Politecnico di Torino
International University

LAVAZZA
TORINO, ITALIA, 1895

AI PERSONAS

Enrico Chen - s337750
Van Thanh Nguyen - s336748
Xiaoning Ma - s337332

Project overview

Politecnico di Torino
International University

OUR GOAL

➔ Build an **AI platform** that enables internal business teams to **interact with AI personas** representing different market segments, in order to e**xplore and test strategies.**

KEY FEATURES

➔ **Automatically organize** unstructured data
➔ **Generate data-driven personas** based on real market data
➔ **Deliver explainable answers** with clear references

BENEFITS

➔ **Reduce time and resources** spent on data preprocessing
➔ Provide **trustworthy and credible** responses
➔ Help teams make more **focused and informed** market strategies

VALUE PROPOSITION

Politecnico di Torino
International University

For **business units** struggling with **evaluating marketing performances, customer understanding, models and ideas testing**, our platform allows **interacting with data-driven AI Personas** representing the different **market segments.**

Research Questions

Extraction

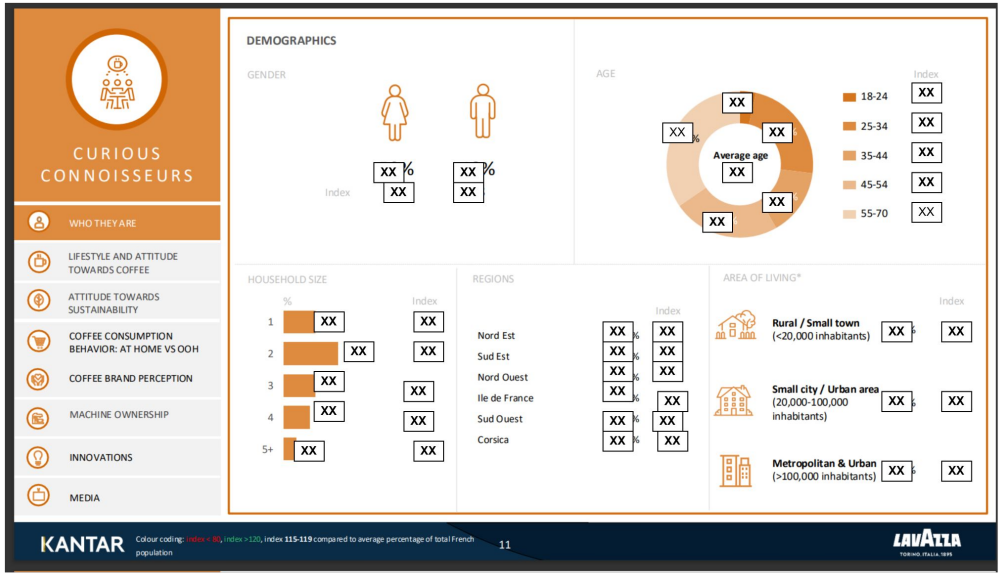
How can we automatically extract accurate data from **complex, visually rich** PDFs while preserving layout semantics?

Personality

How can we infer **latent reasoning traits** (style profiles & value frames) from heterogeneous, noisy raw data to build consistent AI personas?

Grounding

How can we **reduce hallucinations** in language model outputs while maintaining factual grounding?



For privacy and compliance reasons, certain details are omitted from this presentation.

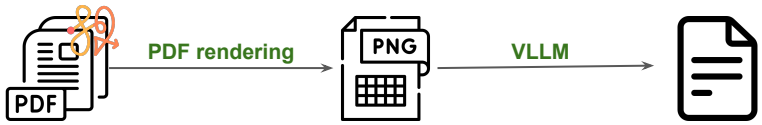
Research Question 1

Question:

How can we automatically extract accurate data from **complex, visually rich** PDFs while preserving layout semantics?

Our solution:

We treat PDFs as visual documents and extract semantic meaning using a VLLM



- Preserves layout and spatial relationships
- Fully automated, no manual rules
- Directly produces schema-consistent data ready for downstream use

Alternative Solutions:

- **Manual annotation**: high accuracy, but low scalability, high costs
- **OCR-based pipelines**: scalable, but limited layout information

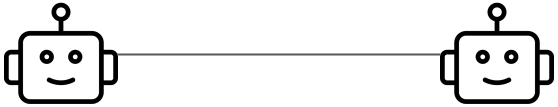
Research Question 2

Question:

How can we infer **latent reasoning traits** (style profiles & value frames) from heterogeneous, noisy raw data to build consistent AI personas?

Our solution:

We use a cooperative multi-agent system in which each agent specializes in a specific stage of reasoning and trait inference.



Agent 1
Extract persona information from raw data

Agent 2
Reason over extracted evidence to infer latent persona traits and priorities

- Scalable and repeatable across datasets
- Reduces subjective human bias
- Produces consistent persona traits at scale

Alternative Solutions:

- **Human review and inference**: high accuracy but slow, subjective, and not scalable
- **Single-model end-to-end inference**: automated, but sensitive to noise and lacking trait consistency

Research Question 3

Question:

How can we **reduce hallucinations** in language model outputs while maintaining factual grounding?

Our solution:

We integrate a RAG system that retrieves relevant evidence before generation and constraints model outputs to the retrieved context.

Orchestrator
Integrate retrieved information into the generation context

RAG
Retrieve relevant facts and indicators from the vector database

- Improves factual accuracy and reduces hallucinations
- Produces explainable, evidence-grounded responses
- Maintains low computational costs

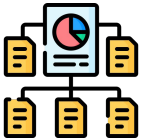
Alternative Solutions:

- **Prompt-based**: lightweight, but unreliable and sensitive to prompt design
- **Pure fine-tuning**: improves fluency, but does not guarantee factual grounding, computationally expensive

Dataset

Source & Scale

- Kantar France 2023
- 4,001 Respondents
- 9 consumer segments
- Rich qualitative + quantitative data



