MULTIDISCIPLINARY SCIENTIFIC RESEARCH
BJMSR VOL 11 NO 2 (2026) P-ISSN 2687-850X E-ISSN 2687-8518
Available online at https://www.cribfb.com
Journal homepage: https://www.cribfb.com/journal/index.php/BJMSR
Published by CRIBFB, USA

DISCOVERING LEARNER PERSONAS IN AI-ASSISTED ENGLISH LANGUAGE LEARNING USING COSINE-BASED CLUSTERING: IMPLICATIONS FOR PERSONALIZED SUPPORT IN GCC CONTEXTS

Marwan Alshar'e [(a)1] *Shaher Elayyan* [(b)] *Abdallah Abualkishik* [(c)] *Khaled Abuhmaidan* [(d)] *Wasin Al Kishri* [(e)]

[(a)]*Associate Professor, Faculty of Computing and IT, Sohar University, Sohar, Oman; E-mail: MShare@su.edu.om*
[(b)]*Assistant Professor, Faculty of Education and Arts, Sohar University, Sohar, Oman; E-mail: SElAyyan@su.edu.om*
[(c)]*Associate Professor, Faculty of Computing and IT, Sohar University, Sohar, Oman; E-mail: AAbualkishik@soharuni.edu.om*
[(d)]*Associate Professor, Faculty of Computing and IT, Sohar University, Sohar, Oman; E-mail: KHmaidan@su.edu.om*
[(e)]*Assistant Professor, Faculty of Computer Studies, Arab Open University, Muscat, Oman; E-mail: wasan.k@aou.edu.om*

ARTICLE INFO

Article History:

Received: 8th October 2025
Reviewed & Revised: 8th October 2025 to 8th April 2026
Accepted: 9th April 2026
Published: 14th April 2026

Keywords:

AI-Assisted Language Learning; English as a Second Language (ESL); Learner Profiling; Unsupervised Learning; Clustering Analysis; Cosine Similarity; Educational Data Mining; Oman Education

JEL Classification Codes:

G32, F65, L66, L25, M41

Peer-Review Model:

External peer review was done through double-blind method.

ABSTRACT

The rapid expansion of artificial intelligence (AI)-assisted English language learning tools has introduced substantial variability in learner outcomes due to differences in behavioural patterns, task engagement, and usage strategies among non-native learners, particularly within Omani educational contexts. This heterogeneity creates a methodological challenge in identifying consistent learner profiles without relying on predefined or subjective labels. This study investigates the effectiveness of unsupervised cosine-based clustering in identifying distinct learner personas in AI-assisted English learning environments. The study utilizes a dataset of 15,000 learner interaction records obtained from Kaggle, incorporating demographic attributes, behavioural features, task modalities, and learning outcome indicators. A structured experimental methodology is employed, beginning with baseline Euclidean K-Means clustering, followed by dimensionality reduction using Singular Value Decomposition (SVD), and subsequent clustering using cosine similarity across multiple algorithms, including K-Means, Gaussian Mixture Models, Agglomerative Clustering, and BIRCH. The results reveal that cosine-based K-Means clustering (k = 6) achieves a Silhouette Score of 0.678 compared to 0.10 for baseline Euclidean clustering, representing an absolute improvement of 0.578 and approximately a sixfold increase in clustering performance. Compared to SVD-based Euclidean clustering (Silhouette = 0.41), cosine similarity improves clustering effectiveness by approximately 65%, while the Davies–Bouldin Index decreases to 0.56 and the Calinski–Harabasz Index increases to 33,074. The findings indicate that cosine-based unsupervised modelling effectively identifies distinct learner personas, demonstrating that learning-gain variations are driven by behavioural interaction patterns rather than usage intensity alone.

INTRODUCTION

Conversational agents, automated grammar checkers, and speech-feedback systems are all artificial intelligence (AI)-powered tools that have become essential for non-native speakers to acquire English (Kasneci et al., 2023). Within the proximities of the Gulf Cooperation Council (GCC), in which English pervades academic circles and the business world, students grow to rely on AI to compensate a lack of authentic communicative practice, as well as to support their self-esteem (Park et al., 2022; Aljehani & Modiano, 2025). Nonetheless, empirical research consistently demonstrates that the use of these tools is not strategic, as learners apply them at any time rather than systematically, resulting in inconsistent learning outcomes and uneven satisfaction (Martín-Moncunill & Alonso Martínez, 2025).

Recent research in school and post-secondary education reveals that, even though AI-assisted language learning tools are becoming increasingly accessible, learners most often use these systems without teacher involvement.

[1]Corresponding author: *ORCID ID: 0000-0001-5187-4902*

https://doi.org/10.46281/bjmsr.v11i2.2852

To cite this article: Alshar'e, M., Elayyan, S., Abualkishik, A., Abuhmaidan, K., & Al Kishri, W. (2026). DISCOVERING LEARNER PERSONAS IN AI-ASSISTED ENGLISH LANGUAGE LEARNING USING COSINE-BASED CLUSTERING: IMPLICATIONS FOR PERSONALIZED SUPPORT IN GCC CONTEXTS. *Bangladesh Journal of Multidisciplinary Scientific Research*, *11*(2), 37-47. https://doi.org/10.46281/bjmsr.v11i2.2852

Consequently, funneling the inputs from AI-generated feedback through trial-and-error or informal peer influence has often been the way of interaction rather than through well-planned instructional strategies. Previous studies have demonstrated that such unassisted interaction may yield diverse learning outcomes, especially in English-as-a-foreign-language settings in GCC countries, including Oman, where classroom exposure is limited and communicative practice is rarely sustained. (Kasneci et al., 2023; Park et al., 2022). There is considerable variation in how learners benefit from AI-supported feedback, mainly due to these conditions. This makes it very important to study the patterns of learner interaction and their learning outcomes before assuming uniform effectiveness.

In recent studies, supervised machine learning methods have been used to predict learning performance and even to prescribe tool use based on prior performance data (Albahli, 2025; Wang et al., 2023). However, these methodologies depend on the presence of clearly defined target labels - the name of an ideal tool. These labels are objectively subjective, situation-specific, and changeable across heterogeneous learner profiles. In the context of language learning with AI, the effectiveness of a specific tool is conditioned by latent behavioural constituents, including engagement patterns, task preferences, and self-regulated learning behaviours, which cannot be explicitly defined (Rosenberg et al., 2022). As a result, a supervisory paradigm risks increasing noise and bias, thereby hindering the extraction of meaningful structural insights. This exposes a methodological gap in discovering stable, interpretable learner profiles in the context of AI-assisted language learning, without reliance on predefined or subjective outcome labels.

