AI Personas: Grounded Conversational Agents from Visually-Rich Segmentation Data

Nguyen Van Thanh
s336748@studenti.polito.it
Politecnico di Torino
Turin, Italy

Chen Enrico
s337750@studenti.polito.it
Politecnico di Torino
Turin, Italy

Ma Xiaoning
s337332@studenti.polito.it
Politecnico di Torino
Turin, Italy

Abstract

This report presents the design and evaluation of Lavazza AI Personas, a system that transforms static customer segmentation studies into dynamic, conversational AI agents. The system addresses the challenge of operationalizing unstructured and visually rich marketing reports by combining a multi-agent architecture with Visual Large Language Models (VLLM) for semantic data extraction and a Retrieval-Augmented Generation framework for factual grounding. Each market segment is represented as an interactive AI Persona, such as the Curious Connoisseurs, enabling marketers to explore and test strategic decisions through dialogue. We evaluate the system across four dimensions: persona extraction, fact extraction, retrieval relevance, and persona authenticity. Experimental results show a 95.3% recall in persona extraction, 97% exact match accuracy in fact extraction, an 82.08% recall rate in context retrieval at $k = 20$, and an expert authenticity score of 4.66/5. These findings demonstrate the adopted approach effectively reduces hallucinations while preserving coherent and authentic segment-specific behavior.

1 Introduction

In the domain of market research, consumer insights are often locked within static, lengthy PDF reports, making dynamic strategy testing and real-time decision-making difficult. While *Large Language Models* (*LLMs*) offer advanced conversational capabilities, they typically lack the domain-specific knowledge and consistent personality traits required to accurately simulate distinct consumer segments. This project introduces an AI Persona System designed to bridge this gap by transforming proprietary customer segmentation data into interactive agents, enabling business teams to engage directly with data-driven personas representing Lavazza's market segments.

To address this challenge, our research focuses on three core objectives: the automated extraction of accurate information from visually rich PDF documents while preserving layout semantics, the inference of latent reasoning traits from heterogeneous and noisy data to construct consistent personas, and the minimization of hallucinations in generated outputs. The proposed solution adopts a multi-agent orchestration framework that leverages *Visual Large Language Models* (*VLLMs*) to extract structured facts and persona indicators from raw documents. These indicators are indexed into a vector database and accessed through a *Retrieval-Augmented Generation* (*RAG*) pipeline, where a dedicated orchestrator manages context retrieval and response generation. By injecting persona-specific style and value traits into the generation prompt, the system ensures that responses remain both factually grounded and authentically aligned with the intended consumer profile.

The remainder of this report is organized as follows. Section 2 reviews related work on persona conditioning and document understanding. Section 3 details the proposed system architecture and extraction pipeline. Section 4 evaluates the system's performance on the Lavazza Curious Connoisseurs dataset. Section 5 concludes with a discussion of limitations and future research directions. To support reproducibility, the full implementation of the proposed system is publicly available at https://github.com/adsp-polito/2025-P3-AI-Personas.

2 Related Work

2.1 Data-Driven AI Personas

Existing work on AI personas generally follows two directions. One line of research relies on manually defined persona descriptions, often designed for role-playing or task-specific interactions. Another line focuses on dynamic user profiling, where persona representations are continuously updated from long-term interaction histories to support personalization [9, 11].

These approaches, however, are not directly applicable to enterprise market analysis, where personas must be derived from static and proprietary datasets that are typically provided as complex, unstructured documents. In such settings, personas are constructed at the level of consumer segments and remain fixed, rather than being refined through ongoing user interactions. In this respect, the approach presented in [8] is closely aligned with our objectives; however, in our case, the underlying datasets are provided in PDF format.

2.2 Knowledge Extraction from Visually-Rich Documents

Automated extraction of structured knowledge from visually rich documents is a common prerequisite for downstream document understanding systems. Traditional OCR-based pipelines are limited in this respect, as they primarily focus on text recognition and often fail to capture the semantic structure encoded in document layouts [10].