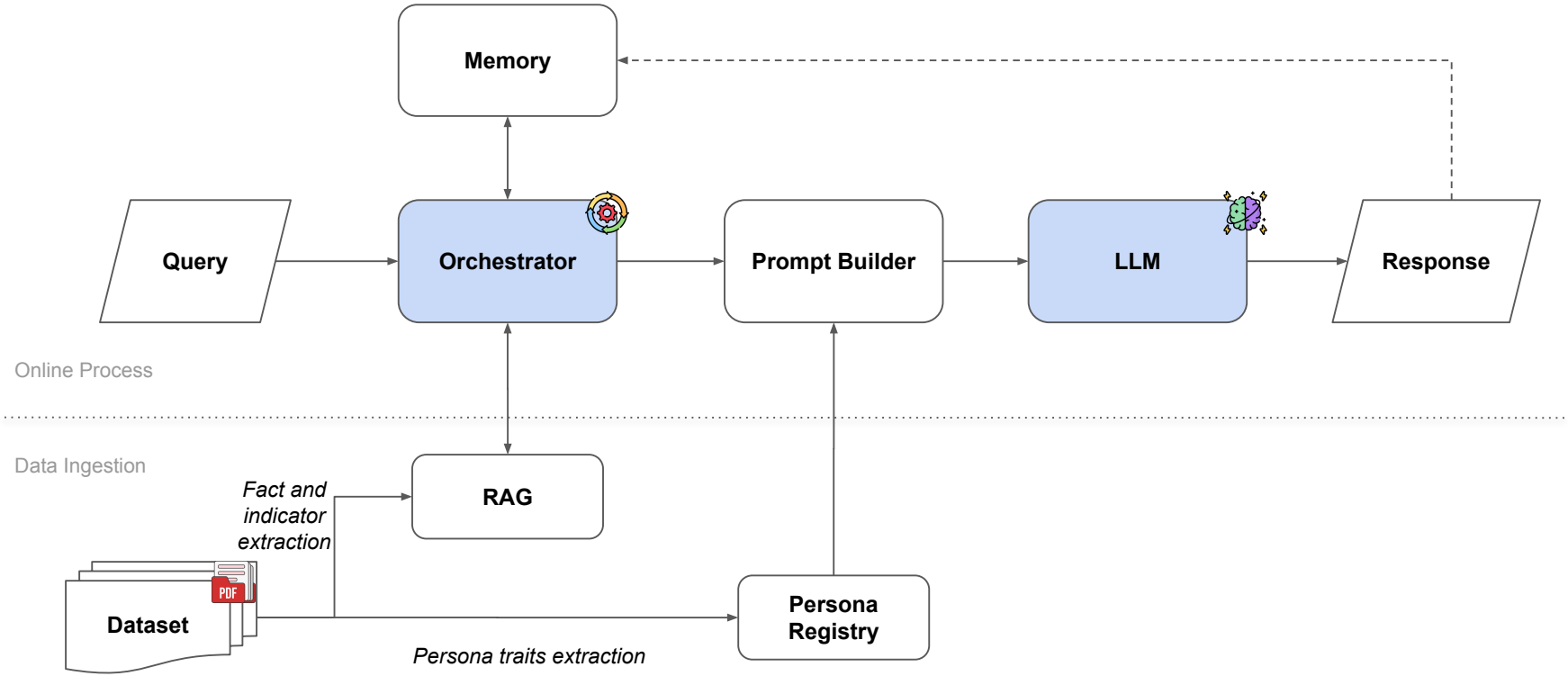
Data Structure

- Multi-modal composition
- Inconsistent layouts across segments
- Mixed granularity
- No unified or machine-readable format

Key Dimensions

- Demographics
- Psychographics
- Coffee attitudes
- Consumption behaviours
- Brand perception & sustainability attitudes

Functional Diagram



Fact data: raw data extracted from dataset (including indicators)
Persona traits data: characteristics of the customer segments (like tone, style, values, preferences)

Method - Persona Extraction

Each page of the dataset

VLLM Extraction

JSON

JSON indicators

VLLM Reasoning

Persona Traits

Prompt tuning

Fine-tuned AI Personas

Chunker and Embedding Model

Database

EXAMPLE

Indicator: “Segment S does not buy plastic packed items”
Trait:

- **style_profile** (how they speak) includes “innovative”
- **value_frame** (what they prioritize) includes “sustainability”
- **reasoning_policies** (decision rules) includes “sustainability concerns”

Method - Fact Extraction

Each page of the dataset

VLLM

Markdown output

Chunker

Chunked output

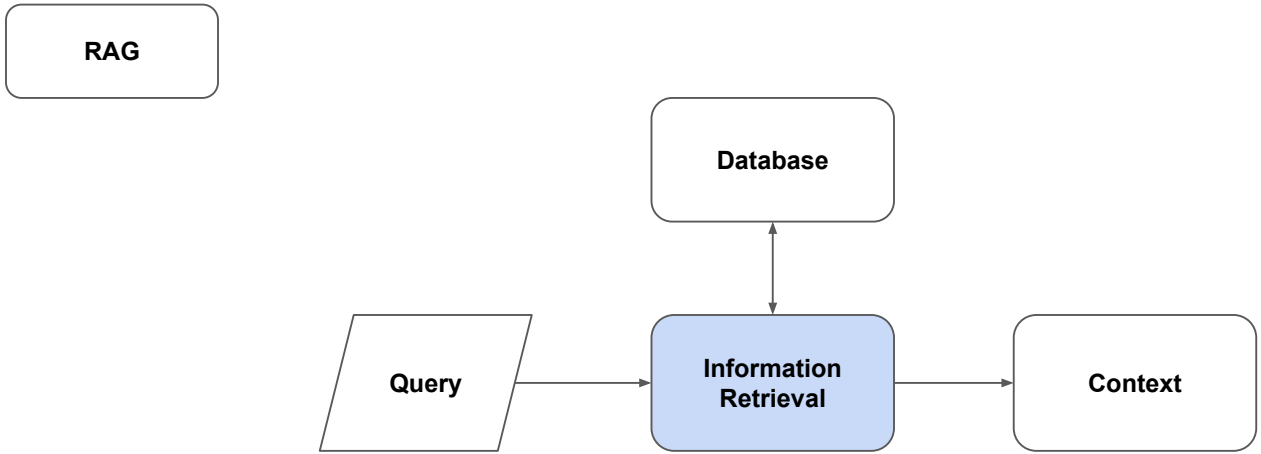
Embedding Model

Database

Fact data:

- raw data from the dataset
- includes indicators
- included for future inputs

Method - RAG



Given the query and customer segment type:

