Exploring large language models for persona construction: a case study in educational ergonomics

Explorando modelos de linguagem de grande escala para a construção de personas: um estudo de caso em ergonomia educacional

Jefferson Lewis Velasco, Júlio Monteiro Teixeira, Eugênio Diaz Merino

large language models, data-driven personas, clustering, information design, ergonomics in education

This study investigates the potential of large language models (LLMs), specifically ChatGPT 4o, to support the creation of data-driven personas using a real-world dataset on ergonomic awareness among educators in Ho Chi Minh City, Vietnam. The proposed workflow combined automated data cleaning, clustering using *k*-prototypes algorithm, statistical testing, and narrative persona generation. Three distinct educator profiles were identified, each differing in age, teaching status, teaching experience, and musculoskeletal discomfort. These personas highlight relevant risk factors and opportunities for targeted ergonomic interventions in educational settings. The process demonstrated that LLMs can lower the technical barriers typically associated with quantitative methods and narrative construction, making persona development more accessible in low-resource environments. Although the dataset was limited in size and depth, and expert validation of statistical procedures was not performed, the approach proved feasible and reproducible. The study suggests that combining clustering algorithms with LLM-supported interpretation can bridge analytical rigor and human-centered storytelling. This hybrid method holds promise for design researchers and practitioners seeking scalable, interpretable, and empathetic tools for understanding user groups in small organizations and businesses. It also opens new directions for applying generative AI in occupational health, education and information design contexts where data interpretation and stakeholder communication are essential.

modelos de linguagem de grande escala, personas orientadas por dados, agrupamento, design da informação, ergonomia na educação

Este estudo investiga o potencial dos modelos de linguagem de grande escala (LLMs), especificamente o ChatGPT 4o, para apoiar a criação de personas orientadas por dados a partir de um conjunto de dados real sobre a conscientização ergonômica entre educadores na cidade de Ho Chi Minh, Vietnã. O fluxo de trabalho proposto combinou limpeza automatizada de dados, agrupamento utilizando o algoritmo k-prototypes*, testes estatísticos e geração narrativa de personas. Foram identificados três perfis distintos de educadores, cada um diferindo em idade, situação de ensino, experiência docente e desconforto musculoesquelético. Essas personas destacam fatores de risco relevantes e oportunidades para intervenções ergonômicas direcionadas em contextos educacionais. O processo demonstrou que os LLMs podem reduzir as barreiras técnicas normalmente associadas a métodos quantitativos e à construção narrativa, tornando o desenvolvimento de personas mais acessível em ambientes com poucos recursos. Embora o conjunto de dados fosse limitado em tamanho e profundidade, e não tenha sido realizada validação especializada dos* *procedimentos estatísticos, a abordagem mostrou-se viável e reprodutível. O estudo sugere que a combinação de algoritmos de agrupamento com interpretação apoiada por LLMs pode estabelecer uma ponte entre o rigor analítico e a narrativa centrada no ser humano. Esse método híbrido apresenta potencial para pesquisadores e profissionais de design que buscam ferramentas escaláveis, interpretáveis e empáticas para compreender grupos de usuários em pequenas organizações e empresas. Também abre novas direções para a aplicação de IA generativa em contextos de saúde ocupacional, educação e design da informação, nos quais a interpretação de dados e a comunicação com as partes interessadas são essenciais.*

1 Introduction

In contemporary design practices, the effective representation of data is relevant not only for analytical clarity but also for fostering empathy toward the people behind the data (Gill et al., 2023). Particularly within the field of information design, the challenge lies in transforming complex and often abstract datasets into meaningful narratives that meet stakeholder expectations and support user-centered decision-making (Alhamadi et al., 2022). Personas – semi-fictional representations of user groups – have long served as a tool to bridge the gap between raw data and empathetic user centered design (Grudin & Pruitt, 2002).

Despite their widespread use, persona construction often relies on heuristic or intuitive approaches, limited qualitative insights and non-optimized quantitative segmentation. While such methods can produce useful archetypes, they frequently lack methodological rigor, resulting in biased or imprecise representations. This lack of precision can misguide design decisions, particularly in data-rich contexts where subtle distinctions between user groups are critical (Jansen et al., 2022).

In this context, the integration of computational techniques – such as clustering algorithms supported by large language models (LLMs) – offers promising alternatives for refining persona-building processes. By leveraging LLMs to support data handling tasks, such as generating Python code for cluster analysis, it becomes possible to construct personas that are data-driven, scalable, and interpretable while reducing the operational complexity of traditional quantitative approaches (Velasco et al., 2025).

To explore this potential, we utilized the dataset "Ergonomic Awareness and MSDs Among Educators", available on Kaggle ([ghret76yh], 2024). This dataset includes responses from 300 educators based in Ho Chi Minh City, Vietnam, capturing demographic variables, ergonomic knowledge, and self-reported experiences of musculoskeletal discomfort ([ghret76yh], 2024). The richness and diversity of the data offer a foundation for identifying meaningful user segments based on both behavioral and experiential factors – making it particularly suitable for testing the application of LLMs in data-driven persona generation.

By applying clustering techniques with the aid of LLMs to this dataset, we aim to assess patterns that can inform the creation of precise personas with less effort. These personas, in turn, may serve as strategic tools in the development of targeted educational programs, workplace policies, and ergonomic interventions tailored to the needs of distinct educator profiles or inspire the use of such procedures in completely different fields. In addition, the workflow may support information design practices by facilitating the development of visual summaries, dashboards, and communication artifacts derived from statistically grounded user profiles. This may help multidisciplinary teams translate analytical outputs into actionable design strategies.

