

# Empowering Persona Creation in Small Organizations

## Evaluating ChatGPT 4o for Clustering and Analysis using PersonaCraft

**Authors:** Velasco, J. L.<sup>1</sup>; Teixeira, J. T.<sup>1</sup>; Peruchini, M.<sup>1</sup>; Modena, G.<sup>1</sup>.

### **Presenters:**

Jefferson Lewis Velasco, MSc - [jeffvelasco.crm@gmail.com](mailto:jeffvelasco.crm@gmail.com)

Júlio Monteiro Teixeira, Dr - [juliomontex@gmail.com](mailto:juliomontex@gmail.com)

**UFSC - Federal University of Santa Catarina<sup>1</sup>**



UNIVERSIDADE FEDERAL  
DE SANTA CATARINA





## Jeff

Jefferson is a PhD candidate in Design at UFSC and a researcher at Lemme Lab. A career civil servant at CIASC since 2010, he currently manages Digital Government projects. His work focuses on data-driven design and data personas to support innovation in public services.

## Areas of Interest:

- Data-Driven Design
- Personas
- Bibliometrics
- Low Resource Contexts



## Julio Teixeira

Adjunct Professor at UFSC, teaching in the Design Program and in the Graduate Programs in Design (PPGD), and Knowledge Engineering (PPGEGC). Leads the Lemme – Digital Innovation Research Group. Postdoctoral Fellow at the Academy of Art University (San Francisco, USA). PhD in Production Engineering from UFSC, with a research period at the University of Wuppertal (Germany). Master's in Design Management (UFSC) and Bachelor's in Design (UDESC).