To overcome this dilemma, unsupervised learning, especially clustering, offers a principled alternative. Clustering avoids imposing preconceived outcome labels by directly identifying latent learner personas from behavioural and performance attributes (Shaffer & Ruis, 2024). The approach can be used to discover homogeneous groups of learners with similar usage preferences, error modes, and learning benefits, thereby enabling personalised outcomes at the persona level rather than individualised outcome prediction. Similarity-based clustering, particularly the use of cosine similarity in reduced latent spaces, has shown strong results on high-dimensional educational and behavioural data, highlighting relationships rather than absolute values (Fan et al., 2023). Hence, the research aims to reveal hidden learner personas in AI-assisted English learning by implementing an unsupervised clustering framework.

In this research, we propose a clustering-based framework for identifying learner personas in the context of AI-enabled English learning, using a large, publicly accessible dataset. Even though the dataset does not explicitly stratify by region, the behavioural patterns we examine are very similar to the obstacles that non-native English learners in the GCC face, which gives the findings broad generalizability. We evaluate baseline clustering, dimensionality reduction via truncated singular value decomposition (SVD), and cosine-based side-trial clustering across a range of algorithms, using popular internal evaluation standards methodically.

The research treats personalization in AI-supported English language learning as an unsupervised learner-profiling problem. It thus does not rely on predetermined or subjective labels of the best tool usage. It models a systematic clustering framework that integrates feature-aware preprocessing, dimensionality reduction via truncated singular value decomposition (SVD), and similarity-based distance modelling to uncover hidden learner behaviour patterns.

By comparing clustering algorithms and internal validation metrics, the study reveals that cosine-based clustering in reduced latent spaces produces more consistent learner personas than traditional Euclidean-based methods. Besides, the results show that, among non-native English learners, learning gains are not always dependent on levels of tool use, indicating distinct learner profiles in the educational contexts of the GCC. Altogether, this work offers a scalable, data-driven framework on which subsequent AI-based personalization and suggestion systems for English language learning will be built.

This work makes three key contributions. First, it demonstrates that cosine-based clustering applied to low-dimensional latent representations significantly outperforms traditional baseline clustering methods that operate directly on the original feature space, highlighting the effectiveness of representation learning in improving clustering quality. Second, it introduces the concept of learnable learner profiles, which capture distinct patterns of learning gain and user satisfaction, enabling a more nuanced understanding of how different learners interact with educational systems. Third, it establishes a strong conceptual foundation for chatbot-driven personalization mechanisms, emphasizing how such systems can guide learners to use AI tools more effectively, particularly within GCC-related educational contexts.

This study explored the use of unsupervised learning techniques to identify learner personas in AI-supported English language learning, with a particular focus on non-native English learners in educational settings across GCC countries.

The subsequent parts of the paper first set out the methodological framework, then present the results of the experiments, and analyze the learner personas discovered. The final part identifies the key insights gained from the work.

LITERATURE REVIEW

Recent research on the use of AI in English language learning has highlighted the growing heterogeneity of interactions between learners and advanced technological artefacts such as chatbots, automated feedback systems, and speech-evaluation systems (Crompton & Burke, 2023; Zhu & Wang, 2024; Lee & Lim, 2023). Though such tools can regularly provide quantifiable improvements in grammar accuracy, fluency, and learner interaction, the degree of improvement varies significantly across learners, with the most pronounced differences observed among non-native English speakers (Alanazi et al., 2025). Latest systematic evidence also suggests that there is a wide range of fluctuations in learner engagement with AI-assisted educational tools, depending on factors such as interaction styles, perceived usefulness and individual strategies for self-regulated learning, thus highlighting the intrinsically diverse nature of AI-mediated learning experiences (D'Mello & Graesser, 2012; Rebolledo-Méndez et al., 2022; Booth et al., 2023; Granström & Oppi, 2025). This heterogeneity has sparked researchers' interest in learning analytics methods that not only identify observable achievement factors but also

reveal the intricacies of students' behaviour.

The literature on supervised predictive models has explored predicting learning paths and prescribing teaching materials (Kaur et al., 2019). However, the methods rely on clear, reliable outcome labels, a condition that is burdensome in open learning settings, where the concept of success is inherently multidimensional and context-specific (Ferguson, 2019; Da Silva et al., 2023). Empirical studies have reported declines in performance and bias when labels are noisy or when learners have divergent objectives (Baker & Hawn, 2022). Previous studies, for instance, have revealed that student engagement correlates very closely with their emotional states, such as confusion, frustration, and flow, which together vary and determine learning behaviour and outcomes in technology-based learning environments (D'Mello & Graesser, 2012; Viberg, Khalil & Baars, 2020). The results demonstrate that supervisory paradigms fall short of accurately reflecting the intertwined cognitive and affective processes that result from AI-assisted learning. As such, the academic community is trending towards unmonitored and investigatory analytics, seeking to support the heterogeneity of learners without labelling them strictly into a few categories (Watson, 2023). Clustering has become a statistical method used to identify learner profiles based on behavioural, cognitive, and affective aspects (Tudor et al., 2025; Asahara et al., 2009). In the context of language learning, scholars have used clustering to group learners based on engagement, error patterns, and self-regulation (Arumugam et al., 2013; Najem, Seghroucheni, & Ziti, 2026). Engagement with digital learning environments can be assessed through interaction-level data, thereby enabling the use of unsupervised modelling techniques that are not limited to single-performance metrics (Booth et al., 2023). These personas make it easier to implement scalable personalization that does not require personal-level predictions. This feature aligns with the expansive, diverse learner populations typical of GCC educational systems (Holi, 2025). Research on intelligent tutoring systems also indicates that meta-affective behaviours (i.e., how learners emotionally and cognitively respond to AI-generated feedback) play a major role in shaping learning outcomes, underscoring the importance of modelling relational patterns rather than absolute usage levels (Rebolledo-Méndez et al., 2022).