Therefore, the main goal of this work is to evaluate the effectiveness of large language models (LLMs) in supporting the development of data-driven personas by generating Python code for cluster analysis, using a dataset on ergonomic awareness among educators. Secondary objectives are:

- To identify and interpret user segments within the dataset;
- To test the ability of LLMs to autonomously generate executable clustering code and guide the persona-building process;
- To explore the potential of combining LLM generated narrative structures with quantitative clustering outcomes;
- To reflect on the implications of integrating LLMs into persona design practices.

In regard to the work of educators, apart from the persona subject by itself, it is our understanding that teaching is a profession marked by significant physical and cognitive demands which, if inadequately addressed, can negatively impact both the health and performance of educators. In contexts where ergonomic awareness is limited and work conditions are suboptimal, the prevalence of musculoskeletal disorders (MSDs) among teachers becomes a critical issue (Solis-Soto et al., 2017).

Given the role educators play in the learning process, it is important to promote working environments that support their physical and mental well-being. Identifying risk profiles and understanding the specific needs of different educator segments can provide a structure for evidence-based interventions (Erick & Smith, 2011). In this sense, the use of data-driven personas offers means to translate complex survey data into actionable insights. Therefore, this study is also motivated by the potential to apply advanced computational tools – particularly large language models – to enhance persona construction and support more targeted, empathetic, and effective strategies for improving occupational health in education.

2 Theory

2.1 Personas

Personas have originated in the field of human-computer interaction and are defined as hypothetical archetypes of real users, developed to support design decisions by embodying user goals, behaviors, as defined by Cooper (2004). This concept was expanded by Garrett (2011) who emphasized personas are tools for synthesizing mental models – integrating expectations, experiences, and behavioral patterns – into accessible, communicative formats. These constructs often integrate factual attributes (such as age and professional background) with personal and attitudinal aspects (like goals and beliefs), forming multimodal narratives that combine written descriptions, visual elements, and contextual scenarios to support empathy and mutual comprehension within cross-functional teams (Jansen et al., 2021).

In addition to their application in user experience design, personas have been adopted in areas such as healthcare, economics, public policy, and education (McGinn & Kotamraju, 2008; Salminen et al., 2020). In these contexts, personas act as tools for translating data into accessible narratives, helping stakeholders interpret complex information, support communication, and coordinate decision-making. When developed from empirical data, data-driven personas can serve as intermediaries that assist in representing user diversity in ways that are easier to interpret and apply.

From a semiotic perspective, Manovich (2002) describes interfaces as symbolic cultural constructs that mediate how people interact with data. Building on this idea, Velasco et al. (2023) suggest that personas can function as narrative interfaces – connecting quantitative data with qualitative interpretation. In this view, data-driven personas go beyond simple outputs of segmentation models and act as communicative constructs that make abstract data more relatable through simplified, human-centered representations.

As the use of personas expands across academic and professional activities, it becomes relevant to examine how these representations are constructed and validated. The methods used to create personas influence their consistency, interpretability, and alignment with ethical and practical goals. The next section discusses how computational techniques, including clustering algorithms and workflows supported by large language models (LLMs), may contribute to the development of more consistent and interpretable personas based on structured data.

2.2 Data-driven persona construction

While personas are often developed from qualitative insights and empirical knowledge, data-driven personas are constructed through the systematic collection and analysis of user data. According to Gaiser et al. (2006), personas that aim to support research or decision-making should be grounded in verifiable data – ideally integrating both qualitative and quantitative sources in order to maintain methodological consistency. Similarly, Grudin and Pruitt (2002) emphasize the importance of deriving personas from actual behavioral patterns captured in structured datasets.

The construction of data-driven personas usually involves the use of clustering algorithms, dimensionality reduction techniques, and other multivariate analysis methods to identify distinct user segments (Salminen et al., 2020). These procedures enable the grouping of individuals with similar characteristics, forming the basis for persona development. However, the application of these techniques often requires specialized knowledge in statistics and access to computational tools, which may limit their use in resource-constrained environments. In such cases, organizations frequently adopt empirical or informal methods for persona creation (Norman, 2008).

These alternative approaches may rely on intuition, assumptions, or internal discussions rather than structured analysis. Although easier to implement, they can result in oversimplified or inconsistent representations of the target audience, with limited ability to reflect the diversity and complexity present in user data (Güneş & Ercömert Görgün, 2022). Inaccurate personas may lead to practical challenges, such as a mismatch between services and user expectations, reinforcement of stereotypes, or suboptimal use of resources in design and communication strategies (Guan et al., 2024). These limitations highlight the relevance of developing more accessible, yet analytically sound, methods for persona construction that balance usability with data integrity.

2.3 Persona construction with clustering algorithms and large language models

The construction of data-driven personas usually relies on segmentation techniques that group users based on shared attributes or behaviors. Among these, clustering algorithms such as *k*-means, hierarchical clustering, and principal component analysis (PCA) are frequently used to identify meaningful user segments from structured datasets (Salminen et al., 2020; Salminen et al., 2021). These techniques allow practitioners to scale persona generation across large samples by automatically classifying individuals into statistically derived profiles. This process addresses key limitations of manual persona creation, including subjectivity, cost, and limited scalability.

