System requirements

InLine Prod

Classification	unclassified
Status	in examination
Program name	InLine Prod
Version	0.1
Date	19. Dezember 2023
Author(s)	Pietro Belfanti, Gabriele Bottinelli
Distribution	

Description

The system requirements describe requirements for the InLine Prod AI Assistant. They are structured according to Hermes 5 together with related system models and prototypes.

Contents

List of Figures

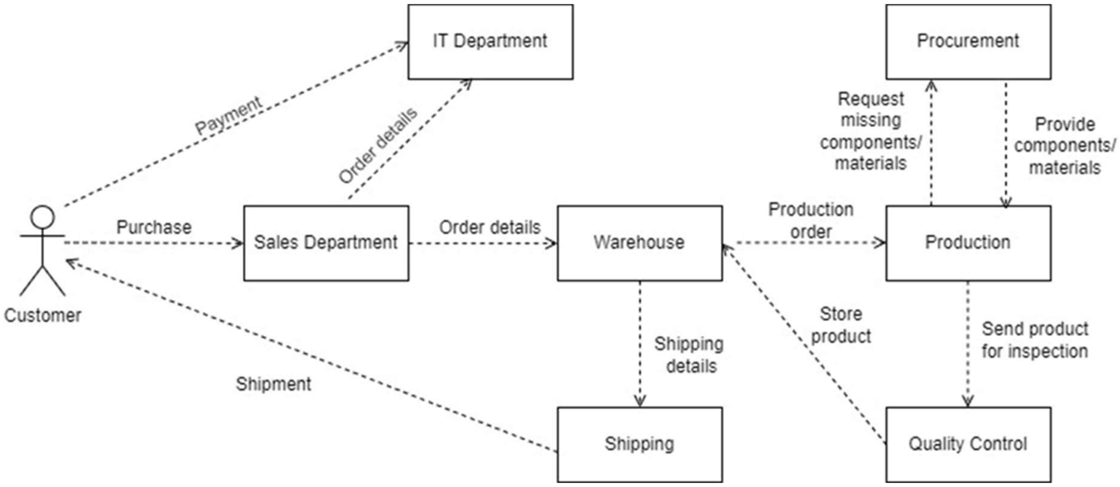
1. System Overview

A high-level overview of the system.

1.1. High-Level Overview

1.1.1. Information System Overview

The system structure, seen in a high-level overview diagram



1.1.2. Main Use Cases and Features

SALES:

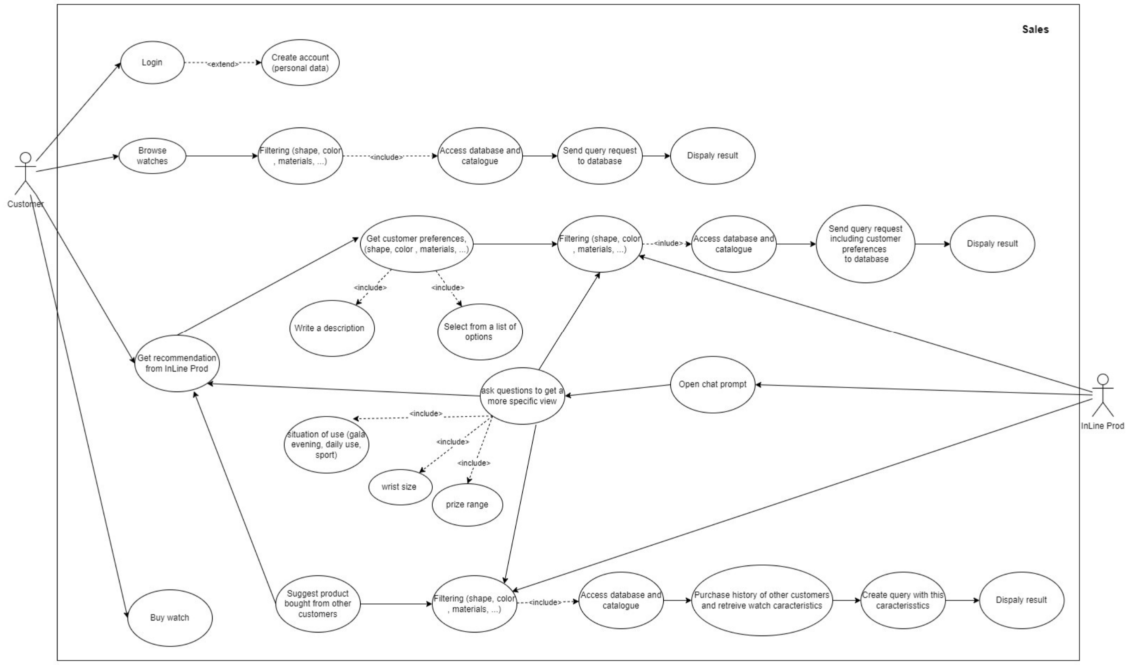

Sales
Customer
Login
<extend>
Create account (personal data)
Browse watches
Filtering (shape, color , materials, ...)
<include>
Access database and catalogue
Send query request to database
Dispaly result
Get customer preferences, (shape, color , materials, ...)
Filtering (shape, color , materials, ...)
<inlude>
Access database and catalogue
Send query request including customer preferences to database
Dispaly result
<include>
<include>
Write a description
Select from a list of options
Get recommendation from InLine Prod
ask questions to get a more specific view
Open chat prompt
InLine Prod
<include>
situation of use (gala evening, daily use, sport)
<include>
wrist size
<include>
prize range
Buy watch
Suggest product bought from other customers
Filtering (shape, color , materials, ...)
<include>
Access database and catalogue
Purchase history of other customers and retreive watch caracteristics
Create query with this caracterisstics
Dispaly result

Figure 1: Sales Model

CUSTOMER SUPPORT:

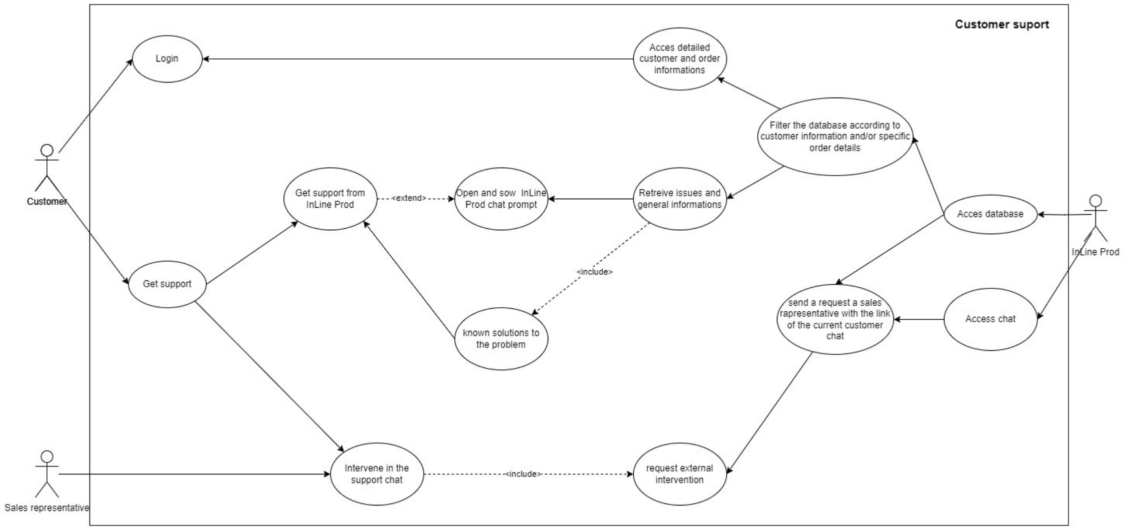

Customer suport
Login
Acces detailed customer and order informations
Customer
Filter the database according to customer information and/or specific order details
Get support from InLine Prod
<extend>
Open and sow InLine Prod chat prompt
Retreive issues and general informations
Acces database
InLine Prod
<include>
Get support
known solutions to the problem
send a request a sales rapresentative with the link of the current customer chat
Access chat
Sales representative
Intervene in the support chat
<include>
request external intervention