To address this limitation, layout-aware vision-language models have been proposed to jointly model textual content and spatial information. Early work such as *LayoutLM* [12] incorporates two-dimensional positional features alongside text embeddings, while subsequent models including *DocLLM* [10] extend this paradigm

Applied Data Science Project, Politecnico di Torino

to generative settings. These approaches enable more faithful extraction of structured information from complex document layouts and form the basis for downstream tasks such as factual retrieval and persona-related attribute inference.

2.3 Inferring Latent Persona Traits

Prior studies have shown that extracting explicit attributes alone is insufficient for constructing coherent and stable personas. Higher-level reasoning traits, such as decision-making priorities or communication styles, often need to be inferred from heterogeneous signals. However, end-to-end inference using a single language model can lead to persona instability, especially under complex or multi-faceted prompts, a phenomenon commonly referred to as persona inconsistency [1].

To address this issue, recent work has explored multi-agent and modular reasoning frameworks, in which different components specialize in distinct stages of the inference process [3]. By decomposing persona construction into separate extraction and reasoning steps, these approaches aim to improve consistency and robustness when modeling latent persona traits.

2.4 Factual Grounding with RAG

RAG has emerged as a widely adopted approach for improving factual accuracy in knowledge-intensive language generation tasks [6]. By conditioning model outputs on retrieved evidence from an external corpus, *RAG*-based systems reduce hallucinations and enable closer alignment between generated responses and source material.

Recent work further highlights the role of dense retrieval models in supporting effective semantic search over large document collections [7]. Beyond improving accuracy, retrieval-based grounding also facilitates transparency and traceability, which are increasingly recognized as critical requirements in enterprise and decision-support applications.

2.5 Persona Authenticity Evaluation

Evaluating persona-driven generation presents challenges that extend beyond standard automatic NLP metrics [4]. Recent studies on persona fidelity emphasize the need for multi-dimensional evaluation protocols that capture both behavioral and factual aspects of generated responses. Commonly considered dimensions include authenticity, which reflects alignment with the target user group, stylistic consistency across interactions, and factual grounding in the underlying data sources [5].

While general evaluation frameworks for large language models continue to evolve [4], these persona-oriented criteria provide a more targeted basis for assessing consistency and reliability in persona-based systems.

3 Methodology

In this section, we first outline the problem statement and then provide a comprehensive overview of the methodology adopted for the design of the proposed solution. Figure 1 represents the entire workflow, from the user query to the generation of the response.

3.1 Problem Statement

This work focuses on enabling dynamic interaction with *AI Personas* that represent distinct market segments, allowing marketers to evaluate marketing performance, improve customer understanding, and test strategic assumptions.

3.1.1 Extraction. The methodology adopted to solve the first research question treats PDFs as visual documents and extracts semantic information using a *VLLM*. This solution offers several advantages over alternative approaches, such as *manual annotation* and *OCR-based pipelines*. In particular, it preserves layout and spatial relationships, operates in a fully automated manner without reliance on manually defined rules, and directly produces schema-consistent results that are immediately usable for downstream tasks. Although *manual annotation* achieves higher accuracy, its low scalability and high costs make it impractical for adoption. Conversely, *OCR-based pipelines*, while scalable, provide limited layout information or introduce noise [13].

3.1.2 Personality. We address this second research question by employing a *multi-agent system* with *cooperative* collaboration pattern, where each agent specializes in a specific stage of reasoning and trait inference. In particular, a first agent extracts persona information from raw data, and a second agent infers latent persona reasoning traits. This solution offers scalability and reproducibility across datasets, while reducing subjective human bias, in contrast to alternative approaches such as *manual techniques* and *single-model end-to-end inference*.

3.1.3 Grounding. Alternative approaches, such as *prompt-based techniques* and *pure fine-tuning* attempt to mitigate hallucinations in language generation; however, they are often either unreliable or computationally expensive.

We integrate the retrieval component of a *Retrieval Augmented Generation (RAG)* system that retrieves relevant evidence before answer generation and constrains model outputs to the retrieved context. This approach not only improves the factual accuracy [6], but also reduces hallucinations, makes responses explainable, and has lower computational costs.

