

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

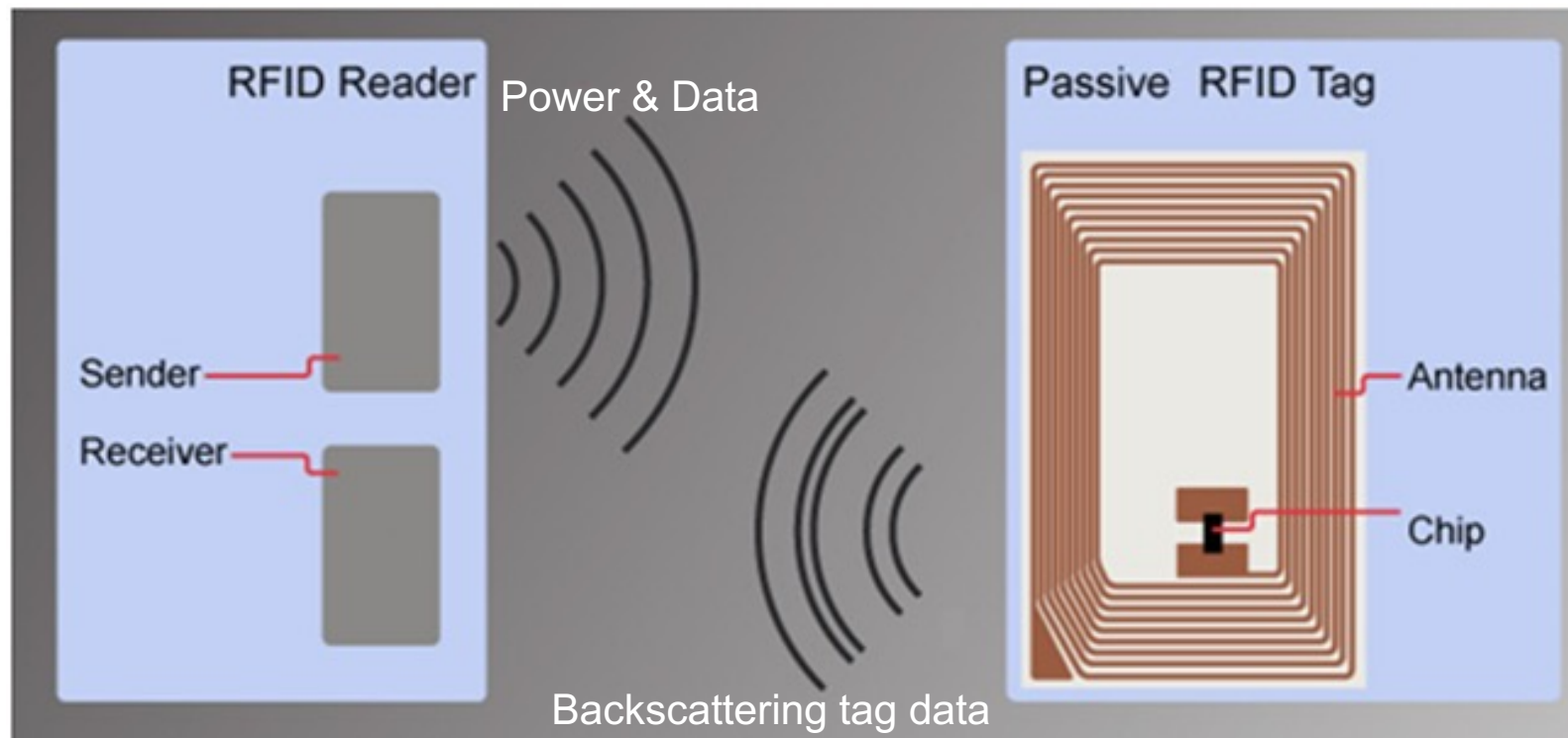
Deng, Xiao Dong
November 29th, 2024

Outline

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

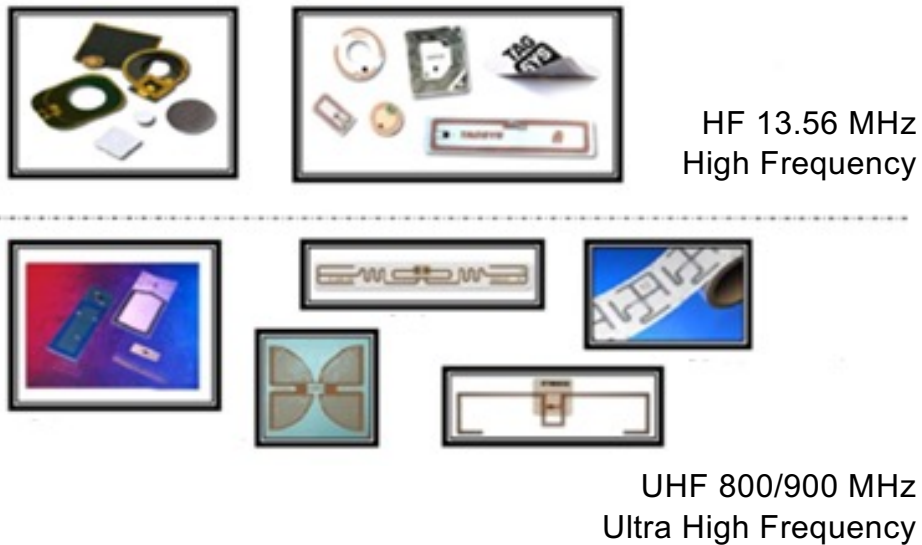
What is RFID?

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



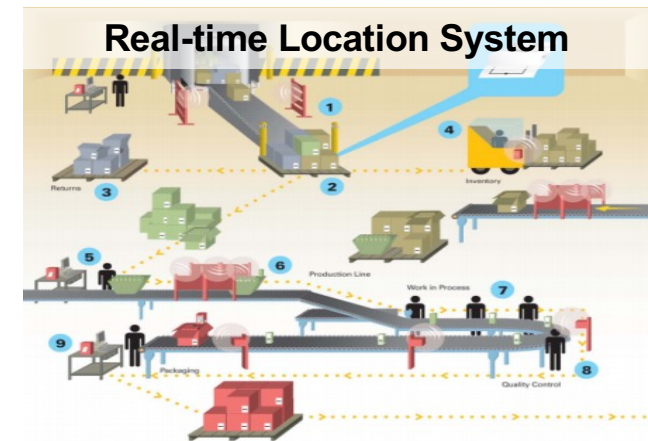
RFID Frequencies

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



	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



Siemens SIMATIC UHF RFID Portfolio



Web UI of Siemens SIMATIC Reader

SIEMENS SIMATIC RF690R

11/21/2024 11:24:47 Device status: Idle ■ English

Tag monitor

Basic settings: Readpoint_1 [1] Antenna 1 Circular ▶ ⏸ ✕ Continuous acquisition Single acquisition

Transponder list

Identified transponders 3 Valid transponders 3 Transponders in the antenna field 1 EPC ID in ASCII format

	EPC ID	Antenna	Power (min)	Power	RSSI	RSSI min	RSSI max	Acquisition cycles	Date/time
<input checked="" type="checkbox"/>	30084009500A600B700C800D30073008	1	20	20	-25.19	-28.20	-25.15	259	11/21/2024 11:22:24.975
<input type="checkbox"/>	30093382DD9014000000001	1	20	20	--	-48.69	-46.44	26	11/21/2024 11:22:24.977
<input checked="" type="checkbox"/>	30083382DD9014000000000	1	20	20	--	-55.79	-54.94	14	11/21/2024 11:22:25.243

Change power 20 dBm

RSSI graph

Recording time 15 [s]