## Areas of Interest:

- Generative AIs
- CX/UX Journey
- Data-Driven Design



## lemme

Area



Digital Innovation

Subjects



CX Journey



Creative Data Analysis



Generative AI

Topics

Wearables

Digital Education

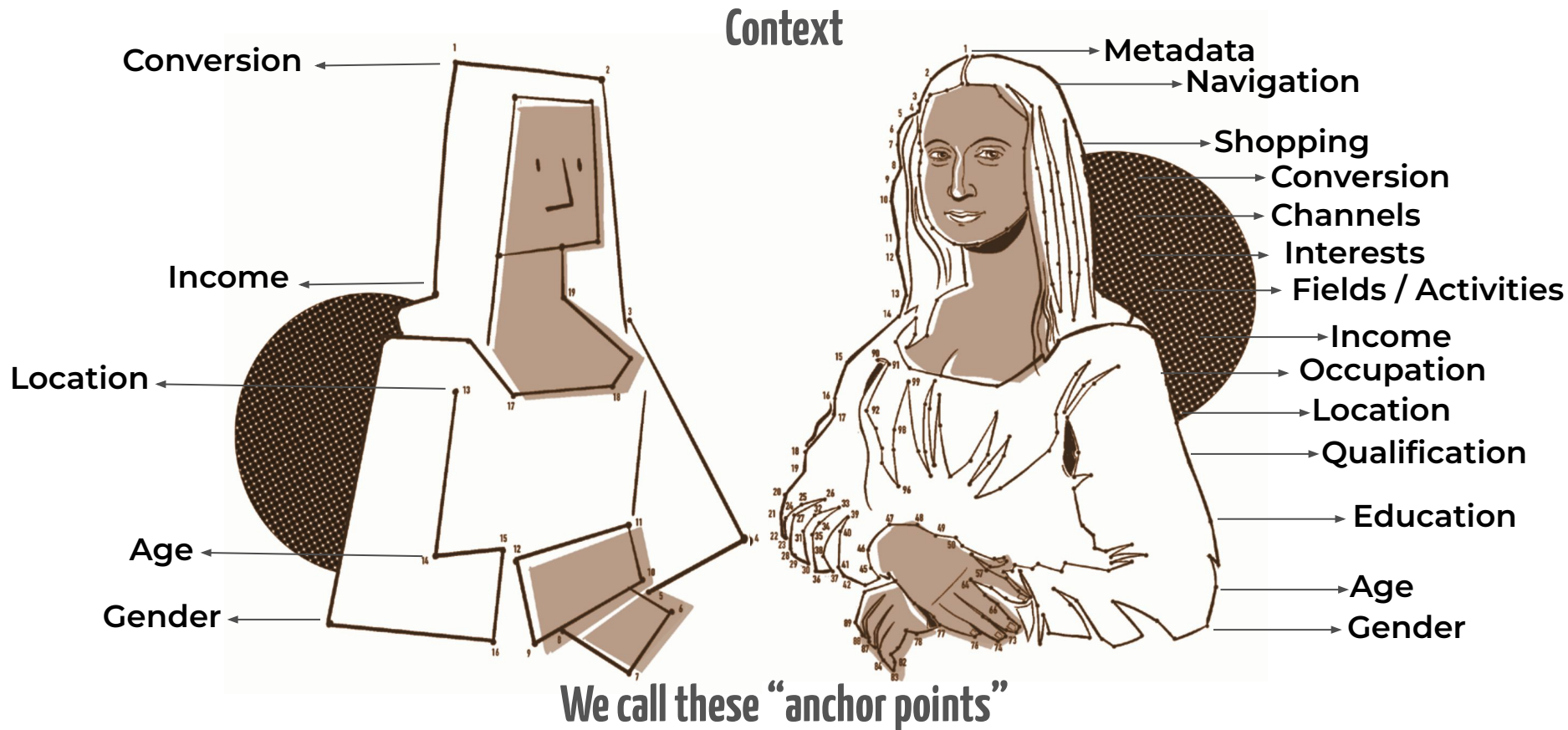
Personas

Visual Management

Digital Business







# Context:

## Problem and Motivation

- **Small organizations face barriers in persona creation:**
  - Lack of data science expertise;
  - Complexity of clustering methods;
  - Limited resources for advanced analysis.
- **Current methods are designed for experts.**
- **Question:** Can AI tools bridge this gap?

## Context:

### Objective

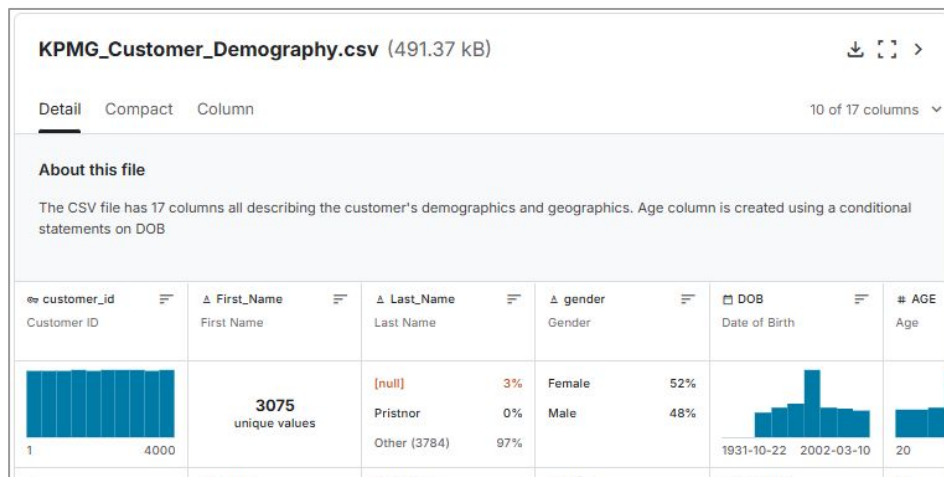
Evaluate how ChatGPT 4o, through the PersonaCraft methodology, can assist non-experts in creating data-driven personas.

### Research Question

How effectively can ChatGPT 4o guide small organizations in executing the core stages of PersonaCraft for persona creation?

## Dataset Selection

The dataset used was a publicly available retail customer dataset from Kaggle, containing 3,908 rows of simulated demographic and behavioral data—ideal for testing segmentation methods in small business contexts due to its simplicity and completeness.

