

LLM 을 사용한 AI 챗봇 연구



201824423 김도훈

201924655 이보원

지도교수 김호원 교수님

목 차

1. 서론	1
1.1. 연구 배경	1
1.2. 연구 목표	2
2. 연구 배경	2
2.1. Fine Tuning	2
2.2. Stranford Alpaca	3
2.3. 데이터셋 구축	4
2.4. 사용자 인터페이스	6
2.5. 시스템 구성도	6
3. 연구 내용	7
3.1. 모델 선정	7
3.2. 모델 Fine Tuning	8
3.3. Web UI	9
4. 연구 결과 분석 및 평가	10
5. 결론 및 향후 연구 방향	11
6. 참고 문헌	12

1. 서론

1.1. 연구 배경

IT 기술의 발전으로 인간의 작업을 돕는 다양한 어플리케이션과 서비스들이 개발되고 있다. 기존의 자동화 시스템은 사전에 작동 로직을 직접 설정해 주어야 했고 챗봇의 경우에도 미리 설정된 질문에 대해서만 답변이 가능하다는 한계가 있었다. 그렇기에 사람과의 소통이 가능한 AI 기술의 필요성이 대두되었고 다양한 자연어 처리 기술에 대한 다양한 시도가 시작되었다. 그 결과로 LLM을 기반으로 한 생성형 AI가 등장하면서 선풍적인 인기를 끌고 있으며 비즈니스, 교육, 의료 등 여러 분야에 활용되고 있다.

대규모 언어 모델 즉, LLM(Large Language Model)은 방대한 양의 데이터를 학습하고 자연어를 비롯한 다양한 유형의 콘텐츠를 이해하고 생성하여 광범위한 작업을 수행하는 모델로서, 자연어 처리 기술의 성능의 발전에 크게 기여하고 있다. OpenAI의 GPT(Generative Pre-trained Transformer), 구글의 LaMDA(Language Model for Dialogue Applications), 삼성의 Gauss 등도 LLM의 한 종류이며 이러한 모델들은 수천억 개에 달하는 파라미터를 포함하는 초거대 모델이라 할 수 있다. LLM은 인간의 텍스트를 이해하고 다음에 올 단어뿐만 아니라 단어와 단어를 연결하는 부분까지도 예측하여 사람이 입력한 것 같은 자연스러운 문장을 생성해 내기 때문에 대화를 기반으로 하는 AI 서비스에 적합하다. 언어 번역, 고객 지원, 창작 활동 등 다양한 분야에서 대화형 AI 서비스가 개발되고 있으며 이러한 서비스들은 점점 더 정교해져 소비자와 더욱 자연스러운 대화를 나눌 수 있게 된다.

현대 사회에서 사람들은 식사 메뉴를 선택하는 데에 있어 다양한 정보와 선택지를 필요로 한다. 특히 바쁜 일상에서 시간을 절약하고 싶은 사람들이나 새로운 맛을 경험하고 싶어 하는 사람들이 늘어감에 따라 사용자에게 맞춤형 음식을 추천해 주는 시스템의 필요성이 커지게 되었다.

기존의 음식 메뉴 추천 시스템은 단순한 필터링 방법을 사용하여 개개인의 선호와 요구 사항을 충분히 반영하지 못하는 한계가 있다. 또한 동일하거나 유사한 음식만을 반복적으로 추천하는 경향이 있어 사용자가 새로운 음식을 시도하는 것이 제한된다는 어려움이 있다.

LLM을 기반으로 한 챗봇은 사용자의 선호도와 취향을 정확하게 파악하여 맞춤형 추천과 다양한 선택지를 제공함으로써 기존 시스템의 한계를 극복하고 사용자의 만족도를 높일 수 있다

1.2. 연구 목표

본 연구에서는 음식 추천 관련 대화형 데이터셋을 구축한 후 LLM을 fine tuning하여 다양한 식사 메뉴와 정보를 제공하는 메뉴 추천 AI 챗봇을 웹 어플리케이션 형태로 개발하고자 한다

2. 연구 배경

2.1. Fine Tuning

Fine Tuning이란 특정 작업이나 도메인에 높은 적합성을 확보하기 위해, 이미 훈련된 대규모 언어 모델에 특정 데이터셋을 사용하여 추가적인 학습을 수행하는 작업을 말한다.