3.2 Persona Extraction

In the *data ingestion* pipeline of figure 1, we can observe that the dataset is processed to populate both *RAG* and *Persona Registry*. As discussed in 3.1.2, we treat each PDF page as an image and extract semantic meaning using a *VLLM*. More precisely:

(1) We provide each page of the dataset to a *VLLM* (*Mistral Medium 3 Instruct*), and the results are structured into a JSON schema. These files capture the persona indicators, which are subsequently chunked, embedded, and stored in a *Vector Database* for retrieval within the *RAG* system when needed. To be more precise, the indicators are the raw data referred to a customer segment
(2) The same indicators of a persona are also given to another instance of the identical *VLLM* to infer the traits of the segment (such as communication style, prioritization criteria, and decision rules). These inferred traits are incorporated into the prompt (*prompt tuning*) to guide the *AI Persona*,

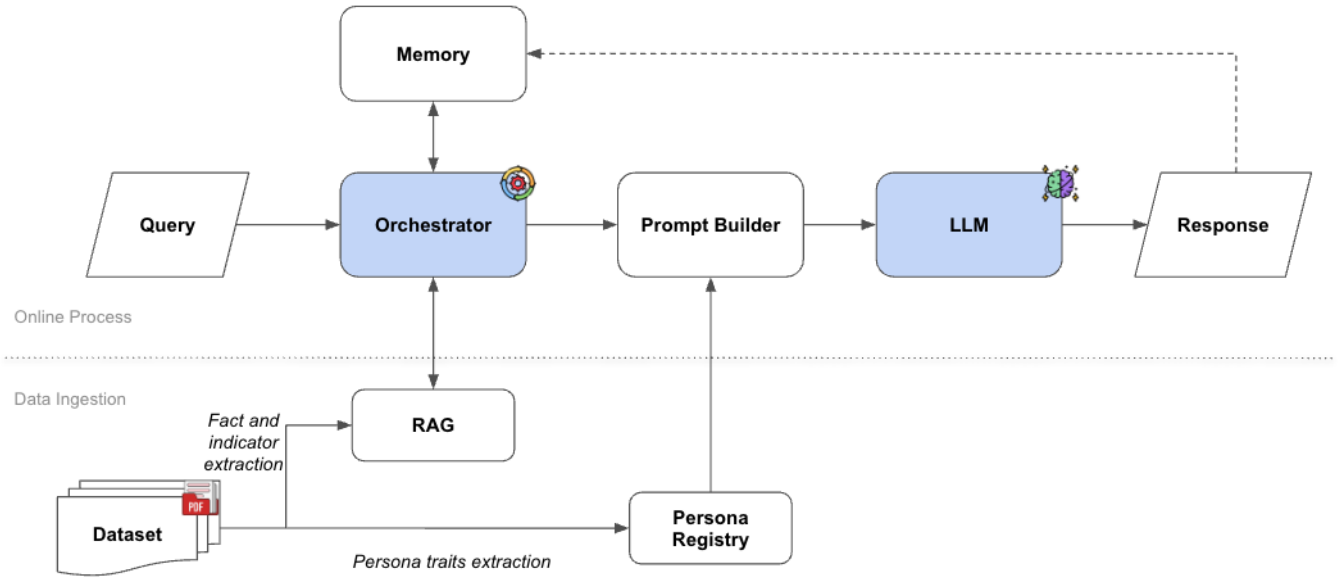

Figure 1: Functional diagram of the proposed solution. It is divided into an online process and a data ingestion process.

without necessitating fine-tuning the models representing the *AI Persona*.