In their systematic review, Salminen et al. (2021) highlight that clustering methods are central to data-driven persona development, particularly in contexts involving large user populations. The use of clustering enables the identification of consistent behavioral patterns that serve as a foundation for persona synthesis. The authors point out that despite the diversity of applied techniques, segmentation remains a core step in constructing statistically grounded and replicable personas. This is also supported by Jenkinson (1994) who argues that segmentation should follow a logic where users are grouped by their similarities rather than segregated by their differences.

Building on these computational approaches, recent advances in generative AI – particularly large language models (LLMs) – introduce new possibilities for persona construction. Amin et al. (2025) identify that LLMs are now being used not only to assist in coding tasks and data processing, but also to automate multiple stages of persona development, including data segmentation, attribute enrichment, and narrative generation. Unlike earlier rule-based systems, LLMs demonstrate interpretive capabilities, allowing them to transform cluster outputs into coherent, context-sensitive persona descriptions.

Furthermore, LLMs offer practical advantages in low-resource environments. By prompting LLMs to generate clustering code, explain clustering results, or compose persona narratives, teams without deep statistical expertise can still engage in computational persona construction. Amin et al. (2025) emphasize that this broadens access to computational persona construction for organizations with limited technical expertise or infrastructure, while also supporting reproducibility and adaptability.

Nonetheless, challenges remain. Salminen et al. (2020) caution that methodological fragmentation and a lack of best practices hinder standardization. Similarly, Amin et al. (2025) note that GenAI-generated personas still face evaluation challenges related to consistency, representativeness, and ethical oversight. Despite these limitations, the convergence of clustering algorithms and LLMs suggests a promising hybrid approach: clustering techniques ensure analytical rigor, while LLMs contribute to interpretability and communication – key requirements in information design and user-centered research.

3 Methods

This study employed an exploratory computational approach to construct data-driven personas with the support of a large language model (LLM), ChatGPT 4o. The workflow followed seven sequential stages integrating dataset selection, clustering analysis, statistical evaluation, and persona narrative construction:

- **Persona purpose**: the first stage defined the objective of the personas as to represent educator profiles with varying ergonomic awareness and susceptibility to musculoskeletal disorders (MSDs).
- **Data selection**: the study used the publicly available dataset "Ergonomic Awareness and MSDs Among Educators" ([ghret76yh], 2024) from Kaggle, containing 300 responses from educators in Ho Chi Minh City, Vietnam. Variables included demographics, ergonomic knowledge indicators, and self-reported MSD symptoms.
- **Data cleaning and clustering technique selection**: ChatGPT 4o was prompted to explore the dataset, identifying variable types, missing values, and distribution patterns. Through iterative guidance, the model recommended the *k*-prototypes algorithm for clustering mixed-type data, combining *k*-means for numerical and *k*-modes for categorical variables (Salminen et al., 2020).
- **Optimal number of clusters**: after selecting the clustering method, ChatGPT 4o was prompted to generate Python code to identify the optimal number of clusters. The code implemented the elbow method to determine the best configuration. All Python code was run locally using Microsoft VS Code.
- **Clustering**: once the number of clusters was defined, the LLM was prompted to generate Python scripts to execute the *k*-prototypes clustering. The clustering process grouped educators into distinct segments based on their characteristics. Cluster outputs included both the assignment of each individual to a cluster and descriptive statistics for each segment.
- **Interpretation and statistical testing**: the resulting clusters were analyzed to identify key differences among groups in order to evaluate

the performance of the clustering process. ChatGPT 4o performed appropriate statistical tests, including the Kruskal-Wallis *H* test and ANOVA test for numerical variables and Chi-Square tests for categorical variables. Beyond generating executable code, the LLM actively supported the interpretation of the PCA outputs, bridging the technical results with human-centered insights.
- **Persona narrative construction**: finally, the clustered data were translated into narrative personas as the LLM generated human-centered descriptions – mini-biographies, key needs, and frustrations – integrating demographic attributes, ergonomic awareness, and MSD risk patterns.

By integrating clustering algorithms with LLM-generated code and narratives, this workflow proposes a reproducible approach to persona construction that reduces the need for advanced programming and statistical knowledge as the LLM enabled a viable transition from raw data to human-centered personas.

ChatGPT 4o was also used for language and text cohesion reviews, as well as a programming assistant during the development of the Python code employed in the experiment. All outputs were reviewed and validated by the authors. The originality, scientific content, methodological decisions, interpretation of results, and responsibility for the manuscript remain exclusively with the authors.

4 Results

As outlined in the Methods section, the initial stages of this study established the purpose of the personas and selected a suitable dataset for analysis. Building upon these, the following results focus on the subsequent steps of the workflow, which were supported by LLM interactions. By presenting these outcomes sequentially, we highlight the practical contributions of LLMs in the process.

The raw dataset was submitted to ChatGPT 4o for inspection and cleaning, including verification of variable types, identification of missing values, and detection and treatment of such inconsistencies. Given the dataset's mix of numerical and categorical variables, the LLM recommended the *k*-prototypes algorithm, a hybrid approach that extends *k*-means to accommodate categorical data by integrating the distance calculations of *k*-modes. This method is consistent with the findings of Salminen et al. (2020) and allowed the grouping of educators into coherent clusters suitable for subsequent persona construction.

After selecting *k*-prototypes as the clustering method, ChatGPT 4o was prompted to generate Python code to identify the optimal number of clusters, as required by *k*-prototypes. The code implemented the elbow method, which evaluates the within-cluster sum of squares (WCSS) to detect the point of diminishing returns as the number of clusters increases. The analysis indicated that three clusters (Figure 1) minimized the increase in clustering cost beyond which adding additional clusters yielded only marginal improvements. This configuration was selected as optimal for the clustering process.

