

PIP INSTALL REQUIREMENTS

```
PS E:\EG_INNOVATIONS_TASK> cd .\rag_application\  
PS E:\EG_INNOVATIONS_TASK\rag_application> py -3.11 -m venv venv  
PS E:\EG_INNOVATIONS_TASK\rag_application> venv\Scripts/activate  
(venv) PS E:\EG_INNOVATIONS_TASK\rag_application> pip install -r requirements.txt  
collecting faiss_cpu==1.12.0 (from -r requirements.txt (line 1))  
Using cached faiss_cpu-1.12.0-cp311-cp311-win_amd64.whl.metadata (5.2 kB)  
collecting fastapi==0.119.0 (from -r requirements.txt (line 2))  
Using cached fastapi-0.119.0-py3-none-any.whl.metadata (28 kB)  
collecting httpx==0.28.1 (from -r requirements.txt (line 3))  
Using cached httpx-0.28.1-py3-none-any.whl.metadata (7.1 kB)
```

Start the application

```
(venv) PS E:\EG_INNOVATIONS_TASK\rag_application> uvicorn app.main:app  
025-10-20 15:20:42 - rag_app - INFO - Application logging initialized.  
025-10-20 15:20:59 - rag_app - INFO - GeminiRAG core initialized for model: gemini-2.5-flash. Ready for live API calls.  
INFO: Started server process [4260]  
INFO: Waiting for application startup.  
025-10-20 15:20:59 - rag_app - INFO - RAG Chatbot API is starting up...  
INFO: Application startup complete.  
INFO: Uvicorn running on http://127.0.0.1:8000 (Press CTRL+C to quit)
```

Q1 CHAT BOT

UPLOAD PDF

REQUEST

POST /chat/upload/pdf Upload And Index Pdf

Parameters

No parameters

Request body required

multipart/form-data

file * required

string(\$binary)

Choose File PlatformsSupported.pdf

Execute

Terminal Process

```

[0] 127.0.0.1:64887 - GET /api/users/john@f19v1.1? HTTP/1.1 200 OK
2023-10-26 10:23:05 - rag_esp - INFO - Scoring PDF text extraction from: qghand.pdf/LettersSupported.pdf
2023-10-26 10:24:11 - rag_esp - INFO - Text extraction complete. Total pages: 34
Extracted text (first 500 chars): Platforms supported
Platforms Supported
Applications/Servers Monitored in Agent-based/Agentless Manner
Operating Systems Support for
Agentless Agent-based Monitoring
Overview
Versions Supported Monitoring
From Windows
Recommended
2008/R2/ Linux AIX HP-UX Solaris
SAP/DB2
Digital Perceptics Technologies
Citrix
Citrix Access
R, J, K, L
Decency
Citrix ADC 10K 11.x and higher
Citrix ADC
P.s. and above
UPS/MPS
Citrix ADM Insight 10 (and above)
Citrix MSP
P.P.P
T.S. and higher
L
2023-10-26 10:24:22 - rag_esp - INFO - Tokenized document into 9 sentences for semantic analysis.
2023-10-26 10:24:28 - rag_esp - WARNING - Semantic chunk exceeded character limit: 12000 chars.
2023-10-26 10:24:28 - rag_esp - WARNING - Semantic chunk exceeded character limit: 4000 chars.
2023-10-26 10:24:28 - rag_esp - INFO - Building FAISS index for 6 documents...
2023-10-26 10:24:29 - rag_esp - INFO - FAISS index build complete.
2023-10-26 10:24:29 - rag_esp - INFO - Vectors now ready. Indexed 6 text chunks.
2023-10-26 10:24:29 - rag_esp - INFO - RAG Web Service successfully indexed: PlatformsSupported.pdf
[0] 127.0.0.1:64887 - POST /chat/qghand/pdf HTTP/1.1 200 OK

```

Output:

Responses

Curl

```
curl -X 'POST' \
  'http://127.0.0.1:8000/chat/upload/pdf' \
  -H 'accept: application/json' \
  -H 'Content-Type: multipart/form-data' \
  -F 'file@PlatformSupported.pdf;type=application/pdf'
```

Request URL

```
http://127.0.0.1:8000/chat/upload/pdf
```

Server response

Code

Details

200

Response body

```
{
  "status": "success",
  "message": "PDF 'PlatformSupported.pdf' uploaded and RAG index successfully built."
}
```

Download

Response headers

```
content-length: 104
content-type: application/json
date: Mon, 20 Oct 2025 09:54:05 GMT
server: uvicorn
```

Answering PDF

Q1) Do you support WebLogic and what versions of WebLogic is supported

Answer: Yes, we support Oracle WebLogic. The supported versions are 5.1, 6.x, 7, 8.x, 9.x, 10.x, 11G, 12.x, and 14c.

POST /chat/query/pdf Query Pdf Rag

Endpoint for GPT-based RAG Chatbot (PDF Source).

Parameters

No parameters

CancelReset

Request body required

application/json

Edit Value | Schema


```
{  "query": "Do you support WebLogic and what versions of WebLogic is supported"}
```

Execute

Clear

```
INFO: 127.0.0.1:64887 - POST /chat/query/pdf HTTP/1.1 200 OK
2025-10-20 15:27:02 - rag_app - INFO - Controller: Received PDF query: Do you support WebLogic and what versions of WebLogic is supported
2025-10-20 15:27:02 - rag_app - INFO - PDF RAG: Executing Retrieve phase for query: Do you support WebLogic and what versions of WebLogic is supported
2025-10-20 15:27:03 - rag_app - INFO - Retrieved 5 relevant chunks from index.
2025-10-20 15:27:03 - rag_app - INFO - GeminiRAG: Preparing prompt for PDF query.
2025-10-20 15:27:07 - rag_app - INFO - Gemini API call succeeded. Response received.
2025-10-20 15:27:07 - rag_app - INFO - PDF RAG response successfully generated via Gemini.
INFO: 127.0.0.1:61083 - "POST /chat/query/pdf HTTP/1.1" 200 OK
```

Server responses

Code	Details
200	<div>Response body</div> <div> <pre>{ "response": "Yes, we support Oracle WebLogic. The supported versions are 5.3, 6.x, 7, 8.x, 9.x, 10.x, 11g, 12.x, and 14c.", "source_type": "PDF" }</pre> <div>  <div>Download</div> </div> </div> <div>Response headers</div> <div> <pre>content-length: 143 content-type: application/json date: Mon, 20 Oct 2025 09:57:02 GMT server: unicorn</pre> </div>

Responses

Code	Description	Links
200	Successful Response	No links

Q2) Do you support IBM Mainframe?

Answer: **No.**

POST /chat/query/pdf Query Pdf Rag

Endpoint for GPT-based RAG Chatbot (PDF Source).

Parameters

No parameters

Cancel Reset

Request body required

application/json

Edit Value Schema

```
{
  "query": "Do you support IBM Mainframe?"
}
```

```

2025-10-20 15:28:33 - rag_app - INFO - Controller: Received PDF query: Do you support IBM Mainframe?
2025-10-20 15:28:33 - rag_app - INFO - PDF RAG: Executing Retrieve phase for query: Do you support IBM Mainframe?
2025-10-20 15:28:33 - rag_app - INFO - Retrieved 5 relevant chunks from index.
2025-10-20 15:28:33 - rag_app - INFO - GeminiRAG: Preparing prompt for PDF query.
2025-10-20 15:28:38 - rag_app - INFO - Gemini API call succeeded. Response received.
2025-10-20 15:28:38 - rag_app - INFO - PDF RAG response successfully generated via Gemini.
INFO: 127.0.0.1:61800 - "POST /chat/query/pdf HTTP/1.1" 200 OK

```

Request URL

http://127.0.0.1:8000/chat/query/pdf

Server response

Code Details

200

Response body

```
{
  "response": "No.",
  "source_type": "PDF"
}
```

Download

Response headers

```
content-length: 38
content-type: application/json
date: Mon, 20 Oct 2025 09:58:33 GMT
server: uvicorn
```

Responses

Code Description Links

Q3) Do you support Oracle Database monitoring and what versions supported?