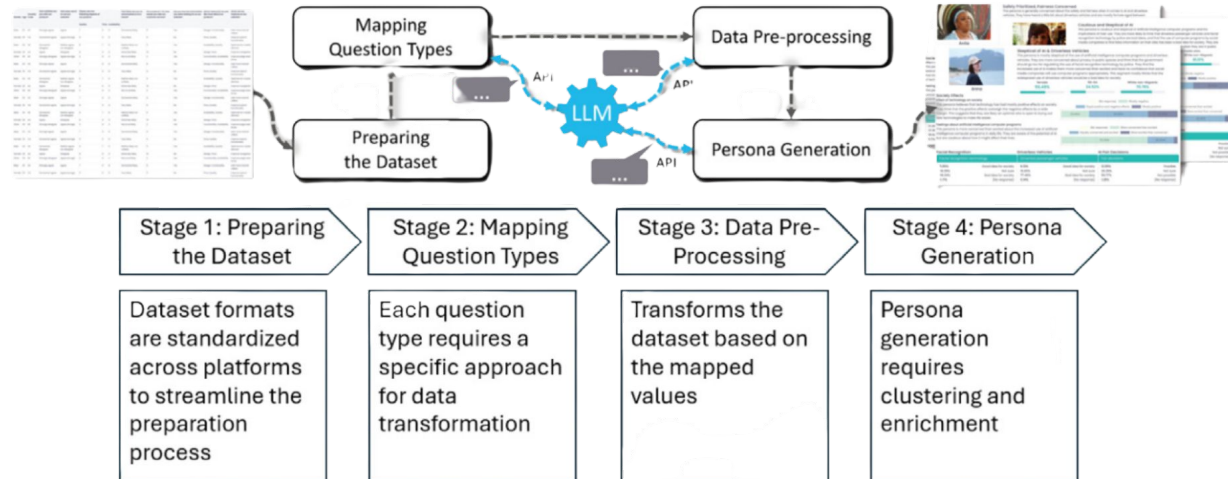


<https://www.kaggle.com/datasets/harishedison/kpmg-customer-demography-cleaned-dataset>



