users. This augmentation increases matrix density and expands the diversity of interaction patterns. From the perspective of GSUs, the enriched matrix enlarges the pool of potential neighbors, allowing users with unusual preferences to establish meaningful relationships with generated profiles without requiring explicit identification.

In the second stage, the augmented rating matrix (i.e., the rating matrix with virtual users or neighbors) is fed into an Autoencoder, which learns a non-linear mapping from high-dimensional rating vectors to a compact latent space through encoding and decoding processes. Unlike traditional factorization methods, the Autoencoder captures complex, non-linear user-item interactions and reconstructs missing ratings based on both observed and augmented signals. The combination of data enrichment and deep representation learning is critical. The augmented matrix provides sufficient data coverage, particularly for users with sparse or irregular behavior, while the Autoencoder leverages this enriched input to learn more expressive latent representations. Although this framework adopts the same opposite user inference mechanism for data augmentation, it fundamentally differs from neighborhood-based methods mentioned in Section 3.2.1. In particular, virtual users are not used for explicit neighbor selection or similarity computation. Instead,

they are incorporated as additional training samples in the augmented rating matrix, and their influence is learned implicitly through the Autoencoder. As a result, the notion of neighborhood is no longer explicitly defined, but is absorbed into the learned latent representations, which justifies categorizing this approach as representation learning rather than neighborhood augmentation.

The effectiveness of this framework is demonstrated through improvements in global evaluation metrics, such as RMSE and MAE, compared with traditional CF and standalone Autoencoder models [11]. These results suggest that enhancing both data density and representation capacity can mitigate the challenges associated with GSUs without requiring explicit detection. However, similar to other detection-free approaches, the evaluation is conducted at the aggregate level, and it remains unclear to what extent the observed gains specifically benefit GSUs as opposed to the general user population.

### 3.2.3. *Semantic Decoupling*

Another detection-free approach reformulates the notion of similarity in recommender systems. López-Nores et al. [31] observed that traditional CF is less effective for GSUs because it depends heavily on user-user or item-item rating overlap to construct neighborhoods. To address this limitation, they proposed Property-Based Collaborative Filtering (PBCF), which represents users and items through their underlying semantic properties, including demographic attributes, health conditions, and item features [31].

Rather than identifying globally similar users, PBCF constructs an explicit mapping between user attributes and item attributes in a shared semantic space. In practice, each user is represented by a set of descriptive properties, such as demographic characteristics or health-related conditions, while each item is described by a corresponding set of features. A matrix is then defined to quantify the relationships between user properties and item properties, where each entry reflects the relevance or compatibility between a specific pair of attributes. Based on this matrix, the system computes an “appeal vector” for a given user by aggregating the contributions of matched attributes, which is subsequently used to estimate the preference score of each item [31]. This process replaces similarity-based neighbor search with attribute-level matching, thereby reducing dependence on shared rating patterns and eliminating the need for explicit neighbor construction.

By operating on individual attributes, the model can produce recommendations even when users exhibit rare or unusual preferences. Consequently, users who would be weakly connected in traditional CF can still be effectively represented without requiring explicit GSU identification. However, the effectiveness of this approach is primarily demonstrated through improvements in application-specific or overall performance metrics, and it remains unclear to what extent these improvements specifically benefit GSUs relative to the general user population.

### 3.2.4. *Critical Insights*

The detection-free strategies discussed above share the common objective of mitigating the GSU problem without explicit identification, although they differ in the signal each introduces and in the mechanism through which that signal is expected to help, as indicated in Table 2. Neighborhood augmentation increases the availability of informative signals, either by synthesizing additional profiles from the rating matrix or by introducing relations observed outside it. Deep representation learning extends this idea by combining data augmentation with non-linear modeling, enabling the capture of complex user–item interactions in a latent space. In contrast, semantic decoupling replaces similarity-based reasoning with feature-level matching, reformulating recommendation as a mapping between user and item attributes. Despite these differences, all approaches shift the emphasis from identifying specific users to improving the robustness of the recommendation process.

An important observation from this paradigm is that GSUs are not treated as a distinct category requiring explicit handling. Instead, they are viewed as a consequence of sparsity, unusual preferences, or weak correlation structures in the data. By enhancing data density, representation capacity, or semantic expressiveness, these methods reduce the conditions under which GSUs adversely affect recommendation performance. This perspective indicates that the grey sheep problem can be addressed through model design rather than through dedicated detection procedures.

Nevertheless, a key limitation remains. With one exception [46], the effectiveness of detection-free strategies is assessed using aggregate performance metrics, and improvements are interpreted as indirect evidence of mitigating GSU-related issues. This evaluation approach does not establish whether the observed gains are specifically associated with GSUs or are driven by improvements for the general user population. As a result, although detection-free methods provide an appealing alternative to explicit detection, their impact on GSUs remains difficult to quantify precisely.

## 4. Evaluation Protocols and Challenges

In this section, we first give a summary of evaluation protocols in existing research related to alleviating GSU issues, and then point out the drawbacks and underlying challenges.

### 4.1. Existing Evaluation Protocols

Before examining the methodological limitations of GSU research, it is necessary to clarify how existing studies design their evaluation protocols. The literature generally follows two distinct paradigms, depending on whether GSUs are explicitly identified or implicitly handled within the model.

