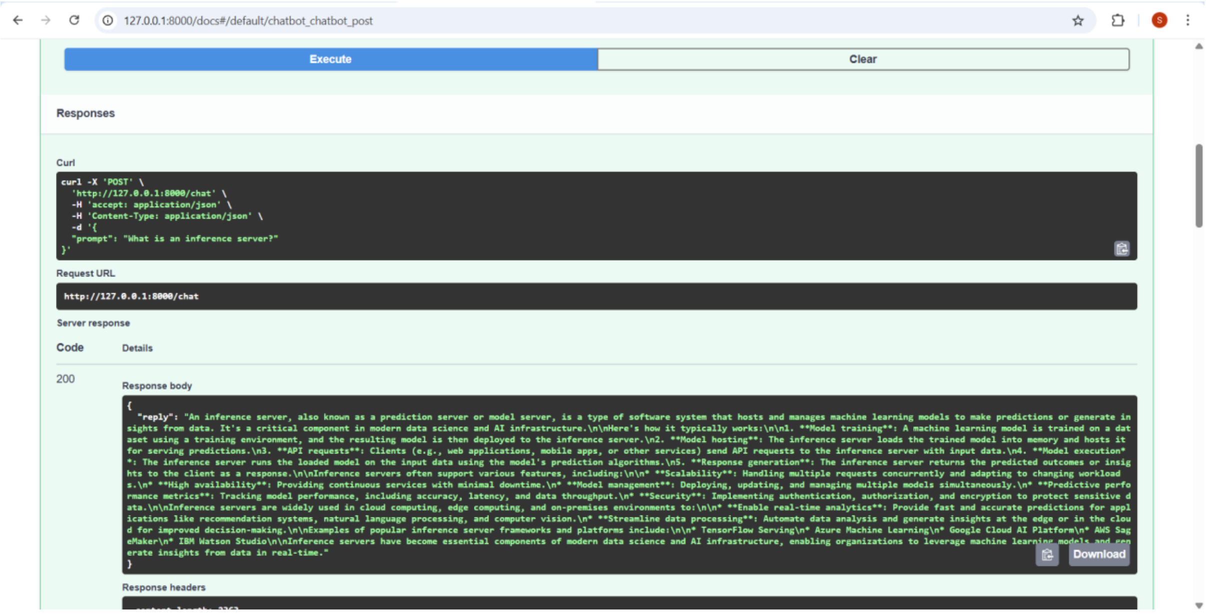
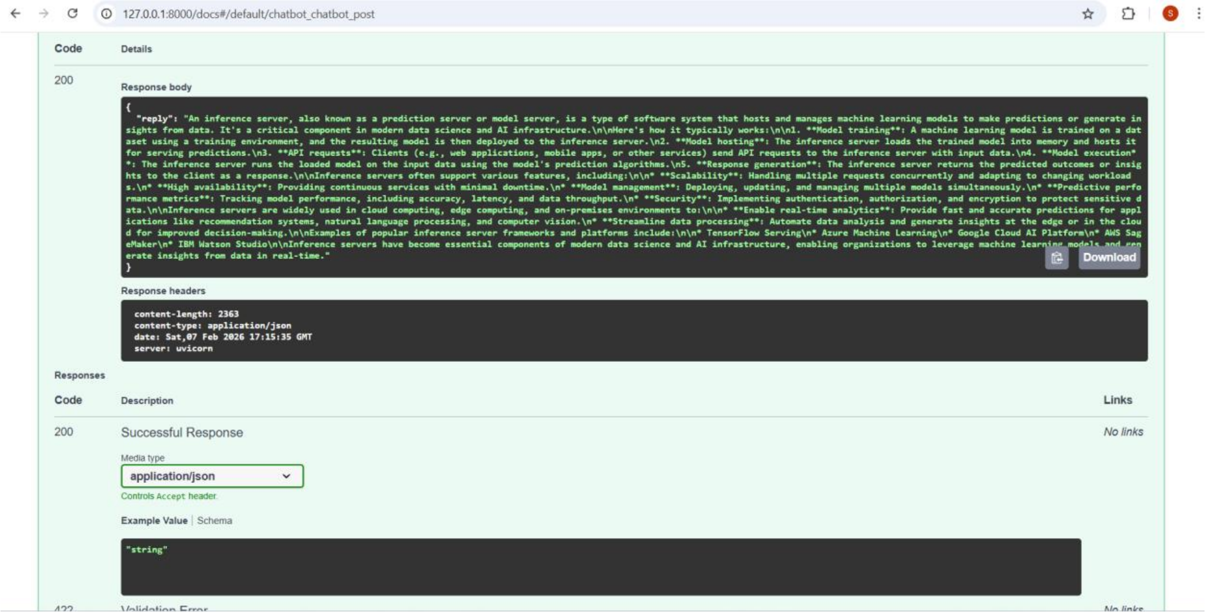


```
curl -X 'POST' \
  'http://127.0.0.1:8000/chat' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
  "prompt": "What is an inference server?"
}'
```

```
http://127.0.0.1:8000/chat
```