3.3 Fact Extraction

The same approach is applied to fact data, representing the raw data. Indicators are quantitative metrics defining the segment (e.g., 45% female), whereas Facts are broader qualitative statements or contextual data found in the report. Fact data encompasses a broader scope than indicators, as the latter constitute a subset of the former. Fact data is extracted to ensure scalability to additional datasets. The main steps are:

- each page of the dataset is given to a *VLLM* (*Mistral Large 3 675B Instruct 2512*), and the results are stored as markdown files
- markdown files are chunked, embedded, and stored in the *Vector Database*. The same embedding model of 3.2 is used

3.4 Orchestrator

The orchestrator constitutes the primary coordination module, responsible for integrating all components involved in processing a user query. Its operations are:

(1) invoking the *input handler* to normalize the user query. As the proposed solution currently processes only textual data, the *input handler* is limited to removing spaces from the input
(2) querying the *memory* module to retrieve past interactions.
(3) accessing the *RAG* system to extract both indicators and fact data, subsequently merging the resulting chunks
(4) filtering past interactions and retrieved chunks to select only the context strictly needed to answer the query
(5) utilizing the *prompt builder* to construct the input prompt
(6) submitting the constructed prompt to an *LLM* to generate a response
(7) updating the *memory* module with the new interaction, encompassing both the query and the generated response

3.5 Memory

This module stores past interactions between the user and the *LLM*. By retaining and incorporating these interactions into subsequent response generation, the *LLM* is able to maintain conversational context, produce more coherent and relevant outputs, and support multi-step interaction. In order to limit the size of the memory, we keep only a limited number of past interactions defined by max_items.

3.6 RAG Retrieval Module

Figure 2 represents the *RAG* retrieval feature adopted in the proposed solution. Given the query and the customer segment type, the *RAG Retrieval Module* retrieves k chunks for indicators and k chunks for fact data, these chunks are returned to the *orchestrator*.

3.7 Prompt Builder

This module is responsible for constructing the input prompt to be submitted to the *LLM* model. This process integrates multiple components to ensure coherent and contextually grounded responses. Specifically, it combines:

- *system prompt*, which encapsulates key persona information including the customer segment, persona name, persona traits, summarized biography, and predefined answer guidelines. The *Persona Registry* serves as the source for this information, providing structured data regarding each persona's characteristics and background

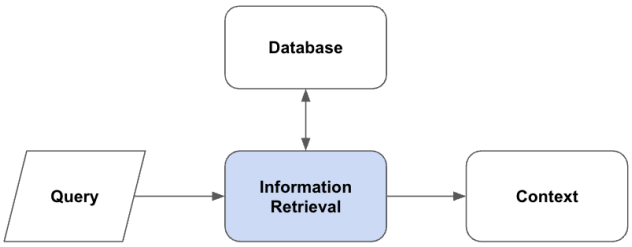

Figure 2: Representation of the *RAG Retrieval Module*. Given a query, relevant information is retrieved from a database, and then the result is used as context

- *filtered conversation history*, which preserves relevant prior interactions to maintain contextual continuity
- *filtered retrieved context*, which supplies contexts from a database pertinent to the current query
- *user query*, representing the user request.

3.8 Persona Registry

This module maintains all personas, enabling their retrieval for conversational use and related processes. It includes the complete set of persona traits to be incorporated into the prompt.

3.9 LLM

This module takes the final prompt from the *prompt builder* and sends it to an instance of *LLM*.

3.10 Application

We also implement an application facilitating direct user interaction with the *AI Personas*, featuring:

- *backend*, it exposes the application layer over *HTTP*, so clients such as *frontend*, *tools*, and other services can interact with the system functionality
- *frontend*, serving as the user interface for interacting with the backend, it has multiple components like: *authentication*, *customer segment selection*, *past chat selection*, *context customization*, *persona name customization*, *persona information*.

3.11 Models and Frameworks

We now provide a complete list of the main models and Frameworks.

(1) **Models**
- *mistralai/mistral-medium-3-instruct*, used for persona indicator extraction and trait reasoning due to its strong reasoning performance and lower computational cost compared to larger models. This model family is widely adopted in enterprise settings and well-suited for production-grade applications
- *mistralai/mistral-large-3-675b-instruct-2512*, used for fact data extraction, as this task presents higher complexity and requires greater accuracy to minimize hallucinations
- *sentence-transformers/all-mpnet-base-v2*, used for embedding given its open-source availability, balanced performance, and efficient inference speed
- *mistral-small-24b-instruct*, used for context filtering and response generation due to its strong conversational and reasoning capabilities, lightweight architecture, and fast inference speed