Time interval: 0 min 15 sec 0 ms

RSSI Valid transponders

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

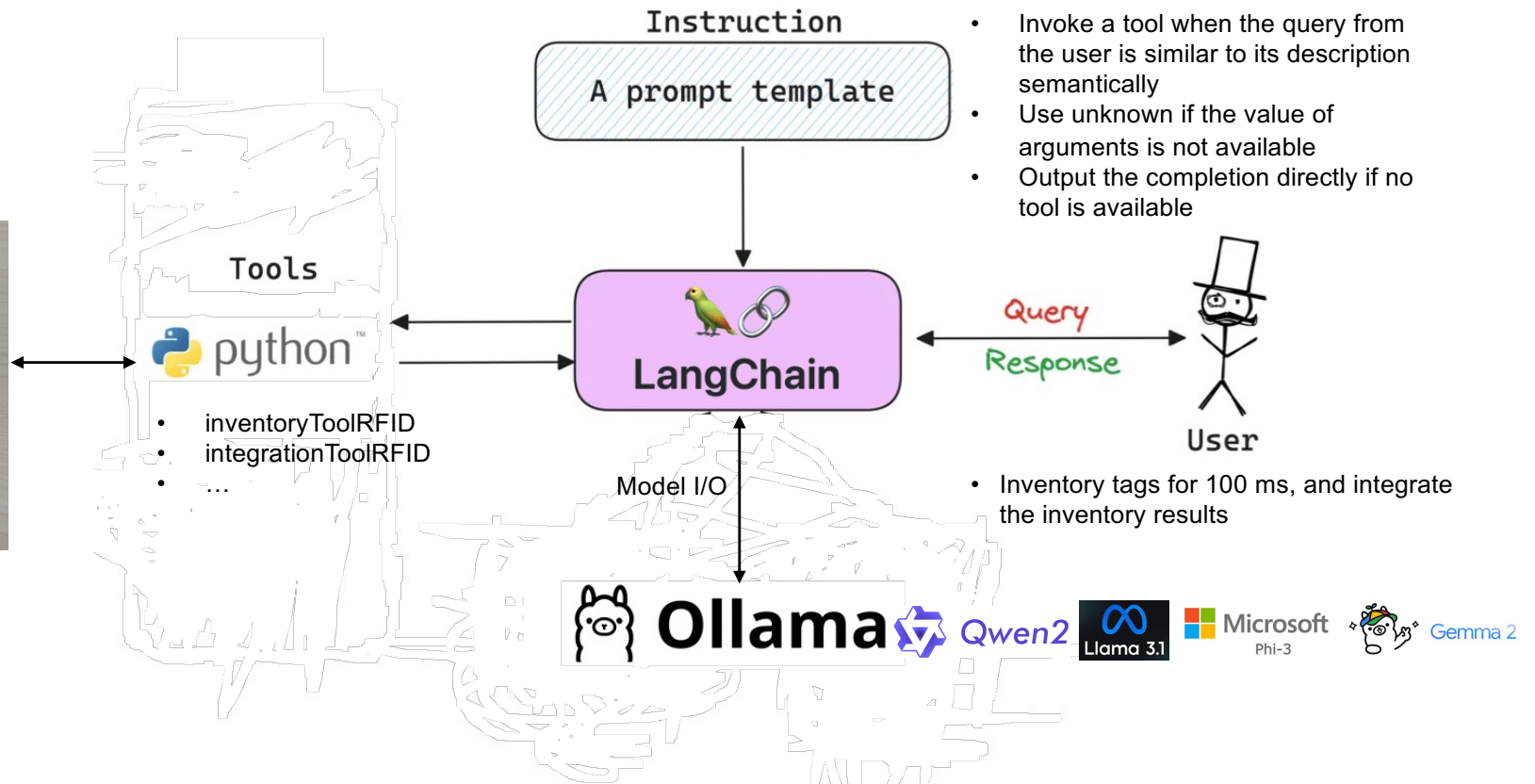
Outline