기존에는 LoRA 방식으로 학습을 진행하려 했지만 Colab Free에서 제공되는 GPU의 성능을 고려 할 때 학습 중 VRAM이 부족한 상황이 생길 수도 있을 것 같아 LoRA보다 메모리 효율적인 QLoRA를 사용하여 학습시키기로 하였다.

QLoRA(Quantized Low Rank Adapters)는 메모리 사용량을 크게 줄이는 fine tuning 접근법으로, 전체 16비트 fine tuning 성능을 유지하면서 낮은 RAM을 가진 GPU에서 더욱 거대한 매개변수를 가진 모델을 fine tuning 할 수 있다. 이를 위해 QLoRA는 몇 가지 기능을 제공한다.

4-bit NormalFloat	QLoRA의 가장 핵심적인 방법론으로 PLM 가중치가 4비트로 양자화된 채 저장되며 그 데이터 타입이 NormalFloat이다.
Double Quantization	양자화 상수를 양자화하여 평균 메모리 설치 공간을 줄인다.
Paged Optimization	GPU가 사용하는 VRAM 페이지를 CPU RAM에도 일부 저장할 수 있게 할 당해주는 기술이다.

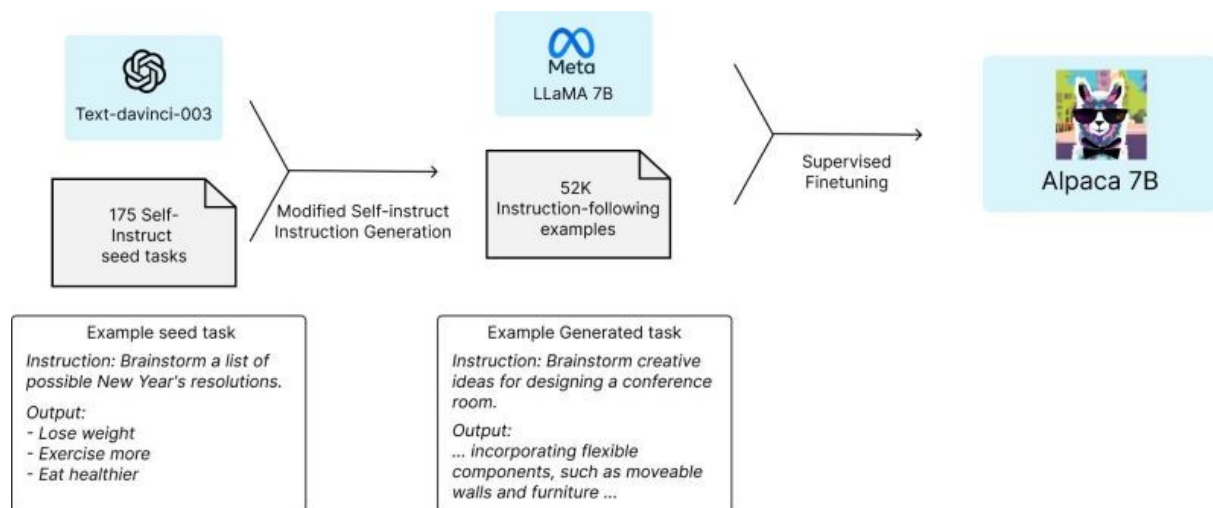
[표 1] QLoRA Mechanism

2.2. Stranford Alpaca

Stanford Alpaca는 Stanford University에서 Meta의 오픈소스 언어 모델인 LLaMA 7B를 fine tuning하여 개발한 모델로써 학습에 많은 리소스가 필요하고 데이터의 양과 품질에 크게 의존하는 기존 대형 언어 모델들의 문제점을 개선하기 위해 개발되었다.

Alpaca는 GPT-3을 사용하여 self-instruction 형식으로 생성된 52K의 instruction-folloing demonstrations를 사용하여 훈련되었으며 self-instruct 평가 세트에서 OpenAI의 GPT-3와 유사하면서 재현이 훨씬이 쉽고 저렴하다는 것을 입증했다.