#### 4.1.1. Detection-Dependent Approaches

In detection-dependent approaches, evaluation is commonly organized as a two-stage process consisting of detection and handling. However, the literature reveals substantial inconsistency in how the detection stage is assessed. Because real-world datasets do not provide ground truth labels for GSUs, evaluation protocols vary significantly across different methodological families. In clustering-based approaches, detection quality is typically evaluated indirectly through downstream performance. For example, Ghazanfar and Prügel-Bennett [15] determine the optimal similarity threshold by identifying the point at which the MAE of the remaining users is minimized, thereby using the performance of the mainstream population as a proxy for validating the detection boundary. In outlier detection approaches, evaluation is often grounded in statistical reasoning. Gras et al. [17] define pseudo-ground truth by identifying users whose prediction errors exceed a predefined quantile, while more recent studies such as Zheng [62] employ hypothesis testing to verify that detected outliers exhibit significantly higher prediction errors than regular users. In classification-based approaches, researchers construct labels using heuristic or statistical rules and evaluate detection performance using standard metrics such as accuracy, precision, recall, and F1-score, as demonstrated by Alabdulrahman and Viktor [2].

Once detection is completed, the handling stage is evaluated using standard recommendation metrics, including MAE, RMSE, and coverage. More rigorous studies adopt segmented evaluation protocols, reporting performance separately for GSUs and for the general user population [15, 2]. Despite this practice, the separation between detection and handling evaluation introduces a conceptual limitation, as improvements in detection accuracy or outlier ranking do not necessarily imply improved recommendation quality for GSUs.

To address the limitations of global evaluation, Zheng et al. [65, 64] proposed a segmented evaluation protocol that explicitly analyzes performance across different user groups. In this framework, the dataset is partitioned into three subsets: the complete dataset, the subset containing only identified GSUs, and the subset containing only non-GSUs. A standard UBCF model is then applied independently to each subset. By comparing prediction errors across these partitions, the protocol enables two forms of validation. First, it allows for statistical verification that the detected GSUs exhibit significantly higher prediction errors than non-GSUs, thereby providing indirect evidence for the validity of the detection process without relying on predefined labels. Second, it quantifies the impact of GSUs on the overall system by comparing the performance of the full dataset with that of the dataset after removing GSUs. Studies such as Tennakoon et al. [53] demonstrate that removing GSUs can improve accuracy for the remaining users, although they also report that this effect reverses on smaller datasets, which indicates that the benefit of removal depends on the setting rather than holding generally.

This segmented evaluation protocol provides a more transparent assessment of GSU-related phenomena by explicitly examining performance differences across user groups. Nevertheless, it still depends on the initial detection results and does not fully resolve the challenges associated with defining and evaluating GSUs in the absence of ground truth.

#### 4.1.2. Detection-Free Approaches

In contrast, detection-free approaches bypass the identification stage entirely. Methods such as opposite user inference [13] and Autoencoder-based representation learning [11] do not produce GSU labels or isolate users, and their effectiveness is assessed through global recommendation metrics, such as MAE or RMSE, computed over the entire dataset. The underlying assumption is that a reduction in overall error indicates improved robustness to users with unusual tastes. Because GSUs are never identified, however, such evaluations cannot verify whether the observed improvements are attributable to this subgroup.

Detection and evaluation need not be coupled in this way. Sánchez-Moreno et al. [46] construct an explicit subset of users with unusual tastes for evaluation purposes only, leaving the recommendation model itself detection-free,

and report performance separately for that subset. Segmented evaluation is therefore available to any method, and its absence in most detection-free work reflects a choice rather than a constraint imposed by the paradigm.

## 4.2. Evaluation Challenges

Despite the widespread adoption of these evaluation paradigms, their validity remains limited. The dependency on constructed labels, the separation between detection and recommendation evaluation, and the absence of user-level analysis introduce significant methodological concerns. Recent studies have also recognized the limitations of current evaluation practices specifically designed for GSU detection, further highlighting the need for more reliable and standardized evaluation frameworks [6]. The following subsections analyze these issues in detail, focusing on the lack of ground truth, inconsistencies in definitions, limitations of evaluation protocols, and biases in commonly used datasets.

### 4.2.1. Lack of Ground Truth

A crucial limitation in GSU research is the absence of ground truth, i.e., we do not have true labels to indicate real GSUs. In real-world datasets such as MovieLens, FilmTrust, or Amazon, no user is inherently labeled as a GSU. This limitation has also been acknowledged by recent work, which notes that commonly used datasets such as MovieLens and FilmTrust do not provide ground-truth labels for GSUs, making objective evaluation particularly challenging [6].

To facilitate evaluation, researchers construct pseudo labels based on heuristic or statistical rules. For instance, Alabdulrahman et al. [2] apply Expectation-Maximization clustering combined with interquartile range thresholds, whereas Kaur and Rani [24] derive labels from similarity matrices based on personality features. This practice introduces the pseudo-labeling problem. Labels generated through predefined rules are subsequently used to train supervised models and are then employed for evaluation. As a result, high detection accuracy often reflects the ability of a model to reproduce the underlying heuristics rather than to identify genuinely unusual user behavior. This creates a circular evaluation process in which the definition and validation of GSUs are not independent.