Answer) **Yes, Oracle Database monitoring is supported. The versions supported are 7, 8, 9, 10G, 11G, 12c (including multi-tenant setup), 19c, and 21c.**

POST /chat/query/pdf Query Pdf Rag

Endpoint for GPT-based RAG Chatbot (PDF Source).

Parameters Cancel Reset

No parameters

Request body required application/json

Edit Value | Schema

```
{
  "query": "Do you support Oracle Database monitoring and what versions supported?"
}
```

```
2025-10-20 15:30:37 - rag_app - INFO - Controller: Received PDF query: Do you support Oracle Database monitoring and what versions supported?
2025-10-20 15:30:37 - rag_app - INFO - PDF RAG: Executing Retrieve phase for query: Do you support Oracle Database monitoring and what versions supported?
2025-10-20 15:30:37 - rag_app - INFO - Retrieved 5 relevant chunks from index.
2025-10-20 15:30:37 - rag_app - INFO - GeminiRAG: Preparing prompt for PDF query.
2025-10-20 15:30:40 - rag_app - INFO - Gemini API call succeeded. Response received.
2025-10-20 15:30:40 - rag_app - INFO - PDF RAG response successfully generated via Gemini.
INFO: 127.0.0.1:62943 - "POST /chat/query/pdf HTTP/1.1" 200 OK
```

Request URL

http://127.0.0.1:8000/chat/query/pdf

Server response

Code	Details
200	<p>Response body</p> <pre>{ "response": "Yes, Oracle Database monitoring is supported. The versions supported are 7, 8, 9, 10g, 11g, 12c (including multi-tenant setup), 19c, and 21c.", "source_type": "PDF" }</pre> <p>Download</p> <p>Response headers</p> <pre>content-length: 176 content-type: application/json date: Mon, 20 Oct 2025 10:00:36 GMT server: uvicorn</pre>

Q4) Can you share the MSSQL server versions supported by eG Enterprise?

Answer) **eG Enterprise supports Microsoft SQL Server versions 7.0, 2000, 2005, 2008, 2012, 2014, 2016, 2017, 2019, and 2022.**

POST /chat/query/pdf Query Pdf Rag

Endpoint for GPT-based RAG Chatbot (PDF Source).

Parameters Cancel Reset

No parameters

Request body required application/json

Edit Value | Schema

```
{
  "query": "Can you share the MSSQL server versions supported by eG Enterprise?"
}
```

```
2025-10-20 15:32:14 - rag_app - INFO - Controller: Received PDF query: Can you share the MSSQL server versions supported by eG Enterprise?
2025-10-20 15:32:14 - rag_app - INFO - PDF RAG: Executing Retrieve phase for query: Can you share the MSSQL server versions supported by eG Enterprise?
2025-10-20 15:32:15 - rag_app - INFO - Retrieved 5 relevant chunks from index.
2025-10-20 15:32:15 - rag_app - INFO - GeminiRAG: Preparing prompt for PDF query.
2025-10-20 15:32:17 - rag_app - INFO - Gemini API call succeeded. Response received.
2025-10-20 15:32:17 - rag_app - INFO - PDF RAG response successfully generated via Gemini.
INFO: 127.0.0.1:63535 - "POST /chat/query/pdf HTTP/1.1" 200 OK
```

Request URL

http://127.0.0.1:8080/chat/query/pdf

Server response

Code	Details
200	<p>Response body</p> <pre>{ "response": "eG Enterprise supports Microsoft SQL Server versions 7.0, 2000, 2005, 2008, 2012, 2014, 2016, 2017, 2019, and 2022.", "source_type": "PDF" }</pre> <p>Download</p>

Q5) Does eG Enterprise supports 'Open Telemetry' ?