The quality of clustering depends on representation learning and the methodological approach to similarity metrics. Dimensionality-reduction methods such as principal component analysis and truncated singular value decomposition have been shown to increase the separability of clusters in high-dimensional educational data by reducing redundancy and noise (Munassar & Al-hobishi, 2025). Furthermore, clustering based on similarity, especially Cosine similarity, has demonstrated better performance than Euclidean distance in instances where the variance in behavioural intensity among learners is wide-ranging (Shirkhorshidi et al., 2015; Mello & Souza, 2021). The cosine similarity is a measure of relationship patterns based on absolute levels, making it highly suitable for analyzing use and interaction data (Venkatesh Sharma et al., 2024). Although these methodological improvements have been made, numerous studies continue to use a single clustering setup and do not conduct systematic comparisons across preprocessing pipelines, similarity measures, or validation metrics (Beautiful; Melchor et al., 2026). Besides, there is a significant amount of literature that ends at the persona level, without addressing how the identified clusters can be converted into processes accessible to learners. To overcome these inadequacies, the latest literature emphasizes the need for empirically grounded frameworks that link clustering outcomes to adaptive guidance mechanisms, such as conversational agents (Chen et al., 2025; Dorneich et al., 2004). Overall, current research favours the use of unsupervised, similarity-aware clustering methods for representing learner diversity in AI-assisted language learning. However, such studies lack thorough pipeline-level evaluation and do not provide clear guidelines for instructional implementation. Hence, this research paper aims to fill these gaps by systematically comparing clustering pipelines and assessing their relevance for AI-guided personalization in English-language learning contexts.

The research sought to examine how learner personas can be discovered through unsupervised learning within AI-assisted English learning environments, specifically targeting non-native English learners in GCC-based educational contexts. The following are the hypotheses of the study:

H_1: Clustering in reduced latent spaces using cosine similarity produces higher internal validity than clustering based on Euclidean distance in the original feature space.
H_2: Learner task involvement intensity does not exhibit a linear relationship with learning gains among non-native English learners using AI-assisted language tools.

MATERIALS AND METHODS

This research outlines a progressive, unsupervised personalisation system for AI-assisted English language learning, in which learner persona discovery is sequentially optimised via representation learning and similarity-aware clustering. In contrast to previous studies that used only one clustering arrangement, the current formulation systematically interrogates baseline clustering, latent-space modelling, and similarity-led grouping to identify the most discriminative learner representations. It is novel in both (i) the structured comparison of clusters of pipelines and not the analysis of algorithms as an independent entity, and (ii) it has explicitly used cosine similarity in reduced latent spaces to model behavioural patterns that are not scale dependent. This design supports powerful learner persona discovery without relying on noisy or subjective target labels. The Method section describes in detail how the study was conducted, including conceptual and operational definitions of the variables used in the study, Different types of studies will rely on different methodologies; however, a complete description of the methods used enables the reader to evaluate the appropriateness of your methods and the reliability and the validity of your results, It also permits experienced investigators to replicate the study, If your manuscript is an update of an ongoing or earlier study and the method has been published in detail elsewhere, you may refer the reader to that source and give a brief synopsis of the method in this section.

Dataset Description

The data used in this investigation were obtained from Kaggle.com, an open data repository, and include 15,000 records of learners studying the English language with AI. The entries will describe the demographics (e.g., age, level of proficiency), behavioral variables (frequency and duration of use), attributes of tasks (essay writing, grammar correction, speaking practice, vocabulary practice), outcome and satisfaction measures (satisfaction of rating and learning gain score), and linguistic performance variables (number of errors before the AI interventions). It is moderately high-dimensional and heterogeneous, with some numerical, categorical, and ordinal characteristics, and is therefore appropriate for assessing clustering pipelines under real learning-analytic conditions. There are no area-specific identifiers; however, the behavioural patterns are typical of non-native English learners, making the results applicable to GCC learning environments.

Methodological Pipeline

Figure 1 summarises the overall methodological pipeline adopted in this study.

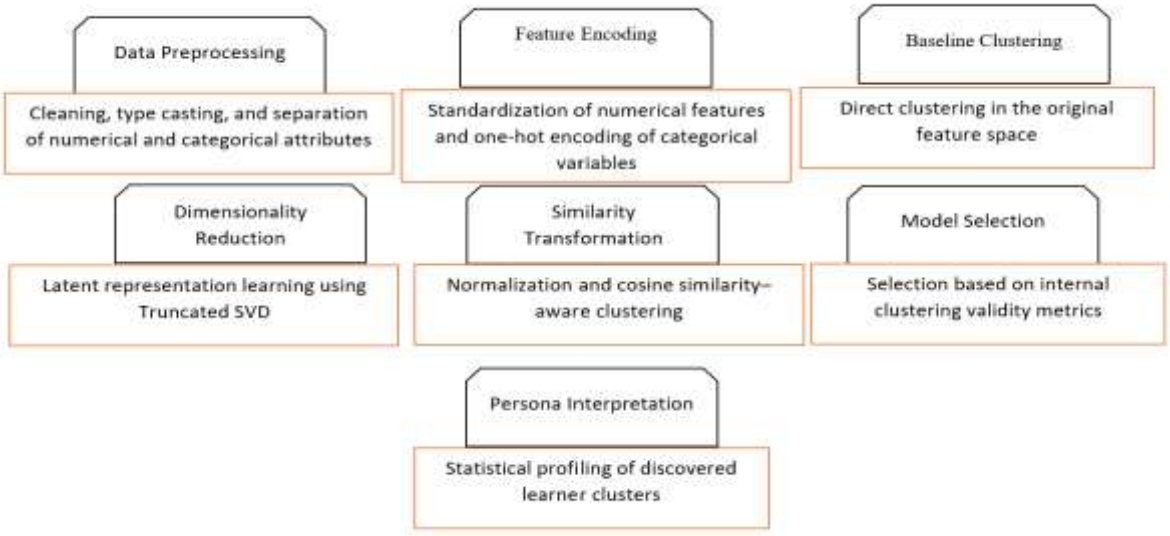

Figure 1. Proposed Clustering-Based Persona Discovery Pipeline

Experimental Design and Configurations

To ensure a fair and systematic comparison, multiple experiments were conducted, each modifying only one methodological component at a time.

Table 1. Experimental Settings and Variations

Experiment	Representation	Distance/ Similarity	Clustering Algorithms	Purpose
Exp-A	Original feature space	Euclidean	K-Means, GMM, Agglomerative, Birch	Baseline clustering
Exp-B	SVD-reduced (2 components)	Euclidean	Same as Exp-A	Effect of dimensionality reduction
Exp-C	SVD-reduced (2 components)	Cosine	K-Means, GMM, Agglomerative, Birch	Similarity-aware clustering
Exp-D	Normalized SVD space	Cosine	K-Means, GMM	Robust persona discovery

Clustering Algorithms

In our exploration, we have methodically examined the various, well-known clustering methods, namely K-Means, Gaussian Mixture Models (GMM), Agglomerative Hierarchical Clustering, and Birch. All these methods have been selected based on the peculiarities of the corresponding methodological paradigms: centroid-based, probabilistic, hierarchical, and scalable clustering. The number of clusters, represented as 'k' in our experiment, could be varied by varying the number of runs of our experiment and, in turn, interrogating the resultant stability and interpretability of each algorithmic method.