Figure 2Customer Support Model

1.2. IT Infrastructure

1.2.1. Components of the IT Infrastructure

The IT system structure, seen in a class diagram

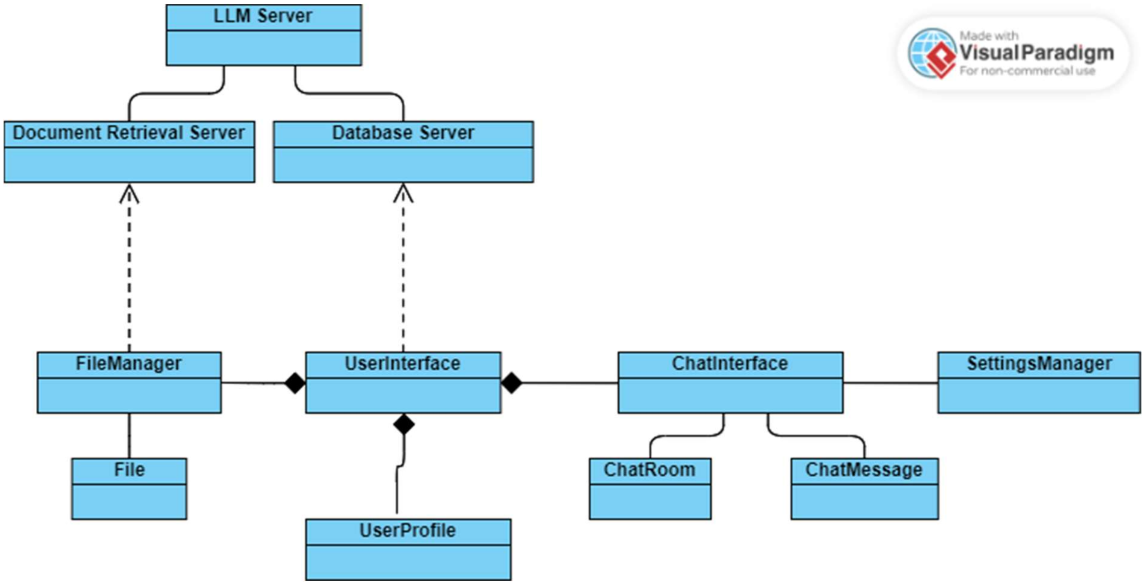


Figure 3: IT class diagram

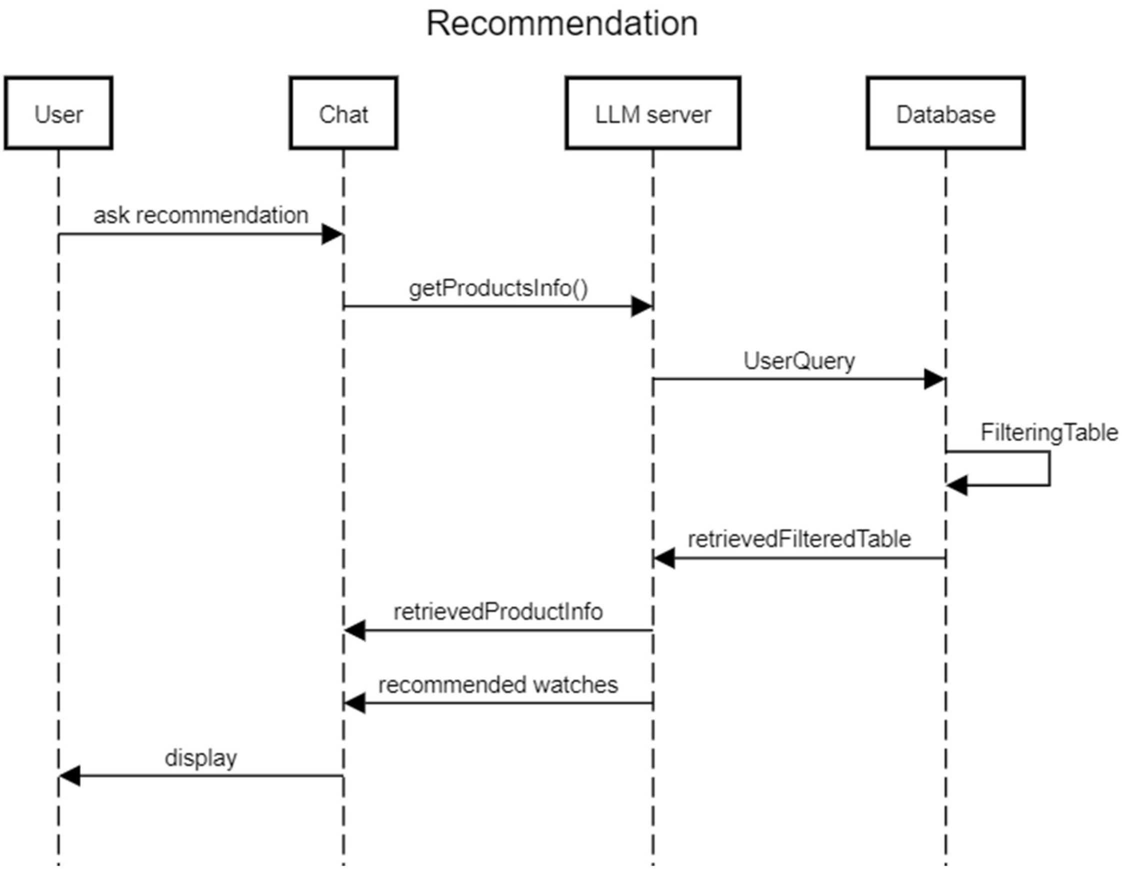


Figure 4: Sequence diagram for recommendation

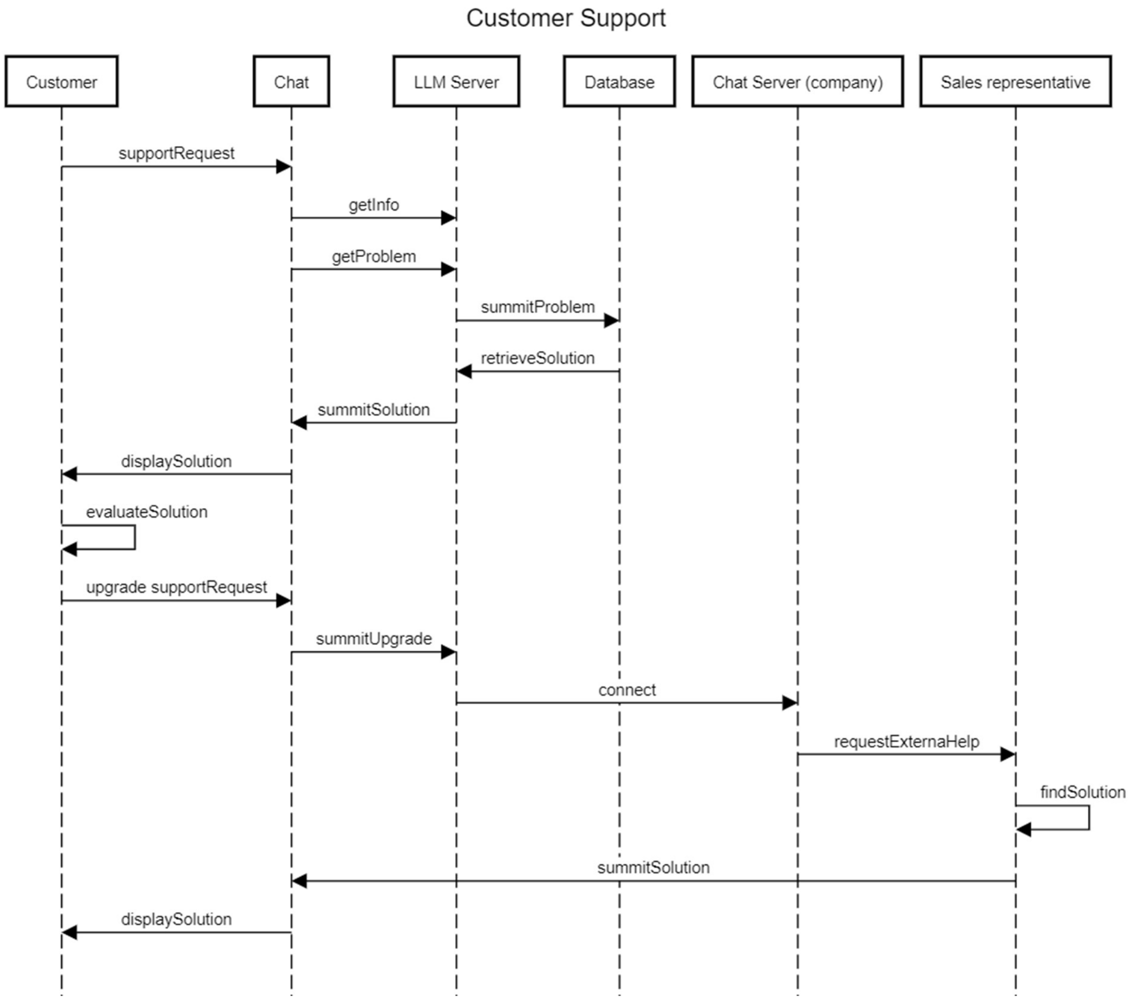


Figure 5: Sequence diagram for customer support

1.2.2. Technical Requirements

Here's an overview of the IT infrastructure needed. We estimate that our company will have around 50 to 100 users.

- **LLM Server**
  - **High-performance CPU (at least 16 cores)**
  - **At least 128 of RAM for handling data with efficiency**
  - **Fast storage like SSD or NVMe, several terabytes**
  - **High-performance network**
  - **LLM framework, libraries, drivers, …**
- **Document Retrieval Server**
  - **High-performance CPU (at least 32 cores)**