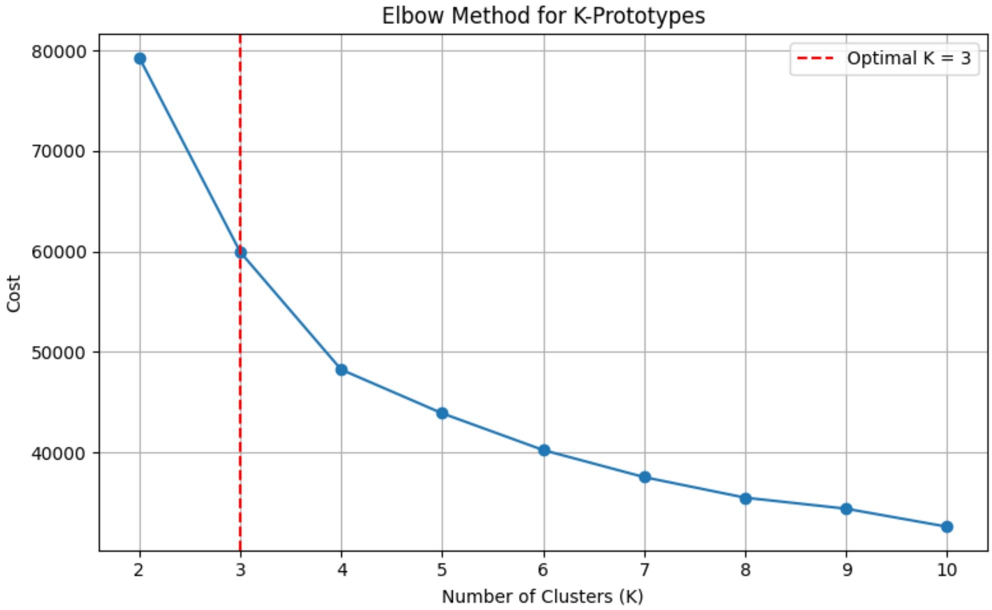

Figure 1 Optimal number of clusters.

With the optimal number of clusters defined, ChatGPT 4o was prompted to generate Python scripts to execute the *k*-prototypes algorithm. This process assigned each educator to one of the three clusters, producing a segmentation that reflected similarities in the sample data. A two-dimensional PCA visualization of the results is presented in Figure 2. The outputs included both individual cluster assignments and descriptive statistics for each group.

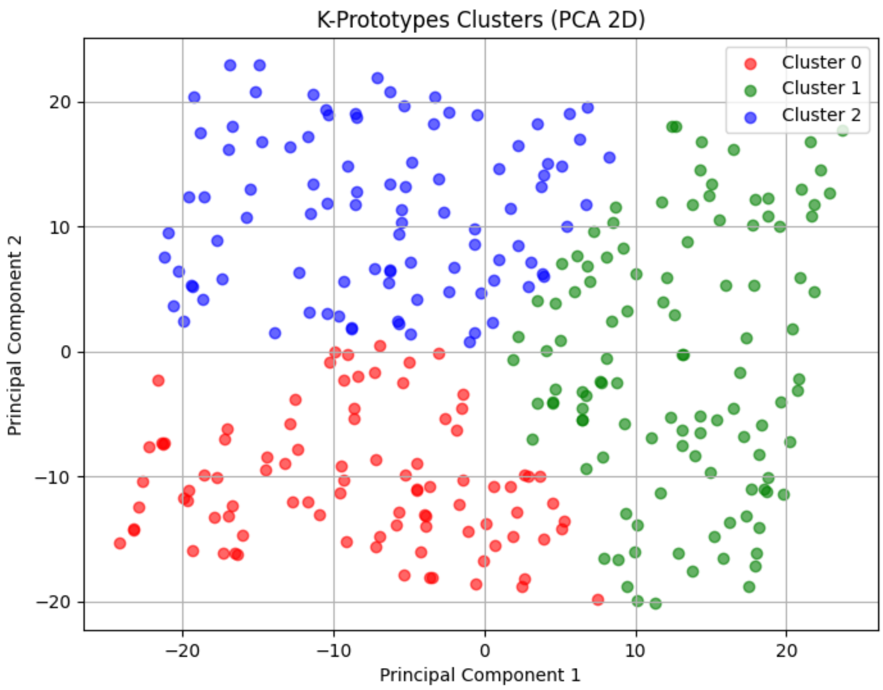

Figure 2 Graphic representation of clusters.

Following the clustering process, the three segments were evaluated to identify the variables most responsible for their differentiation. ChatGPT 4o performed Kruskal-Wallis *H* tests and one-way ANOVA for numerical variables, and Chi-Square tests for categorical variables. These tests confirmed that differences among clusters were statistically significant for age, teaching status, and years of teaching experience, finding meaningful patterns rather than random variation.

The detailed outcomes of these tests are summarized in Table 1, highlighting age, teaching status and years of teaching experience as the most discriminative variables. These results provided the analytical groundwork for translating the clusters into narrative personas.

Table 1 Statistical testing for cluster variation.

