ChatRFID: : a tool-calling AI-agent for controlling RFID devices

Deng, Xiao Dong

November 29th, 2024

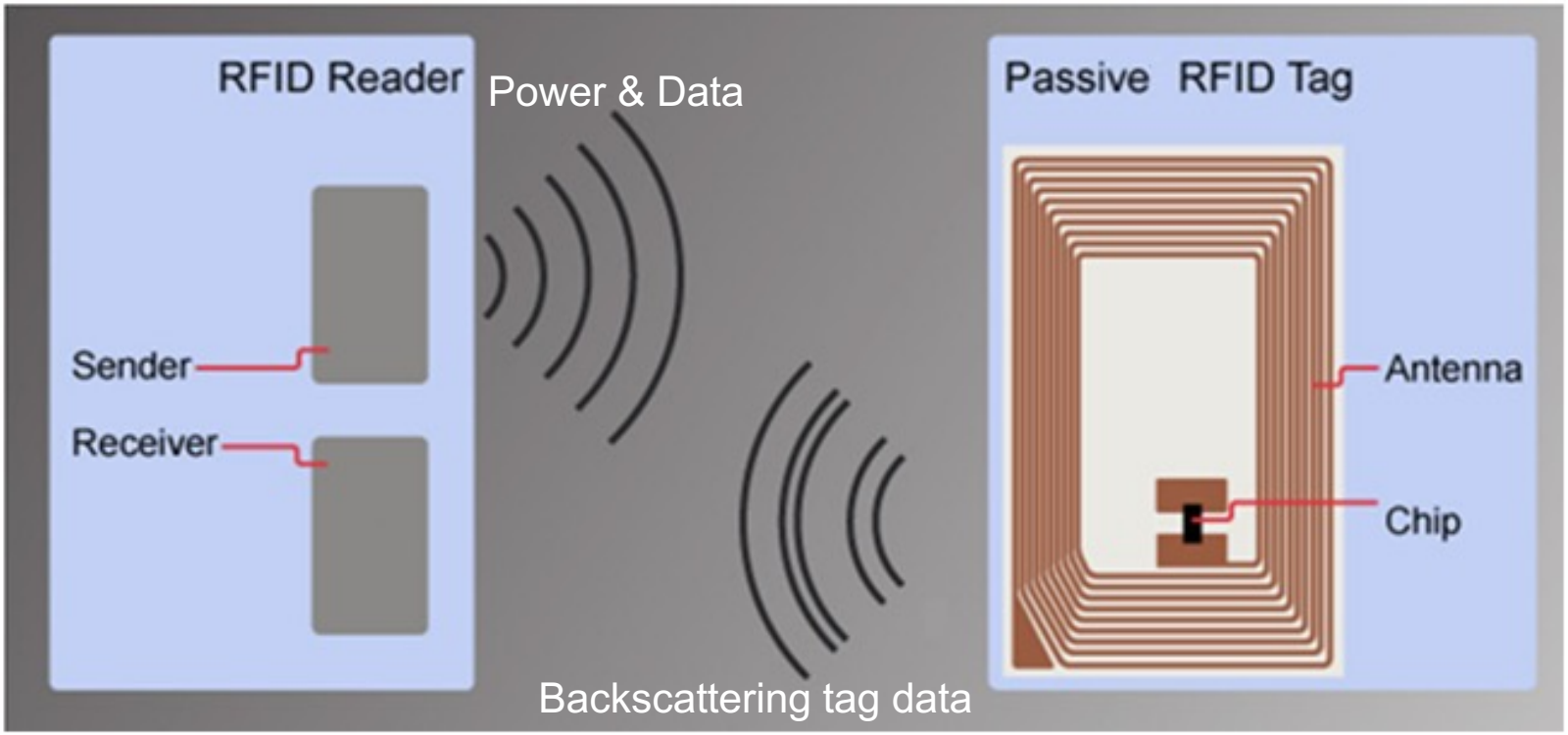
SIEMENS

Outline

- What is RFID?
- ChatRFID: a tool-calling AI-agent for controlling RFID devices
- Fine-tune the LLM of ChatRFID with open-source projects

What is RFID?

