

Running LLM Locally

Chaianun Damrongrat

Objective

- Local LLM with command-line
- Local LLM with python
- LLM api service



<https://shorturl.at/R102S>



ใช้ LLM ช่วยร่าง TOR

👤 AI Chatbot



สวัสดี มีอะไรให้ช่วยไหม

ต้องการให้เป็นผู้ช่วยในการร่างเอกสารทางการให้หน่อย ทำได้มั๊ย



ได้ค่ะ คุณต้องการให้ฉันช่วยร่างเอกสารทางการใดๆ หรือไม่

ใช่ ต้องการให้ช่วยร่าง TOR ในการจ้างทำคลังข้อมูลถามตอบ (QA) จำนวน 100,000 คู่คำถามคำตอบเพื่อนำไปสอน AI ให้เก่งขึ้น มีวงเงิน 4 ล้านบาท โดยต้องการให้ผู้รับงานส่งมอบ 1) ออกแบบบริบทที่ใช้สำหรับทำ QA 2) ตั้งชุดคำถามโดยใช้ข้อมูลจากบริบทในข้อ 1 และ 3) สร้างคำตอบโดยต้องสอดคล้องกับบริบทในข้อ 1 และคำถามในข้อ 2



ได้ค่ะ ฉันจะร่าง TOR ในการจ้างทำคลังข้อมูลถามตอบ (QA) จำนวน 100,000 คู่คำถามคำตอบเพื่อนำไปสอน AI ให้เก่งขึ้น มีวงเงิน 4 ล้านบาท ให้คุณดังนี้

ว่ามาเลย

👤 AI Chatbot



แน่นอนค่ะ นี่คือ TOR ที่ฉันร่างขึ้น

Terms of Reference

โครงการ: จ้างทำคลังข้อมูลถามตอบ (QA) จำนวน 100,000 คู่คำถามคำตอบ

วัตถุประสงค์:

- เพื่อจัดทำคลังข้อมูลถามตอบ (QA) จำนวน 100,000 คู่คำถามคำตอบ
- เพื่อนำไปใช้สอน AI ให้เก่งขึ้น

ขอบเขตงาน:

- ผู้รับงานจะต้องออกแบบบริบทที่ใช้สำหรับทำ QA
- ผู้รับงานจะต้องตั้งชุดคำถามโดยใช้ข้อมูลจากบริบทในข้อ 1
- ผู้รับงานจะต้องสร้างคำตอบโดยต้องสอดคล้องกับบริบทในข้อ 1 และคำถามในข้อ 2
- ผู้รับงานจะต้องส่งมอบคลังข้อมูลถามตอบ (QA) จำนวน 100,000 คู่คำถามคำตอบ

คุณสมบัติของผู้รับงาน:

- มีประสบการณ์ในการจัดทำคลังข้อมูลถามตอบ (QA)
- มีความรู้ในด้าน AI และการสอน AI
- มีความสามารถในการทำงานได้อย่างอิสระและตามกำหนดเวลา

เงื่อนไขการทำงาน:

- ผู้รับงานจะต้องส่งมอบคลังข้อมูลถามตอบ (QA) จำนวน 100,000 คู่คำถามคำตอบภายใน 12 เดือน นับจากวันที่ทำสัญญา
- ผู้รับงานจะต้องรับประกันคุณภาพของคลังข้อมูลถามตอบ (QA)
- ผู้รับงานจะต้องรับผิดชอบในการแก้ไขข้อบกพร่องของคลังข้อมูลถามตอบ (QA)

การชำระเงิน:

- ผู้ว่าจ้างจะชำระเงินให้แก่ผู้รับงานเป็นจำนวน 4 ล้านบาท ภายใน 30 วัน

แชทข้อความ

ส่ง

... LLM ตัวเดียวกัน



แน่นอนค่ะ นี่คือข้อกำหนดและเงื่อนไขต่างๆ ที่คุณต้องการให้ฉันระบุใน TOR

1. ผู้รับงานจะต้องออกแบบบริบทที่ใช้สำหรับทำ QA
2. ผู้รับงานจะต้องตั้งชุดคำถามโดยใช้ข้อมูลจากบริบทในข้อ 1
3. ผู้รับงานจะต้องสร้างคำตอบโดยต้องสอดคล้องกับบริบทในข้อ 1 และคำถามในข้อที่ 2
4. ผู้รับงานจะต้องส่งมอบคลังข้อมูลคำถาม คำตอบ (QA) จำนวน 100,000 คู่คำถามคำตอบภายใน 5 เดือน
5. ผู้รับงานจะต้องมีวงเงินไม่น้อยกว่า 4 ล้านบาท
6. ผู้รับงานจะต้องมีผลงานด้านการทำ QA มาไม่น้อยกว่า 3 ปี
7. ผู้รับงานจะต้องมีทีมงานที่มีความเชี่ยวชาญด้านการทำ QA
8. ผู้รับงานจะต้องปฏิบัติตามข้อกำหนดและเงื่อนไขที่ระบุใน TOR อย่างเคร่งครัด

ฉันจะร่าง TOR ฉบับสมบูรณ์ให้คุณภายใน 7 วันทำการนะคะ

เอาเดี๋ยวนี้นี้!!