Variable	Type	Test	Statistic	*P*-value	Interpretation
Gender	Categorical	Chi-Square	0,6487037795	0,7229957828	Not significant
Age	Numerical	ANOVA	278,5283721	0	Highly significant
Nationality	Categorical	Chi-Square	4,30269154	0,6357868101	Not significant
Marital status	Categorical	Chi-Square	2,740068024	0,8406920595	Not significant
Number of children	Ordinal/Discrete	Kruskal-Wallis	4,697678522	0,09547992517	Not significant
Education level	Categorical	Chi-Square	7,207513295	0,3020822565	Not significant
Teaching status	Categorical	Chi-Square	16,82383034	0,009952973041	Highly significant
School type	Categorical	Chi-Square	1,366831182	0,5048895477	Not significant
District location	Categorical	Chi-Square	4,679015863	0,7912674643	Not significant
Years of teaching experience	Numerical	ANOVA	185,3227671	0	Highly significant
Subjects taught	Ordinal/Discrete	Kruskal-Wallis	0,4843820067	0,7849062435	Not significant
Weekly teaching hours	Numerical	ANOVA	0,8996848733	0,4078032431	Not significant
Hours spent standing per day	Ordinal/Discrete	Kruskal-Wallis	1,246860522	0,5361023091	Not significant
Hours spent sitting per day	Ordinal/Discrete	Kruskal-Wallis	2,610074493	0,271162435	Not significant
Awareness of ergonomics definition	Categorical	Chi-Square	6,073241573	0,1937449098	Not significant
Awareness of acceptable sitting posture duration	Categorical	Chi-Square	5,11545331	0,275655083	Not significant
Understanding of back pain from lifting loads	Categorical	Chi-Square	9,164415238	0,05711872457	Not significant
Importance of neck/head rest in chairs	Categorical	Chi-Square	2,432516064	0,6567592083	Not significant
Belief in standing too much causing back pain	Categorical	Chi-Square	6,698649383	0,1526963299	Not significant
Knowledge of age as a risk factor	Categorical	Chi-Square	4,535238179	0,3383895687	Not significant
Awareness of gender susceptibility	Categorical	Chi-Square	2,747787084	0,6008787631	Not significant
Ergonomic equipment provided	Categorical	Chi-Square	6,149563213	0,1882537571	Not significant
Neck pain frequency	Categorical	Chi-Square	14,02953569	0,08099882037	Not significant
Neck pain severity	Categorical	Chi-Square	8,771354967	0,1868511003	Not significant
Interference with work due to neck pain	Categorical	Chi-Square	3,882187104	0,4221848265	Not significant

The fact that age, teaching status, and years of experience emerged as the most discriminative variables not only through statistical testing but also as highlighted in the reviewed literature reinforces the robustness of the analytical process.

Finally, the cluster summaries were submitted to ChatGPT 4o to generate persona descriptions. Each persona incorporated a concise biography, along with needs and frustrations, reflecting demographic attributes, ergonomic awareness, and musculoskeletal risk profiles. This process produced three distinct personas, each representing a cluster, that provided an interpretable and actionable representation of the educator segments. As a sample, a synthesis of the first persona descriptive analysis is presented in Table 2. This persona is represented in Figure 3.

Table 2 Descriptive analysis of a cluster (first persona).

Variable	top	%	mean	std. dev.	min	max
Gender	male	52,17%				
Age			36,98913043	8,407864666	22	54
Nationality	american	35,87%				
Marital status	married	30,43%				
Number of children			2,304347826	1,419945289	0	4
Education level	phd	32,61%				
Teaching status	internship	36,96%				
School type	private (international/bilingual)	51,09%				
District location	district 2	25,00%				
Years of teaching experience			7,826086957	4,960522606	1	19
Subjects taught			2,815217391	1,366228046	1	5
Weekly teaching hours			31,02173913	6,403944974	20	40
Hours spent standing per day			5,206521739	2,799285794	1	10
Hours spent sitting per day			5,293478261	2,903344201	1	10
Awareness of ergonomics definition	FALSE	60,29%				
Awareness of acceptable sitting posture duration	TRUE	50,72%				
Understanding of back pain from lifting loads	TRUE	57,14%				
Importance of neck/head rest in chairs	FALSE	51,72%				
Belief in standing too much causing back pain	TRUE	53,03%				
Knowledge of age as a risk factor	FALSE	51,56%				
Awareness of gender susceptibility	TRUE	57,38%				
Ergonomic equipment provided	TRUE	56,06%				
Neck pain frequency	1-2 times	22,83%				
Neck pain severity	slightly uncomfortable	31,52%				
Interference with work due to neck pain	slightly interfered	42,39%				

The personas are briefly described below, according to the data interpretation provided by the LLM. Slight changes were made by the authors to increase consistency, which highlights the need for human supervision of the LLM analysis.

- **Persona 1**: Michael Johnson is a 37-year-old American teacher working as an internship teacher at a private bilingual school in District 2. He teaches middle- and high-school science and works approximately 40 hours per week, spending most of his time standing. Michael experiences mild neck discomfort once or twice a week, which only slightly interferes with his teaching. He understands basic ergonomic principles and lifting risks but tends to underestimate age-related factors.
- **Persona 2**: Trần Thị Lan is a 57-year-old Vietnamese teacher with over 30 years of experience, currently teaching full-time at a private bilingual school in District 3. She works about 30 hours per week, alternating between long periods of standing and sitting. Lan experiences severe neck pain several times a day, which substantially interferes with her teaching. Despite having access to ergonomic equipment, she rarely uses it and has limited awareness of posture and lifting risks.
- **Persona 3**: David Miller is a 36-year-old widowed American teacher working part-time at a public school in District 4. With 10 years of experience and a 25-hour weekly workload, he spends most of his time seated. He experiences rare and mild neck discomfort, which does not interfere with his teaching. David understands lifting and gender-related risks but does not perceive standing or aging as significant ergonomic concerns.