  - **At least 128GB to 256GB of RAM for managing document indices.**
  - **Indexing software**
  - **High-bandwidth network connection for serving document retrieval requests.**
  - **Index storage size depends on the number of documents indexed and their size. I can estimate an amount from 20 to 100+ terabytes**
  - **Separate storage for database related data**
- **Database Server**
  - **High-performance CPU (at least 16 cores)**
  - **At least 64GB of RAM to manage and access the database efficiently**
  - **SSD for high-speed storage (1-10 TB)**
  - **DBMS for document storage and document retrieval**
  - **High-performance network to serve the documents to the LLM server**
- **Other**
  - **Robust security protocols**
  - **Firewall**
  - **Data encryption**
  - **High-speed connection with sufficient bandwidth to allow multiple users simultaneously**

1.3. Planning Studies

1.3.1. Technology Evaluation

Natural Language **Processing** (NLP):

- Challenges:
  - **Ambiguity and Context**: Understanding nuances, context, and ambiguities in user queries or text-based interactions.
  - **Model Training and Accuracy**: Developing accurate models for language understanding, requiring vast and diverse datasets and continuous refinement.
  - **Multilingual Support**: Handling multiple languages and dialects for global reach.

Machine Learning (ML) Models (LLM - Language Models):

- Challenges:
  - **Model Complexity and Size**: Managing large-scale models with millions or billions of parameters, requiring substantial computational power and memory.
  - **Model Bias and Ethics**: Addressing biases in training data that might lead to biased outputs or responses.
  - **Continuous Learning**: Enabling models to adapt and learn from new data in real-time.