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

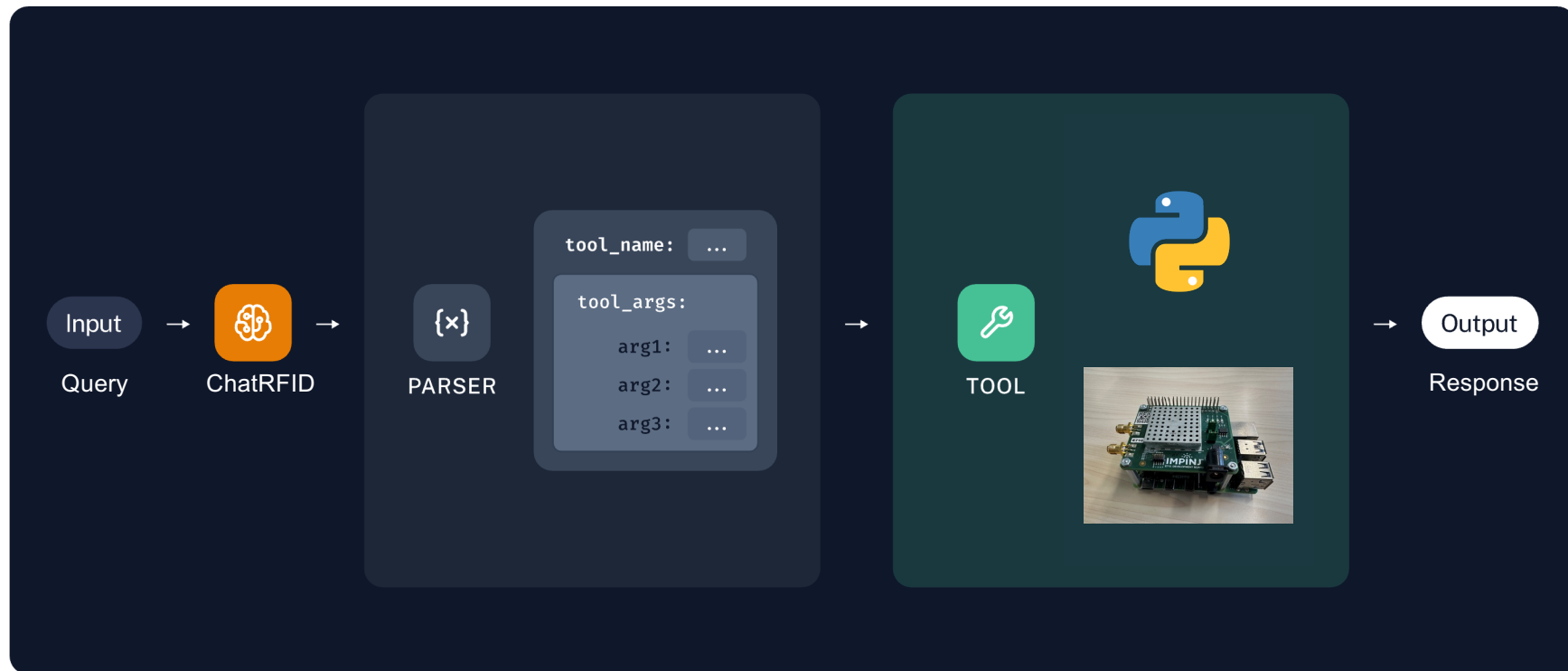
ChatRFID: a tool-calling AI-Agent for controlling RFID devices