- Retrieve k chunks for indicators
- Retrieve k chunks for fact data
- Return the chunks to the orchestrator

Method - Memory and Persona Registry

Politecnico di Torino
International University

Memory

- It stores the past max_items interactions
- It enables multi-turn conversations where the persona can reference previous messages
- It is updated at each new interaction

Persona Registry

- It is a catalog of segments
- It provides the correct persona to chat with together with its traits

Method - Prompt Builder

Prompt Builder

The prompt builder is the module that constructs the prompt to be sent to the LLM, it combines:

- **System prompt** (persona segment, name of the persona, persona traits, summary biography, answer guidelines). **Persona registry** is used to get the information about the persona (customer segment, name, traits, biography)
- **Filtered conversation history**
- **Retrieved filtered context**
- **User query**

Method - Orchestrator

The orchestrator is the main coordination module that wires together all the components to process a user query:

- It calls **input handler** to normalize the user query
- It calls the **memory** module to get past interactions
- It calls the **RAG** to extract indicators and fact data, and merges the 2k chunks
- It filters past interactions and 2k chunks
- It calls the **prompt builder** to build the prompt
- It sends the prompt to an LLM to get the answer
- It updates the **memory** with this new interaction (query and answer)

Method - Application

Back-end

(REST APIs)

Handles system operations and provides results to the frontend

Streamlit

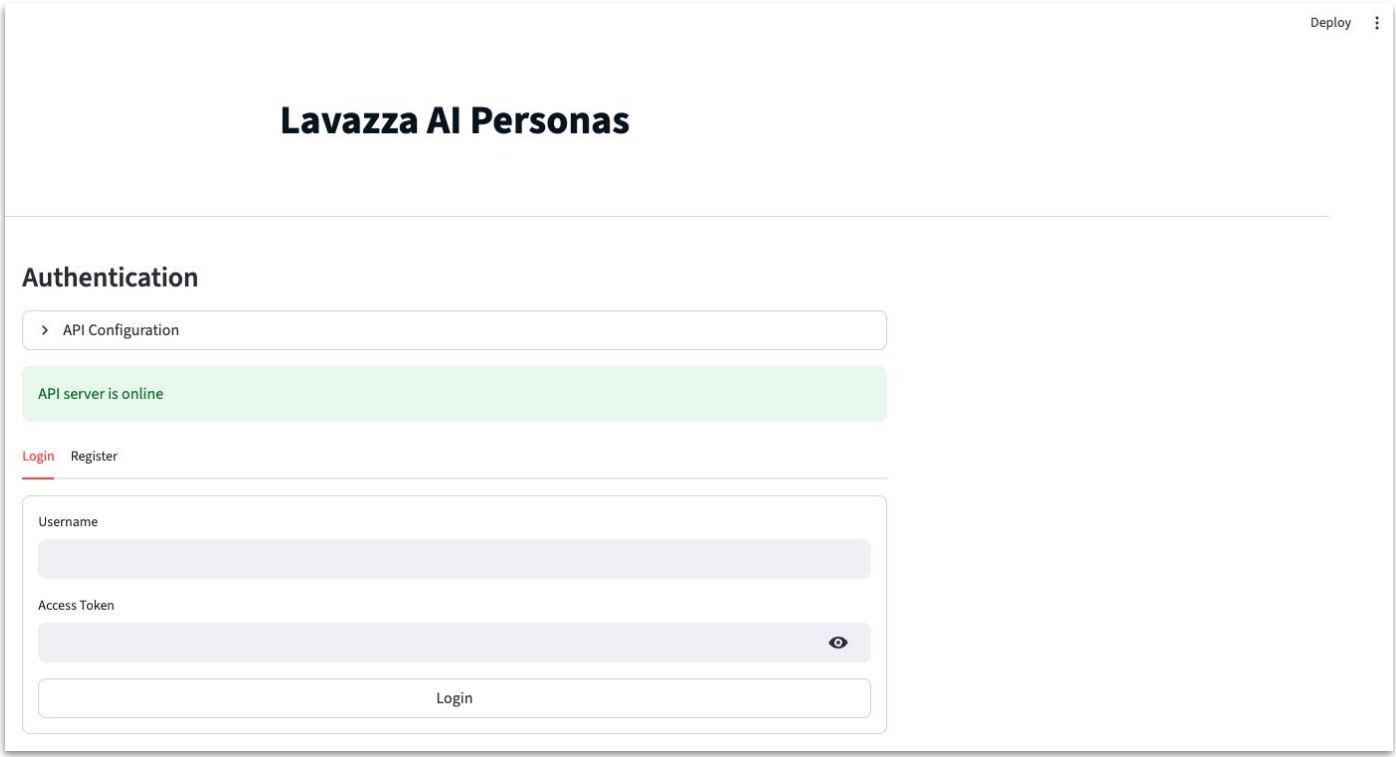
Front-end

User interface for interacting with the backend, built with Streamlit

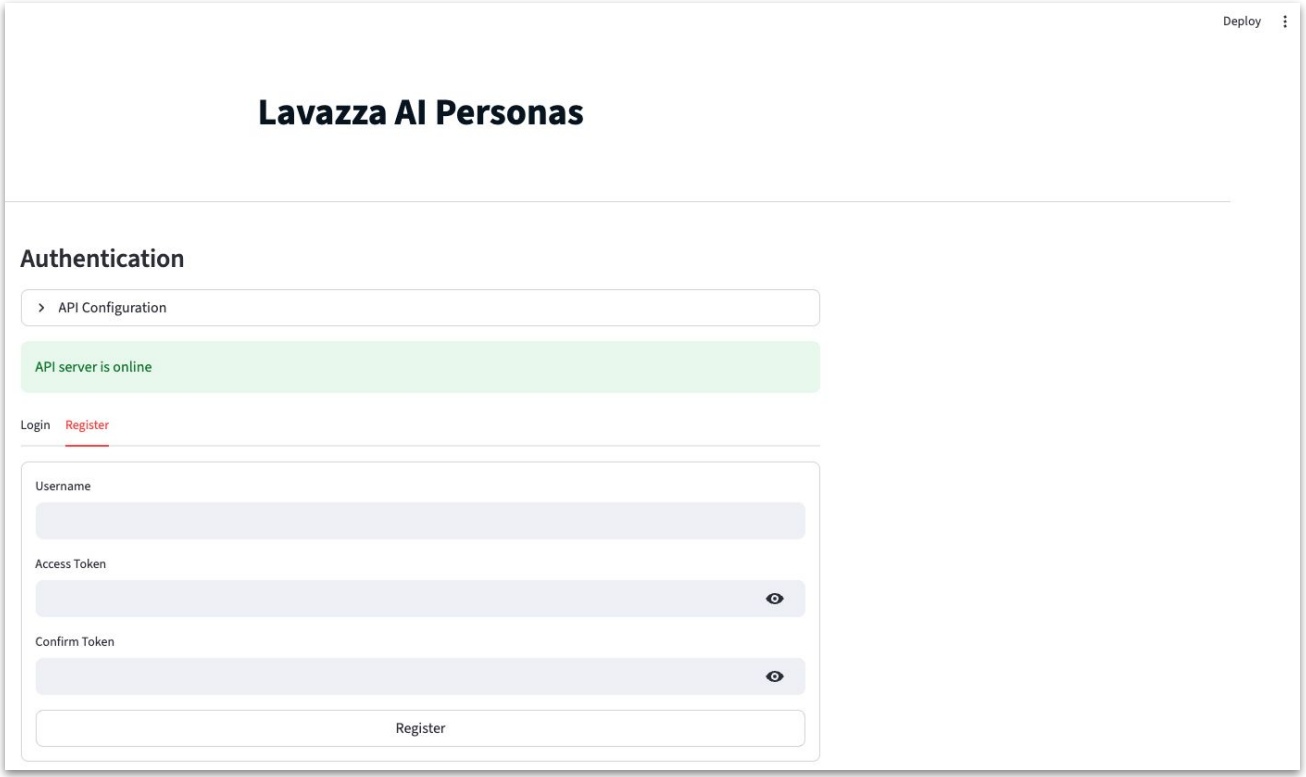
Features implemented

1. Authentication
2. Customer segment selection
3. Past chat selection
4. Context customization
5. Persona name customization
6. Persona information

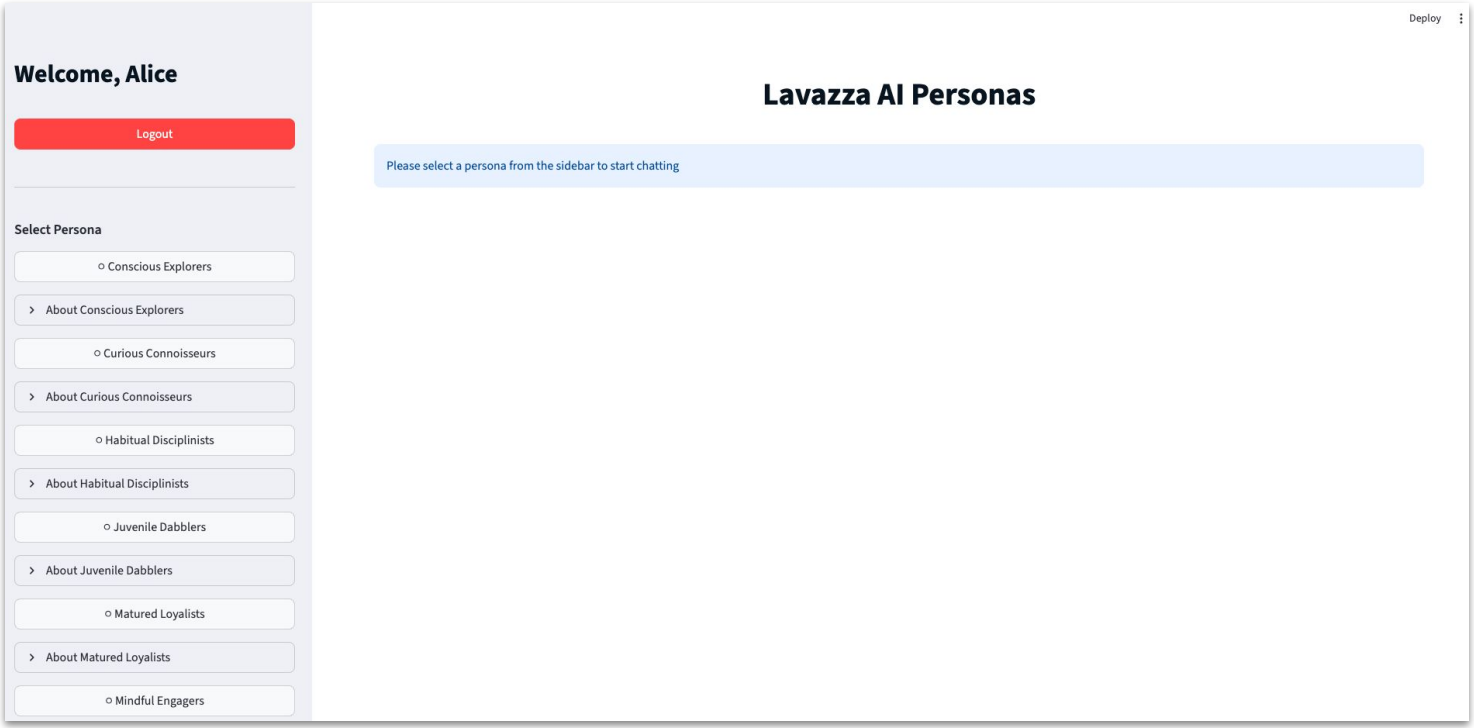
UI - Login


Deploy
Lavazza AI Personas
Authentication
API Configuration
API server is online
Login
Register
Username
Access Token
Login