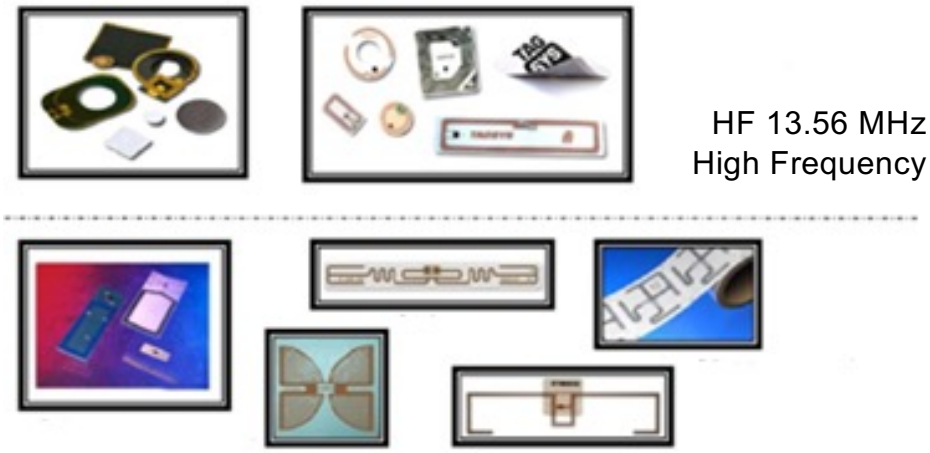
Radio Frequency Identification (RFID) is the use of radio waves to read and capture information stored on a tag attached to an object.



SIEMENS

RFID Frequencies

UHF has a unified standard, higher data rate, longer range, lower cost and is suitable for item tagging.



HF 13.56 MHz
High Frequency

UHF 800/900 MHz
Ultra High Frequency

	HF	UHF
Standard	Multiple competing standards	Single worldwide Gen2 standard
Data Speed & Range	HF-based NFC (Near-Field Communication) for secure payment	20X the range and speed of HF
Cost	Labels, cards, inlays cost $0.5 - 2	Labels cost $0.05-0.15 in 2017
Use case	Used in access control, ticketing, payment	Separate EPC (Electronic Product Code) and Tag ID for item tagging