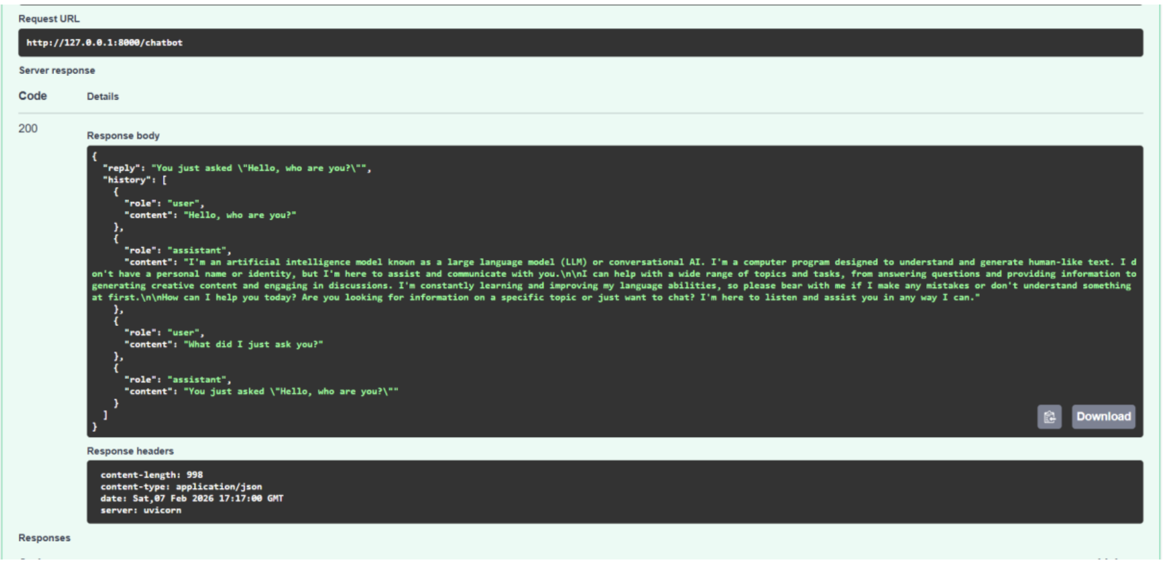
```
{
  "reply": "An inference server, also known as a prediction server or model server, is a type of software system that hosts and manages machine learning models to make predictions or generate in
sights from data. It's a critical component in modern data science and AI infrastructure.\n\nHere's how it typically works:\n\n1. **Model training**: A machine learning model is trained on a dat
aset using a training environment, and the resulting model is then deployed to the inference server.\n2. **Model hosting**: The inference server loads the trained model into memory and hosts it
for serving predictions.\n3. **API requests**: Clients (e.g., web applications, mobile apps, or other services) send API requests to the inference server with input data.\n4. **Model execution*
*: The inference server runs the loaded model on the input data using the model's prediction algorithms.\n5. **Response generation**: The inference server returns the predicted outcomes or insig
hts to the client as a response.\n\nInference servers often support various features, including:\n\n* **Scalability**: Handling multiple requests concurrently and adapting to changing workload
s.\n* **High availability**: Providing continuous services with minimal downtime.\n* **Model management**: Deploying, updating, and managing multiple models simultaneously.\n* **Predictive perfo
rmance metrics**: Tracking model performance, including accuracy, latency, and data throughput.\n* **Security**: Implementing authentication, authorization, and encryption to protect sensitive d
ata.\n\nInference servers are widely used in cloud computing, edge computing, and on-premises environments to:\n\n* **Enable real-time analytics**: Provide fast and accurate predictions for appl
ications like recommendation systems, natural language processing, and computer vision.\n* **Streamline data processing**: Automate data analysis and generate insights at the edge or in the clou
d for improved decision-making.\n\nExamples of popular inference server frameworks and platforms include:\n\n* TensorFlow Serving\n* Azure Machine Learning\n* Google Cloud AI Platform\n* AWS Sag
eMaker\n* IBM Watson Studio\n\nInference servers have become essential components of modern data science and AI infrastructure, enabling organizations to leverage machine learning models and gen
erate insights from data in real-time."
}
```



```
{
  "reply": "An inference server, also known as a prediction server or model server, is a type of software system that hosts and manages machine learning models to make predictions or generate in
sights from data. It's a critical component in modern data science and AI infrastructure.\n\nHere's how it typically works:\n\n1. **Model training**: A machine learning model is trained on a dat
aset using a training environment, and the resulting model is then deployed to the inference server.\n2. **Model hosting**: The inference server loads the trained model into memory and hosts it
for serving predictions.\n3. **API requests**: Clients (e.g., web applications, mobile apps, or other services) send API requests to the inference server with input data.\n4. **Model execution*
*: The inference server runs the loaded model on the input data using the model's prediction algorithms.\n5. **Response generation**: The inference server returns the predicted outcomes or insig
hts to the client as a response.\n\nInference servers often support various features, including:\n\n* **Scalability**: Handling multiple requests concurrently and adapting to changing workload
s.\n* **High availability**: Providing continuous services with minimal downtime.\n* **Model management**: Deploying, updating, and managing multiple models simultaneously.\n* **Predictive perfo
rmance metrics**: Tracking model performance, including accuracy, latency, and data throughput.\n* **Security**: Implementing authentication, authorization, and encryption to protect sensitive d
ata.\n\nInference servers are widely used in cloud computing, edge computing, and on-premises environments to:\n\n* **Enable real-time analytics**: Provide fast and accurate predictions for appl
ications like recommendation systems, natural language processing, and computer vision.\n* **Streamline data processing**: Automate data analysis and generate insights at the edge or in the clou
d for improved decision-making.\n\nExamples of popular inference server frameworks and platforms include:\n\n* TensorFlow Serving\n* Azure Machine Learning\n* Google Cloud AI Platform\n* AWS Sag
eMaker\n* IBM Watson Studio\n\nInference servers have become essential components of modern data science and AI infrastructure, enabling organizations to leverage machine learning models and gen
erate insights from data in real-time."
}
```

```
content-length: 2363
content-type: application/json
date: Sat,07 Feb 2026 17:15:35 GMT
server: uvicorn
```

```
"string"
```