아래 그림은 Alpaca 모델을 얻는 방법에 대해서 설명하고 있다. 데이터를 self-instruction method를 통해 생성하여 비용을 크게 감축하였다.



[그림 1] Stranford Alpaca 학습 과정

2.3. 데이터셋 구축

데이터셋은 그림 2와 같이 Instruction과 Input, Output으로 구성하여 제작한다.

```
{
  "instruction": "Identify the odd one out.",
  "input": "Twitter, Instagram, Telegram",
  "output": "Telegram"
},
```

[그림 2] Alpaca 데이터셋 예시

기존에는 학습 데이터 수집을 위해 배달 플랫폼이나 리뷰 사이트를 크롤링하려 했지만 음식 추천에 관련한 대화 데이터가 부족하여 ChatGPT API를 이용한 Self-Instruct를 활용하여 데이터셋을 직접 제작하기로 하였다.

Self-Instruct는 모델이 스스로 학습 데이터를 생성하고 그 데이터를 사용하여 훈련하는 방식으로, 대량의 instruction 데이터셋을 생성하기 위한 목적으로 사용하는 프레임워크이다. 사람이 작성한 Instruction-output 쌍으로 이루어진 Instruction seed를 gpt의 프롬프트에 입력하면 모델이 이를 기반으로 instruction set을 생성한다. 이 과정을 여러 번 반복하면 언어 모델의 fine tuning에 사용할 수 있는 대량 학습 데이터셋이 생성된다. 이를 통해 사람이 만든 학습 데이터의 품질과 양에 크게 의존하여 다양성과 창의성을 제한받을 수 있다는 문제점을 극복할 수 있다. 본 연구에서는 gpt-3.5 모델 중 가장 효과적이고 비용 효율적인 gpt-3.5-turbo 모델의 API를 이용하여 self-instruct를 수행할 것이다.

```
def encode_prompt(prompt_instructions):
    """Encode multiple prompt instructions into a single string."""
    prompt = open("./prompt.txt").read() + "\n"

    for idx, task_dict in enumerate(prompt_instructions):
        (instruction, input, output) = task_dict["instruction"], task_dict["input"], task_dict["output"]
        instruction = re.sub(r"\s+", " ", instruction).strip().rstrip(":")
        input = "<noinput>" if input.lower() == "" else input
        prompt += f"###\n"
        prompt += f"{idx + 1}. Instruction: {instruction}\n"
        prompt += f"{idx + 1}. Input:\n{input}\n"
        prompt += f"{idx + 1}. Output:\n{output}\n"
    prompt += f"###\n"
    prompt += f"{idx + 2}. Instruction:"
    return prompt
```

▲ 프롬프트 인코딩

```
{
  "id": "seed_task_0",
  "name": "recommend_menu",
  "instruction": "편의점 간식 추천해줘",
  "instances": [
    {
      "input": "",
      "output": "젤리나 초콜릿은 어떠세요?"
    }
  ],
  "is_classification": false
}
```

▲ Instruction seed 예시

제작한 데이터셋은 다양한 머신러닝 프레임워크를 제공하는 인공지능 플랫폼 huggingface에 저장하고 관리한다. huggingface에서 지원하는 서비스는 표2와 같다.

서비스	기능
트랜스포머 라이브러리	파이썬으로 짜여진 오픈소스 패키지이며 인공지능 칩 레이어 2단에 위치하며 연산 처리 지원
허깅 페이스 허브	호스팅,버전관리 지원
데이터셋	빅 데이터 프로세싱 라이브러리
인터페이스 API	직접적이고 단순화된 모델 처리 지원

[표 2] huggingface 지원 서비스

instruction string · lengths	input string · classes	output string · lengths
 4 302	 25 values	 9 903
삼계탕은 어떤 재료로 만드나요?		삼계탕은 닭고기, 대추, 밤, 마늘, 생강 등을 사용하여 맑은 육...
닭가슴살 요리법 좀 알려줄래?		닭가슴살을 사용한 요리는 다양한데, 그 중에서 간단한 방법으로...
부대찌개 만드는 법 좀 알려줘		부대찌개 만드는 법은 물에 된장과 고추장, 고기를 넣고 삶아 육...
샐러드 어떤 재료로 만들어야 맛있을까		샐러드는 식재료를 잘 고르는 것이 중요한데, 양상추, 토마토, ...
떡은 어떻게 만드는지 알려줘		떡 만드는 방법은 밥을 갈아서 반죽을 만든 뒤 원하는 모양으로 ...
계란 후라이는 어떻게 만드는 건가요?		계란 후라이는 달걀을 틀에 넣고 기름을 두루 덮어 중약불에서 익...