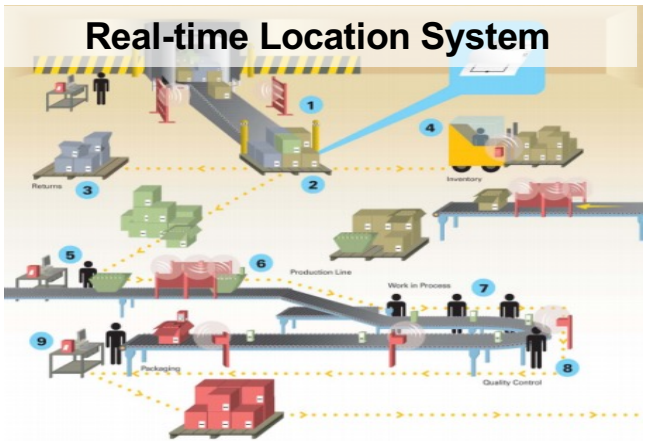
UHF RFID Applications

Manufacturing

Logistics & Supply Chain Visibility

Materials Management

Real-time Location System



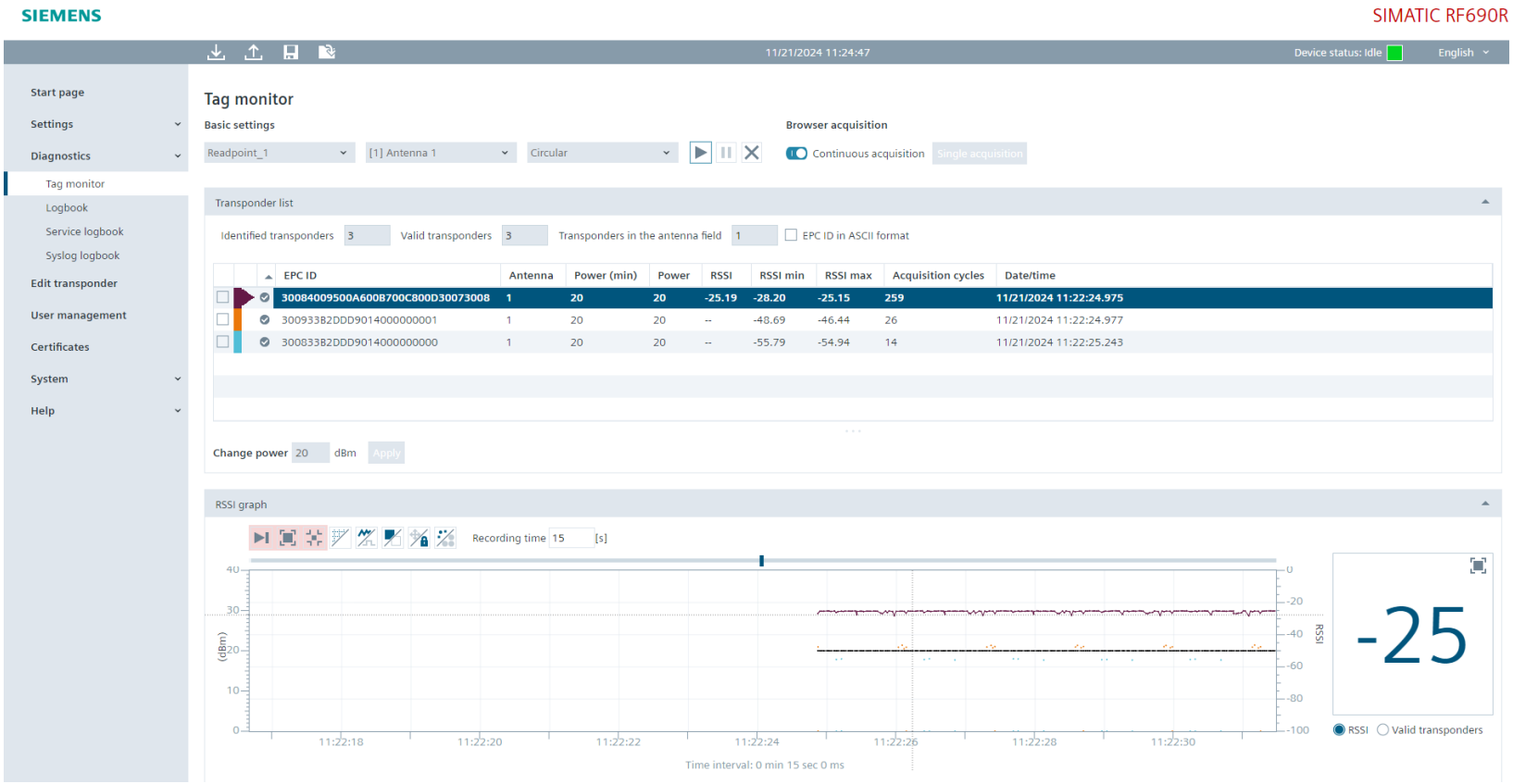
Deng, Xiao Dong, FT RPD CED INW - CN

SIEMENS

Siemens SIMATIC UHF RFID Portfolio

SIEMENS

Web UI of Siemens SIMATIC Reader



SIEMENS

LLM for RFID

- LLM can access information or perform actions on RFID
- Interpret LLM outputs for RFID
- Build and run LLM on the local machine
- Fine-tuning on own data