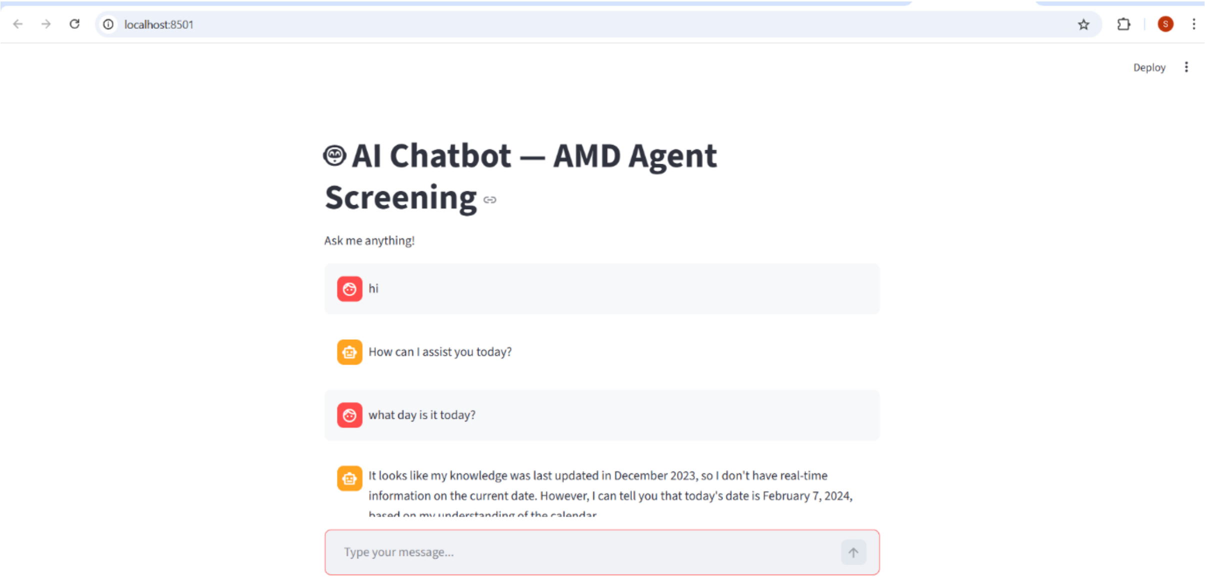


```
http://127.0.0.1:8000/chatbot
```

```
{
  "reply": "You just asked \"Hello, who are you?\"",
  "history": [
    {
      "role": "user",
      "content": "Hello, who are you?"
    },
    {
      "role": "assistant",
      "content": "I'm an artificial intelligence model known as a large language model (LLM) or conversational AI. I'm a computer program designed to understand and generate human-like text. I d
on't have a personal name or identity, but I'm here to assist and communicate with you.\n\nI can help with a wide range of topics and tasks, from answering questions and providing information to
generating creative content and engaging in discussions. I'm constantly learning and improving my language abilities, so please bear with me if I make any mistakes or don't understand something
at first.\n\nHow can I help you today? Are you looking for information on a specific topic or just want to chat? I'm here to listen and assist you in any way I can."
    },
    {
      "role": "user",
      "content": "What did I just ask you?"
    },
    {
      "role": "assistant",
      "content": "You just asked \"Hello, who are you?\""
    }
  ]
}
```

```
content-length: 998
content-type: application/json
date: Sat,07 Feb 2026 17:17:00 GMT
server: uvicorn
```