### 4.2.2. Inconsistent Definitions Across Studies

The lack of ground truth is further compounded by inconsistencies in how GSUs are defined across studies. The criteria adopted differ in the source of evidence from which unusualness is inferred. Some are derived from the structure of user-user similarity, either through the statistical characteristics of a user's similarity distribution [65] or through the position of a user relative to cluster boundaries [15]. Others draw on information external to the rating matrix, such as psychographic attributes derived from personality models [51]. A further criterion has emerged in domains where explicit ratings are unavailable, defining unusual users through the popularity of the items they consume rather than through their relations to other users [46]. These criteria are not equivalent, and need not select the same users. Listeners who share an interest in the same niche items may be highly similar to one another while all having low popularity scores, whereas a user with idiosyncratic preferences across both popular and obscure items may not be identified by a popularity-based criterion at all. As a consequence, the classification of a user as a GSU may vary significantly across methods.

A particularly problematic case arises when GSUs are defined through performance-based criteria. Some studies associate GSUs with users exhibiting high prediction error, such as elevated RMSE [17]. This definition introduces a dependence on the underlying model. When a more expressive model is applied, such as a non-linear Autoencoder [11], prediction error for these users may decrease, leading to a change in classification. In this context, the notion of GSUs becomes model-dependent rather than reflecting intrinsic user characteristics.

### 4.2.3. Limitations in Evaluation Protocols

Evaluation protocols in GSU research often exhibit a mismatch between detection and recommendation objectives. Some studies report detection metrics such as precision, recall, or F1-score when pseudo labels are available, while others focus on recommendation metrics such as MAE or RMSE. These two aspects are frequently confused, even though accurate detection does not necessarily lead to improved recommendation quality. A recent study by Boualaoui et al. [6] summarizes the dominant evaluation practices in GSU research, including MAE, RMSE, coverage, and user-group-specific metrics such as Inc(MAE) and Imp(MAE), originally introduced by Rashidi et al. [41] to quantify, respectively, the MAE increase for GSUs and the MAE improvement for the remaining users. Their analysis further shows that most studies evaluate GSU detection indirectly through recommendation performance rather than through independent validation of user labels.

Another concern arises from the use of global evaluation metrics to assess methods that remove GSUs. Some approaches report improvements in overall RMSE after excluding identified GSUs from the dataset [51]. Although such results indicate reduced average error, they do not reflect system performance for the excluded users. From both practical and application perspectives, removing challenging users to improve aggregate metrics is inconsistent with the goal of providing personalized recommendations for all users. More appropriate protocols should report performance separately for GSUs and for the general user population.

#### 4.2.4. Dataset Bias

Empirical evaluation in GSU research is also influenced by dataset bias. Most studies rely on a limited set of benchmark datasets, such as MovieLens and FilmTrust [15, 65]. While these datasets are widely used, they may not adequately capture the diversity and complexity of user preferences present in other domains, including health-aware systems [31], music domains or specialized e-commerce applications.

In addition, these datasets represent static snapshots of user behaviors. In practical settings, user preferences evolve over time. A user with sparse initial interactions may temporarily exhibit unusual tastes due to limited data, but may later develop stable preference patterns as more interactions are observed, as depicted by Figure 1. Static evaluation fails to capture such temporal dynamics and may incorrectly treat transient behavior as persistent grey sheep characteristics. As a result, the absence of temporal modeling limits the ability of current evaluation frameworks to distinguish between short-term irregularity and long-term preference divergence.

## 5. Future Research Directions

Building on the limitations identified in existing detection and evaluation frameworks, future research on GSUs should move beyond static, rating-centric paradigms toward more dynamic, data-rich, and user-aware modeling strategies. The following directions outline key opportunities for advancing the understanding and handling of GSUs in modern recommender systems.

### 5.1. Benchmark Datasets and Domain Diversity

A critical bottleneck in GSU research is the lack of standardized benchmark datasets specifically designed for evaluating users with unusual tastes. Existing datasets, such as MovieLens and FilmTrust, are not constructed with GSU analysis in mind and do not provide reliable annotations or sufficient diversity in user behavior. Future work should focus on developing benchmark datasets that explicitly capture unusual preference patterns, either through controlled simulation or curated real-world data. Such datasets should support consistent evaluation protocols, enable reproducibility, and facilitate fair comparison across methods.

Extending this effort beyond entertainment is not simply a matter of collecting data from more areas, since domains differ in the aspect of behavior through which greyness is expressed. The literature reviewed in this survey already spans e-commerce [51, 24], music [46], news [53] and health-aware recommendation [31], yet in each case the domain is treated as an incidental property of the dataset rather than as something that shapes what a GSU is. Table 3 summarizes how the characteristics of GSUs differ across these domains, why detection methods developed on movie ratings do not transfer directly, and which research directions follow.

The requirements listed in Table 3 are not compatible with one another. A benchmark that preserves session boundaries for music serves little purpose in news, and one that records product hierarchies for e-commerce says nothing about clinical constraints. Healthcare differs further in that the cost of an unsuitable recommendation is asymmetric, which makes the strategy of isolating GSUs without serving them, identified in Section 3.1.4, difficult to justify in that setting. This suggests that a single benchmark for GSU research is unlikely to be adequate, and that a small family of domain-specific benchmarks, each recording the signals that make greyness observable in that domain, would be more appropriate.

### 5.2. Local Similarity Modeling

Existing GSU research commonly assumes that GSUs exhibit uniformly low similarity with the rest of the population, leading to the absence of reliable neighbors. However, this assumption may oversimplify real-world user behavior. In practice, GSUs may align with other users on a subset of items, such as the most popular or widely consumed items, while diverging significantly on less common or niche items. This results in heterogeneous similarity patterns, where users agree on some items but differ overall.