Outline

- What is RFID?
- ChatRFID: a tool-calling AI-agent for controlling RFID devices
- Fine-tune the LLM of ChatRFID with open-source projects

ChatRFID: a tool-calling AI-Agent for controlling RFID devices

RFID Reader

Instruction

A prompt template

- Invoke a tool when the query from the user is similar to its description semantically
- Use unknown if the value of arguments is not available
- Output the completion directly if no tool is available

Tools

python

- inventoryToolRFID
- integrationToolRFID
- …

LangChain

Query

Response

User

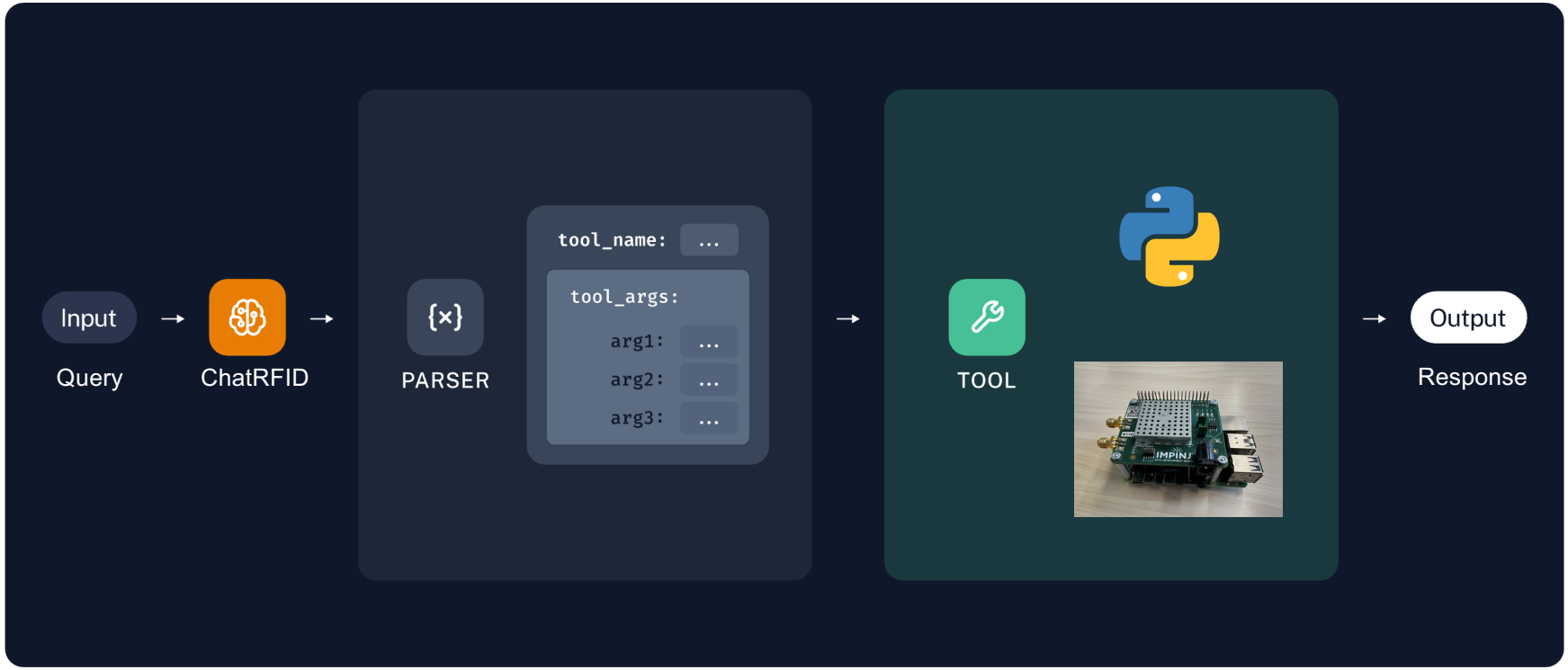
Model I/O

- Inventory tags for 100 ms, and integrate the inventory results

Ollama Qwen2 Llama 3.1 Microsoft Phi-3 Gemma 2

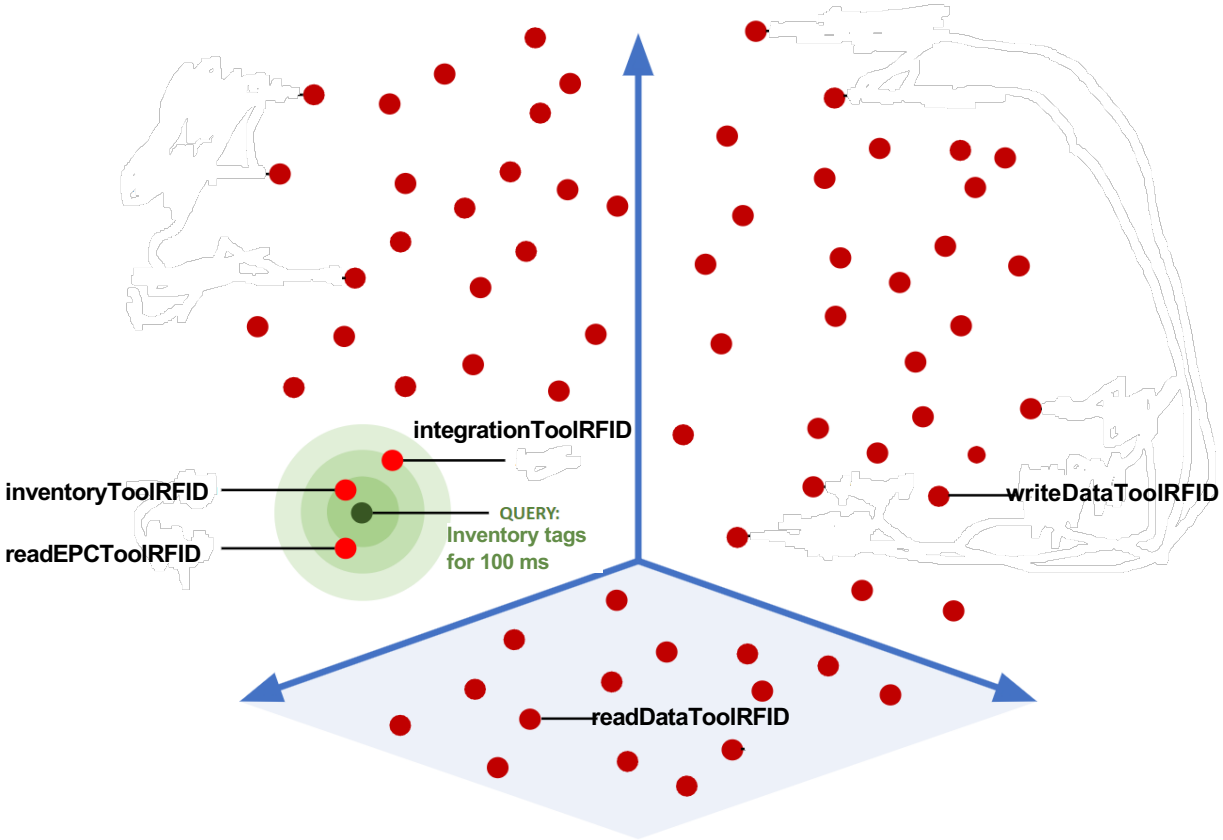
SIEMENS

Tool Calling



SIEMENS

Which Bound Tool Should be Called? Sematic Similarity Task




```
inventoryToolRFID = {
    'name': 'get_inventory',
    'description': 'Inventory tags for a duration of time',
    'parameters': {
        'type': 'object',
        'propeties': {
            'duration': {
                'type': 'number',
                'description': 'A duration of time in millisecond e.g. 500',
            }
        },
        'required': ['duration'],
    },
}

integrationToolRFID = {
    'name': 'tag_integration',
    'description': 'Integrate the inventory results after inventory tags for a duration of time',
    'parameters': {
        'type': 'object',
        'propeties': {
            'duration': {
                'type': 'number',
                'description': 'A duration of time in millisecond e.g. 500',
            }
        },
        'required': ['duration'],
    },
}
```