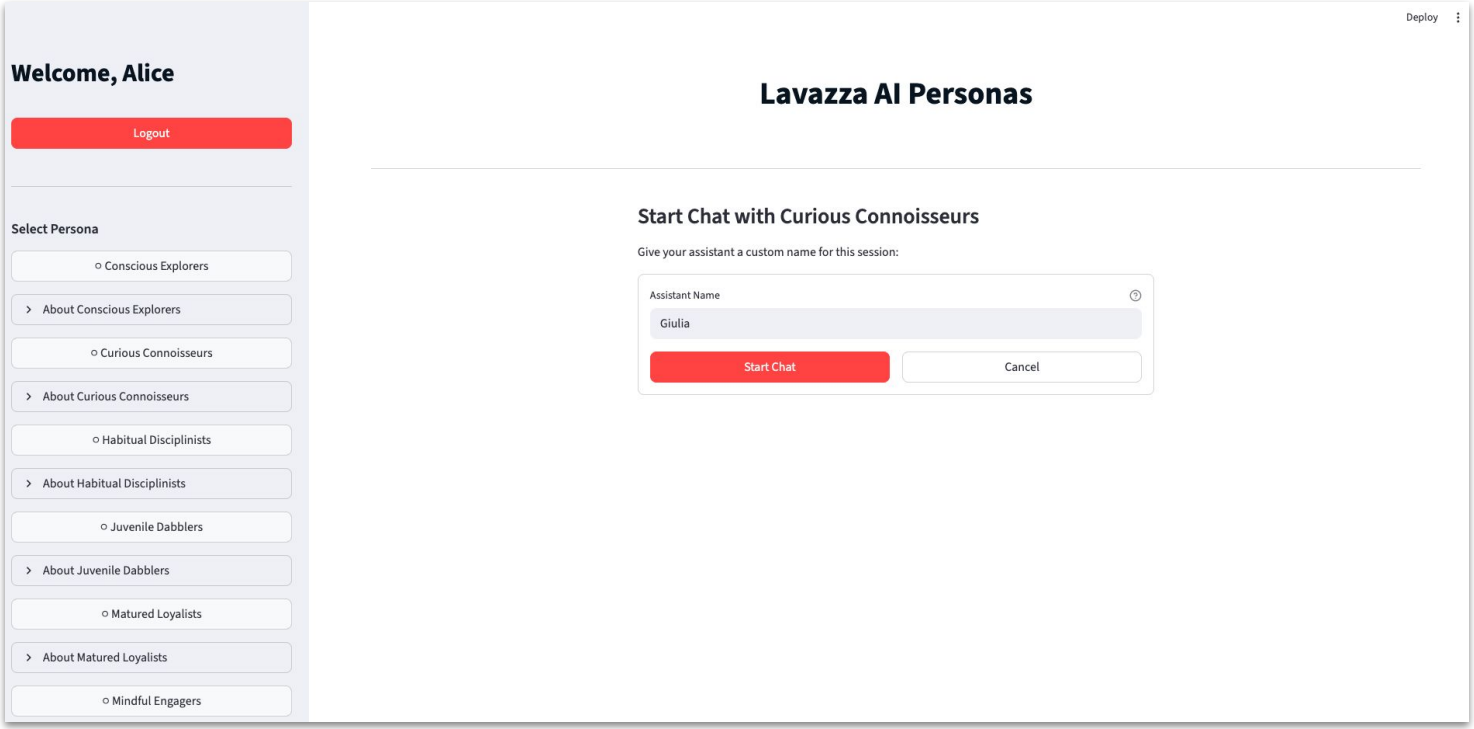
UI - Registration



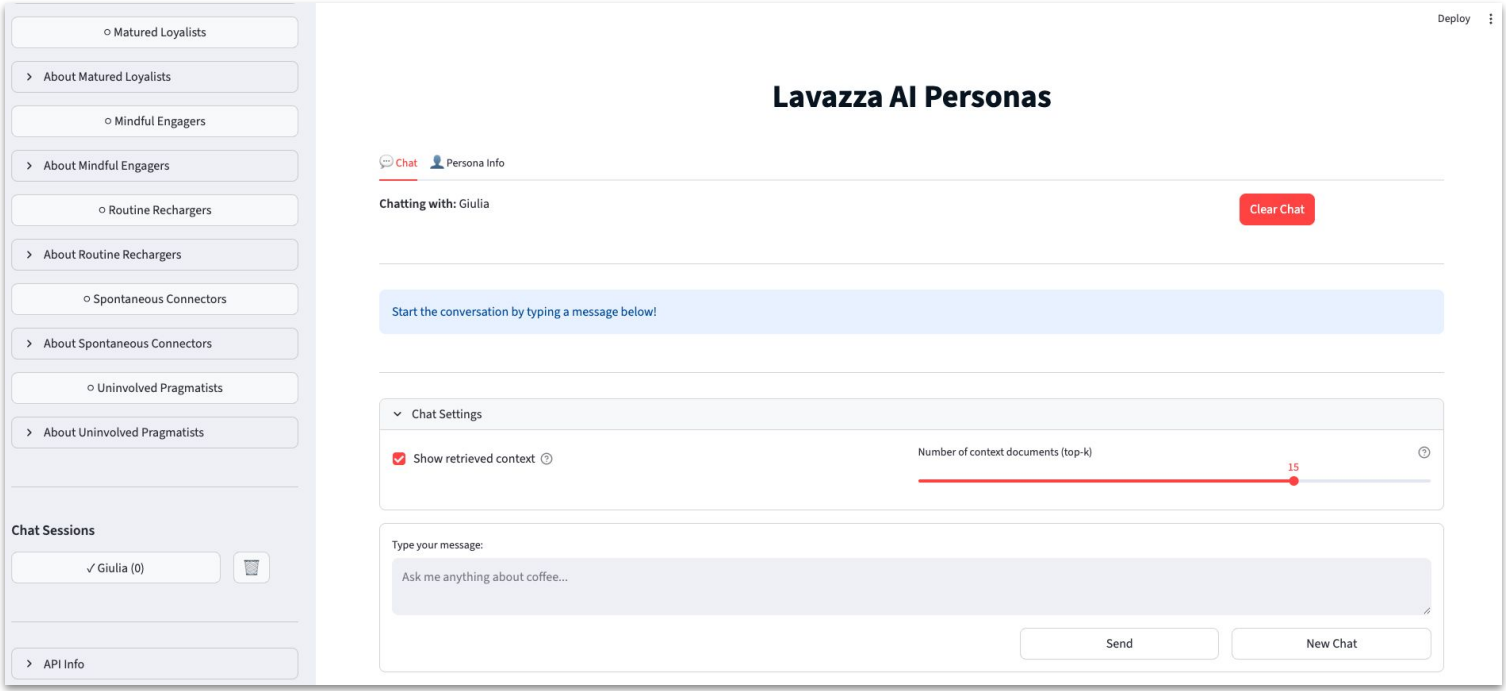
UI - Homepage


Deploy
Welcome, Alice
Logout
Select Persona
Conscious Explorers
About Conscious Explorers
Curious Connoisseurs
About Curious Connoisseurs
Habitual Disciplinists
About Habitual Disciplinists
Juvenile Dabblers
About Juvenile Dabblers
Matured Loyalists
About Matured Loyalists
Mindful Engagers
Lavazza AI Personas
Please select a persona from the sidebar to start chatting