Answer) No

POST /chat/query/pdf Query Pdf Rag

Endpoint for GPT-based RAG Chatbot (PDF Source).

Parameters Cancel Reset

No parameters

Request body required application/json

Edit Value | Schema

```
{
  "query": "Does eG Enterprise supports 'Open Telemetry' ?"
}
```

```

2025-10-20 15:34:03 - rag_app - INFO - Controller: Received PDF query: Does eG Enterprise supports 'Open Telemetry' ?
2025-10-20 15:34:03 - rag_app - INFO - PDF RAG: Executing Retrieve phase for query: Does eG Enterprise supports 'Open Telemetry' ?
2025-10-20 15:34:03 - rag_app - INFO - Retrieved 5 relevant chunks from index.
2025-10-20 15:34:03 - rag_app - INFO - GeminiRAG: Preparing prompt for PDF query.
2025-10-20 15:34:06 - rag_app - INFO - Gemini API call succeeded. Response received.
2025-10-20 15:34:06 - rag_app - INFO - PDF RAG response successfully generated via Gemini.
INFO: 127.0.0.1:54765 - "POST /chat/query/pdf HTTP/1.1" 200 OK

```

Request URL

http://127.0.0.1:8000/chat/query/pdf

Server response

Code	Details
200	<div>Response body</div> <pre>{ "response": "No", "source_type": "PDF" }</pre> <div>Download</div>

LIVE API: Measures the CPU Utilization and Free Space (API: /api/realtime_metrics_source)

GET /api/realtime_metrics_source Realtime Metrics Source

Simulates the external API endpoint. It returns the current host system metrics (CPU, Memory, Disk) in the requested legacy JSON structure, dynamically generated via psutil.

Parameters

No parameters

Execute Clear

Responses

Curl

```
curl -X 'GET' \
  'http://127.0.0.1:8000/api/realtime_metrics_source' \
  -H 'accept: application/json'
```

Request URL

http://127.0.0.1:8000/api/realtime_metrics_source

```

llama_perf_context_print: prompt eval time = 15211.46 ms / 380 tokens ( 39.77 ms per token, 25.14 tokens per second)
llama_perf_context_print: total time = 18069.31 ms / 548 tokens
llama_perf_context_print: graphs reused = 290
2025-10-23 10:29:29 - rag_app - INFO - Offline LLM RAG generation succeeded.
INFO: 127.0.0.1:56717 - "POST /api/query/api HTTP/1.1" 200 OK
2025-10-23 10:30:05 - rag_app - INFO - Controller: Successfully fetched real-time system metrics.
INFO: 127.0.0.1:64670 - "GET /api/realtime_metrics_source HTTP/1.1" 200 OK
2025-10-23 10:35:07 - rag_app - INFO - Controller: Successfully fetched real-time system metrics.
INFO: 127.0.0.1:49743 - "GET /api/realtime_metrics_source HTTP/1.1" 200 OK

```

CodeDetails

200

Response body

```
Value: 7.221
}
},
{
  "Drive availability ": [
    {
      "unit": "-",
      "priority": "normal",
      "value": "Available"
    }
  ]
},
{
  "Free space ": [
    {
      "unit": "MB",
      "priority": "normal",
      "value": "335855"
    }
  ]
},
{
  "CPU Utilization": {
    "lastMeasurementTime": "Oct 23, 2025 10:35:07",
    "value": "15.2",
    "unit": "%",
  }
}
}
```

Download

Response headers

Query Chat Bot: (API: /api/query/api)

Q1) What is the free space for the server

Answer) **The free space for the server is 335855 MB, which is available. The context JSON provides the last measurement time for the server, which is Oct 23, 2025 10:37:25. The value for CPU Utilization is 15.2%, which is within the normal range of 0% to 100%. Memory Usage is also within the normal range of 0% to 100%. Therefore, the system metrics monitor should state that the server has free space of 335855 MB.**