Male	52.17%
37 years old	36.99 (mean)
American	35.87%
Married	30.43%
PhD	32.61%
Bilingual Private School	51.09%
8 years of teaching exp.	7.82 (mean)
40h/week contract	31.02 (mean)
Neck discomfort 1–2 times/week	22.83%
Slightly uncomfortable	31.52%
Slightly interferes with work	22.83%
Equipped with ergonomic furniture	42.39%

Michael Johnson

Persona 1

Michael Johnson is 37 years old and lives in Ho Chi Minh City with his wife and two young children. Originally from the United States, he has been teaching sciences in Vietnam for eight years and currently works as an internship teacher at a private bilingual school in District 2, serving middle- and high-school students from both local and expatriate families.

He works about 40 hours a week, spending most of his day standing in classrooms and labs. The school provides ergonomic furniture, which helps him manage mild neck discomfort that appears once or twice a week. Michael understands ergonomic principles and lifting risks, but he tends to underestimate age-related factors. His ergonomic challenges only slightly interfere with his teaching.

Outside of work, Michael enjoys cycling along the Saigon River and spending time with his children at local community events.

Needs

- Adjustable ergonomic furniture to reduce neck strain;
- Opportunities for professional growth in international and bilingual education.
- Health monitoring and preventive guidance to avoid the progression of musculoskeletal discomfort.
- Family-friendly work-life balance to spend quality time with his family

Frustrations

- Mild neck pain after long teaching sessions Limited awareness of age-related risks, which could lead to overlooking preventive measures.
- Physical fatigue from balancing active teaching and lab supervision in the same day.
- Classroom layouts that may not always support optimal ergonomic posture.
- Administrative or teaching tasks that require prolonged sitting and standing time.

Figure 3 Persona 1 – Michael Johnson.

Table 3 Persona comparison by most distinctive variables.

Variable	Persona 1	Persona 2	Persona 3
Age	35	57	36
Teaching status	Internship	Full-time	Part-time
Years of teaching experience	8	20	30

Finally, the proposed workflow is summarized on Figure 4:

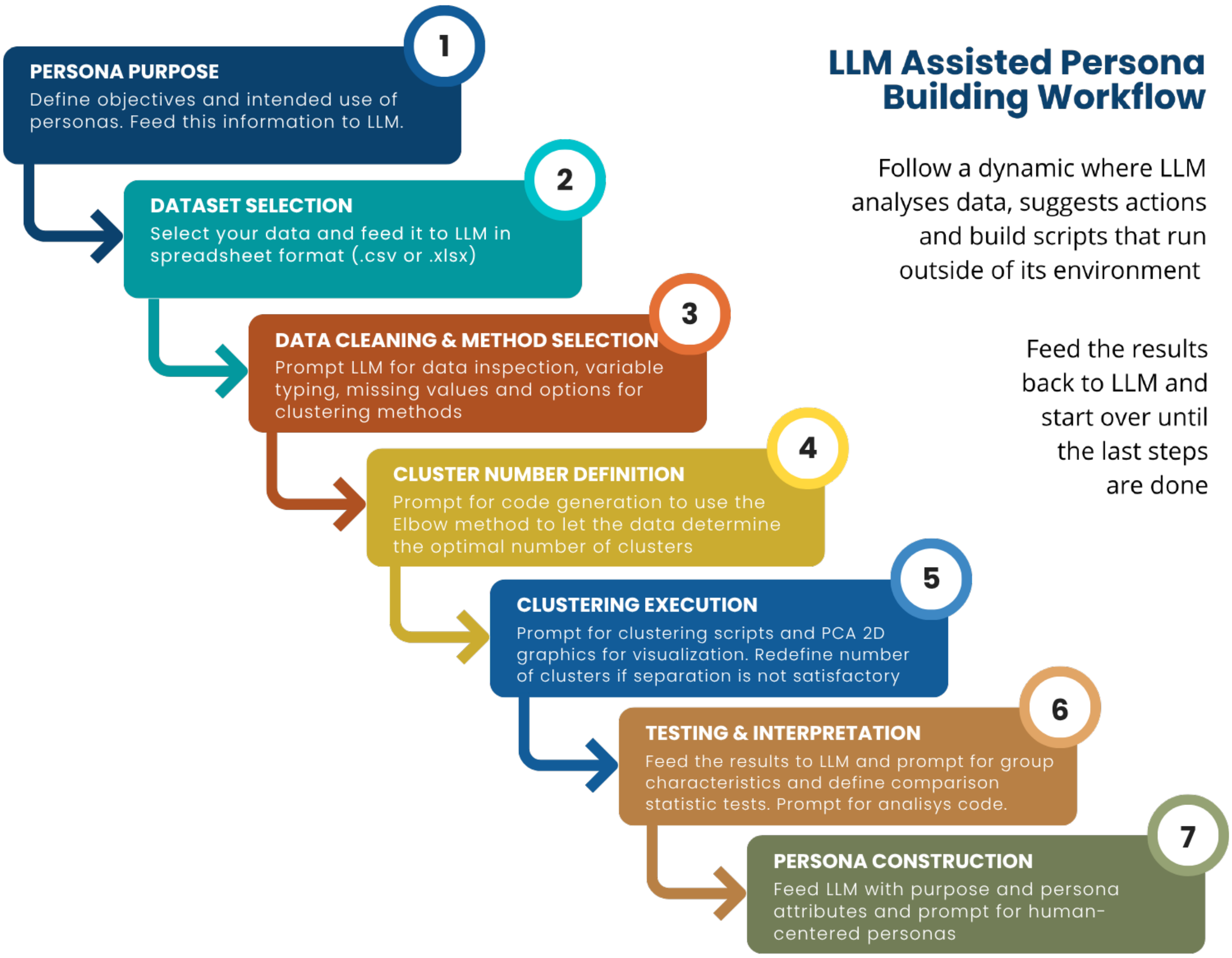

Figure 4 LLM assisted persona building workflow.