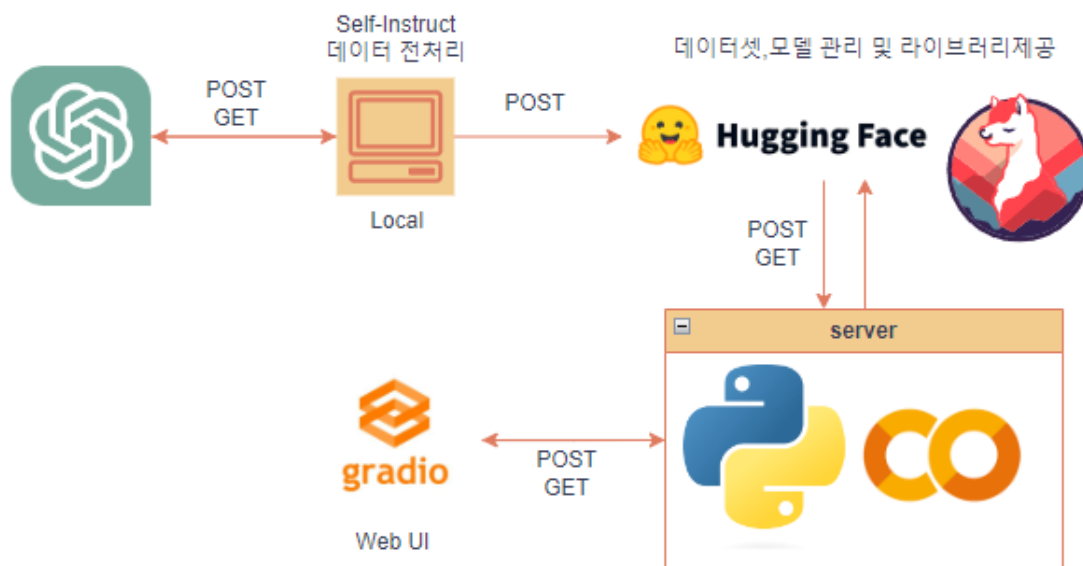
[그림 3] 데이터셋 preview

2.4. 사용자 인터페이스

Web UI를 제작하기 위해 Gradio 라이브러리를 사용하기로 하였다. Gradio는 머신 러닝 모델 사용자 인터페이스를 만드는 데 사용되는 Python 오픈 소스 패키지이다. 웹 어플리케이션을 만들 때 자바스크립트, HTML, CSS 등으로 인터페이스를 만들고 Django, Flask 등으로 서버를 만들어 배포해야 하지만 Gradio를 사용하면 이런 복잡한 과정 없이 배포가 가능하다. 파이썬으로 만든 머신러닝이나 데이터 사이언스 프로그램을 웹 어플리케이션으로 만들어주며 PyTorch, TensorFlow와 같은 기계 학습 프레임워크와 호환된다. 또한 기계 학습에 최적화된 UI 컴포넌트를 제공하여 쉽고 빠르게 인터페이스를 생성할 수 있다. 앱을 빌드하면 최대 72시간 사용할 수 있는 public URL을 제공한다.

2.5. 시스템 구성도

개발 환경은 구글이 제공하는 클라우드 기반 Jupyter Notebook 환경인 Google Colab을 이용하였고 최종 시스템 구성도는 아래와 같다.



[그림 4] 시스템 구성도

3. 연구 내용

3.1. 모델 선정

fine tuning을 진행할 베이스 모델로 오픈소스 모델 KoAlpaca를 선정했다. KoAlpaca는 Stanford Alpaca 모델을 학습한 방식과 동일한 방식으로 학습을 진행한 Alpaca 모델이며 백본 모델로(한국어) Polyglot-ko 모델을 사용하였다.