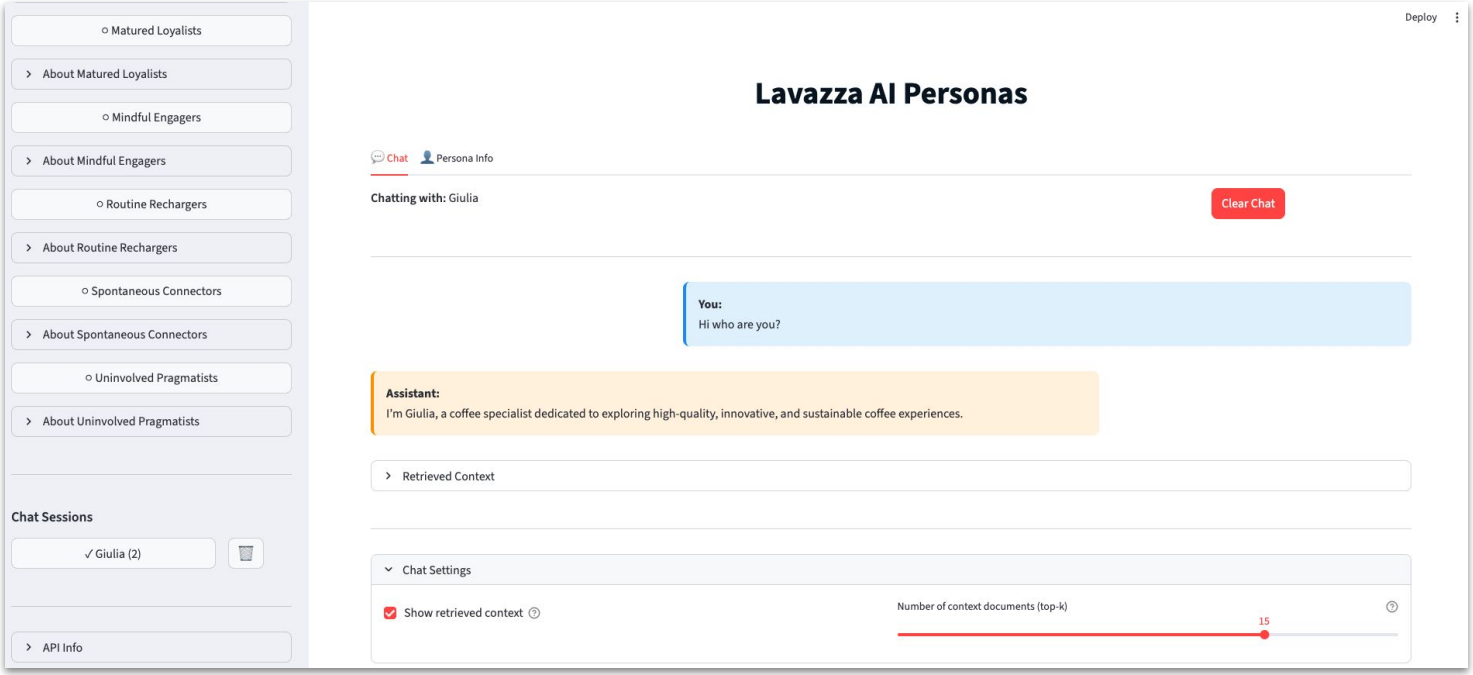
UI - Persona name customization



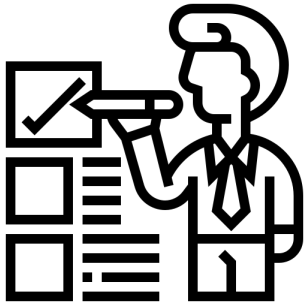
UI - Chat Management



UI - Chat Interaction



Evaluation



Persona Extraction

Is persona information correctly extracted?

Fact Extraction

Is extracted fact data correct?

Retrieval Relevance

Is the retrieved information relevant?

Authenticity Evaluation

Do personas reflect real consumer segments?

Evaluation - Persona Extraction

Approach

Compare persona extraction outputs with manually annotated ground truth.

Metrics

➔	**Persona Detection Rate**	Are all target personas identified?
➔	**Metrics Recall**	Is the required information complete?
➔	**Metrics Precision**	Is the extracted information accurate?

*Persona Detection Rate = (Correctly Detected Personas) / (Total Ground Truth Personas)
*Metrics Recall = (Correct Metrics) / (Total Ground Truth Metrics)
*Metrics Precision = (Correct Metrics) / (All Extracted Metrics)

Evaluation - Persona Extraction

Configuration

- **Models**:
  - **Indicators extractor:** mistralai/mistral-medium-3-instruct (Temperature = 0)
  - **Persona traits reasoning**: mistralai/mistral-medium-3-instruct (Temperature = 0)
- **Test dataset:**
  - **Scope**: 23 pages focused on the Curious Connoisseurs segment
  - **Ground Truth**: 1,051 metrics manually extracted and validated from the PDF

Result

	Score	Interpretation
Persona Detection Rate	100%	All personas correctly identified (23/23)
Metrics Recall	95.30%	Very high coverage of ground-truth metrics (1002/1051)
Metrics Precision	96.80%	Minimal noise in extracted metrics (1002/1035)

Evaluation - Fact Extraction

Approach

- Manually verify numeric and factual data against source PDFs.
- Run extraction and check exact matches.
- A result is correct only if ≥80% textual overlap with the extracted markdown.

Metrics

➔ **Exact Match Accuracy** what fraction of expected facts were extracted correctly?