**Table 3**

Domain-specific characteristics of GSUs and corresponding research directions

| Domain                 | Characteristics of GSUs                                                      | Why standard detection fails                                                                          | Research directions                                                                                             |
|------------------------|------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| E-commerce<br>[51, 24] | Fine-grained interests; marked preference for long-tail products             | Purchases are sparse and non-repeating, so item overlap is near zero even for similar users           | Local similarity defined over product hierarchies and attributes; sequential modeling of purchase behavior      |
| Music<br>[46]          | Preferences span multiple genres; strong temporal and session-level dynamics | Aggregating repeated plays into a static profile removes the sequence in which deviation is expressed | Time-aware and session-level modeling; separating long-tail consumption from genuine preference divergence      |
| News<br>[53]           | Rapidly shifting interests; strong recency effects                           | Items expire within hours, so two users rarely interact with the same article                         | Deviation defined over topics or entities rather than items; shorter windows for assessing preference stability |
| Healthcare<br>[31]     | Preferences constrained by conditions, health status and risk                | Deviation may reflect medical necessity rather than unusual taste                                     | Semantic and demographic representations; privacy-aware detection; risk-sensitive evaluation                    |

Such observations suggest that modeling GSUs through global similarity [56] measures may be insufficient. Future research should explore localized similarity modeling, where user similarity is analyzed within specific item subsets, contexts, or interaction patterns [44]. For example, segmenting items into popular and niche groups, or modeling similarity as a function of item characteristics, may provide a more accurate representation of user relationships. This perspective enables identifying situations in which GSUs behave similarly to the mainstream and when they deviate, offering a more fine-grained understanding of their impact on recommendation quality.

Incorporating local similarity into recommendation models may also improve robustness by allowing systems to selectively rely on reliable signals while discounting inconsistent interactions. Despite its potential, this direction has only begun to be explored [7, 5], and it represents a promising avenue for developing more realistic and effective modeling strategies.

### 5.3. Temporal Modeling of User Behavior

Current GSU research treats user behaviors or characteristics as static ones, ignoring the temporal evolution of preferences as illustrated by Figure 1. In practice, user interests change over time, and behaviors that appear unusual at one stage may become consistent as more interactions are observed [25, 40, 47]. This suggests that GSUs should be modeled as a dynamic state rather than a fixed user category. Future research should incorporate temporal modeling techniques, such as sequential recommendation models or time-aware embeddings, to capture the evolution of user preferences. This perspective enables distinguishing between temporary anomalies caused by data sparsity and persistent preference divergence, leading to more accurate and adaptive recommendation strategies.

### 5.4. GSUs in Specialized Recommender Systems

Most existing studies on GSUs are grounded in traditional recommender systems with explicit ratings, where greyiness is expressed as a single scalar summarizing a user's relationship to the rest of the population. This formulation already becomes difficult to apply when only implicit feedback is available, since most detection methods reviewed in Section 3.1 rely on pairwise correlations that cannot be computed in the absence of ratings. Studies in such domains have consequently adopted alternative criteria, as when Sánchez-Moreno et al. [46] identify users with unusual tastes in a music setting through the average popularity of the artists they listen to, which is a popularity-based rather than a similarity-based criterion, and one that does not necessarily select the same users as similarity-based criteria do, as indicated in Section 4.2.2. Modern recommender systems present further conditions under which the standard formulation is inadequate. The following discussion considers three such settings and, in each case, examines how the notion of a GSU changes and which existing detection methods cease to apply.

One promising application lies in multi-criteria recommendations [36, 68, 70], where users rate items along several aspects of the items, such as location, room cleanliness, room size, value for money in the context of hotel bookings. Greyiness is no longer a scalar property, since a user may agree with the majority on some criteria while diverging sharply on others. A traveler who shares mainstream opinions about location and cleanliness but consistently assigns

unusually low importance to room size would be averaged into the mainstream by any method operating on overall ratings, yet is precisely the kind of user whose recommendations are difficult to produce. This suggests representing greyness as a vector defined over criteria and detecting users whose deviation is concentrated in particular dimensions. It also raises the question of whether such users require an alternative model or only a re-weighting of the criteria used to compute similarity, which cannot be answered within the single-rating formulation. An alternative is to rank items across criteria without aggregating them into a single score [67], which would locate greyness in the ranking behavior rather than in the aggregation weights. Recent work on multi-criteria recommendation distinguishes two architectural paradigms [70]. Multi-stage architectures separate criteria-level prediction from preference aggregation, whereas single-stage architectures model the criteria jointly. This distinction suggests that greyness may reside in either stage, since a user may rate individual criteria in an entirely conventional manner while combining them through an unusual aggregation, a pattern that overall ratings cannot reveal.

Context-aware recommender systems introduce a different scenario. These systems assume that user preferences vary across contextual situations (e.g., time, location, weather, companion, etc.), so that the same user may make different choices depending on the context in which a decision is made [66, 34]. A viewer who watches action films alone, for instance, may prefer animated titles when watching with children. Greyness then becomes a property of the user-context pair rather than of the user alone, since a user may align with others in one context and deviate in another. A viewer whose choices are conventional when watching with family but unusual when watching alone would appear moderately atypical once contexts are pooled, which is an artifact of aggregation rather than a description of behavior. Detection therefore needs to operate on contextual similarity [60], and the appropriate handling strategy may be to switch models only in those contexts where the user deviates, rather than to isolate the user permanently.