Database Management Systems:

- Challenges:
  - **Scalability**: Scaling databases to handle increasing data volumes efficiently and ensuring high performance.
  - **Data Security and Privacy**: Implementing robust security measures to safeguard sensitive data from breaches or unauthorized access.
  - **Data Integration**: Ensuring seamless integration and synchronization between different databases or data sources.

Cloud Computing and Infrastructure:

- Challenges:
  - **Resource Management**: Optimizing resource allocation in cloud environments to handle varying workloads and cost-effectiveness.
  - **Latency and Connectivity**: Managing latency issues and ensuring uninterrupted connectivity, especially in distributed systems.
  - **Compliance and Regulations**: Adhering to data privacy regulations and compliance standards in different regions.

Document Retrieval Systems:

- Challenges:
  - **Document Indexing**: Efficiently indexing and searching through a large volume of documents while maintaining accuracy and speed.
  - Scalability: Ensuring scalability to handle growing document repositories without compromising retrieval performance.

  - Relevance and Ranking: Providing relevant and accurate search results based on user queries or system requirements.

Hardware Infrastructure (Server Architecture):

- Challenges:
  - Resource Allocation: Optimizing server configurations for AI workloads, balancing CPU, GPU, and memory requirements.
  - Scalability and Redundancy: Designing resilient architectures that scale seamlessly and provide redundancy for fault tolerance.
  - Energy Efficiency: Addressing power consumption and heat dissipation concerns, especially with high-performance computing.

1.3.2. Feasibility of Use Cases

The primary objective is to establish a comprehensive framework outlining the various actions a customer can undertake when engaging in the online purchase of a watch. Additionally, the goal is to provide the necessary resources for receiving appropriate assistance both post-purchase and during the buying process.

2. Detailed Requirements

2.1. Functional Requirements

2.1.1. Category 1: Sales

These use cases / features concern …

ID	C1-F1	Source	Model, Technology Evaluation, IT Infrastructure, …	Author		Date		Status	approved
Name	Filtering								
Description	Create a query to filter the database and catalogue with specific characteristics								
Acceptance criteria	Company with up-to-date database, fast internet, high-performance server, LLM server								
Importance [1]	4	Urgency [2]	4	Risk [3]	2	Outlay [4]	3		

ID	C1-F2	Source	Model, Technology Evaluation, IT Infrastructure, …	Author		Date		Status	approved
Name	Chat								
Description	Open an interactive chat with the customer to ask specific questions to be used as filters in the query								
Acceptance criteria	LLM server, chat option, internet connection,								
Importance [1]	4	Urgency [2]	4	Risk [3]	2	Outlay [4]	3		