KoAlpaca는 Polyglot-ko 12.8B 기반 모델과 Polyglot-ko 5.8B기반 모델을 제공하고 있지만 12.8B 기반 모델의 경우 크기가 크기 때문에 학습 시간이 오래 걸리고 효과적인 fine tuning을 위해 많은 학습 데이터가 필요하므로 연구 개발 환경인 Colab의 저장 용량과 GPU RAM을 고려하여 Polyglot-ko 5.8B 기반 모델을 사용하기로 하였다.

Hyperparameter	Value
$n_{parameters}$	5,885,059,072
n_{layers}	28
d_{model}	4096
d_{ff}	16,384
n_{heads}	16
d_{head}	256
n_{ctx}	2,048
n_{vocab}	30,003 / 30,080
Positional Encoding	Rotary Position Embedding (RoPE)
RoPE Dimensions	64

[그림 5] Polyglot-ko-5.8B 하이퍼파라미터 값

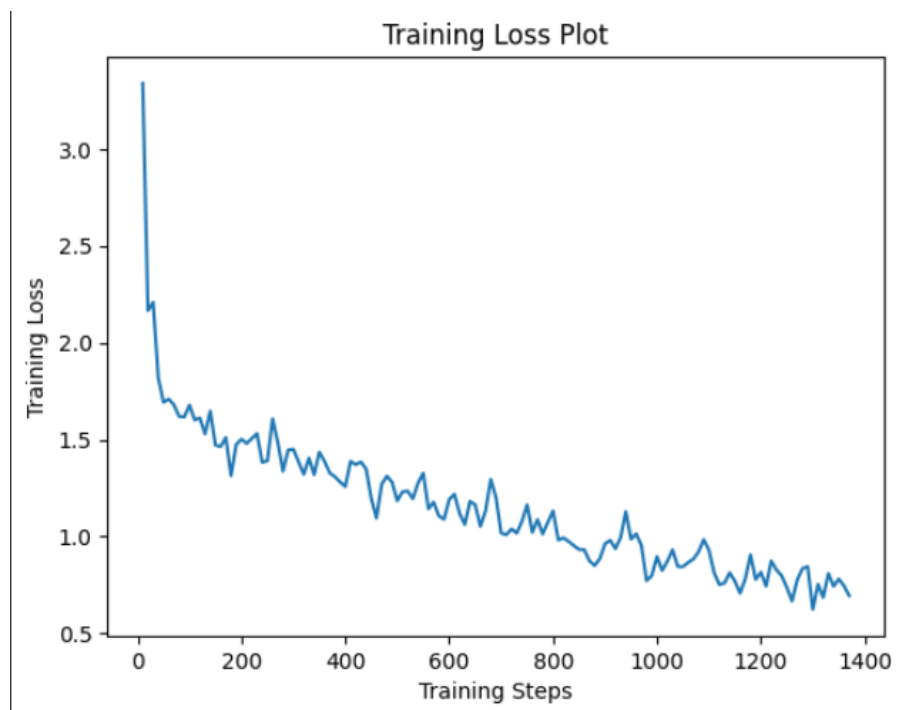
3.2. 모델 Fine Tuning

fine tuning은 Google Colab에서 무료로 제공되는 Tesla T4 GPU 런타임 환경에서 진행하였고 GPU RAM을 고려하여 QLoRA 방식으로 학습시켰다.

Self-instruct로 수집한 1100여개의 데이터셋을 사용하여 batch size는 8, max step은 1372로 약 10 epoch 학습시켰으며, 10 step마다 training loss 체크 로그를 남기고 이를 토대로 그래프를 작성하였다. 학습 하이퍼 파라미터와 학습 결과는 아래와 같다.

