Ollama

Fast campus

Ollama 설치

https://ollama.com/

원하는 모델 다운로드

아래의 예시는 EEVE

- HF: https://huggingface.co/yanolja/EEVE-Korean-Instruct-10.8B-v1.0
- GGUF: https://huggingface.co/teddylee777/EEVE-Korean-Instruct-10.8B-v1.0-gguf

Modelfile 로부터 커스텀 모델 생성하기

- 모델을 임포트하기 위해 ModelFile을 먼저 생성해야 합니다. 자세한 정보는 ModelFile 관련 공식 문서에서 확인할 수 있습니다.

> Modelfile

```
FROM ggml-model-Q5_K_M.gguf

TEMPLATE """{{- if .System }}
<s>{{ .System }}</s>
{{- end }}
<s>Human:
{{ .Prompt }}</s>
```

```
<s>Assistant:
"""

SYSTEM """A chat between a curious user and an artificial
intelligence assistant. The assistant gives helpful,
detailed, and polite answers to the user's questions."""

PARAMETER stop <s>
PARAMETER stop </s>
```

모델 생성

모델 생성

```
ollama pull solar
```

```
ollama create EEVE-Korean-10.8B -f Modelfile
```

모델 리스트 확인

```
ollama list
```

```
(base)  teddy@Teddyui-MacBookPro  ~/Dev/ollama/gguf  ollama list
NAME                            ID              SIZE    MODIFIED
EEVE-Korean-10.8B:latest        5dd1a4c1b923    7.7 GB  4 seconds ago
gemma:7b                        430ed3535049    5.2 GB  4 weeks ago
llama2:13b                      d475bf4c50bc    7.4 GB  2 months ago
llama2:latest                   78e26419b446    3.8 GB  2 months ago
llava:7b                        8dd30f6b0cb1    4.7 GB  4 weeks ago
mistral:latest                  61e88e884507    4.1 GB  2 months ago
mixtral:latest                  7708c059a8bb    26 GB   4 weeks ago
mymistral:latest                f3f3a01e0440    4.1 GB  2 months ago
mymodel:latest                  c8c0410374ad    3.8 GB  2 months ago
```

```
ollama run EEVE-Korean-10.8B:latest
```

서버 port 바꾸기

```
OLLAMA_MODELS=~/.ollama/models OLLAMA_HOST=127.0.0.1:11434
ollama serve
```