By contrast, multi-objective recommendations raise a further question. These systems optimize several objectives at once, such as accuracy, diversity, novelty and coverage, and typically select a single compromise solution along the Pareto frontier [69]. Whether a user counts as problematic then depends on which objective is being considered. A listener who consumes only long-tail content is difficult to serve under an accuracy objective, yet is among the easiest to satisfy under novelty or coverage, since recommending unpopular items is precisely what suits them. Greyness in this setting is therefore not a property of the user but of the pairing between the user and the objective, which is a sharper form of the model dependence discussed in Section 2.2. A further difficulty is that the compromise solution is selected for the population as a whole, so that users whose preferred trade-off lies far from the collective one are absorbed into an aggregate that does not represent them. This suggests investigating whether such users are better served by a different point on the frontier rather than by a different model, a possibility that the current formulation of the GSU problem does not accommodate.

These examples indicate that the conceptual questions raised in Section 2 become more difficult, not less, in specialized settings. Extending GSU research to these settings requires more than applying existing detectors to new data, since what counts as a GSU is defined differently in each case.

### 5.5. Multi-Modal User Modeling

Traditional GSU research relies heavily on numerical rating data, which provides a limited view of user preferences. The increasing availability of multi-modal data, including textual reviews, images, browsing behavior, and interaction logs, offers new opportunities for understanding GSUs. By integrating multiple data sources, recommender systems can construct richer user representations that go beyond sparse rating matrices [54, 30]. This may help alleviate the challenges associated with GSUs by uncovering latent preference signals that are not captured by ratings alone. Future work should explore multi-modal fusion techniques to improve the robustness of recommendation models in the presence of diverse and unusual user behaviors.

### 5.6. LLM-Based User Understanding

Recent advances in large language models (LLM) provide new opportunities for modeling user preferences at a semantic level. Unlike traditional methods that rely on numerical similarity, LLM-based approaches can interpret user intent, contextualize preferences, and generate structured representations from unstructured data such as reviews or interaction histories [59, 71]. This capability is particularly relevant for GSUs, whose preferences may not align with dominant patterns but can still be meaningfully understood through semantic reasoning. Future research should investigate how LLM can be integrated into recommender systems to enhance user understanding, generate personalized explanations, and improve recommendation quality for GSUs.

## 5.7. User-Centric Evaluation

Finally, future research should move toward user-centric evaluation protocols [39, 10] that explicitly assess the performance of recommender systems on different user groups. Existing evaluation practices largely rely on global metrics, which obscure the performance on GSUs due to their minority status. More informative evaluation should include subgroup analysis, reporting performance separately for GSUs and non-GSUs, as well as robustness measures that assess model behavior under sparse or unusual conditions. In addition, evaluation frameworks should incorporate fairness considerations, ensuring that improvements in overall performance do not come at the expense of minority users. Developing standardized, user-aware evaluation protocols will be essential for advancing the practical relevance of GUS research.

## 5.8. Grey Sheep Items

While existing research primarily focuses on GSUs, the concept can be naturally extended to the item side. Grey sheep items refer to items that exhibit weak or inconsistent similarity relationships with other items, making them difficult to model using standard CF techniques, especially in item-based CF.

In many cases, such behavior may arise from data sparsity, particularly during the early stages of an item's lifecycle when only limited interactions are available. As more ratings accumulate, these items are more likely to establish reliable similarity connections with other items, reducing their grey sheep characteristics. However, not all grey sheep items are temporary. Certain items may consistently appeal to niche audiences or exhibit highly polarized feedback, leading to persistent irregular similarity patterns even in dense datasets.

This perspective suggests that grey sheep phenomena are not limited to users, but may reflect a broader issue related to irregular interaction structures in recommender systems. Future research could investigate how to model such items explicitly, for example by distinguishing between temporary sparsity-driven effects and persistent atypical patterns, and by developing item-aware strategies that complement existing user-centric approaches.

## 6. Conclusion

This paper presents a comprehensive review of GSUs in recommender systems, covering their conceptual foundations, detection methodologies, handling strategies, evaluation challenges, and future research directions. Existing approaches can be broadly categorized into *detection-dependent strategies*, which explicitly identify GSUs prior to applying specialized handling, and *detection-free strategies*, which aim to improve model robustness without relying on explicit user categorization. While these methods offer different perspectives on addressing unusual user behavior, they collectively highlight the complexity of modeling users whose preferences deviate from dominant patterns. In particular, this work revisits the classical similarity-based interpretation and provides a refined conceptual and dynamic understanding of GSUs beyond static and model-dependent definitions.

More fundamentally, this review suggests that the grey sheep phenomenon should not be interpreted as an inherent user type, but rather as a consequence of the interaction between user behavior, data properties, and model design. In particular, the assumption of globally consistent user similarity is often violated in practice, where users may exhibit heterogeneous similarity patterns across different subsets of items. This perspective reframes GSUs as a modeling challenge rather than a fixed user category.