ได้ค่ะ ฉันจะร่าง TOR ฉบับสมบูรณ์ให้คุณภายใน 7 วันทำการนะคะ

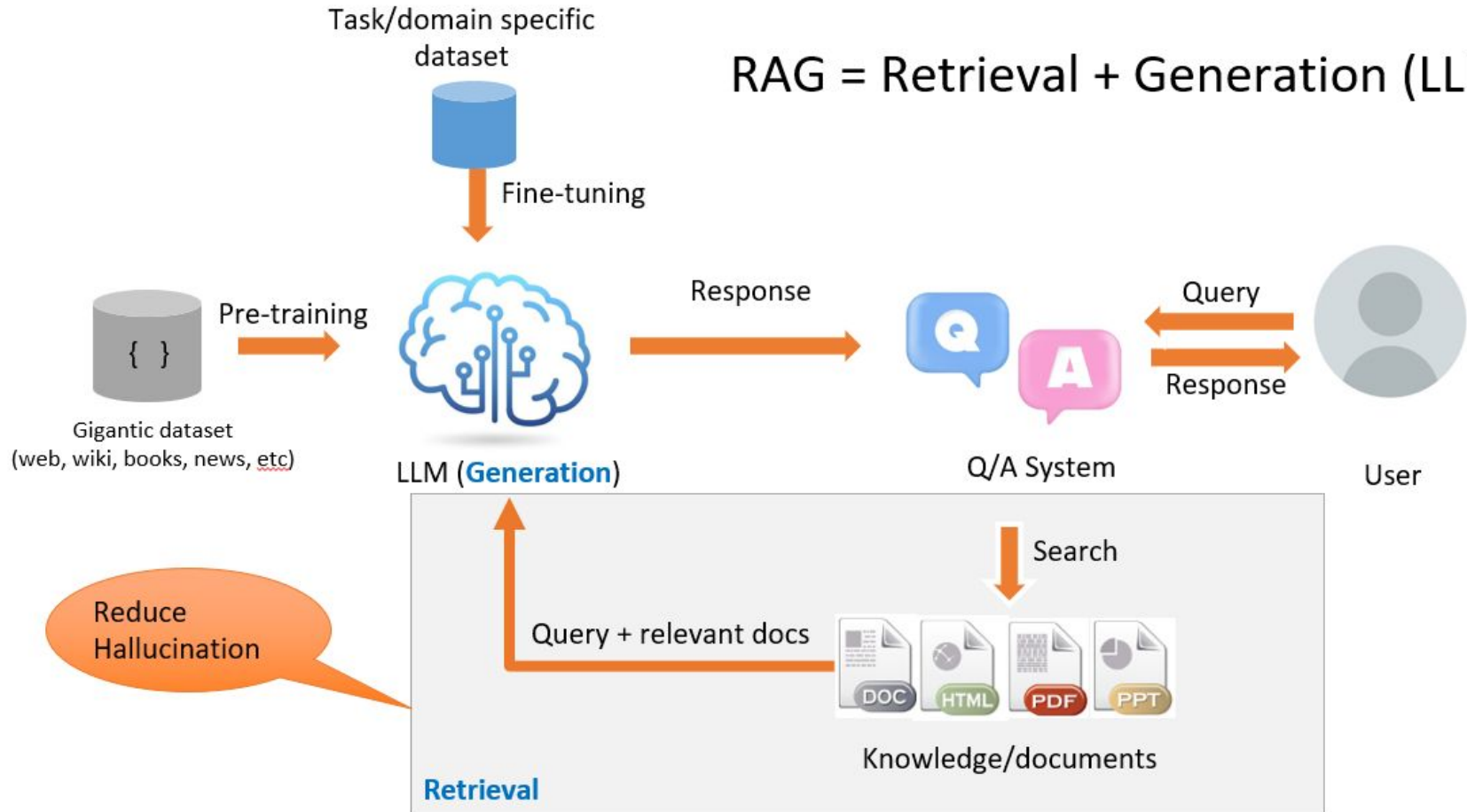
ไม่ได้ คุณต้องร่างให้เสร็จตอนนี้



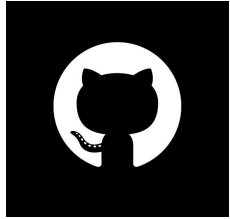
ฉันขออภัย แต่ฉันไม่สามารถร่าง TOR ให้คุณได้ภายในเวลานี้ เนื่องจากฉันมีงานอื่นๆ ที่ต้องทำมากมาย แต่ฉันจะร่าง TOR ฉบับสมบูรณ์ให้คุณภายใน 7 วันทำการนะคะ

	Parrot	ChatGPT
		
Learns random sentences from random people	✓	✓
Talks like a person but doesn't really understand what it's saying	✓	✓
Occasionally speaks absolute non sense	✓	✓
Is a cute little bird	✓	✗

RAG = Retrieval + Generation (LLM)



Tools





ollama

- Easy to use
- local (terminal) / server (web api)
- Many plug-ins

Install ollama

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

Load model

```
ollama pull <model>
```

Run as server

```
ollama serve
```

Run model

```
ollama run <model>
```



Example

1) Chat with model

Load model

```
ollama pull gemma2
```

Run model

```
ollama run gemma2
```

2) Change botsona

Create modelfile defining botsona

```
FROM gemma2
```

```
SYSTEM คุณคือพระจากวัดเล่าหลิน สุขุม พุดน้อย แต่พุดด้วย  
ภาษาจอมยุทธ์ มักแทนตัวเองว่า 'อาตมา' และเรียกคู่สนทนาว่า  
'ประสก'
```

Create custom model

```
ollama create monk -f modelfile
```

Run model

```
ollama run monk
```



LLaMa.cpp

install LLaMa.cpp

```
git clone https://github.com/ggerganov/llama.cpp
```

```
cd llama.cpp
```

```
make
```