(2) **Frameworks and Libraries**
- *FastAPI*, used for implementation of *RESTful API* endpoints with *Python* based on standard *Python* type hints and providing high-performance request handling
- *Streamlit*, used for development of an interactive web-based frontend, enabling rapid prototyping
- *uvicorn*, used as an *Asynchronous Server Gateway Interface* server for *FastAPI*, enabling asynchronous request handling, making it ideal for high-concurrency needs of AI chat and streaming responses
- *OpenAI*, used as interface for invoking *LLMs* and *VLLMs* exposed through API-based services
- *LangChain (with integrations)*, *Sentence-Transformers*, used to orchestrate pipelines, integrate embedding models, and perform document chunking

4 Experiment

We evaluate four critical components of the AI Personas system: persona extraction, fact extraction, retrieval relevance, and persona authenticity. Each component is tested using manually annotated ground truth data and standardized metrics.

4.1 Dataset

4.1.1 Source Document. All experiments are grounded in the *Lavazza Customer Segmentation Analysis* report, a 222-page proprietary market research document. The report combines structured survey data with qualitative insights and rich visual content, including charts, tables, and infographics. It spans multiple analytical dimensions, such as demographic composition (e.g., age, gender, income, education), coffee consumption habits, purchasing behaviors, lifestyle attitudes, sustainability preferences, and brand awareness metrics. This heterogeneity makes the document a realistic and challenging benchmark for vision-based extraction and retrieval systems.

4.1.2 Test Subset. The evaluation focuses on a subset of 23 pages dedicated to the *Curious Connoisseurs* segment, a quality-oriented and explorative group of coffee consumers of the French market.

4.1.3 Ground Truth Datasets. Four ground truth datasets are constructed for evaluating the system:

(1) **Persona Extraction:** 1,051 manually validated metrics extracted across the 23 selected pages.
(2) **Fact Extraction:** 467 validated text snippets with exact wording preserved.
(3) **Retrieval Relevance:** 31 evaluation questions covering demographics, behaviors, attitudes, and brand-related attributes.
(4) **Authenticity:** the same 31 questions paired with expert-defined expected responses.

4.2 Component Evaluations

4.2.1 Persona Extraction.

Table 1: Persona Extraction Performance

Metric	Score
Persona Detection Rate	100% (23/23)
Metrics Recall	95.3% (1,002/1,051)
Metrics Precision	96.8% (1,002/1,035)

Objective. Extract structured persona indicators (demographics, behaviors, values) from PDF pages.

Methodology. The evaluation follows a controlled three-step process. First, representative pages related to the Curious Connoisseurs segment are manually annotated to enumerate all expected personas, indicators, and quantitative metrics. Second, the automated extraction pipeline is executed on the same pages. Finally, the system output is systematically compared against the manual annotations to measure coverage and accuracy.

Matching Logic. A system-extracted metric is considered a match if the indicator and statement labels each have at least 70% word overlap with the ground truth, the numeric value matches exactly, and the unit is semantically equivalent (e.g., % vs. percentage, index vs. idx).

Evaluation Metrics. Performance is quantified using three complementary metrics: *Persona Detection Rate*, measuring whether all expected personas are identified; *Metric Recall*, capturing how many ground truth metrics are successfully extracted; and *Metric Precision*, assessing the proportion of extracted metrics that are correct.

Experimental Setup. Both extraction and persona reasoning are performed using the *mistral-medium-3-instruct* model with the temperature parameter set to 0 to ensure deterministic and consistent extraction behavior. The evaluation spans the 23-page Curious Connoisseurs subset, with 1,051 manually validated metrics serving as ground truth.

Results and Discussion. The results summarized in Table 1 show that the system accurately identifies all personas and recovers most expected indicators with minimal false positives. Overall, these findings confirm that the vision-LLM pipeline performs with near-human accuracy, effectively handling visually complex and heterogeneous documents.

4.2.2 *Fact Extraction.*

Objective. Convert PDF pages to clean, structured markdown for RAG indexing.