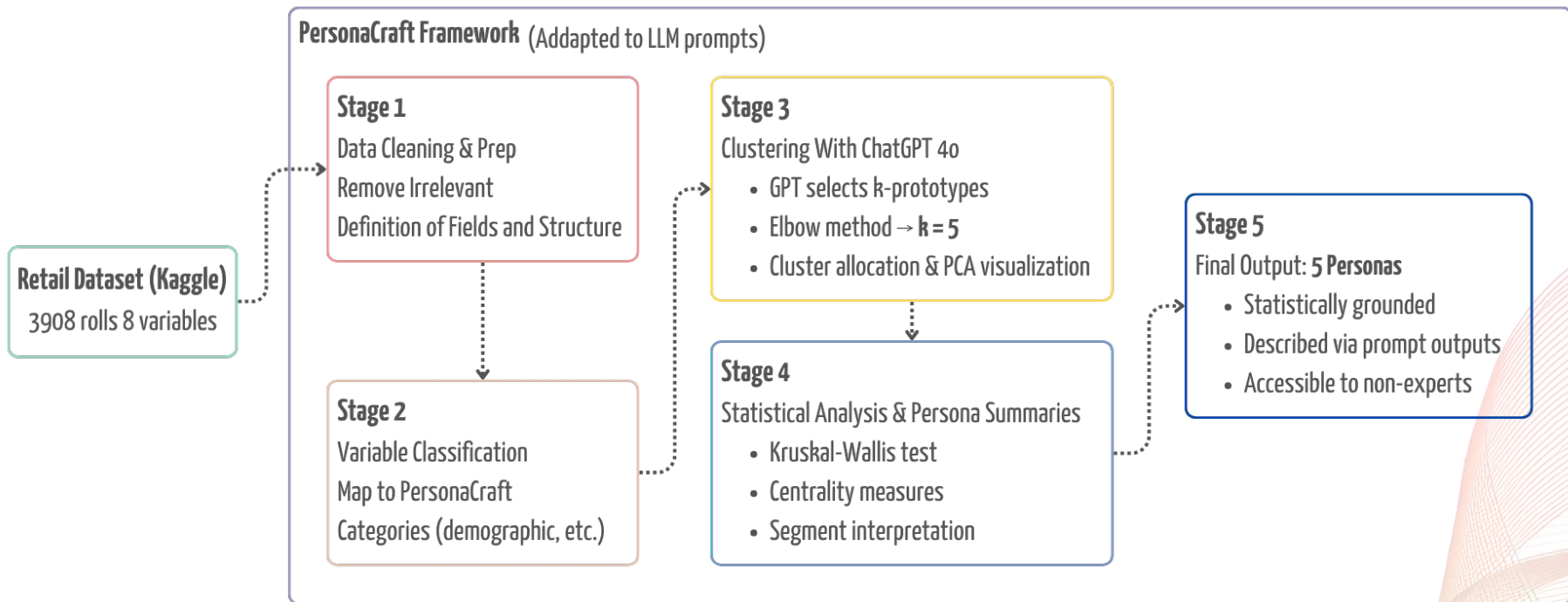
## PersonaCraft Framework

PersonaCraft is a structured framework that guides the creation of data-driven personas through defined stages, combining clustering techniques with descriptive analysis to generate reliable user segments.



PersonaCraft: A novel persona generation methodology that leverages LLMs (Jung et al., 2025)

## Workflow



## Dataset Overview

Variable Group	Description	Variables	Headings
Personal Identification	Used to uniquely identify or describe a person.	customer_id	Customer ID
Demographics	Attributes related to socioeconomic status, age, and personal traits.	age	Age
		gender	Gender
		job_industry_category	Job Industry Category
		wealth_segment	Wealth Segment
Purchasing Behavior	Data on purchases or transaction-related behavior.	past_3_years_bike_related_purchases	Bike Purchases Last 3 Years
Customer Tenure	Duration of the customer's relationship.	tenure	Customer Tenure
Assets & Ownership	Indicators of asset ownership	owns_car	Car Ownership
Location Information	Geographic location of the customer.	state	State

# Methodology

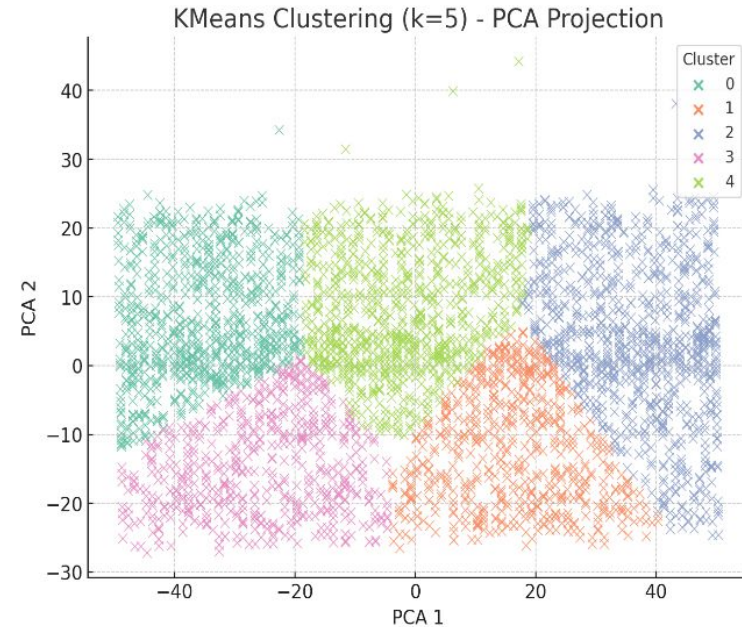
## Data Preparation

- **Removed irrelevant variables:**  
(names, addresses, IDs, etc.);
- **Classified variables into 5 groups:**  
(Personal Identification, Demographics, Purchasing Behavior, Customer Tenure, Assets & Ownership and Location Information);
- **Retained numeric variables in original form for statistical accuracy;**
- **Prompts adapted to guide GPT through each PersonaCraft stage.**

# Methodology

## Clustering

- **Clustering Algorithm:**  
k-prototypes (mix of categorical + numeric);
- **Cluster Number Decision:**  
Elbow Method → Best at **k=5**;
- **Output:**  
Excel file with cluster labels & PCA visualization of clusters.