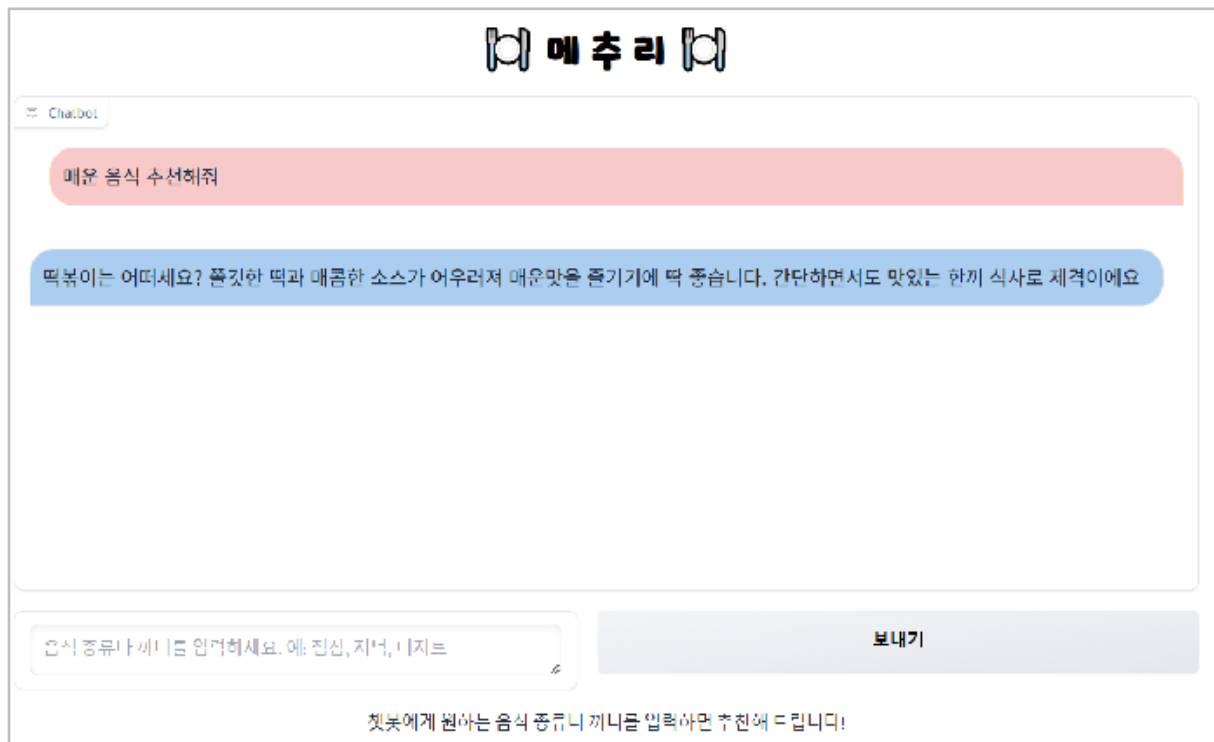
하이퍼 파라미터	설정값
per_device_train_batch_size	8
gradient_accumulation_steps	1
max_steps	1372
learning_rate	1e-4
logging_steps	10

[표 3] 학습 하이퍼 파라미터



[그림 6] training loss plot

3.3. Web UI



[그림 7] 인터페이스 디자인

어플리케이션의 이름은 '메추리'이며 사용자가 채팅창에 "매운 요리 추천해줘", "아침에 간단하게 먹을 수 있는 음식 추천해줘"와 같이 자신의 상황이나 추천받고 싶은 음식의 종류를 입력하면 챗봇이 그에 맞는 음식을 추천해준다. 답변이 마음에 들지 않으면 다시 챗봇에게 요청하여 다른 음식을 추천받을 수 있다.

```
def gen(x):
    q = f"### 질문: {x}\n\n### 답변:"
    # print(q)
    gened = model.generate(
        **tokenizer(
            q,
            return_tensors='pt',
            return_token_type_ids=False
        ).to('cuda'),
        max_new_tokens=200,
        early_stopping=True,
        do_sample=True,
        eos_token_id=2,
    )
    print(tokenizer.decode(gened[0]))
```

▲ 답변 생성 함수

4. 연구 결과 분석 및 평가

베이스 모델의 낮은 파라미터 값으로 인해 높은 추론 능력을 기대할 순 없지만 음식 추천 질문에는 어느 정도 자연스러운 답변을 생성하였다. 그러나 복잡하고 세밀한 답변을 요구하는 질문에는 엉뚱한 답변을 생성하거나 불완전한 문장을 생성하기도 하였다. 답변의 max_length값을 200으로 설정해 두었을 때 평균 응답 시간은 13~15초를 기록하였다.