5 Discussion and conclusions

The personas developed in this study provide a structured procedure to understand the data and the ergonomic awareness among individuals in the sample. The clusters revealed that age, teaching status, and years of experience were the most discriminative variables, aligning with established literature on musculoskeletal risk factors in educational settings (Tahernejad et al., 2024). Persona 2, for instance, represents well older educators with frequent neck pain, illustrating the need for targeted ergonomic interventions, while Personas 1 and 3 reflect lower risk profiles with limited ergonomic concern. These distinctions demonstrate the potential of data-driven personas to uncover patterns that may remain implicit in raw survey data and this alignment between empirical results and established findings in the field strengthens the credibility of the personas and demonstrates the methodological soundness of the approach. By converging with previously reported ergonomic risk factors, the identified variables provide additional validation that the data-driven clustering process yields meaningful and interpretable educator profiles.

This study contributes to the ongoing discussion on the use of large language models (LLMs) across the stages of data-driven persona creation. The findings demonstrate that LLMs can support tasks ranging from clustering and statistical analysis to narrative generation, effectively lowering the technical entry barrier for such processes. The proposed workflow provides an operational framework for implementing persona design in settings with limited resources or technical expertise, positioning LLMs as valuable bridges between computational analysis and empathetic communication.

This integration may foster broader adoption of data-driven persona methods in both design research and professional practice. This hybrid process also illustrates the value of combining human judgment with AI-assisted analysis, enabling a smoother transition from statistical clustering to narrative personas. From an information design perspective, the workflow may also support the creation of clearer visual narratives and decision-support artifacts, such as dashboards, persona cards, and intervention planning tools based on clustered user data.

Despite demonstrating the feasibility of LLM-assisted persona creation, this study is subject to limitations. Rather than being framed solely as a limitation, the dataset size can be interpreted within the paradigm of small-data learning, where smaller but well-controlled samples have been shown to produce robust and interpretable results. This perspective aligns with recent studies demonstrating that carefully curated datasets with fewer than one hundred cases may yield surprisingly accurate insights, particularly when supported by rigorous statistical testing and structured workflows (Suzuki, 2022). Moreover, only a small subset of the collected variables showed statistically significant differences between clusters, restricting the richness of persona differentiation. While ChatGPT 4o facilitated automated clustering and statistical testing, the absence of expert-led validation during stages such as cluster interpretation and test selection suggests opportunities to strengthen the methodological rigor of the workflow.

To address these limitations, future research should prioritize the use of larger and more diverse datasets, ideally incorporating additional ergonomic, behavioral, and contextual variables to support better segmentation and the inclusion of expert oversight, especially in the execution of clustering methods and statistical tests, would strengthen the validity and reproducibility of the results. These enhancements would broaden the applicability of LLM-assisted persona workflows in professional design practice and academic research alike.

Finally, the study successfully met its objectives by demonstrating the feasibility of using LLMs to support the key stages of data-driven persona creation. The resulting personas captured meaningful differences among educator profiles and validated the potential of combining algorithmic segmentation with LLM-generated descriptions. While limitations such as a small dataset and lack of expert validation remain, the study presents a replicable and accessible workflow that can inform future applications and research.

References

Alhamadi, M., Alghamdi, O., Clinch, S., & Vigo, M. (2022). Data quality, mismatched expectations, and moving requirements: The challenges of user-centred dashboard design. *Proceedings of the Nordic Human-Computer Interaction Conference (NordiCHI '22)*, Article 42, 1–14. https://doi.org/10.1145/3546155.3546708

Cooper, A. (2004). *Inmates are running the asylum: Why high tech products drive us crazy and how to restore the sanity* (2nd ed.). Sams Publishing.

Erick, P. N., & Smith, D. R. (2011). A systematic review of musculoskeletal disorders among school teachers. *BMC Musculoskeletal Disorders*, *12*(1), Article 260. https://doi.org/10.1186/1471-2474-12-260

Gaiser, B., Panke, S. & Arnold, P. (2006). Community design: The personas approach. In T. Reeves & S. Yamashita (Eds.), *Proceedings of E-Learn 2006 – World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (pp. 520–525). Association for the Advancement of Computing in Education (AACE). https://www.learntechlib.org/primary/p/23740/

Garrett, J. J. (2011). *The elements of user experience: User-centered design for the Web and beyond* (2nd ed.). New Riders.

[ghret76yh]. (2024). *Ergonomic awareness and MSDs among educators*. [Data set]. Kaggle. https://www.kaggle.com/datasets/ghret76yh/ergonomic-awareness-and-msds-among-educators

Gill, C., Umstead, K., Behnam Asl, S., & Mahtani, R. (2023). Data visualisation as a tool for public engagement and empathy building. In I. Ordóñez, P. Sustersic, L. Buck, H. Grierson, & E. Bohemia (Eds.), *Responsible innovation for global co-habitation. Proceedings of the 25th International Conference on Engineering and Product Design Education (E&PDE 2023)* (pp. 607–612). The Design Society. Institution of Engineering Designers. https://doi.org/10.35199/EPDE.2023.102

Grudin, J., & Pruitt, J. (2002). Personas, participatory design and product development: An infrastructure for engagement. In T. Binder, J. Gregory, & I Wagner (Eds.), *PDC 02 Proceedings of the Participatory Design Conference* (pp. 144–152). CPSR.