Methodology. We first manually collect key facts and numbers from the original PDF pages to create a reference list. The automated system then converts those same pages into markdown text. A fact counts as correctly extracted if at least 80% of its words appear in the markdown output, showing that the main meaning and wording are preserved.

Key Metrics. The main metric is *Exact Match Accuracy*, measuring the percentage of ground truth facts correctly found in the generated markdown.

Table 2: Fact Extraction Performance

Metric	Score
Overall Accuracy	97.0% (453/467)

Experimental Setup. The fact extraction pipeline is implemented using the *mistral-large-3-675b-instruct-2512* vision–language model, with the temperature parameter fixed at 0 to ensure deterministic and reproducible extraction behavior. Evaluation is performed on a 23-page subset corresponding to the *Curious Connoisseurs* segment, using 467 manually extracted and validated snippets as ground truth.

Results and Discussion. As shown in Table 2, the system achieves high overall accuracy, demonstrating that the vision-to-markdown conversion process is reliable and effective for transforming complex PDF documents into structured, machine-readable formats suitable for downstream RAG applications.

4.2.3 *Retrieval Relevance.*

Objective. This evaluation measures how effectively the *RAG retrieval module* retrieves relevant context in response to persona-specific queries.

Methodology. We ask the system 31 questions related to the target persona. For each question, the system retrieves the top-k=25 most relevant text chunks. A human evaluator then reviews these chunks and marks each one as either relevant or not relevant. Using these judgments, we calculate *Precision@K* and *Recall@K* for $k = 3, 5, 10, 20$, and 25. Precision@K measures the proportion of retrieved chunks among the top k that are actually relevant, reflecting how accurate the system's highest-ranked results are. Recall@K, on the other hand, measures the proportion of all relevant chunks that appear within the top k retrieved results, indicating how completely the system captures the available relevant information.

Experimental Setup. The retrieval relevance evaluation is conducted using the *all-mpnet-base-v2* embedding model. The full 222-page *Customer Segmentation Analysis* report is segmented into 712 overlapping text chunks, each containing approximately 1,200 characters with a 50-character overlap to preserve contextual continuity. The evaluation focuses on 31 manually curated questions covering demographics, behaviors, attitudes, and brand-related attributes specific to the Curious Connoisseurs segment.

Results and Discussion. Table 3 shows consistently high precision across all values of k, indicating effective noise filtering. However, lower recall at small k reflects fragmentation of relevant information across chunks. Optimal performance is reached at $k = 20$, where both precision and recall exceed 82%, providing the *LLM* with the context necessary for complete answers.

4.2.4 *Authenticity Evaluation.*

Objective. Validate that AI responses authentically represent the consumer segment.

Methodology. To evaluate authenticity, a set of questions (such as, "What is your age?") are created for a specific persona. The AI

Table 3: Retrieval Performance

Retrieval Window	Precision@k	Recall@k
$k = 3$	89.25%	14.87%
$k = 5$	88.39%	23.35%
$k = 10$	86.13%	45.98%
k = 20	**82.58%**	**82.08%**
$k = 25$	80.09%	100.00%

Table 4: Persona Authenticity Performance

Criteria	Score
Authenticity	4.66/5
Style Alignment	4.74/5
Factual Grounding	4.44/5

persona then generates complete responses to each question. These responses are reviewed by domain experts, who rate them on a 1–5 scale based on three key criteria: authenticity, style alignment, and factual grounding. Authenticity measures how well the response reflects the target persona, style alignment assesses whether the tone and vocabulary are appropriate, and factual grounding evaluates whether the response is supported by accurate data or citations.

Experimental Setup. The authenticity evaluation uses the *mistral-small-24b-instruct* model for context filtering with the temperature fixed at 0 to ensure fully deterministic and reproducible selection of relevant context. The same model is used for response generation with a temperature of 0.2, allowing limited controlled variability to produce more natural and coherent outputs while preserving factual consistency. Text embeddings are produced using the *all-mpnet-base-v2* model. The dataset consists of 222 pages from the Customer Segmentation Analysis PDF, segmented into 712 text chunks of 1, 200 characters each with a 50-character overlap. Ground truth data includes 31 questions covering demographics, behavior, attitudes, and brand-related attributes, all focused on the Curious Connoisseurs segment.