# Methodology

## Statistical Testing

- **Applied Kruskal-Wallis test to check variable significance between clusters.**
- **Significant variables:**
  - Purchase Frequency ( $p < 0.01$ );
  - Age ( $p < 0.01$ );
  - Customer Tenure ( $p < 0.01$ ).

## Prompt Sample

### Initial Prompt

Hi, I am trying to segment a dataset for data persona creation. I have already analysed variables and determined the best approach is to use k-prototypes. But in order to do that, first I must define the optimal number of clusters for my dataset. Can you help me?

### Response

Below is a step-by-step guide (with sample code) on how you can determine the optimal number of clusters for **k-prototypes**. The most common approach is an "Elbow Method" based on the **cost** (sometimes called the within-cluster distance) returned by the **k-prototypes** model.

In the k-prototypes algorithm:

1. **Numeric attributes** contribute to the distance with the usual Euclidean term.
2. **Categorical attributes** contribute to the distance using simple matching (Hamming) distance.

The **cost** attribute of a fitted k-prototypes model is a sum of these distances over the entire dataset. By running k-prototypes multiple times with different values of **kkk** (number of clusters), then plotting or examining how the cost changes, you can look for an "elbow" point to help choose a suitable **k**.

### Example Interaction:

A non-expert user engages ChatGPT 4o through natural language. The model interprets the request, provides clear guidance, and enables the user to complete the task successfully.

The initial prompt and response are followed by a conversation with detailed instructions that lead to the completion of the task.

# Results

## Persona Summaries: 5 Personas Identified

- **Cluster 1:** Older (52), Low purchases, Long tenure;
- **Cluster 2:** Young (33), High purchases, Short tenure (New, engaged buyers);
- **Cluster 3:** Middle-aged (48), Highest purchase rate, Stable tenure;
- **Cluster 4:** Very young (31), Low purchases, Short tenure;
- **Cluster 5:** Older (52), Medium purchases, Longest tenure.

## Results

### Clusters Descriptions (using variable frequencies and centrality measurements)

Variable	Values	C1	C2	C3	C4	C5
Gender	Female	468	328	484	295	462
	Male	373	317	471	281	429
Job Industry Category	Manufacturing	173	144	185	107	187
	Financial Services	162	131	179	122	172
	n/a	142	102	165	96	150
	Health	122	86	161	83	144
	Retail	86	66	85	52	69
	Property	46	35	66	46	73
	Entertainment	35	21	30	21	29
	IT	29	26	39	27	29
	Agriculture	25	25	29	10	24
	Telecommunications	21	9	16	12	14
Wealth Segment	Mass Customer	401	328	498	275	449
	High Net Worth	229	162	227	157	220
	Affluent Customer	211	155	230	144	222

Variable	Values	C1	C2	C3	C4	C5
Car Ownership	Yes	432	338	472	280	449
	No	409	307	483	296	442
State	New South Wales	456	330	530	316	457
	Victoria	214	155	229	147	253
	Queensland	171	160	196	113	181
Age	Mean	52,11	32,95	47,86	31,25	52,55
	Median	51	32	47	30	51
	Mode	45	27	44	28	44
Bike Purchases Last 3 Years	Mean	14,06	65,38	85,90	23,45	47,22
	Median	14	66	87	25	47
	Mode	2	68	98	27	53
Customer Tenure	Mean	12,34	7,45	11,41	6,82	13,05
	Median	12	6	12	5	13
	Mode	11	1	12	2	18

# Results

## Key Insights

- **Numeric variables are critical** — especially purchase frequency and age;
- **Demographic labels** (like gender or job industry) **don't differentiate clusters**;
- **GPT successfully executed technical steps via prompting without coding**;
- **The process significantly lowers the barrier for persona creation in small organizations.**



## Conclusion & Future Work

### Conclusion

- ChatGPT 4o can guide non-experts through persona creation tasks;
- Effective for small organizations lacking analytics teams.