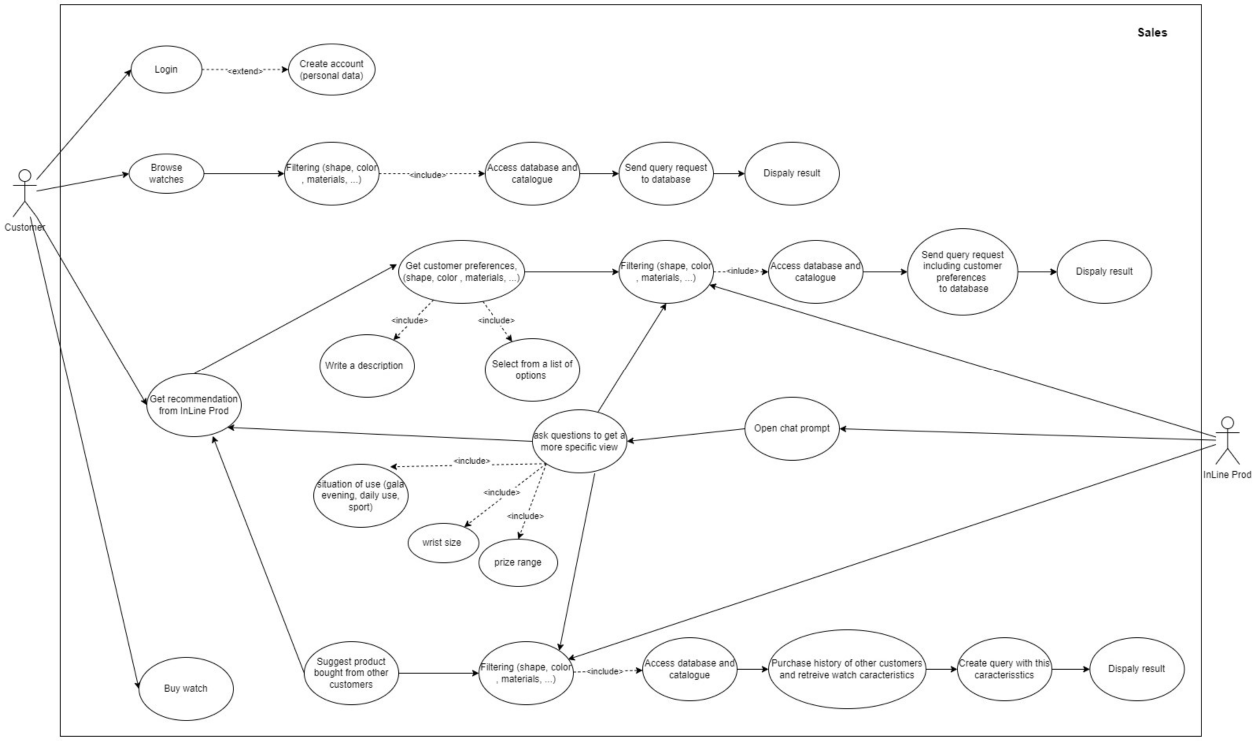

Sales
Customer
Login
<extend>
Create account (personal data)
Browse watches
Filtering (shape, color , materials, ...)
<include>
Access database and catalogue
Send query request to database
Dispaly result
Get customer preferences, (shape, color , materials, ...)
Filtering (shape, color , materials, ...)
<inlude>
Access database and catalogue
Send query request including customer preferences to database
Dispaly result
<include>
<include>
Write a description
Select from a list of options
Get recommendation from InLine Prod
ask questions to get a more specific view
Open chat prompt
InLine Prod
<include>
situation of use (gala evening, daily use, sport)
<include>
wrist size
<include>
prize range
Buy watch
Suggest product bought from other customers
Filtering (shape, color , materials, ...)
<include>
Access database and catalogue
Purchase history of other customers and retreive watch caracteristics
Create query with this caracterisstics
Dispaly result

Figure 6: Sales Models

2.1.2. Category 2: Support

These use cases / features concern …

ID	C2-F1	Source	Model, Technology Evaluation, IT Infrastructure, …	Author		Date		Status	approved
Name	Sending email								
Description	Request external assistance from the sales representative by means of an e-mail with an attached link to the current chat with the customer								
Acceptance criteria	LLM server, email credential for AI, internet connection,								

Importance [1]	4	Urgency [2]	4	Risk [3]	2	Outlay [4]	3

ID	C1-F1	Source	Model, Technology Evaluation, IT Infrastructure, …	Author		Date		Status	approved
Name	Filtering for solution								
Description	Create a query to filter the database and catalogue according to customer information and/or specific order details. Suggest know solution to the problem related to the query filter.								
Acceptance criteria	Company with up-to-date database, fast internet, high-performance server, LLM server, login credential, order information								
Importance [1]	4	Urgency [2]	4	Risk [3]	2	Outlay [4]	3		