Request body required

application/json

Edit Value | Schema

```
{
  "query": "what is the free space for the server"
}
```

Execute

Clear


```

2025-10-23 10:37:25 - rag_app - INFO - Controller: Successfully retrieved context via HTTP from http://127.0.0.1:8000/api/realtime_metrics_source.
2025-10-23 10:37:25 - rag_app - INFO - Offline RAG: Preparing prompt for metrics query.
llama.generate: 178 prefix-match hit, remaining 251 prompt tokens to eval
llama_perf_context_print:    load time =    9913.97 ms
llama_perf_context_print: prompt eval time =   2453.27 ms /   251 tokens (    9.77 ms per token,   102.31 tokens per second)
llama_perf_context_print:    eval time =   8072.92 ms /   126 runs (   64.07 ms per token,    15.61 tokens per second)
llama_perf_context_print:    total time =  10887.36 ms /   377 tokens
llama_perf_context_print:    graphs reused =        121
2025-10-23 10:37:36 - rag_app - INFO - Offline LLM RAG generation succeeded.
INFO:    127.0.0.1:59241 - "POST /api/query/api HTTP/1.1" 200 OK

```

200

Response body

```
{
  "response": "The free space for the server is 335855 MB, which is available. The context JSON provides the last measurement time for the server, which is Oct 23, 2025 10:37:25. The value for CPU Utilization is 15.2%, which is within the normal range of 0% to 100%. Memory Usage is also within the normal range of 0% to 100%. Therefore, the system metrics monitor should state that the server has free space of 335855 MB.",
  "source_type": "API"
}
```

Download

Q2) What is the CPU Utilization for the server

Ans) "The CPU Utilization metric for the server is \"22.0\", which is reported as \"22.0\" with a priority of \"normal\". The unit is \"%\", indicating that the metric is expressed as a percentage. The specific requested metric is not explicitly present in the JSON, but it is present in the context.",

Edit Value | Schema

```
{
  "query": "What is the CPU Utilization for the server"
}
```

Execute Clear

Responses

Curl

```
curl -X 'POST' \
  'http://127.0.0.1:8000/api/query/api' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
    "query": "What is the CPU Utilization for the server"
  }'
```

Request URL

```
http://127.0.0.1:8000/api/query/api
```

```

INFO: 127.0.0.1:53984 - "GET /api/realtime_metrics_source HTTP/1.1" 200 OK
2025-10-23 10:39:10 - rag_app - INFO - Controller: Successfully retrieved context via HTTP from http://127.0.0.1:8000/api/realtime_metrics_source.
2025-10-23 10:39:10 - rag_app - INFO - Offline RAG: Preparing prompt for metrics query.
llama.generate: 179 prefix-match hit, remaining 250 prompt tokens to eval
llama_perf_context_print: load time = 9913.97 ms
llama_perf_context_print: prompt eval time = 2456.58 ms / 250 tokens ( 9.83 ms per token, 101.77 tokens per second)
llama_perf_context_print: eval time = 4776.03 ms / 67 runs ( 71.28 ms per token, 14.03 tokens per second)
llama_perf_context_print: total time = 7358.54 ms / 317 tokens
llama_perf_context_print: graphs reused = 64
2025-10-23 10:39:18 - rag_app - INFO - Offline LLM RAG generation succeeded.
INFO: 127.0.0.1:53983 - "POST /api/query/api HTTP/1.1" 200 OK

```

Responses

Curl

```
curl -X 'POST' \
  'http://127.0.0.1:8000/api/query/api' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
    "query": "What is the CPU Utilization for the server"
  }'
```