*Exact Match Accuracy = (Exactly Matching Values) / (Total Ground Truth Values)

Evaluation - Fact Extraction

Configuration

- ➔ **Models**:
  - ◆ **PDF to Markdown Conversion:** mistralai/mistral-large-3-675b-instruct-2512 (Temperature = 0)
- ➔ **Test dataset:**
  - ◆ **Scope**: 23 pages focused on the Curious Connoisseurs segment
  - ◆ **Ground Truth**: 467 validated metric and statement snippets manually extracted from the PDF

Result

	Score	Interpretation
Exact Match Accuracy	97%	High extraction accuracy with minimal deviation from ground truth

Evaluation - Retrieval Relevance

Approach

- Create persona-specific test queries.
- Retrieve top-K documents (K = 3, 5, 10, 20).
- Manually label relevance for retrieval documents.

Metrics

➔ **Precision@K** How many of the top K results are actually correct.

➔ **Recall@K** How many of the total correct results were successfully found within the top K.

*Precision@K = (Relevant docs in top-K) / K
*Recall@K = (Relevant docs in top-K) / (Total relevant docs)

Evaluation - Retrieval Relevance

Politecnico di Torino International University

Configuration

- ➔ **Models**:
  - ◆ **Embedding:** sentence-transformers/all-mpnet-base-v2
- ➔ **RAG setup**:
  - ◆ **Chunk Size**: 1,200 characters
  - ◆ **Chunk Overlap**: 50 characters
  - ◆ **Input**: 222-page split to 712 total text chunks
- ➔ **Test dataset:**
  - ◆ **Ground Truth**: 31 evaluation questions (e.g. “What are your preferred brands?”, …) focused on the Curious Connoisseurs segment
  - ◆ Top-25 retrieval per query; relevant documents manually labeled

- ➔ Optimal performance occurs at **k=20**, supplying sufficient context for complete responses.

	Score	Interpretation
Precision@3	89.25%	Minimal noise at the top
Precision@5	88.39%	Still high quality
Precision@10	86.13%	relevance remains dominant
Precision@20	**82.58%**	**Precision stabilizes at broader window**
Recall@3	14.87%	Very limited early coverage
Recall@5	23.35%	Most relevant docs missed early.
Recall@10	45.98%	Roughly half captured in top 10.
Recall@20	**82.08%**	**Most relevant documents recovered.**

Result

Evaluation - Persona Authenticity

Approach

- Generate persona-based responses to predefined questions.
- Have domain experts score them on authenticity, style alignment, and factual grounding, each on a 1–5 scale.

Metrics

➔	**Expert Authenticity Score**	Average authenticity ratings.
➔	**Style Alignment Score**	Average all style alignment ratings.
➔	**Factual Grounding Score**	Average all factual grounding ratings.

Evaluation - Authenticity Results

Politecnico di Torino International University

Configuration

- **Models**:
  - **Filter context model:** mistralai/mistral-small-24b-instruct (Temperature = 0)
  - **Generation model:** mistralai/mistral-small-24b-instruct (Temperature = 0.2)
- **RAG Setup:**
  - **Chunk Size**: 1,200 characters (~300–350 tokens)
  - **Chunk Overlap**: 50 characters (to preserve context continuity)
  - **Input**: 222-page to 712 total text chunks
- **Test dataset:**
  - **Ground Truth**: 31 evaluation questions focused on the Curious Connoisseurs segment

Result

	Score	Interpretation
Expert Authenticity Score	4.66/5	Persona behavior is largely authentic.
Style Alignment Score	4.74/5	Style is mostly consistent, with minor persona drift.
Factual Grounding Score	4.44/5	Responses are well grounded.

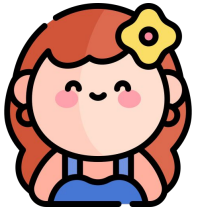
Manage

Enrico

- System design
- Semantic fact extraction
- Fact ingestion & indexing
- RAG retrieval logic
- Frontend development
- Monitoring & evaluation
- Communication & demo

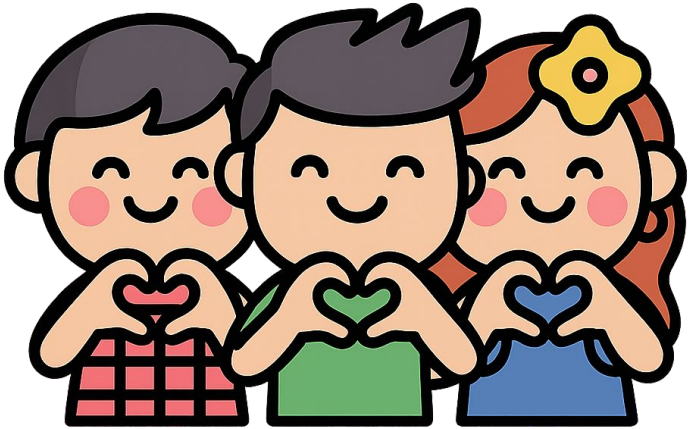
Thanh

- Project conception & initiation
- Data foundation
- Core layer
- Persona semantic extraction
- AI persona prompt tuning
- Backend development
- Evaluation: persona and fact data extraction



Xiaoning

- Management
- Application layer design
- AI persona common trait extraction
- Persona inference & serving
- Evaluation dataset construction
- Evaluation: Retrieval performance and persona authenticity analysis
- Quality assessment