SIEMENS

UI of ChatRFID

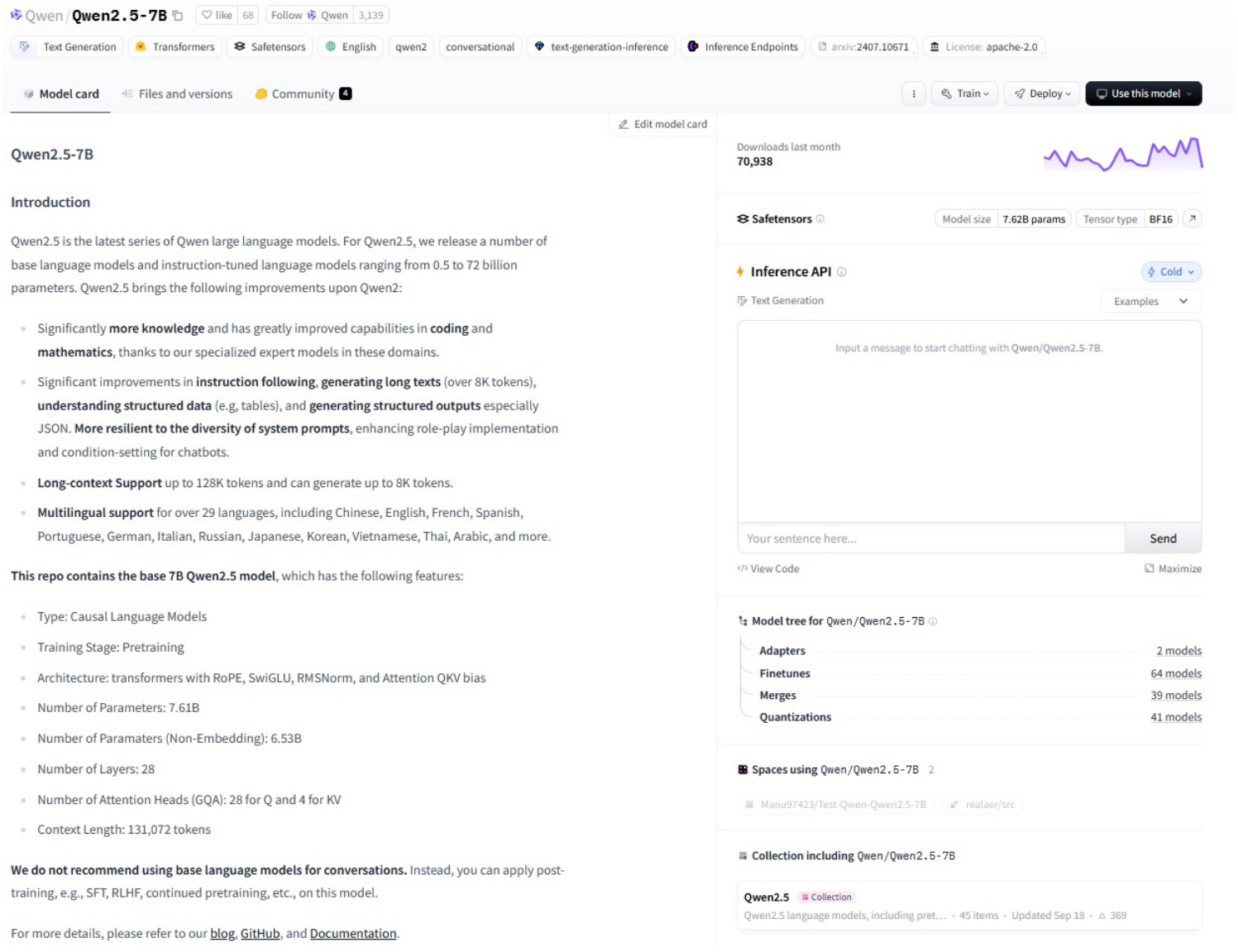
```
pi@raspi-11-lite-rfid:~ $ docker run -it --rm --network host --log-driver fluentd --name chatrfid chatrfid:v1.0
Enter your prompt : what is RFID?
----------------------
user prompt: what is RFID?
LLM return: RFID stands for Radio Frequency Identification. It's a technology that uses radio waves to automatically identify mobile objects, like vehicles or equipment, without any human
interaction and in an efficient manner. This technology can be used across various applications such as inventory management systems, access control systems, tracking of assets, and more.


Enter your prompt : Inventar-Tags für 200 ms
----------------------
user prompt: Inventar-Tags für 200 ms
Function call return: [{"id":1,"count":1,"timestamp":"2024-11-21 04:00:32.535534","Epc":"300933b2ddd9014000000001"},{"id":1,"count":2,"timestamp":"2024-11-21 04:00:32.535730","Epc":"300840
09500a600b700c800d30073008"},{"id":1,"count":3,"timestamp":"2024-11-21 04:00:32.535860","Epc":"300833b2ddd9014000000000"},{"id":1,"count":4,"timestamp":"2024-11-21 04:00:32.547525","Epc":"
30084009500a600b700c800d30073008"},{"id":1,"count":5,"timestamp":"2024-11-21 04:00:32.547679","Epc":"300833b2ddd9014000000000"},{"id":1,"count":6,"timestamp":"2024-11-21 04:00:32.547792","
Epc":"300933b2ddd9014000000001"},{"id":1,"count":7,"timestamp":"2024-11-21 04:00:32.559561","Epc":"300833b2ddd9014000000000"},{"id":1,"count":8,"timestamp":"2024-11-21 04:00:32.559696","Ep
c":"30084009500a600b700c800d30073008"},{"id":1,"count":9,"timestamp":"2024-11-21 04:00:32.559821","Epc":"300933b2ddd9014000000001"},{"id":1,"count":10,"timestamp":"2024-11-21 04:00:32.5709
80","Epc":"30084009500a600b700c800d30073008"},{"id":1,"count":11,"timestamp":"2024-11-21 04:00:32.571124","Epc":"300833b2ddd9014000000000"},{"id":1,"count":12,"timestamp":"2024-11-21 04:00
:32.571250","Epc":"300933b2ddd9014000000001"},{"id":1,"count":13,"timestamp":"2024-11-21 04:00:32.583119","Epc":"30084009500a600b700c800d30073008"},{"id":1,"count":14,"timestamp":"2024-11-