Results and Discussion. Table 4 summarizes the evaluation outcomes. The system achieves strong performance across all criteria, with particularly high scores in style alignment (4.74/5) and authenticity (4.66/5). Factual grounding scores slightly lower (4.44/5), indicating remaining room for improvement in constraining factual accuracy. Overall, the results confirm effective persona expression, while highlighting the need for tighter grounding mechanisms. The evaluation is reviewed and validated by a Lavazza expert tutor.

5 Conclusion

In this work, we present a complete chatbot system that allows marketers to interact with an *LLM* representing a specific customer segment, thereby supporting the assessment of alternative marketing strategies. Our proposed solution also addresses the challenge of automatically extracting semantic information from visually complex PDF documents by leveraging *VLLMs*. By integrating a chatbot architecture with a *RAG Retrieval Module*, a memory module, and a Persona Registry, the architecture employs prompt-tuning to guide the *LLM* in generating responses grounded in retrieved evidence, historical interactions, and predefined persona traits. The design improves factual grounding, contextual coherence, and response explainability. Overall, the solution demonstrates satisfactory performance across persona and fact data extraction, as well as in retrieval accuracy and authenticity evaluation.

Despite these encouraging results, several opportunities for improvement remain. As future work, persona indicator and fact data extraction can be enhanced by subdividing each PDF page into smaller semantic regions and providing these localized sections to the *VLLM*, rather than processing entire pages at once, the retrieval can also be improved as in [2]. Additionally, instead of relying on prompt-tuning to induce persona-specific behavior, future extensions could explore fine-tuning *LLMs* using *PEFT* techniques. While such approaches require more computational resources, they may yield more stable and consistent persona representations.

References

[1] Razan Baltaji, Babak Hemmatian, and Lav R. Varshney. 2024. Persona Inconstancy in Multi-Agent LLM Collaboration: Conformity, Confabulation, and Impersonation. arXiv:2405.03862 [cs.AI] https://arxiv.org/abs/2405.03862

[2] Chandana Cheerla. 2025. Advancing Retrieval-Augmented Generation for Structured Enterprise and Internal Data. arXiv:2507.12425 [cs.CL] https://arxiv.org/abs/2507.12425

[3] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model based Multi-Agents: A Survey of Progress and Challenges. arXiv:2402.01680 [cs.CL] https://arxiv.org/abs/2402.01680

[4] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating Large Language Models: A Comprehensive Survey. arXiv:2310.19736 [cs.CL] https://arxiv.org/abs/2310.19736

[5] Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful Persona-based Conversational Dataset Generation with Large Language Models. arXiv:2312.10007 [cs.CL] https://arxiv.org/abs/2312.10007

[6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL] https://arxiv.org/abs/2005.11401

[7] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084 [cs.CL] https://arxiv.org/abs/1908.10084

[8] Muhammed Rizwan, Lars Carlsson, and Mohammad Loni. 2025. PersonaBOT: Bringing Customer Personas to Life with LLMs and RAG. arXiv:2505.17156 [cs.CL] https://arxiv.org/abs/2505.17156

[9] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. arXiv:2406.01171 [cs.CL] https://arxiv.org/abs/2406.01171

[10] Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2023. DocLLM: A layout-aware generative language model for multimodal document understanding. arXiv:2401.00908 [cs.CL] https://arxiv.org/abs/2401.00908

[11] Tiannan Wang, Meiling Tao, Ruoyu Fang, Huilin Wang, Shuai Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2024. AI PERSONA: Towards Life-long Personalization of LLMs. arXiv:2412.13103 [cs.CL] https://arxiv.org/abs/2412.13103

[12] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining (KDD '20)*. ACM, 1192–1200. doi:10.1145/3394486.3403172

[13] Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. 2025. OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation. arXiv:2412.02592 [cs.CV] https://arxiv.org/abs/2412.02592