Request URL

http://127.0.0.1:8000/api/query/api

Server response

Code	Details
200	<div>Response body</div> <pre>{ "response": "The CPU Utilization metric for the server is \"22.0%\", which is reported as \"22.0\" with a priority of \"normal\". The unit is \"%\", indicating that the metric is expressed as a percentage. The specific requested metric is not explicitly present in the JSON, but it is present in the context.", "source_type": "API" }</pre> <div>Response headers</div> <pre>content-length: 329 content-type: application/json date: Thu, 23 Oct 2025 05:09:09 GMT server: unicorn</pre>

Responses

Q3) What is the memory usage for the server

Ans) "The system metrics monitor for the server 10.200.2.192:7077:eG Manager is as follows:\n\n- Disk Space:\n - Disk Space:\n - lastMeasurementTime: Oct 23, 2025 10:40:36\n - C:\n - Used space:\n - unit: MB\n - priority: warning\n - value: 6708\n - Drive availability:\n - unit: -\n - priority: normal\n - value: Available\n - Free space:\n - unit: MB\n - priority: normal\n - value: 335855\n\n- CPU Utilization:\n - lastMeasurementTime: Oct 23, 2025 10:40:36\n - value: 22.5\n - unit: %\n - priority: normal\n\n- Memory Usage:\n - lastMeasurementTime: Oct 23, 2025 10:40:36\n - value: 6708\n - unit: MB\n - priority: warning\n\nThe memory usage for the server 10.200.2.192:7077:eG Manager is 6708 MB."

Request body required

application/json

Edit Value | Schema

```
{
  "query": "What is the memory usage for the server "
}
```

ExecuteClear

LOADING

Responses

```
INFO: 127.0.0.1:53144 - "GET /api/realtime_metrics_source HTTP/1.1" 200 OK
2025-10-23 10:40:36 - rag_app - INFO - Controller: Successfully retrieved context via HTTP from http://127.0.0.1:8000/api/realtime_metrics_source.
2025-10-23 10:40:36 - rag_app - INFO - Offline RAG: Preparing prompt for metrics query.
llama.generate: 178 prefix-match hit, remaining 250 prompt tokens to eval
llama_perf_context_print: load time = 9913.97 ms
llama_perf_context_print: prompt eval time = 2570.43 ms / 250 tokens ( 10.28 ms per token, 97.26 tokens per second)
llama_perf_context_print: eval time = 19926.43 ms / 316 runs ( 63.06 ms per token, 15.86 tokens per second)
llama_perf_context_print: total time = 22867.42 ms / 566 tokens
llama_perf_context_print: graphs reused = 305
2025-10-23 10:40:59 - rag_app - INFO - Offline LLM RAG generation succeeded.
INFO: 127.0.0.1:53143 - "POST /api/query/api HTTP/1.1" 200 OK
```

Curl

```
curl -X 'POST' \
  'http://127.0.0.1:8000/api/query/api' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
    "query": "What is the memory usage for the server "
  }'
```

Request URL

```
http://127.0.0.1:8000/api/query/api
```

Server response

CodeDetails

200

Response body

```
{
  "response": "The system metrics monitor for the server 10.200.2.152:7877:e6 Manager is as follows:\n\n Disk Space:\n - Disk Space:\n - lastMeasurementTime: Oct 23, 2025 10:40:36\n - C:\n - Used space:\n - unit: MB\n - priority: warning\n - value: 6708\n - Drive availability:\n - unit: \n - priority: normal\n - value: Avail\n - Free space:\n - unit: MB\n - priority: normal\n - value: 335855\n\n CPU Utilization:\n - lastMeasurementTime: Oct 23, 2025 10:40:36\n - value: 22.5\n - un\n - priority: normal\n\n Memory Usage:\n - lastMeasurementTime: Oct 23, 2025 10:40:36\n - value: 6708\n - unit: MB\n - priority: warning\n\nThe memory usage for the server 10.200.2.152:7877:e6 Manager is 6708 MB.",
  "source_type": "API"
}
```

Download

Response headers

```
content-length: 824
content-type: application/json
date: Thu, 23 Oct 2025 05:10:33 GMT
server: uvicorn
```