21 04:00:32.583386","Epc":"300833b2ddd9014000000000"},{"id":1,"count":15,"timestamp":"2024-11-21 04:00:32.583599","Epc":"300933b2ddd9014000000001"},{"id":1,"count":16,"timestamp":"2024-11-
21 04:00:32.596995","Epc":"300933b2ddd9014000000001"},{"id":1,"count":17,"timestamp":"2024-11-21 04:00:32.597159","Epc":"300833b2ddd9014000000000"},{"id":1,"count":18,"timestamp":"2024-11-
21 04:00:32.597298","Epc":"30084009500a600b700c800d30073008"},{"id":1,"count":19,"timestamp":"2024-11-21 04:00:32.609020","Epc":"300833b2ddd9014000000000"},{"id":1,"count":20,"timestamp":"
2024-11-21 04:00:32.609152","Epc":"300933b2ddd9014000000001"},{"id":1,"count":21,"timestamp":"2024-11-21 04:00:32.609289","Epc":"30084009500a600b700c800d30073008"},{"id":1,"count":22,"time
stamp":"2024-11-21 04:00:32.621141","Epc":"300933b2ddd9014000000001"},{"id":1,"count":23,"timestamp":"2024-11-21 04:00:32.621403","Epc":"30084009500a600b700c800d30073008"},{"id":1,"count":
24,"timestamp":"2024-11-21 04:00:32.621735","Epc":"300833b2ddd9014000000000"},{"id":1,"count":25,"timestamp":"2024-11-21 04:00:32.633332","Epc":"300833b2ddd9014000000000"},{"id":1,"count":
26,"timestamp":"2024-11-21 04:00:32.633526","Epc":"30084009500a600b700c800d30073008"},{"id":1,"count":27,"timestamp":"2024-11-21 04:00:32.633654","Epc":"300933b2ddd9014000000001"},{"id":1,
"count":28,"timestamp":"2024-11-21 04:00:32.645397","Epc":"300833b2ddd9014000000000"},{"id":1,"count":29,"timestamp":"2024-11-21 04:00:32.645538","Epc":"30084009500a600b700c800d30073008"},
{"id":1,"count":30,"timestamp":"2024-11-21 04:00:32.645664","Epc":"300933b2ddd9014000000001"},{"id":1,"count":31,"timestamp":"2024-11-21 04:00:32.656908","Epc":"300833b2ddd9014000000000"},
{"id":1,"count":32,"timestamp":"2024-11-21 04:00:32.668178","Epc":"300833b2ddd9014000000000"},{"id":1,"count":33,"timestamp":"2024-11-21 04:00:32.680893","Epc":"30084009500a600b700c800d300