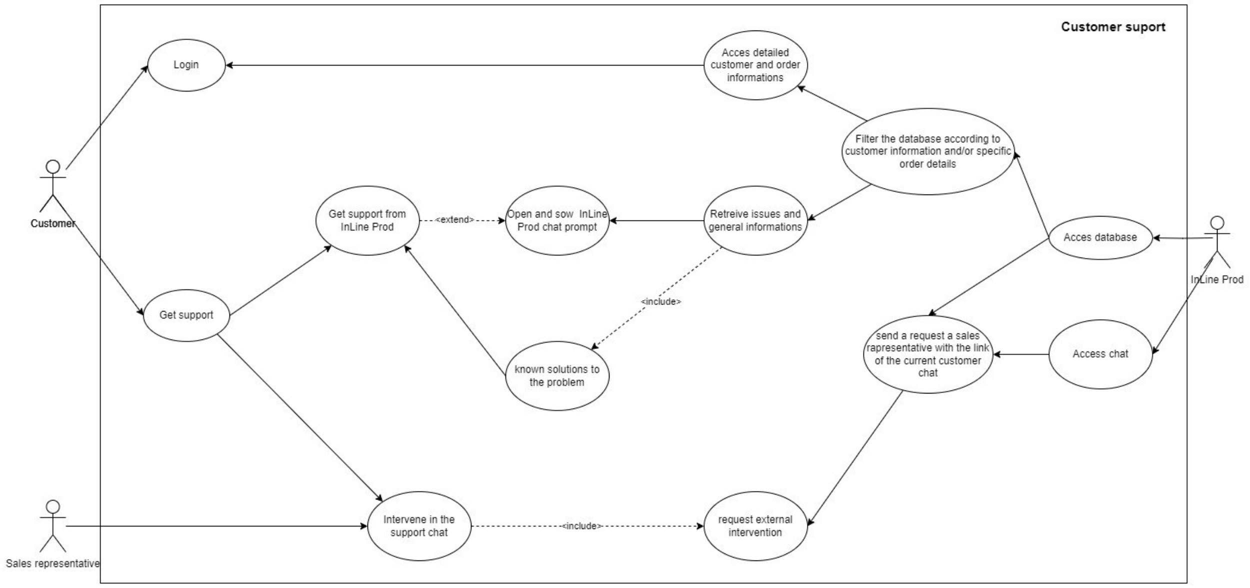

Customer suport
Login
Acces detailed customer and order informations
Filter the database according to customer information and/or specific order details
Customer
Get support from InLine Prod
<extend>
Open and sow InLine Prod chat prompt
Retreive issues and general informations
Acces database
InLine Prod
<include>
Get support
send a request a sales rapresentative with the link of the current customer chat
Access chat
known solutions to the problem
Intervene in the support chat
<include>
request external intervention
Sales representative

Figure 7: Support Model

2.2. Non-Functional Requirements

2.2.1. Category 1

These use cases / features concern …

ID	C1-NF1	Source	Model, Technology Evaluation, IT Infrastructure, …	Author		Date		Status	approved
Name	Data Encryption								
Category	Security								
Description	Customer data and interactions handled by the AI system must be encrypted using industry-standard protocols to ensure the confidentiality and integrity of sensitive information.								
Acceptance criteria									
Importance[1]	5	Urgency[2]	4	Risk[3]	1	Outlay[4]	2		

ID	C1-NF2	Source	Model, Technology Evaluation, IT Infrastructure, …	Author		Date		Status	approved
Name	Error Handling								
Category	Reliability								
Description	The system should have effective error-handling mechanisms in place to gracefully manage and recover from unexpected errors or disruptions.								
Acceptance criteria									

Importance [1]	5	Urgency [2]	4	Risk [3]	1	Outlay [4]	2

2.3. User Interface Prototype

This is a really simple version of our prototype, that shows how a client can ask for recommendation, and in case, request extra help from the help desk.
https://marvelapp.com/project/5946002/screen/83224075

2.4. Relevance Criteria

Each requirement is described with …

- Importance: 5 = mandatory implementation; 4 = very important; 3 = important; 2 = normal; 1 = not important
- Urgency: 5 = must be implemented immediately, 4 = very urgent, 3 = urgent, 2 = normal, 1 = not urgent
- Risk/critical nature: 5 = unacceptable risk, 4 = very high risk, 3 = medium risk, 2 = low risk, 1 = no risk whatsoever
- Outlay: 5 = unacceptable outlay, 4 = very high outlay, 3 = high, 2 = reasonable, 1 = negligible or no outlay