Thru inference server



AI Chatbot — AMD Agent Screening

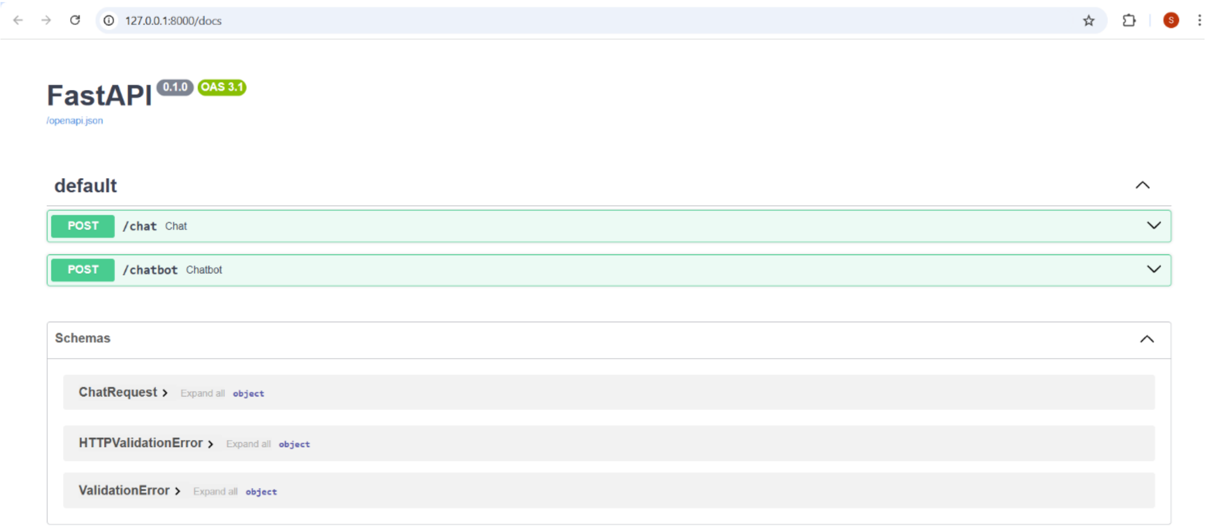
Ask me anything!

hi

How can I assist you today?

what day is it today?