A key finding of this survey is that the primary challenges in GUS research do not arise solely from algorithm design, but from the lack of reliable evaluation frameworks. The absence of ground truth, inconsistent definitions across studies, and dependence on global performance metrics limit the ability to accurately assess the effectiveness of existing methods. In many cases, improvements are inferred indirectly, making it difficult to determine whether GSUs themselves truly benefit from proposed techniques.

The analysis also reveals a shift in research paradigms, from early approaches focused on isolating GSUs and applying alternative models, toward more recent strategies that integrate GSUs within unified frameworks through data augmentation, representation learning, and semantic modeling. These developments suggest that the grey sheep problem may be better addressed by enhancing the flexibility and expressiveness of recommendation models rather than by treating GSUs as a separate category. This shift is further supported by the observation that user similarity is often context-dependent rather than globally consistent, highlighting the importance of localized similarity modeling in understanding and addressing GSUs.

Looking forward, advancing GUS research requires a transition toward dynamic, multi-source, and user-centric perspectives. Incorporating temporal modeling, multi-modal data, and semantic understanding through LLM has the

potential to provide richer representations of GSUs. At the same time, the development of benchmark datasets and standardized evaluation protocols will be essential for enabling consistent and meaningful comparison across methods. Ultimately, this work argues that advancing GSU research requires moving beyond static similarity assumptions toward dynamic, context-aware, and user-centric modeling paradigms, which are essential for building more inclusive and robust recommender systems.

## CRediT authorship contribution statement

**Yong Zheng:** Conceptualization of this study, Writing, Editing.

**Declaration of competing interest.** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Declaration of usage of GenAI.** During the preparation of this work the authors used ChatGPT in order to assist with language editing. Figure 1 was generated using ChatGPT following the authors' instructions. All content generated with these tools was verified against the original publications, reviewed and edited by the authors, who take full responsibility for the content of the published article.

**Funding acknowledgement.** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Data availability.** No data were created or analyzed in this study.

## References

- [1] Aggarwal, C.C., 2016. Content-based recommender systems, in: *Recommender systems: The textbook*. Springer, pp. 139–166.
- [2] Alabdulrahman, R., Viktor, H., 2021. Catering for unique tastes: Targeting grey-sheep users recommender systems through one-class machine learning. *Expert systems with applications* 166, 114061.
- [3] Bejaoui, G., Ayachi, R., 2020. A trust-based clustering approach for identifying grey sheep users, in: *International Conference on Digital Economy*, Springer. pp. 76–88.
- [4] Bellogín, A., Said, A., de Vries, A.P., 2014. The magic barrier of recommender systems—no magic, just ratings, in: *International Conference on User Modeling, Adaptation, and Personalization*, Springer. pp. 25–36.
- [5] Boualaoui, B., Zellou, A., Ajallouda, L., 2026a. Gtmedoids: A new grey sheep users detection approach. *International Journal of Advanced Computer Science & Applications* 17, 934.
- [6] Boualaoui, B., Zellou, A., Ajallouda, L., 2026b. Towards a new evaluation protocol for grey sheep user detection approaches, in: *The 3rd International Conference on Artificial Intelligence and Smart Applications (AISA'25)*, Volume 1: Artificial Intelligence, IoT, and Smart Applications, Springer Nature Switzerland. pp. 13–25.
- [7] Bouchra, B., Ahmed, Z., Lahbib, A., 2026. Detect grey sheep users for recommender system: A new clustering approach, in: *2026 6th International Conference on Innovative Research in Applied Science, Engineering and Technology*, IEEE. pp. 1–6.
- [8] Chen, Z., Gan, W., Wu, J., Hu, K., Lin, H., 2025. Data scarcity in recommendation systems: A survey. *ACM Transactions on Recommender Systems* 3, 1–31.
- [9] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M., et al., 1999. Combining content-based and collaborative filters in an online newspaper, in: *Proceedings of ACM SIGIR workshop on recommender systems*, Berkley, CA, USA. pp. 1853–1870.
- [10] Cremonesi, P., Garzotto, F., Turrin, R., 2013. User-centric vs. system-centric evaluation of recommender systems, in: *Ifip conference on human-computer interaction*, Springer. pp. 334–351.
- [11] El Youbi El Idrissi, L., Akharraz, I., El Ouazizi, A., Ahaitouf, A., 2024. Enhanced collaborative filtering: Combining autoencoder and opposite user inference to solve sparsity and gray sheep issues. *Computers* 13, 275.
- [12] Fazziki, A., El, Y., Alami, E., Elhassouni, M., Benbrahim, M., 2024. Enhancing collaborative filtering: Addressing sparsity and gray sheep with opposite user inference. *Journal of Theoretical and Applied Information Technology* 15.
- [13] Fazziki, A.E., El Aissaoui, O., El Alami, Y.E.M., Alloui, Y.E., Benbrahim, M., 2019. A new collaborative approach to solve the gray-sheep users problem in recommender systems, in: *2019 Third International Conference on Intelligent Computing in Data Sciences*, IEEE. pp. 1–4.
- [14] Ghazanfar, M.A., Prugel-Bennett, A., 2011. Fulfilling the needs of gray-sheep users in recommender systems, a clustering solution, in: *2011 International Conference on Information Systems and Computational Intelligence*.
- [15] Ghazanfar, M.A., Prugel-Bennett, A., 2014. Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems. *Expert Systems with Applications* 41, 3261–3275.
- [16] Ghorbani, S., Novin, A.H., 2016. An introduction on separating gray-sheep users in personalized recommender systems using clustering solution. *International Journal of Computer Science and Software Engineering* 5, 14.
- [17] Gras, B., Brun, A., Boyer, A., 2016. Identifying grey sheep users in collaborative filtering: a distribution-based technique, in: *Proceedings of the 2016 conference on user modeling adaptation and personalization*, pp. 17–26.