### Future Work

- Test with real, richer datasets;
- Combine LLM output with stakeholder validation;
- Explore automation of prescriptive tasks (e.g., customer journey design),

# References:

- [1] E. L. Melnic, "How to strengthen Customer Loyalty, using Customer Segmentation?", Bulletin of the Transilvania University of Brasov. Series V: Economic Sciences, pp. 51–60, dez. 2016. [retrieved: March, 2025].
- [2] J. Brickey, S. Walczak, and T. Burgess, "Comparing Semi-Automated Clustering Methods for Persona Development", IEEE Transactions on Software Engineering, vol. 38, n° 3, pp. 537–546, May 2012, doi: 10.1109/TSE.2011.60. [retrieved: March, 2025].
- [3] J. A. Jansen, A. Manukyan, N. A. Khoury, and A. Akalin, "Leveraging large language models for data analysis automation," PLOS ONE, vol. 20, no. 2, p. e0317084, Feb. 2025, doi: 10.1371/journal.pone.0317084. [retrieved: March, 2025].
- [4] S.-G. Jung, J. Salminen, K. K. Aldous, and B. J. Jansen, "PersonaCraft: Leveraging language models for data-driven persona development", International Journal of Human-Computer Studies, vol. 197, p. 103445, mar. 2025, doi: 10.1016/j.ijhcs.2025.103445. [retrieved: March, 2025].
- [5] E. Harish, "KPMG Customer Demography Cleaned Dataset", Kaggle. Accessed on: March 19, 2025. [Online]. Available at: <https://www.kaggle.com/datasets/harishedison/kpmg-customer-demography-cleaned-dataset>. [retrieved: March, 2025].
- [6] J. (Jen) McGinn and N. Kotamraju, "Data-driven persona development," in Proceedings of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08, Florence, Italy: ACM Press, 2008, pp. 1521–1524. doi: 10.1145/1357054.1357292. [retrieved: March, 2025].
- [7] E. Ditton, A. Swinbourne, and T. Myers, "Selecting a clustering algorithm: A semi-automated hyperparameter tuning framework for effective persona development," Array, vol. 14, p. 100186, Jul. 2022, doi: 10.1016/j.array.2022.100186. [retrieved: March, 2025].
- [8] F. Lanfermann, T. Rios, and S. Menzel, "Large Language Model-assisted Clustering and Concept Identification of Engineering Design Data," in 2024 IEEE Conference on Artificial Intelligence (CAI), Singapore, Singapore: IEEE, Jun. 2024, pp. 788–795. doi: 10.1109/CAI59869.2024.00150. [retrieved: March, 2025].
- [9] N. Arora, I. Chakraborty, and Y. Nishimura, "AI–Human Hybrids for Marketing Research: Leveraging Large Language Models (LLMs) as Collaborators," Journal of Marketing, vol. 89, no. 2, pp. 43–70, Mar. 2025, doi: 10.1177/00222429241276529. [retrieved: March, 2025].
- [10] T. Goel, O. Shaer, C. Delcourt, Q. Gu, and A. Cooper, "Preparing Future Designers for Human-AI Collaboration in Persona Creation," in Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work, Oldenburg Germany: ACM, Jun. 2023, pp. 1–14. doi: 10.1145/3596671.3598574. [retrieved: March, 2025].

# Thank you!

## Empowering Persona Creation in Small Organizations

### Evaluating ChatGPT 4o for Clustering and Analysis using PersonaCraft

#### Presenters:

Jefferson Lewis Velasco, MSc - [jeffvelasco.crm@gmail.com](mailto:jeffvelasco.crm@gmail.com)

Júlio Monteiro Teixeira, Dr - [juliomontex@gmail.com](mailto:juliomontex@gmail.com)

UFSC - Federal University of Santa Catarina<sup>1</sup>