❧ 에 추 리 ❧

Chatbot

닭고기 요리 추천해줘

닭갈비는 어떠세요? 닭갈비는 닭고기를 갈비처럼 양념에 재워 볶는 요리로, 매콤하고 달콤한 맛이 특징입니다. 닭갈비는 집에서 쉽게 만들 수 있어 가정에서도 맛있게 먹을 수 있는 요리입니다. 집에서 만들 때는 양념을 잘 배합하여 달콤한 맛을 살릴 수 있습니다. 밖에서 먹을 때는 닭갈비를 대표하는 도시인 춘천이나 닭갈비 축제가 열리는 도시인 춘천에서 먹는 것이 유명합니다. 축제나 공연을 즐기면서 맛있는 닭갈비를 즐길 수 있습니다.

음식 종류나 끼니를 입력하세요. 예: 점심, 저녁, 디저트

보내기

챗봇에게 원하는 음식 종류나 끼니를 입력하면 추천해 드립니다!

차가운 디저트 추천해주세요

차가운 디저트로는 아이스크림, 아이스 커피, 아이스 티, 아이스 에이드, 아이스 스무디 등이 있습니다. 아이스크림은 바닐라, 초콜릿, 딸기, 민트 초콜릿, 바닐라 에센스 등 다양한 맛으로 준비되어 있습니다. 아이스 커피는 에스프레소에 얼음을 넣어 만들며, 에스프레소 자체도 차가운 음료에 잘 어울립니다. 아이스 티는 홍차에 얼음을 넣어 끓여서 만들며, 아이스 에이드는 탄산수와 과일 주스, 시럽 등을 섞어서 만듭니다. 아이스 스무디는 아이스크림과 과일, 견과류 등을 블렌더에 갈아 만들 수 있습니다. 이런 차가운 디저트들은 상큼하고 시원한 맛으로 입맛을 돋웁니다.

[그림 8] 실행 결과

5. 결론 및 향후 연구 방향

본 과제에서는 LLM을 사용해 메뉴 추천 챗봇을 개발하였다.

대화형 데이터셋을 구축하기 위해 ChatGPT API를 이용한 Self-Instruct 방식으로 데이터를 수집하였고 구축한 데이터셋으로 기반 모델인 KoAlpaca에 QLoRA 방식 Fine Tuning을 실행하였다. 또한 모델을 웹 앱 형태로 빌드하고 배포하기 위해 Gradio 라이브러리를 사용하였다.

이처럼 이번 과제를 통해 LLM fine tuning 과정 전체를 경험해보았다. 음식의 레시피나 영양정보에 대한 답변에선 부정확한 정보를 많이 출력하는 등 처음 계획했던 음식 추천 전문 AI로 상용화 하기엔 무리가 있지만 음식 추천에 대한 질문에 한해서는 꽤 자연스러운 답변을 생성하는 성과를 거두었다. 컴퓨팅 자원과 시간의 한계로 낮은 파라미터의 모델과 작은 데이터셋을 사용했기 때문에 파라미터 개수가 더 많은 모델과 더 거대한 데이터셋을 사용한다면 더욱 자연스럽게 전문적인 챗봇을 만들 수 있을 것으로 기대된다.

6. 참고 문헌

- [1] Stranford University. Alpaca: A Strong, Replicable Instruction-Following Model Rohan Taori* and Ishaan Gulrajani* and Tianyi Zhang* and Yann Dubois* and Xuechen Li* and Carlos Guestrin and Percy Liang and Tatsunori B. Hashimoto
<https://crfm.stanford.edu/2023/03/13/alpaca.html>
- [2] Cornell University. QLoRA: Efficient Finetuning of Quantized LLMs Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Luke Zettlemoyer
- [3] Cornell University. Self-Instruct: Aligning Language Models with Self-Generated Instructions. Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, Hannaneh Hajishirzi. <https://arxiv.org/abs/2212.10560>
- [3] Beomi / KoAlpaca [Online]. <https://github.com/Beomi/KoAlpaca>
- [4] EleutherAI / polyglot [Online]. <https://github.com/EleutherAI/polyglot>
- [6] <https://gist.github.com/Beomi>