73008"},{"id":1,"count":34,"timestamp":"2024-11-21 04:00:32.681037","Epc":"300833b2ddd9014000000000"},{"id":1,"count":35,"timestamp":"2024-11-21 04:00:32.681182","Epc":"300933b2ddd90140000
00001"},{"id":1,"count":36,"timestamp":"2024-11-21 04:00:32.689931","Epc":"300833b2ddd9014000000000"},{"id":1,"count":37,"timestamp":"2024-11-21 04:00:32.699886","Epc":"300833b2ddd90140000
00000"},{"id":1,"count":38,"timestamp":"2024-11-21 04:00:32.714350","Epc":"300933b2ddd9014000000001"},{"id":1,"count":39,"timestamp":"2024-11-21 04:00:32.714491","Epc":"30084009500a600b700
c800d30073008"},{"id":1,"count":40,"timestamp":"2024-11-21 04:00:32.714618","Epc":"300833b2ddd9014000000000"}]


Enter your prompt : intégrer les résultats de l'inventaire après les balises d'inventaire pendant 500 ms
----------------------
user prompt: intégrer les résultats de l'inventaire après les balises d'inventaire pendant 500 ms
Function call return: {'30084009500a600b700c800d30073008': 27, '300933b2ddd9014000000001': 33, '300833b2ddd9014000000000': 41}
```

Outline

- What is RFID?
- ChatRFID: a tool-calling AI-agent for controlling RFID devices
- Fine-tune the LLM of ChatRFID with open-source projects



https://huggingface.co/Qwen/Qwen2.5-7B

llama.cpp

Roadmap / Project status / Manifesto / ggml

Inference of Meta's LLaMA model (and others) in pure C/C++

Recent API changes

- Changelog for `libllama` API
- Changelog for `llama-server` REST API

Hot topics

- **Introducing GGUF-my-LoRA** #10123
- Hugging Face Inference Endpoints now support GGUF out of the box! #9669
- Hugging Face GGUF editor: discussion | tool

https://github.com/ggerganov/llama.cpp

Ollama

Discord

Get up and running with large language models.

macOS

Download

Windows

Download

Linux

```
curl -fsSL https://ollama.com/install.sh | sh
```

Manual install instructions

Docker

The official Ollama Docker image `ollama/ollama` is available on Docker Hub.

Libraries

- ollama-python
- ollama-js

Quickstart

To run and chat with Llama 3.2:

```
ollama run llama3.2
```

https://github.com/ollama/ollama