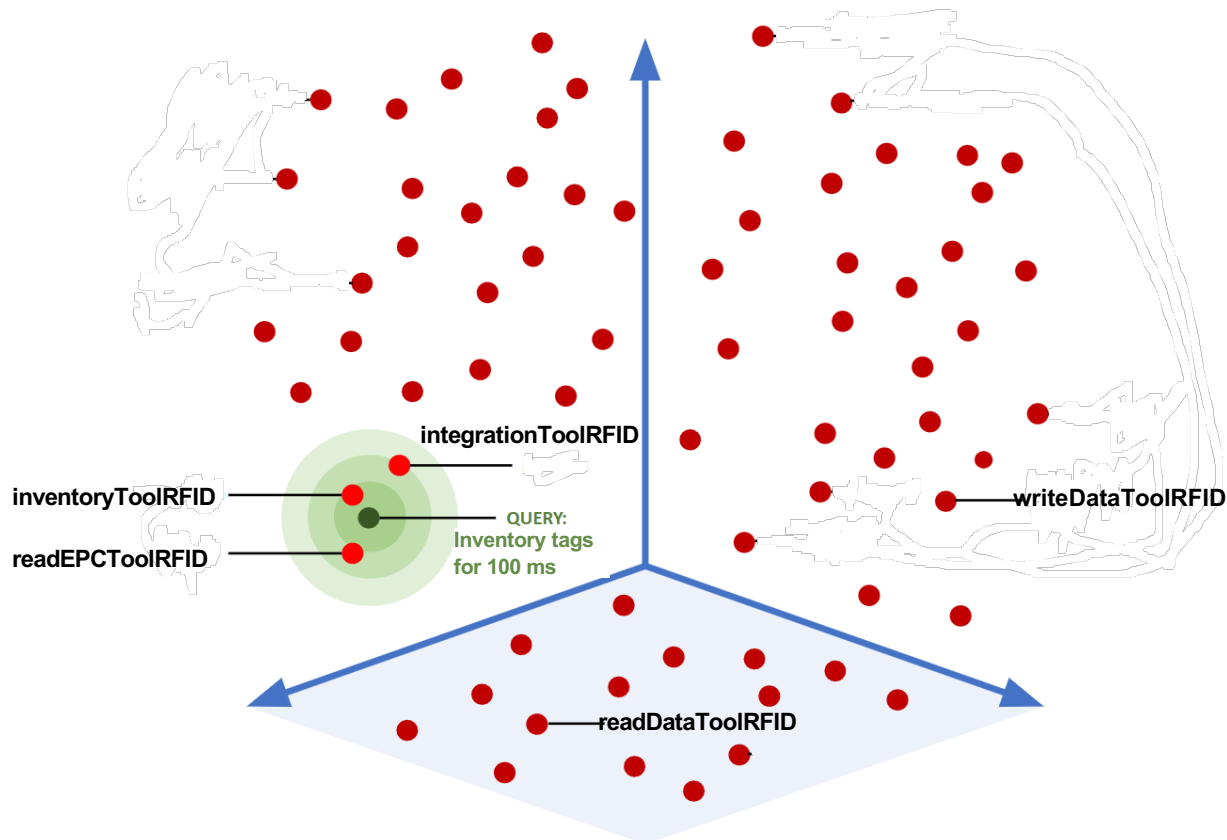
RFID Reader



Tool Calling



Which Bound Tool Should be Called? Sematic Similarity Task



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

UI of ChatRFID

```
pi@raspi-l1-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":"3008409500a600b700c800d30073008"}, {"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","Epc":"30084009500a600b700c800d30073008"}, {"id":1,"count":9,"timestamp":"2024-11-21 04:00:32.559821","Epc":"300933b2ddd9014000000001"}, {"id":1,"count":10,"timestamp":"2024-11-21 04:00:32.570980","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,"timestamp":"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":"30084009500a600b700c800d30073008"}, {"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":"300933b2ddd9014000000001"}, {"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":"300833b2ddd9014000000000"}, {"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":"30084009500a600b700c800d30073008"}, {"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

Qwen2.5-7B

Introduction

Qwen2.5 is the latest series of Qwen large language models. For Qwen2.5, we release a number of base language models and instruction-tuned language models ranging from 0.5 to 72 billion parameters. Qwen2.5 brings the following improvements upon Qwen2:

- Significantly **more knowledge** and has greatly improved capabilities in **coding and mathematics**, thanks to our specialized expert models in these domains.
- Significant improvements in **instruction following**, **generating long texts** (over 8K tokens), **understanding structured data** (e.g. tables), and **generating structured outputs** especially JSON. **More resilient to the diversity of system prompts**, enhancing role-play implementation and condition-setting for chatbots.
- Long-context Support** up to 128K tokens and can generate up to 8K tokens.
- Multilingual support** for over 29 languages, including Chinese, English, French, Spanish, Portuguese, German, Italian, Russian, Japanese, Korean, Vietnamese, Thai, Arabic, and more.

This repo contains the base 7B Qwen2.5 model, which has the following features:

- Type: Causal Language Models
- Training Stage: Pretraining
- Architecture: transformers with RoPE, SwiGLU, RMSNorm, and Attention QKV bias
- Number of Parameters: 7.61B
- Number of Parameters (Non-Embedding): 6.53B
- Number of Layers: 28
- Number of Attention Heads (GQA): 28 for Q and 4 for KV
- Context Length: 131,072 tokens

We do not recommend using base language models for conversations. Instead, you can apply post-training, e.g., SFT, RLHF, continued pretraining, etc., on this model.

For more details, please refer to our [blog](#), [GitHub](#), and [Documentation](#).

Edit model card

Downloads last month
70,938



Safetensors Model size 7.62B params Tensor type BF16

Inference API Cold

Text Generation Examples

Input a message to start chatting with Qwen/Qwen2.5-7B.

Your sentence here... Send

View Code Maximize

Model tree for Qwen/Qwen2.5-7B

Adapters	2 models
Finetunes	64 models
Merges	39 models
Quantizations	41 models

Spaces using Qwen/Qwen2.5-7B 2

- Manu97423/Test-Qwen-Qwen2.5-7B
- realaer/src

Collection including Qwen/Qwen2.5-7B

Qwen2.5 Collection
Qwen2.5 language models, including pret... - 45 items - Updated Sep 18 - 369

llama.cpp

The logo for LLaMA C++ is displayed on a dark background. The text "LLaMA" is in white, and the "C++" is in orange. The "C" is stylized with a flame-like shape above it.

license [MIT](#)  Server [passing](#) conan [b3542](#)

[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>