Guan, K. W., Salminen, J., Jung, S.-G., & Jansen, B. J. (2024). Leveraging personas for social impact: A review of their applications to social good in design. *International Journal of Human–Computer Interaction*, *40*(19), 5569–5584. https://doi.org/10.1080/10447318.2023.2247568

Güneş, S., & Ercömert Görgün, Ç. (2022). The ethical dimension of the persona concept. *Gazi University Journal of Science Part B: Art Humanities Design and Planning*, *10*(2), 147–158. https://izlik.org/JA62HZ48JS

Jansen, B. J., Jung, S.-G., Nielsen, L., Guan, K. W., & Salminen, J. (2022). How to create personas: Three persona creation methodologies with implications for practical employment. *Pacific Asia Journal of the Association for Information Systems*, *14*(3), 1–28. https://doi.org/10.17705/1pais.14301

Jansen, B. J., Jung, S.-G., Salminen, J., Guan, K. W., & Nielsen, L. (2021). Strengths and weaknesses of persona creation methods: Guidelines and opportunities for digital innovations. *Proceedings of the 54th Hawaii International Conference on System Sciences*, 4971–4980. https://doi.org/10.24251/HICSS.2021.604

Jenkinson, A. (1994). Beyond segmentation. *Journal of Targeting, Measurement and Analysis for Marketing*, *3*(1), 60–72.

Manovich, L. (2002). *The language of new media*. MIT Press.

McGinn, J. (Jen), & Kotamraju, N. (2008). Data-driven persona development. In M. Burnett, M. F. Costabile, T. Catarci, B. de Ruyter, D. Tan, M. Czerwinski, & A. Lund (Eds.), *Proceedings of the 26th Annual CHI Conference on Human Factors in Computing Systems (CHI '08)* (pp. 1521–1524). The Association for Computing Machinery. https://doi.org/10.1145/1357054.1357292

Norman, D. (2008, November 17). *Ad-hoc personas & empathetic focus*. jnd.org. https://jnd.org/ad-hoc-personas-empathetic-focus/

Salminen, J., Guan, K., Jung, S.-G., Chowdhury, S. A., & Jansen, B. J. (2020). A literature review of quantitative persona creation. In R. Bernhaupt, F. Mueller, J. McGrenere, A. Cockburn, P. Bjørn, & S. Zhao (Eds.), *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)* (pp. 1–14). The Association for Computing Machinery. https://doi.org/10.1145/3313831.3376502

Salminen, J., Guan, K., Jung, S.-G., & Jansen, B. J. (2021). A survey of 15 years of data-driven persona development. *International Journal of Human–Computer Interaction*, *37*(18), 1685–1708. https://doi.org/10.1080/10447318.2021.1908670

Solis-Soto, M. T., Schön, A., Solis-Soto, A., Parra, M., & Radon, K. (2017). Prevalence of musculoskeletal disorders among school teachers from urban and rural areas in Chuquisaca, Bolivia: A cross-sectional study. *BMC Musculoskeletal Disorders*, *18*(1), Article 425. https://doi.org/10.1186/s12891-017-1785-9

Suzuki, K. (2022). Small data deep learning for lung cancer detection in CT. *Proceedings of the IEEE 8th International Conference on Big Data Computing Service and Applications (BigDataService 2022)*, 114–118. https://doi.org/10.1109/BigDataService55688.2022.00025

Tahernejad, S., Hejazi, A., Rezaei, E., Makki, F., Sahebi, A., & Zangiabadi, Z. (2024). Musculoskeletal disorders among teachers: A systematic review and meta-analysis. *Frontiers in Public Health*, *12*. https://doi.org/10.3389/fpubh.2024.1399552

Velasco, J. L., Teixeira, J. M., & Pinto, A. L. (2025). Validating personas for better communication: A structured model for low-resource contexts. *Iberoamerican Journal of Science Measurement and Communication*, *5*(3), 1–12. https://doi.org/10.47909/ijsmc.236

Velasco, J., Picalho, A., Fonseca, R., Viola, H., Teixeira, J., Fadel, L., & Pinto, M. (2023). Data driven personas: Uma interface de mediação entre dados e usuários [Data driven personas: An interface for mediation between data and users]. In R. Pereira, G. A. Bondan, L. Z. Wolski, C. Menegali, A. E. da Costa, N. Alemsan, L. Cunico, I. W. Reis, A. F. Biglieri, D. Sell, F. A. P. Fialho, L. M. Spinosa, & N. dos Santos (Eds.), *Perspectivas em engenharia, mídias e gestão do conhecimento* (Vol. 4, pp. 193–213). Arquétipos.

About the authors

Jefferson Lewis Velasco
jefferson.velasco@posgrad.ufsc.br
Universidade Federal de Santa Catarina
Florianópolis, SC

Júlio Monteiro Teixeira
julio.teixeira@ufsc.br
Universidade Federal de Santa Catarina
Florianópolis, SC

Eugênio Diaz Merino
eugenio.merino@ufsc.br
Universidade Federal de Santa Catarina
Florianópolis, SC

Submission date/*Artigo recebido em*: 3/2/2026
Approvement date/*Artigo aprovado em*: 25/5/2026