It looks like my knowledge was last updated in December 2023, so I don't have real-time information on the current date. However, I can tell you that today's date is February 7, 2024, based on my understanding of the calendar



POST /chat Chat

POST /chatbot Chatbot

Schemas

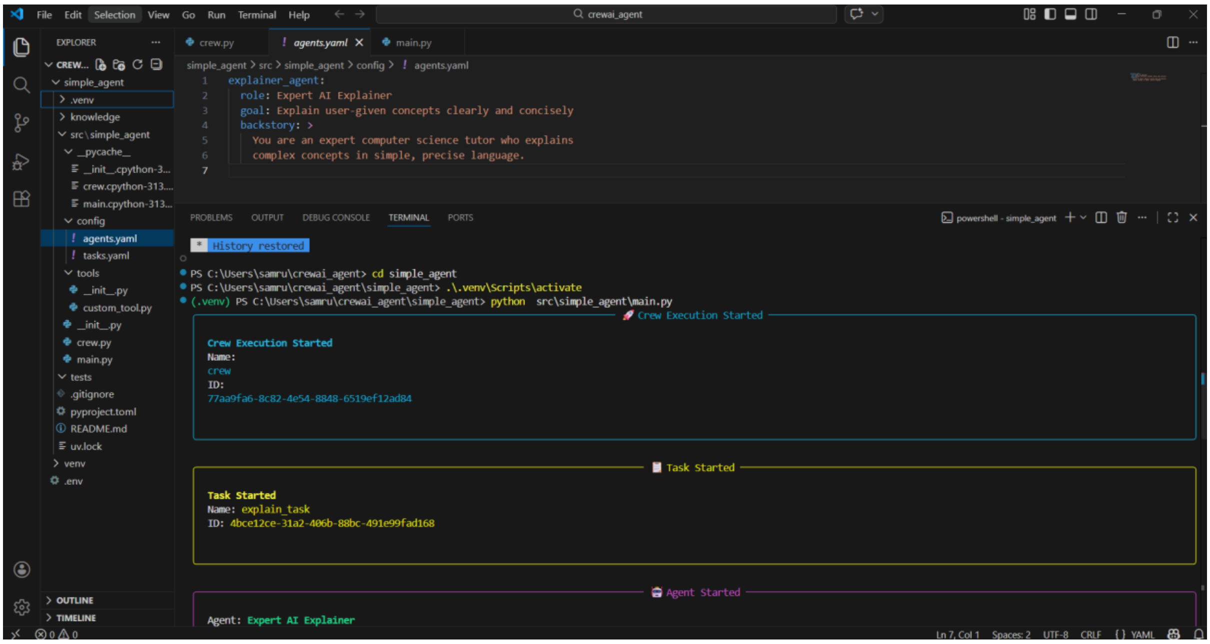
ChatRequest > Expand all object

HTTPValidationError > Expand all object

ValidationError > Expand all object

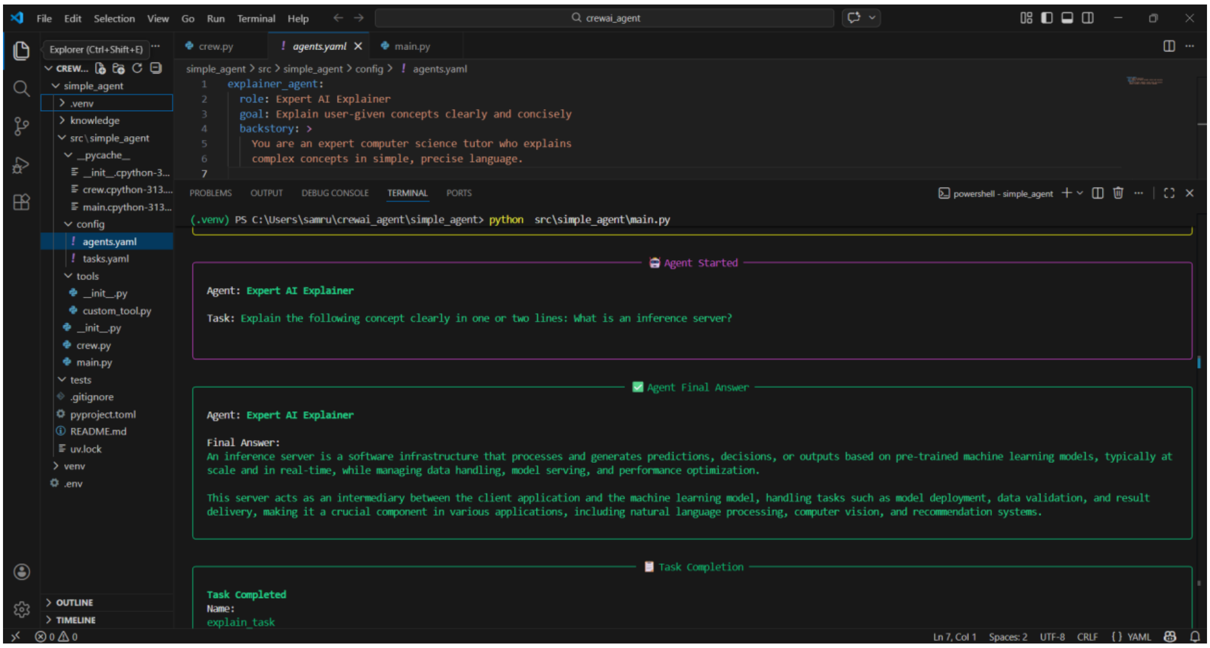
These run locally: yet to deploy.

>> crewai agent



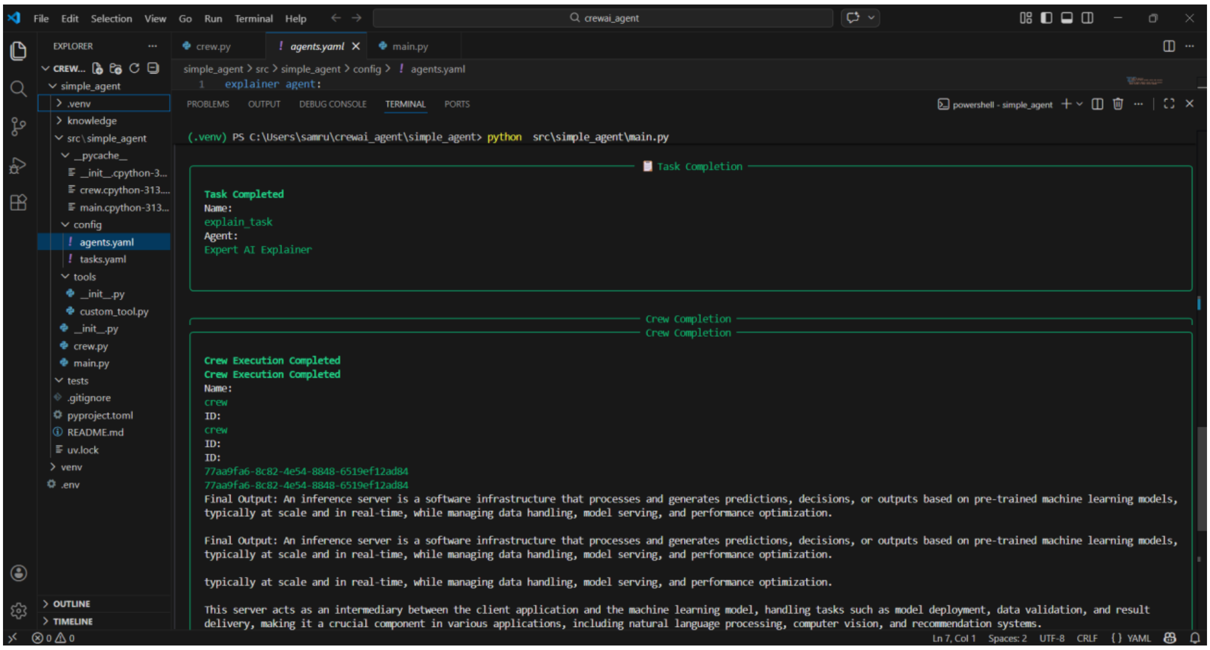
```yaml
explainer_agent:
  role: Expert AI Explainer
  goal: Explain user-given concepts clearly and concisely
  backstory: >
    You are an expert computer science tutor who explains
    complex concepts in simple, precise language.
```

```
* History restored

PS C:\Users\samru\crewai_agent> cd simple_agent
PS C:\Users\samru\crewai_agent\simple_agent> .\.venv\Scripts\activate
(.venv) PS C:\Users\samru\crewai_agent\simple_agent> python  src\simple_agent\main.py
                                        🚀 Crew Execution Started

  Crew Execution Started
  Name:
  crew
  ID:
  77aa9fa6-8c82-4e54-8848-6519ef12ad84


                                        📋 Task Started

  Task Started
  Name: explain_task
  ID: 4bce12ce-31a2-406b-88bc-491e99fad168


                                        🤖 Agent Started

  Agent: Expert AI Explainer
```