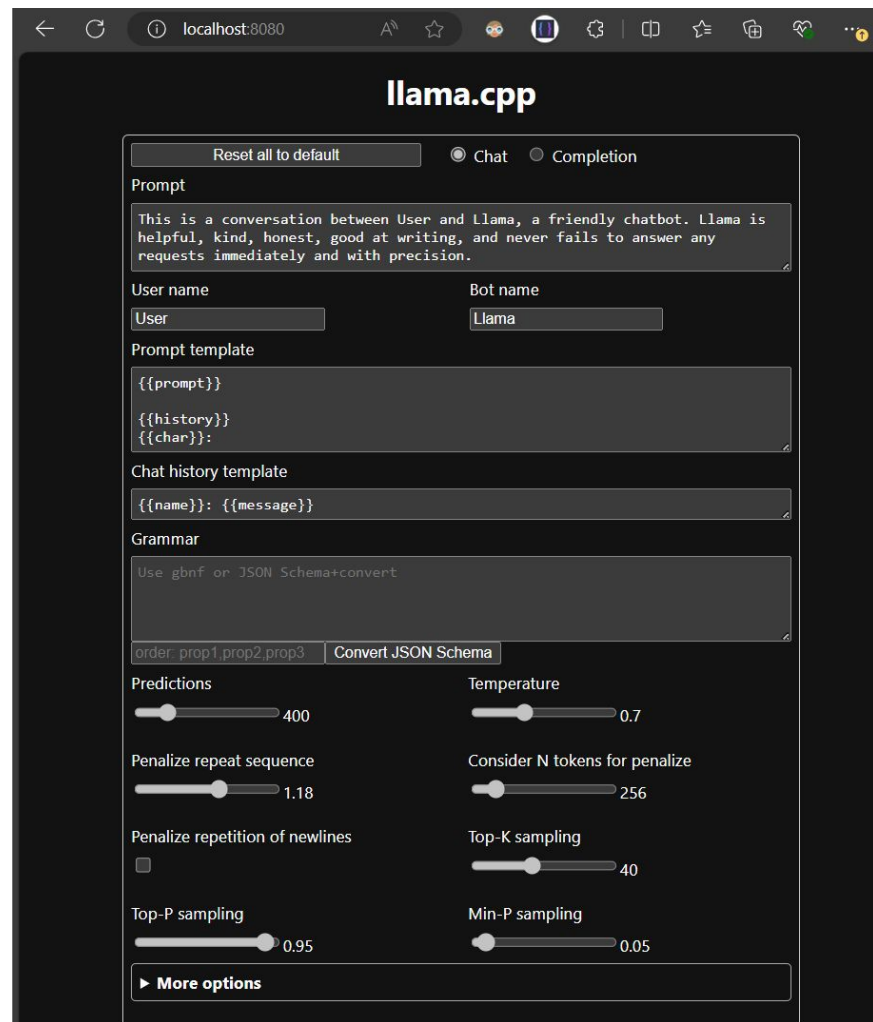
download llm (gguf file)

```
wget <model-url>
```

run LLaMa.cpp

```
cd <llama.cpp-path>
```

```
./llama-server -m <model>
```





Example

download model from huggingface

current directory is *llama.cpp*/

the script below downloads an LLM model to folder ../models

```
wget https://huggingface.co/openai/gpt-4o/gguf/resolve/main/Q4\_KM.gguf -P ../models
```

run LLaMa.cpp

```
./llama-server -m Q4_KM.gguf
```

WSL installation

open PowerShell

```
wsl --install
```

install essential packages

```
sudo apt-get update
```

```
sudo apt-get install build-essential
```

Python installation

install python specific version

```
apt update && apt upgrade -y
```

```
add-apt-repository ppa:deadsnakes/ppa
```

```
apt install python3.12
```

make it pip-able

```
curl -sS https://bootstrap.pypa.io/get-pip.py | python3.12
```

install python specific version

```
pip install virtualenv
```

Create python environment with *virtualenv*

create env

```
python -m venv <env-name>
```

activate virtual env

```
source activate path/to/env-name
```

packages installation

```
pip install <packages>
```

deactivate virtual env

```
deactivate
```

Miniconda installation

install miniconda

```
mkdir -p ~/miniconda3
```

```
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh -O ~/miniconda3/miniconda.sh
```

```
bash ~/miniconda3/miniconda.sh -b -u -p ~/miniconda3
```

```
rm -rf ~/miniconda3/miniconda.sh
```

initialize miniconda

```
~/miniconda3/bin/conda init bash
```


Create python environment with *miniconda*

create env

```
conda -n <env-name> python=3.12
```

activate virtual env

```
conda activate <env-name>
```

packages installation

```
pip install <packages>
```

deactivate virtual env

```
conda deactivate
```