Evaluation Metrics

Since no ground-truth labelling is available in our data set, we have relied solely on internal validity indicators of clustering. These measures were used throughout the experimental settings to give a stable platform of comparison.

Table 2. Evaluation Metrics Used in All Experiments

Metric	Purpose	Interpretation
Silhouette Score	Cluster separation and compactness	Higher is better
Davies–Bouldin Index	Intra-/inter-cluster similarity	Lower is better
Calinski–Harabasz Score	Cluster dispersion ratio	Higher is better
Cluster Size Distribution	Stability and balance	Avoids degenerate clusters

All these metrics aim to ensure that the chosen clustering settings yield high-quality learner personas that are distinct, compact, and interpretable.

Summary

The effects of representation learning and similarity modelling on clustering quality are isolated through the proposed methodology, which employs a strictly controlled experimental approach. Not only does this design facilitate transparent model selection, but it also provides a strong basis for further personalization mechanisms, including chatbot-mediated guidance for learners, which are discussed further within the framework of this study.

RESULTS

Experimental Setup

All the experiments were run on the Kaggle-based AI-assisted English learning dataset, which consisted of 15, 000 records of learners. The clustering was performed according to a strictly regulated protocol, in which only a single methodological element varied across runs to provide a fair comparative evaluation. Numerical characteristics were normalized; categorical characteristics were one-hot encoded; and internal validity scores were used to assess clustering performance without reference to ground-truth labels.

Baseline Clustering Results (Exp -A)

The baseline clustering was first applied directly in the original feature space using Euclidean distance. Table 3 shows that none of the clustering algorithms produced strong separation, and all yielded Silhouette Scores below 0.15, indicating high overlap among learner groups.

The results obtained support the idea that raw feature values are insufficient to describe significant learner profiles in mixed behavioural contexts.

Table 3. Baseline clustering performance (Exp-A)

algo	k	n_clusters	n_noise	silhouette	calinski_harabasz	davies_bouldin	extra
KMeans	2	2	0	0.096457257	1570.268058	2.97994193	inertia=104316.35
MiniBatchKMeans	2	2	0	0.103352899	1353.404863	3.150562217	
GaussianMixture	2	2	0	0.029335613	248.2901187	4.502641702	bic=-502092.86
AgglomerativeWard	2	2	0	0.152115955	1095.177784	2.339721969	
Spectral	2	2	0	-0.063639037	89.27144857	2.966965905	
Birch	2	2	0	0.129796324	1183.891593	2.544306357	
KMeans	3	3	0	0.081074636	1346.315942	2.7175886	inertia=97697.21
MiniBatchKMeans	3	3	0	0.080305111	1322.356431	2.73880917	
GaussianMixture	3	3	0	0.037987166	304.0835618	4.373405792	bic=-692310.92
AgglomerativeWard	3	3	0	0.080404054	1051.815279	2.86566872	
Spectral	3	3	0	-0.044405478	377.1120273	4.128491639	
Birch	3	3	0	0.064856834	1001.23869	3.040419217	
KMeans	4	4	0	0.088621992	1410.552152	2.379857863	inertia=89876.50
MiniBatchKMeans	4	4	0	0.087395749	1379.913297	2.419195912	

Figure 2 presents the K-Means elbow curve for selecting the number of clusters based on the inertia. This value shows that the within-cluster inertia decreases with increasing clusters, with a pronounced elbow at k=4-6. Further on, the returns to inertia reduction decrease, which, in turn, implies a realistic trade-off between model complexity and achieving a small cluster structure.

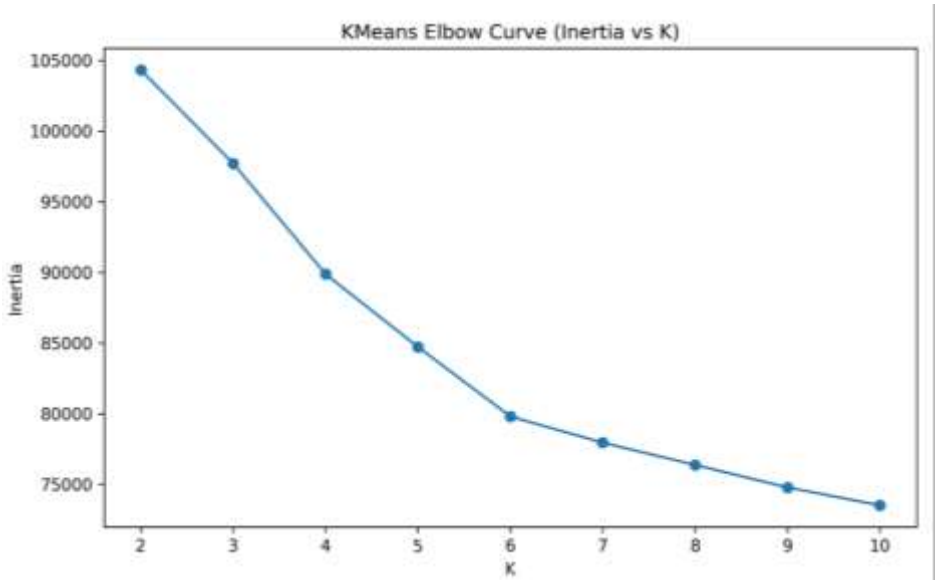

Figure 2. K-Means elbow curve for cluster number selection using inertia

Figure 3 compares silhouette scores for baseline clustering algorithms. This value provides a comparative study of silence scores across the various algorithms used for baseline clustering with different cluster numbers, highlighting the general weak separability under the Euclidean distance. KMeans and MiniBatchKMeans achieve slightly better scores than their counterparts; however, the very poor scores across all methods indicate that more advanced similarity modelling is highly needed.