- [18] Gras, B., Brun, A., Boyer, A., 2017. Can matrix factorization improve the accuracy of recommendations provided to grey sheep users?, in: 13th international conference on Web Information Systems and Technologies, pp. 88–96.
- [19] He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., Wang, M., 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation, in: Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, pp. 639–648.
- [20] He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.S., 2017. Neural collaborative filtering, in: Proceedings of the 26th international conference on world wide web, pp. 173–182.
- [21] Idrissi, N., Zellou, A., 2020. A systematic literature review of sparsity issues in recommender systems. *Social Network Analysis and Mining* 10, 15.
- [22] Jindal, H., Agarwal, S., Sardana, N., 2018. Powkmeans: a hybrid approach for gray sheep users detection and their recommendations. *International Journal of Information Technology and Web Engineering* 13, 56–69.
- [23] Kaur, B., Rani, S., 2021. Identification of gray sheep using different clustering algorithms, in: Proceedings of the Second International Conference on Information Management and Machine Intelligence, Springer. pp. 211–217.
- [24] Kaur, B., Rani, S., 2023. Are the customers receiving exact recommendations from the e-commerce companies? towards the identification of gray sheep users using personality parameters. *International Journal of Performability Engineering* 19, 425.
- [25] Koren, Y., 2009. Collaborative filtering with temporal dynamics, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 447–456.
- [26] Koren, Y., Bell, R., Volinsky, C., 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 30–37.
- [27] Koren, Y., Rendle, S., Bell, R., 2021. Advances in collaborative filtering. *Recommender systems handbook*, 91–142.
- [28] Lika, B., Kolomvatsos, K., Hadjiefthymiades, S., 2014. Facing the cold start problem in recommender systems. *Expert systems with applications* 41, 2065–2073.
- [29] Liu, C., Jiang, D., Cai, Y., Li, H., 2026. A review of deep learning and large language models for cold start problem in recommender systems. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 16, e70068.
- [30] Liu, Q., Hu, J., Xiao, Y., Zhao, X., Gao, J., Wang, W., Li, Q., Tang, J., 2024. Multimodal recommender systems: A survey. *ACM Computing Surveys* 57, 1–17.
- [31] López-Nores, M., Blanco-Fernández, Y., Pazos-Arias, J.J., Gil-Solla, A., 2012. Property-based collaborative filtering for health-aware recommender systems. *Expert Systems with Applications* 39, 7451–7457.
- [32] Lops, P., De Gemmis, M., Semeraro, G., 2010. Content-based recommender systems: State of the art and trends. *Recommender systems handbook*, 73–105.
- [33] Malik, S., Bansal, M., 2023. A new hybrid approach to resolve the gray sheep users problem in recommender systems. *International Journal of Recent Research Aspects* 10, 27–35.
- [34] Mateos, P., Bellogín, A., 2024. A systematic literature review of recent advances on context-aware recommender systems. *Artificial Intelligence Review* 58, 20.
- [35] McCrae, J., Piatek, A., Langley, A., 2004. Collaborative filtering. <http://www.imperialviolet.org>.
- [36] Monti, D., Rizzo, G., Morisio, M., 2021. A systematic literature review of multicriteria recommender systems. *Artificial intelligence review* 54, 427–468.
- [37] Noreen, F., Naqvi, Q.U.A., Rasool, A., Razaq, N., Abbas, F., Alam, B., 2026. Mitigating the grey sheep problem in collaborative filtering via hybrid machine learning architectures. *Journal of Engineering, Science and Technological Trends* 3, 24–32.
- [38] Panda, D.K., Ray, S., 2022. Approaches and algorithms to mitigate cold start problems in recommender systems: a systematic literature review. *Journal of Intelligent Information Systems* 59, 341–366.
- [39] Pu, P., Chen, L., Hu, R., 2011. A user-centric evaluation framework for recommender systems, in: Proceedings of the fifth ACM conference on Recommender systems, pp. 157–164.
- [40] Rabi, I., Salim, N., Da’u, A., Osman, A., 2020. Recommender system based on temporal models: a systematic review. *Applied Sciences* 10, 2204.
- [41] Rashidi, R., Khamforoosh, K., Sheikahmadi, A., 2020. An analytic approach to separate users by introducing new combinations of initial centers of clustering. *Physica A: statistical mechanics and its applications* 551, 124185.
- [42] Rashidi, R., Khamforoosh, K., Sheikahmadi, A., 2022. Proposing improved meta-heuristic algorithms for clustering and separating users in the recommender systems. *Electronic Commerce Research* 22, 623–648.
- [43] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J., 1994. Grouplens: An open architecture for collaborative filtering of netnews, in: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, pp. 175–186.
- [44] Rosa, R.E.V.D.S., Guimarães, F.A.S., da Silva Mendonça, R., de Lucena, V.F., 2020. Improving prediction accuracy in neighborhood-based collaborative filtering by using local similarity. *IEEE Access* 8, 142795–142809.
- [45] Samir, D., El Reheem, E.A., Darwish, S.M., Madbouly, M.M., 2026. A neutrosophic clustering approach to handle recommendation uncertainty for gray sheep users. *Scientific Reports*.
- [46] Sánchez-Moreno, D., López Batista, V.F., Muñoz Vicente, M.D., Sánchez Lázaro, Á.L., Moreno-García, M.N., 2024. Social network community detection to deal with gray-sheep and cold-start problems in music recommender systems. *Information* 15, 138.
- [47] Sánchez-Moreno, D., Zheng, Y., Moreno-García, M.N., 2020. Time-aware music recommender systems: Modeling the evolution of implicit user preferences and user listening habits in a collaborative filtering approach. *Applied Sciences* 10, 5324.
- [48] Sarwar, B., Karypis, G., Konstan, J., Riedl, J., 2001. Item-based collaborative filtering recommendation algorithms, in: Proceedings of the 10th International World Wide Web Conference, pp. 285–295.
- [49] Singh, M., 2020. Scalability and sparsity issues in recommender datasets: a survey. *Knowledge and Information Systems* 62, 1–43.
- [50] Solanki, S., Batra, S., 2015. Recommender system using collaborative filtering and demographic characteristics of users. *Int. J. Recent Innov. Trends Comput. Commun* 3, 4735–4741.