THANK YOU

Enrico Chen - s337750
Van Thanh Nguyen - s336748
Xiaoning Ma - s337332

LAVAZZA

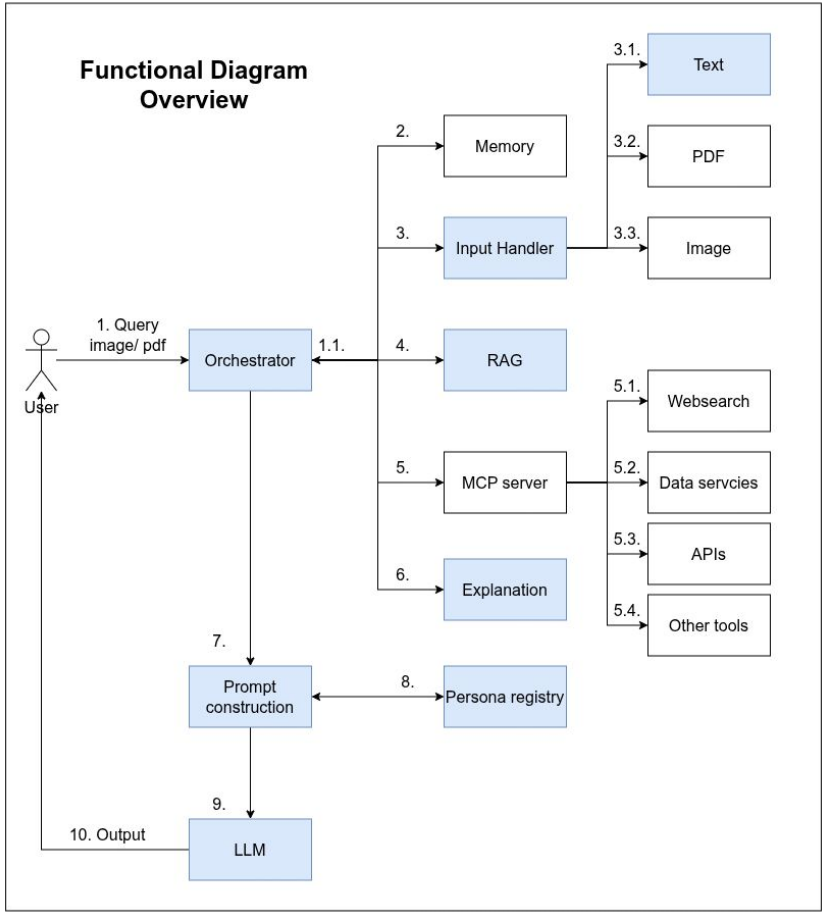
TORINO, ITALIA, 1895

Politecnico di Torino

International University

Appendix

Design



1. User Query Submission: User sends a query with optional attached files (image, PDF, etc.) to the Orchestrator.
1.1 Orchestrator Analysis: The Orchestrator analyzes the query and attachments to decide which services should be used.
2. Memory Integration : Extract useful information from chat history.
3. Input Preprocessing: Inputs are preprocessed before passing to the model.
3.1 Text Input: Normalize text to make it easier to handle in later steps.
3.2 PDF Input: Parse, process, and extract meaningful information from PDF files.
3.3 Image Input: Process images and extract valuable information.
4. Context Retrieval (RAG System): Use the query and relevant input information to retrieve context (e.g., market data) via a RAG system.
5. Tool Selection & MCP Server Requests

- Decide which tools should be used to enrich the context.
- Send requests to the MCP server to gather corresponding context.

5.1 Web Search: Extract updated information from the internet (trends, real-time data, missing internal data, etc.).
5.2 Database Query: Retrieve useful data from internal or external databases.
5.3 External APIs: Call APIs to obtain additional information.
5.4 Other Tools: Use calculators, simulators, weather data extractors, or other utilities to enrich context.
6. Explanation: The explanation module will explain in detail the thought process of the reasoning model and the data used for the thinking process (citations from RAG).
7. Prompt Construction: The Orchestrator aggregates useful context and passes it to Prompt Construction.
8. Persona Selection : Apply the selected Persona profile, including: Demographics, Behavior Data, Transactional Data, …
9. LLM : Route to a particular LLM model, passing the enriched prompt and context.
10. Model Response: Generate a response with: Specific personality,Tone, Linguistic style of the Persona

AI Personas Extraction & Fine-Tuning

Indicators (VLLM extraction output)

Example indicator shape (JSON):

```
{
  Indicator: {
    sources: {
      "url: {"https://www.pclavazza.com/nead/20/Indicator"
    },

    statements: {
      'statement: individual insights within an indicator, sostons and influences'},
      'metrics' {'marks whethr a svisually emphasiz: 'index', '%', 'count', 'rank'},
      'influene': flags whether a statement shapes tone or stance, 'smex': 'sources'}
    }
  }
}
```

AI Personas Extraction & Fine-Tuning

Politecnico di Torino
International University

Traits (reasoning output)

Persona Blueprint (Traits)

style_profile

how they speak: tone, formality, directness, emotional flavour, criticality, verbosity, preferred structures, example phrases.

value_frame

what they prioritize: priority_rank (sustainability, price, etc.), novelty seeking, brand loyalty, health concern, description.

reasoning_policies

purchase_advice, product_evaluation, information_processing, content_filters (biases, rules, praise/criticism triggers, trust, disclaimers).

Example JSON Structure

```
// How the model should "speak"
style_profile: (
  tone_adjectives: string[],      // ["Curious", "confident", "quality-focused", "pragmatic", ...],
  formality_level: "low" | "medium" | "high",
  directness: "very_direct" | "balanced" | "hedged",
  emotional_flavour: "neutral" | "enthasiastic" | "cool_detached" | "warm_reflective",
  criticality_level: "high" | "medium" | "low",
  verbosity_preference: "concise" | "detailed" | "varies_by_question",
  preferred_structures: string[],  // ["bullet_point", "clear_rsadeoffs", "pros_cons", "step_by_step"]
  typical_register_examples: string[]// short example phrases in target style
},

// What they care about – used to biss recommendations / reasoning
value_frame: (
   priority_rank: string[],       // e.g. ["quality", "convenience", "sustainability", "price"],
   sustainability_orientation: "high" | "medium" | "low",
   price_sensitivity: "high" | "medium" | "low",
   price_sensitivity: "high" | "medium" | "low",
   novelty_seeking: "high" | "medium" | "low"
),
}
```

Design - Risks Analysis

Technical Risks

- Hallucinations and inaccurate responses: mitigate with RAG system
- Insufficient critical thinking: mitigate with RAG and prompt engineering
- Opacity: mitigate with RAG
- Inconsistent or generic personality: mitigate by fine-tuning (in case of limited resource use PEFT, smaller models, RAG with few-shot prompting)
- Performance evaluation difficulty

Design - Risks Analysis

Governance and Security Risks

- Privacy and compliance with AI Act and GDPR
- Proprietary data protection
- System integration difficulty with existing systems and infrastructure

Design - Risks Analysis

Data and Other Risks

- Data integration difficulty
- Data quality and bias
- Over relying on AI Personas