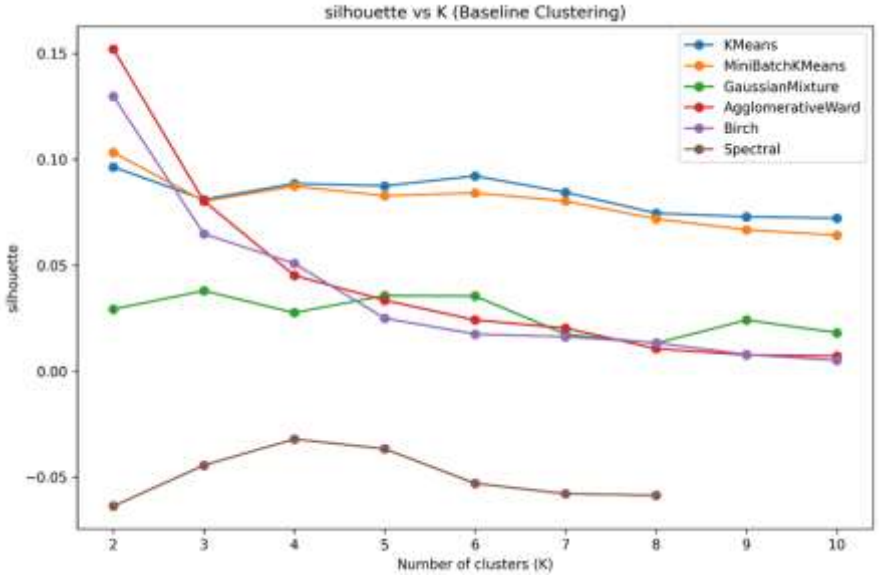

Figure 3. Silhouette score comparison across baseline clustering algorithms

The effect of dimensionality reduction can be determined by examining the difference between two or more datasets, with the reduction being over either one or both dimensions.

Effect of Dimensionality Reduction (Exp-B)

The effect of dimensionality reduction can be quantified by comparing two or more datasets, with the reduction applied to either one or both planes.

To reduce redundancy and noise in the raw feature space, Truncated Singular Value Decomposition (SVD) was used to learn low-dimensional latent representations. According to Table 4, the SVD-based clustering yielded a significant increase in Silhouette scores, ranging from about 0.35 to 0.41. This improvement was only slightly affected by the algorithm used to create the clusters and the number of clusters.

The observed enhancement provides empirical evidence that learners' behaviour is coordinated by a small number of latent factors rather than a set of observable characteristics.

Table 4. SVD-based clustering results (Euclidean distance)

setting	algo	k	silhouette	calinski	davies	n_clusters
svd2_weighted	GaussianMixture	2	0.406451	11526.62	0.964509	2
svd2_weighted	KMeans	2	0.404083	11550.09	0.97122	2
svd2_weighted	GaussianMixture	3	0.379059	9790.607	0.900384	3
svd2_weighted	AgglomerativeWard	2	0.371476	10662.23	1.028107	2
svd2_weighted	KMeans	3	0.354723	10694.22	0.979571	3
svd2_weighted	KMeans	4	0.335077	10472.08	0.916496	4
svd2_weighted	KMeans	7	0.330995	10128.46	0.857216	7
svd2_weighted	GaussianMixture	4	0.327642	9579.234	0.956074	4
svd2_weighted	KMeans	6	0.32649	9981.607	0.914784	6
svd2_weighted	KMeans	10	0.322146	9731.823	0.865344	10
svd2_weighted	KMeans	5	0.320809	10048.27	0.94323	5
svd2_weighted	GaussianMixture	6	0.320707	9583.868	0.917903	6
svd2_weighted	KMeans	8	0.320474	9964.544	0.860181	8
svd2_weighted	GaussianMixture	7	0.31986	9543.839	0.869068	7
svd2_weighted	GaussianMixture	10	0.318948	9475.834	0.850608	10
svd2_weighted	GaussianMixture	5	0.316428	9577.732	0.954197	5
svd2_weighted	GaussianMixture	8	0.311244	9502.363	0.868627	8
svd2_weighted	KMeans	9	0.309111	9576.793	0.885076	9
svd2_weighted	GaussianMixture	9	0.292283	8744.738	0.916221	9
svd2_weighted	AgglomerativeWard	3	0.2691	7959.501	0.977745	3

Cosine similarity-based clustering (Exp -C)

We followed up on the initial experiments by adding cosine similarity to the normalized singular-value decomposition (SVD) representations, thereby highlighting the directionality of behaviour rather than its magnitude. This methodological refinement led to a significant increase in clustering efficacy. Table 5 shows that the Cosine + KMeans algorithm (k=6) had

a Silhouette Score of 0.678, which is significantly higher than the Silhouette Scores for all configurations based on Euclidean metrics.
Figure 4 also outlines the extreme superiority of cosine similarity over a spectrum of cluster cardinalities.

Table 5. Cosine-based clustering comparison

space	algo	k	silhouette	calinski	davies	n_clusters
cosine	KMeans	6	0.677987193	33074.80613	0.560959211	6
cosine	KMeans	10	0.673335171	51061.12224	0.525604254	10
cosine	KMeans	4	0.670873524	23050.78144	0.625390313	4
cosine	KMeans	9	0.670542072	46368.6944	0.535122142	9
cosine	GaussianMixture(on_normalized)	10	0.668678288	50788.23836	0.527943976	10
cosine	KMeans	5	0.667075818	27985.10761	0.607402958	5
cosine	KMeans	7	0.666608699	36753.07373	0.55710218	7
cosine	KMeans	8	0.665411856	41336.02567	0.546710232	8
cosine	KMeans	3	0.665000495	19285.8085	0.702422171	3
cosine	GaussianMixture(on_normalized)	9	0.659692772	45560.24466	0.5392808	9
cosine	GaussianMixture(on_normalized)	8	0.65527984	40542.25065	0.545910603	8
cosine	GaussianMixture(on_normalized)	3	0.654104703	18747.42998	0.692990572	3
cosine	GaussianMixture(on_normalized)	6	0.654075254	31870.43043	0.562795215	6
cosine	KMeans	2	0.654003724	15992.31498	0.87260987	2
cosine	GaussianMixture(on_normalized)	2	0.653988332	15988.92907	0.871987878	2
cosine	Birch(on_normalized)	2	0.652330288	15893.86099	0.874224996	2
cosine	GaussianMixture(on_normalized)	4	0.649099914	22344.40698	0.63161208	4
cosine	Birch(on_normalized)	7	0.645177784	22321.70562	0.692009723	4
cosine	Birch(on_normalized)	6	0.645177784	22321.70562	0.692009723	4
cosine	Birch(on_normalized)	4	0.645177784	22321.70562	0.692009723	4