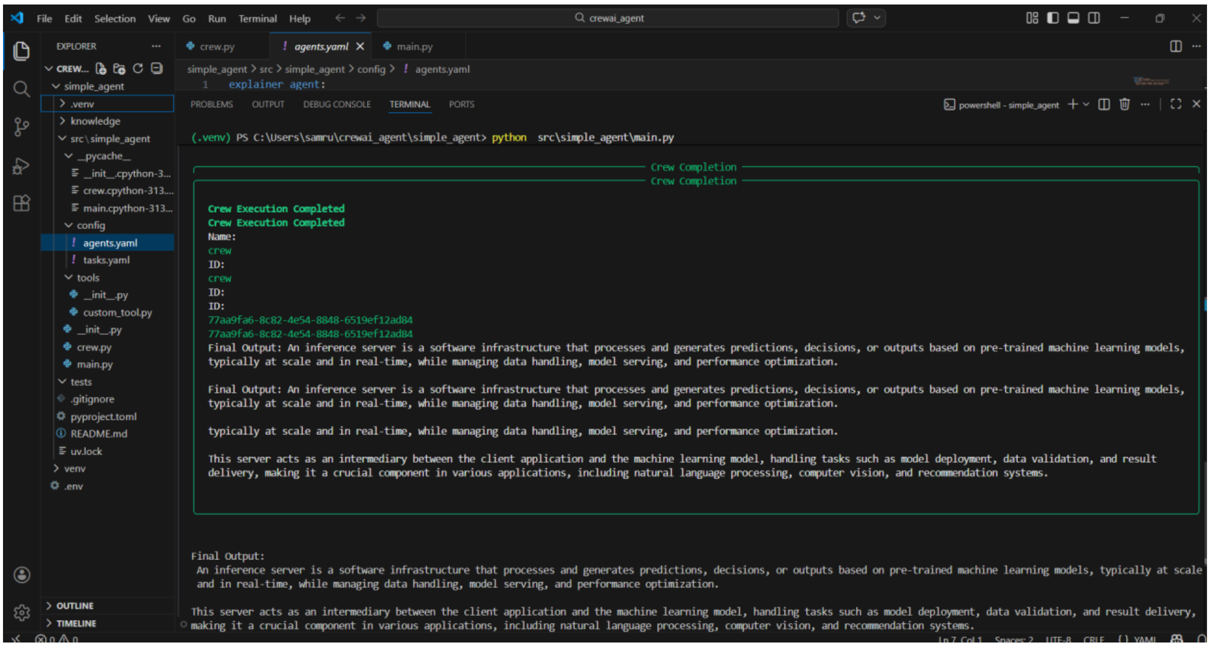
```yaml
explainer_agent:
  role: Expert AI Explainer
  goal: Explain user-given concepts clearly and concisely
  backstory: >
    You are an expert computer science tutor who explains
    complex concepts in simple, precise language.
```

```
(.venv) PS C:\Users\samru\crewai_agent\simple_agent> python  src\simple_agent\main.py

                                        🤖 Agent Started

  Agent: Expert AI Explainer

  Task: Explain the following concept clearly in one or two lines: What is an inference server?


                                        ✅ Agent Final Answer

  Agent: Expert AI Explainer

  Final Answer:
  An inference server is a software infrastructure that processes and generates predictions, decisions, or outputs based on pre-trained machine learning models, typically at
  scale and in real-time, while managing data handling, model serving, and performance optimization.

  This server acts as an intermediary between the client application and the machine learning model, handling tasks such as model deployment, data validation, and result
  delivery, making it a crucial component in various applications, including natural language processing, computer vision, and recommendation systems.


                                        📋 Task Completion

  Task Completed
  Name:
  explain_task
```



```
explainer agent:
```

```
(.venv) PS C:\Users\samru\crewai_agent\simple_agent> python  src\simple_agent\main.py

╭──────────────────────────── 📋 Task Completion ────────────────────────────╮
│                                                                             │
│  Task Completed                                                             │
│  Name:                                                                      │
│  explain_task                                                               │
│  Agent:                                                                     │
│  Expert AI Explainer                                                        │
│                                                                             │
╰─────────────────────────────────────────────────────────────────────────────╯

╭────────────────────────────── Crew Completion ──────────────────────────────╮
╭────────────────────────────── Crew Completion ──────────────────────────────╮
│                                                                             │
│  Crew Execution Completed
│  Crew Execution Completed
│  Name:
│  crew
│  ID:
│  crew
│  ID:
│  ID:
│  77aa9fa6-8c82-4e54-8848-6519ef12ad84
│  77aa9fa6-8c82-4e54-8848-6519ef12ad84
│  Final Output: An inference server is a software infrastructure that processes and generates predictions, decisions, or outputs based on pre-trained machine learning models,
│  typically at scale and in real-time, while managing data handling, model serving, and performance optimization.
│