- [51] Srivastava, A., Bala, P.K., Kumar, B., 2020. New perspectives on gray sheep behavior in e-commerce recommendations. *Journal of Retailing and Consumer Services* 53, 101764.
- [52] Su, X., Khoshgoftaar, T.M., 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence* 2009, 421425.
- [53] Tennakoon, A., Gamlath, N., Kirindage, G., Ranatunga, J., Haddela, P., Kaveendri, D., 2020. Hybrid recommender for condensed sinhala news with grey sheep user identification, in: 2020 2nd International Conference on Advancements in Computing, IEEE. pp. 228–233.
- [54] Truong, Q.T., Salah, A., Lauw, H., 2021. Multi-modal recommender systems: Hands-on exploration, in: *Proceedings of the 15th ACM Conference on Recommender Systems*, pp. 834–837.
- [55] Wang, X., He, X., Wang, M., Feng, F., Chua, T.S., 2019. Neural graph collaborative filtering, in: *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 165–174.
- [56] Wei, S., Ye, N., Zhang, S., Huang, X., Zhu, J., 2012. Collaborative filtering recommendation algorithm based on item clustering and global similarity, in: 2012 Fifth international conference on business intelligence and financial engineering, IEEE. pp. 69–72.
- [57] Yu, J., Rashidi, R., 2026. Hybridfireflykmeans: a novel approach to enhance collaborative filtering through optimized user clustering and gray sheep user separation. *Cluster Computing* 29, 301.
- [58] Yuan, H., Hernandez, A.A., 2023. User cold start problem in recommendation systems: A systematic review. *IEEE access* 11, 136958–136977.
- [59] Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., Wen, Z., Wang, F., Zhao, X., Tang, J., et al., 2024. Recommender systems in the era of large language models (LLMs). *IEEE Transactions on Knowledge and Data Engineering* 36, 6889–6907.
- [60] Zheng, Y., 2022. Context-aware collaborative filtering using context similarity: An empirical comparison. *Information* 13, 42.
- [61] Zheng, Y., 2024. An empirical comparison of outlier detection methods for identifying grey-sheep users in recommender systems, in: *International Conference on Advanced Information Networking and Applications*, Springer. pp. 62–73.
- [62] Zheng, Y., 2025. Using outlier detection to identify grey-sheep users in recommender systems: a comparative study. *Computers, Materials, & Continua* 83, 4315.
- [63] Zheng, Y., 2026. A comprehensive review of rating imputation in recommender systems: From data completion to inference-oriented missing-data estimation. *Computers, Materials & Continua* doi:10.32604/cmc.2026.084278.
- [64] Zheng, Y., Agnani, M., Singh, M., 2017a. Identification of grey sheep users by histogram intersection in recommender systems, in: *International Conference on advanced data mining and applications*, Springer. pp. 148–161.
- [65] Zheng, Y., Agnani, M., Singh, M., 2017b. Identifying grey sheep users by the distribution of user similarities in collaborative filtering, in: *Proceedings of the 6th ACM Annual Conference on Research in Information Technology*, pp. 1–6.
- [66] Zheng, Y., Mobasher, B., 2018. *Context-Aware Recommendations*. World Scientific Publishing. pp. 173–202.
- [67] Zheng, Y., Wang, D., 2022a. Multi-criteria ranking: Next generation of multi-criteria recommendation framework. *Ieee Access* 10, 90715–90725.
- [68] Zheng, Y., Wang, D., 2023. Multi-criteria decision making and recommender systems, in: *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 181–184.
- [69] Zheng, Y., Wang, D.X., 2022b. A survey of recommender systems with multi-objective optimization. *Neurocomputing* 474, 141–153.
- [70] Zheng, Y., Wang, D.X., 2026. A narrative review of multi-criteria recommender systems: From perspectives of multi-criteria decision making. *Neurocomputing* , 134673.
- [71] Zhu, X., Wang, Y., Gao, H., Xu, W., Wang, C., Liu, Z., Wang, K., Jin, M., Pang, L., Weng, Q., et al., 2025. Recommender systems meet large language model agents: A survey. *Foundations and Trends® in Privacy and Security* 7, 247–396.