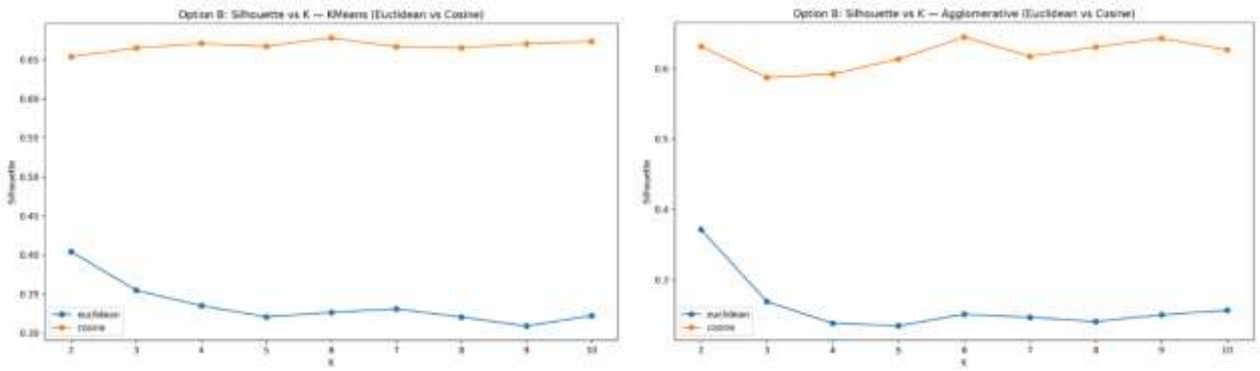

Figure 4. Silhouette score comparison: Euclidean vs Cosine (KMeans)

Learner Persona Analysis (Exp-C, Selected Model)
Using the chosen Cosine + KMeans (k=6) setting, we have considered learner personas based on their behavioural and outcome features. The cluster-wise means of the key variables are summarised in Table 6, whereas Table 7 displays the distribution of task types in the clusters.

The study's findings indicate significant differences in learning achievement, satisfaction, and task preference; thus, the personas identified are quite interpretable.

Table 6. Cluster-wise mean learner characteristics

Cluster	Age (years)	Usage Duration (minutes)	Error Before AI	Satisfaction Rating	Learning Gain Score
0	24.68	20	18.09	3.64	0.19
1	24.81	20.49	17.89	3.62	0.46

The learning-gain score is also significantly higher in Cluster 1. However, the age, duration of use, and levels of satisfaction are similar. Therefore, the identified increase in learning performance is explained by patterns of behavioural engagement, rather than exposure time.

Table 7. Task-type distribution per cluster

Cluster	Essay Writing	Grammar Correction	Speaking Practice	Vocabulary Practice
0	2937	2436	1901	2415
1	1613	1356	1026	1316

Cluster 0 exhibits a more robust pattern of engagement across the gamut of task modalities, implying greater exposure to diverse modalities. However, in opposition, Cluster 1, although with a relatively lower volume of activities, has already been shown to achieve higher learning gains, which reinforces the hypothesis that learning efficiency is not necessarily proportional to task frequency.

Comparative Summary of Experiments

To shed light on the improvements, Table 8 outlines the best setup derived at the next stage of the experiment. This table illustrates the advantage of transitioning from the raw-feature clustering algorithm to a cosine-based similarity model.

To provide a detailed description of the experimental path, Table 8 shows the most efficient clustering configuration obtained at the key experimental stages. The comparison is made between baseline clustering, which uses bare feature representations (Exp -A), clustering in a dimensionally reduced space via singular value decomposition (Exp -B), and clustering using cosine similarity (Exp -C).

This comparative study highlights incremental methodological improvements, including the transition to angular similarity modelling, which significantly increased cluster compactness and separability in place of the magnitude-based Euclidean distance. Despite the baseline experiment providing a reference point for learner grouping, subsequent stages show that preprocessing procedures and the choice of similarity metrics are central to identifying significant learner personas.

In particular, the SVD enhanced heterogeneity and redundancy of features of the feature space (Exp-'B'), leading to moderate improvements in silhouette coefficients. However, the greatest improvement occurs with Exp-C, where cosine similarity helps cluster based on behavioural concordance rather than the absolute magnitude of features. It thus provides the best silhouette coefficient and an optimally balanced Davies-Bouldin index.

Table 8. Summary of Best Clustering Results Across Experimental Stages

Experiment	Method Description	Algorithm	k	Silhouette	Davies–Bouldin	Calinski–Harabasz
Exp-A	Baseline clustering (raw standardized features, Euclidean distance)	KMeans	2	0.1	2.98	1570.27
Exp-B	SVD-reduced clustering (weighted features, Euclidean distance)	Gaussian Mixture	2	0.41	0.96	11526.62
Exp-C	Cosine-based clustering (normalized feature space)	KMeans	6	0.68	0.56	33074.81

Overall, Table 8 diagrammatically outlines the methodological approach followed in this work and provides empirical justification for considering cosine-based similarity the ultimate baseline clustering methodology for profiling learners who do not acquire English as a native speaker.

DISCUSSIONS

The results of this study clearly show that discovering learner personas through unsupervised methods is a sound, methodologically sound substitute for the current supervised prediction paradigm in the context of AI-assisted English-language learning. Baseline clustering in the raw feature space yielded only weak segregation, confirming that raw magnitudes further obscured the heterogeneity of learner behaviours and the intensity of their use.

This pattern is consistent with the existing literature on learning analytics, which reports high scale-dependence in distance calculations that frequently overwhelm latent behavioural structures present in learning data.

The addition of latent representation learning via truncated singular-value decomposition significantly improved clustering fidelity, demonstrating that a small number of underlying factors play a stronger role in determining learner behaviour than a single observable attribute. These latent dimensions almost certainly represent composite aspects such as consistency of engagement, task focus, and outcome orientation, features that are particularly relevant to non-native English learners in GCC settings, who experience learning shaped by time limits, instructional heterogeneity, and unequal exposure to English outside the classroom.

The greatest improvement occurred when cosine similarity was applied to normalized latent representations. Cosine similarity, in contrast to Euclidean distance, highlights the directional profile of learner behaviour rather than the magnitude of the vectors, as in Euclidean distance. This subtlety is particularly relevant in the context of AI-supported learning, where individuals may differ significantly in session duration or engagement frequency yet display similar instructional patterns. The greater effectiveness of cosine-based clustering, in its turn, hints at the fact that a learner's personality is better defined in terms of relative behavioural patterns than the absolute levels of activity, which is a discovery, the implications of which are immediately applicable to personalization frameworks, which should focus on learning approaches, not just on their intensity.

The emerging personas had sellable differences in the dimensions of learning gain, satisfaction, and task preferences, which strengthened the pedagogical credibility of the clustering findings. This avoidance of simplistic high-or-low-performance categories for specific students allows the persona-centred approach to scale personalization, aligning cohort groups with tailor-made instructional resources. This methodology is especially conducive to GCC educational settings, where student groups often face a common language barrier but differ in their learning behaviours and AI tool use.

On the methodological level, the research emphasizes the importance of questioning entire clustering pipelines rather than focusing on disparate algorithms per se. The fact is that representation learning and similarity modelling have a stronger substantive impact on clustering quality than the algorithmic choice itself. In practice, the results provide a solid empirical basis for future learner-centric systems that convert persona knowledge into practical advice.