│  Final Output: An inference server is a software infrastructure that processes and generates predictions, decisions, or outputs based on pre-trained machine learning models,
│  typically at scale and in real-time, while managing data handling, model serving, and performance optimization.
│
│  typically at scale and in real-time, while managing data handling, model serving, and performance optimization.
│
│  This server acts as an intermediary between the client application and the machine learning model, handling tasks such as model deployment, data validation, and result
│  delivery, making it a crucial component in various applications, including natural language processing, computer vision, and recommendation systems.
```



```
explainer agent:
```

```
(.venv) PS C:\Users\samru\crewai_agent\simple_agent> python  src\simple_agent\main.py

╭────────────────────────────── Crew Completion ──────────────────────────────╮
╭────────────────────────────── Crew Completion ──────────────────────────────╮
│                                                                             │
│  Crew Execution Completed
│  Crew Execution Completed
│  Name:
│  crew
│  ID:
│  crew
│  ID:
│  ID:
│  77aa9fa6-8c82-4e54-8848-6519ef12ad84
│  77aa9fa6-8c82-4e54-8848-6519ef12ad84
│  Final Output: An inference server is a software infrastructure that processes and generates predictions, decisions, or outputs based on pre-trained machine learning models,
│  typically at scale and in real-time, while managing data handling, model serving, and performance optimization.
│
│  Final Output: An inference server is a software infrastructure that processes and generates predictions, decisions, or outputs based on pre-trained machine learning models,
│  typically at scale and in real-time, while managing data handling, model serving, and performance optimization.
│
│  typically at scale and in real-time, while managing data handling, model serving, and performance optimization.
│
│  This server acts as an intermediary between the client application and the machine learning model, handling tasks such as model deployment, data validation, and result
│  delivery, making it a crucial component in various applications, including natural language processing, computer vision, and recommendation systems.
│                                                                             │
╰─────────────────────────────────────────────────────────────────────────────╯


Final Output:
 An inference server is a software infrastructure that processes and generates predictions, decisions, or outputs based on pre-trained machine learning models, typically at scale
 and in real-time, while managing data handling, model serving, and performance optimization.

This server acts as an intermediary between the client application and the machine learning model, handling tasks such as model deployment, data validation, and result delivery,
making it a crucial component in various applications, including natural language processing, computer vision, and recommendation systems.
```