Overall, this exploration confirms the suitability of cosine-based unsupervised clustering as a reliable reference point for modelling learners in AI-aided English learning, and creates momentum to continue extending the similarity formulations and promoting interactive personalization processes.

CONCLUSIONS

This research aimed to investigate unsupervised learning of learner personas in AI-assisted English-language learning, focusing on non-native English learners in educational environments in GCC countries. Unlike prediction methodologies that require overt ground-truth data, the framework proposed in this paper uses clustering techniques to learn latent behavioural patterns from learners' interaction logs. This method is particularly responsive to the practical characteristics of AI-supported learning platforms, in which pedagogical productions are developed over time, are multidimensional, and often cannot be explicitly annotated.

Since the starting point is baseline clustering in the raw feature space, the study found that the raw behavioural features yield weak, unstable cluster structures. By applying incremental methodological steps, including dimensionality reduction via truncated singular value decomposition (SVD) and a cosine-based similarity kernel, the overall clustering fidelity has been significantly improved. The final cosine-based clustering configuration demonstrated strong internal validity across multiple evaluation measures, thus providing intuitive, understandable personas with noticeable differences in learning improvement, satisfaction levels, and task preferences.

The experimental results confirm that learners' behavioural patterns in AI-based language learning environments are mostly determined by relative interaction patterns rather than by the magnitude of their use. This observation, therefore, supports the use of direction-sensitive similarity measures in educational clustering projects and highlights the central importance of representation learning in building advanced learner models. Pragmatically, the persona taxonomy identified in this study provides a scaffoldable framework for personalization**,** curricular differentiation, and instructional decision-making, especially in multilingual educational contexts common in GCC institutions. Eventually, this paper establishes a robust empirical basis for the concept of unsupervised learner modelling in AI-enhanced English learning. It demonstrates that intelligently designed clustering pipelines can yield valuable educational information without relying on supervised annotations.

Despite its virtues, this investigation has several caveats that warrant mention. First of all, the statistics on which the experimental analyses were based were retrieved from a publicly available Kaggle repository rather than from actual data-gathering by GCC institutions. Although the behavioural attributes are typical of the AI-supported setting of English learning, they may not fully capture the cultural, institutional, and curricular peculiarities that define GCC learners. Specifically, the absence of authentic, large-scale learner interaction data from Omani schools, especially for Cycle Two of the Basic Education system, has been the primary reason the framework could not be empirically validated using real classroom data from Oman. Second, the validation regimen was restricted to intrinsic clustering fidelity indices such as the Silhouette, Calinski-Harabasz and Davies-Bouldin, and thus provided objective measures of structural integrity but not direct measures of pedagogic efficacy or longer learning curves. Outside verification through professional commentary or performance results over time was not achieved at this stage. Third, this study narrowed down to this type of clustering based on aggregated learner behaviour and thus left out the temporal aspects of clustering, namely the progression through clusters or the ebbs and flows of learning strategies over time.

Moreover, the current model does not incorporate linguistic content analysis or fine-grained error diagnostics, both of which could further improve the granularity of persona interpretability.

Future studies will further develop this framework across a continuum of relevant orientations. First and foremost, the clustering channel will be enhanced by more elaborate similarity measures, including weighted and adaptive cosine measures, to refine learner-persona discrimination. Second, incorporating an interactive chatbot-based simulator will be the next step in transforming the identified personas into the working learning interventions. This would project cluster affiliation into instantaneous learning suggestions, adaptive task selection, and explicable feedback tailored to particular learners. Third, future research will incorporate temporal modelling to track how learners transition between personas, thus supporting early responses and tailored learning path optimization. Finally, data gathered by GCC universities and colleges will be used to empirically validate the proposed paradigm, thereby increasing its generalizability and pragmatic implications.

Author Contributions: Conceptualization**,** M.A. and S.E.; Methodology, M.A.; Software, M.A.; Validation, M.A. and A.A.; Formal Analysis, M.A.; Investigation, M.A. and S.E.; Resources, M.A.; Data Curation, M.A.; Writing – Original Draft Preparation, M.A.; Writing – Review & Editing, M.A., S.E., A.A., K.A., and W.A.K.; Visualization, M.A.; Supervision, M.A.; Project Administration, M.A.; Funding Acquisition, M.A. Authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Ethical review and approval were waived for this study because the research does not involve vulnerable groups or sensitive issues.

Funding: The authors received funding for this research from MoHERI.

Acknowledgements: The research leading to these results has received funding from the Ministry of Higher Education, Research and Innovation (MoHERI) of the Sultanate of Oman under BFP. MoHERI Block Funding Agreement No.: MoHERI/BFP/SU/2024.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process: During the preparation of this work, the author(s) used Grammarly for proofreading and spell checking since the Authors are not native speakers. All intellectual content, analysis, and interpretations were produced solely by the authors. After using this AI tool/service, the author(s) reviewed and edited the content as needed, taking full responsibility for the publication's content.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

Alanazi, M., Soh, B., Samra, H., & Li, A. (2025). The Influence of Artificial Intelligence Tools on Learning Outcomes in Computer Programming: A Systematic Review and Meta-Analysis. *Computers, 14*(5), 185. https://doi.org/10.3390/computers14050185

Albahli, S. (2025). Advancing Sustainable Educational Practices Through AI-Driven Prediction of Academic Outcomes. *Sustainability, 17*(3), 1087. https://doi.org/10.3390/su17031087

Aljehani, K., & Modiano, M. (2025). The impact of English medium instruction in the Gulf: A comparative study of KSA and UAE. *Cogent Education, 12*(1), 2479402. https://doi.org/10.1080/2331186X.2025.2479402

Arumugam, N., Rafik-Galea, S., Mello, G. D., & Dass, L. C. (2013). Cultural Influences on Group Learning in an ESL Classroom. *Review of European Studies, 5*(2), 81-89. https://doi.org/10.5539/res.v5n2p81

Asahara, A., Sato, A., & Maruyama, K. (2009). Evaluation of Trajectory Clustering Based on Information Criteria for Human Activity Analysis. 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware, 329–337. https://doi.org/10.1109/MDM.2009.65

Baker, R. S., & Hawn, A. (2022). Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education, 32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Booth, B. M., Bosch, N., & D'Mello, S. (2023). Engagement detection and its applications in learning: A selective review. *Proceedings of the IEEE, 111*(9), 1026–1046. https://doi.org/10.1109/JPROC.2023.3309560

Chen, D.-L., Aaltonen, K., Lampela, H., & Kujala, J. (2025). The Design and Implementation of an Educational Chatbot with Personalized Adaptive Learning Features for Project Management Training. *Technology, Knowledge and Learning, 30*(2), 1047–1072. https://doi.org/10.1007/s10758-024-09807-5

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education, 20*, 22. https://doi.org/10.1186/s41239-023-00392-8

Da Silva, F. L., Slodkowski, B. K., Da Silva, K. K. A., & Cazella, S. C. (2023). A systematic literature review on educational recommender systems for teaching and learning: Research trends, limitations and opportunities. *Education and Information Technologies, 28*(3), 3289–3328. https://doi.org/10.1007/s10639-022-11341-9

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

Dorneich, M., Whitlow, S., Ververs, P. M., Carciofini, J., & Creaser, J. (2004). Closing the Loop of an Adaptive System with Cognitive State. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 48*(3), 590–594. https://doi.org/10.1177/154193120404800367

Fan, C.-I., Shie, C.-H., Tseng, Y.-F., & Huang, H.-C. (2023). An Efficient Data Protection Scheme Based on Hierarchical ID-Based Encryption for MQTT. *ACM Transactions on Sensor Networks, 19*(3), 1–21. https://doi.org/10.1145/3570506

Ferguson, R. (2019). Ethical Challenges for Learning Analytics. *Journal of Learning Analytics, 6*(3), 25–30. https://doi.org/10.18608/jla.2019.63.5

Granström, M., & Oppi, J. (2025). *Student engagement with AI tools in learning: Evidence from recent educational contexts*. *Frontiers in Education*, 10, 1298456. https://doi.org/10.3389/feduc.2025.1688092

Holi, H. I. (2025). In-Class EMI Challenges Arising in an Arabian Gulf Engineering Programme. *SAGE Open, 15*(3), 21582440251367125. https://doi.org/10.1177/21582440251367125

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kaur, M., Dhalaria, M., Sharma, P. K., & Park, J. H. (2019). Supervised Machine-Learning Predictive Analytics for National *Quality of Life Scoring. Applied Sciences, 9*(8), 1613. https://doi.org/10.3390/app9081613

Lee, K.-A., & Lim, S.-B. (2023). Designing a Leveled Conversational Teachable Agent for English Language Learners. *Applied Sciences, 13*(11), 6541. https://doi.org/10.3390/app13116541

Martín-Moncunill, D., & Alonso Martínez, D. (2025). Students' Trust in AI and Their Verification Strategies: A Case Study at Camilo José Cela University. *Education Sciences, 15*(10), 1307. https://doi.org/10.3390/educsci15101307

Melchor, F., Conejero, J. M., Fernández-García, A. J., Sánchez-Figueroa, F., & Rodríguez-Echeverría, R. (2026). An empirical evaluation of clustering processes for early detection of university dropout. *International Journal of Data Science and Analytics, 22*, 25. https://doi.org/10.1007/s41060-025-00965-y

Mello, F. L. D., & Souza, S. A. D. (2021). Decision Maker Profiling Using Their Mental Behavior Pattern. *Frontiers in Psychology*, 12, 667255. https://doi.org/10.3389/fpsyg.2021.667255

Munassar, N. M. A., & Al-hobishi, M. A. A. (2025). Dimensionality Reduction Techniques in Big Data and Their Impact on E-Learning. *Journal of Science and Technology, 30*(7), 12–28. https://doi.org/10.20428/jst.v30i7.3002

Najem, K., Seghroucheni, Y. Z., & Ziti, S. (2026). Behavioral clustering for adaptive learning: A data-driven alternative to static learning style models. *International Journal of Information and Education Technology, 16*(1), 196–204. https://doi.org/10.18178/ijiet.2026.16.1.2494

Park, S., Kim, S.-Y., Lee, H., & Kim, E. G. (2022). Professional development for English-medium instruction professors at Korean universities. *System*, 109, 102862. https://doi.org/10.1016/j.system.2022.102862

Rebolledo-Méndez, G., Huerta-Pacheco, S., Baker, R. S., & du Boulay, B. (2022). *Meta-affective behaviour within an intelligent tutoring* system. *International Journal of Artificial Intelligence in Education, 32*(1), 81–112.

https://doi.org/10.1007/s40593-021-00247-1

Rosenberg, J. M., Schultheis, E. H., Kjelvik, M. K., Reedy, A., & Sultana, O. (2022). Big data, big changes? The technologies and sources of data used in science classrooms. *British Journal of Educational Technology, 53*(5), 1179–1201. https://doi.org/10.1111/bjet.13245

Shaffer, D. W., & Ruis, A. R. (2024). Theories All the Way Across: The Role of Theory in Learning Analytics and the Case for Unified Methods. In K. Bartimote, S. K. Howard, & D. Gašević (Eds.), Theory Informing and Arising from Learning Analytics (pp. 187–201). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-60571-0_12

Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. PLOS ONE, 10(12), e0144059. https://doi.org/10.1371/journal.pone.0144059

Tudor, I., Holenko Dlab, M., Đurović, G., & Horvat, M. (2025). Using Clustering Techniques to Design Learner Personas for GenAI Prompt Engineering and Adaptive Interventions. Electronics, 14(11), 2281. https://doi.org/10.3390/electronics14112281

Venkatesh Sharma, K., Ayiluri, P. R., Betala, R., Jagdish Kumar, P., & Shirisha Reddy, K. (2024). Enhancing query relevance: Leveraging SBERT and cosine similarity for optimal information retrieval. International Journal of Speech Technology, 27(3), 753–763. https://doi.org/10.1007/s10772-024-10133-5

Viberg, O., Khalil, M., & Baars, M. (2020). *Self-regulated learning and learning analytics in online learning environments: A review of empirical research*. *Computers & Education*, 156, 103878. https://doi.org/10.1016/j.compedu.2020.103878

Wang, S., Ren, J., & Bai, R. (2023). A semi-supervised adaptive discriminative discretization method that improves the discrimination power of regularised naive Bayes. *Expert Systems with Applications*, 225, 120094. https://doi.org/10.1016/j.eswa.2023.120094

Watson, D. S. (2023). On the Philosophy of Unsupervised Learning. *Philosophy & Technology, 36*(2), 28. https://doi.org/10.1007/s13347-023-00635-6

Zhu, M., & Wang, C. (2024). A Systematic Review of Artificial Intelligence in Language Education from 2013 to 2023: Current Status and Future Implications. https://doi.org/10.2139/ssrn.4684304

Publisher's Note: CRIBFB stays neutral about jurisdictional claims in published maps and institutional affiliations.