

# 프로젝트 기획서

SKN2기 : 사재민, 진정현, 장정원, 서종호

2024년 09월 30일

## 컴퓨터 부품 추천 대화형 시스템

### 프로젝트 개요

#### 배경 :

현재 대부분의 추천 시스템은 다음과 같은 한계를 가지고 있음

##### 1. 모델 업데이트의 제약 :

- 기존 모델은 학습된 시점 이후의 새로운 데이터를 활용하지 못해, 최신 트렌드나 선호도를 반영하기 어렵다.

##### 2. 개인화 데이터 활용 부족 :

- LLM은 학습 시 포함되지 않은 사용자 데이터를 사용할 수 없어, 개별 사용자 맞춤형 추천을 제공의 한계가 있다.

##### 3. 파인 튜닝의 비용과 시간문제 :

- LLM을 사용한 추천 시스템은 성능을 유지하기 위해 지속적인 파인 튜닝이 필요, 이 과정은 많은 시간과 비용이 소요

#### 목적 :

- \*\*대형 언어 모델(LLM)\*\*과 협업 필터링을 결합하여 사용자와의 대화를 통해 실시간으로 적합한 상품을 추천할 수 있는 시스템을 구축하는 것
- 이 시스템은 최신 데이터와 개인화된 사용자 정보를 실시간으로 반영, 비용과 시간이 적게 드는 모델을 구현하는 것을 목표로 함

#### 기대 효과 :

##### 1. 최신 데이터 반영 :

- 파인 튜닝 없이도 새로운 데이터를 실시간으로 학습하여 최신 트렌드와 사용자 선호도를 반영한 상품 추천 가능

##### 2. 개인화된 추천 제공 :

- 사용자 개별 데이터를 활용해 보다 정교하고 맞춤형 추천 제공

##### 3. 비용 및 시간 절감 :

- 지속적인 파인 튜닝이 필요 없는 모델 구조로, 시간과 비용을 크게 절감

# 데이터

## 데이터 출처

🔗 Amazon Reviews'23

## 데이터 설명

McAuley Lab에서 수집한 2023년도 대규모 아마존 리뷰 데이터셋

### 포함하는 정보:

- 1. User Reviews (평점, 내용, 유용성 평가 등)
- 2. Item Metadata (제품 설명, 가격, 원본 이미지 등)
- 3. Links (유저-아이템 / 함께 산 아이템 그래프)

### 사이즈:

Year	#Review	#User	#Item	#R-Token	#M-Token	#Domain	Timespan
2023	571.54M	54.51M	48.19M	30.14B	30.78B	33	May'96 - Sep'23

### 카테고리:

Category	#User	#Item	#Rating	#R-Token	#M-Token
All_Beauty	632.0K	112.6K	701.5K	31.6M	74.1M
Amazon_Fashion	2.0M	825.9K	2.5M	94.9M	510.5M
Appliances	1.8M	94.3K	2.1M	92.8M	95.3M
Arts_Crafts_and_Sewing	4.6M	801.3K	9.0M	350.0M	695.4M
Automotive	8.0M	2.0M	20.0M	824.9M	1.7B
Baby_Products	3.4M	217.7K	6.0M	323.3M	218.6M
Beauty_and_Personal_Care	11.3M	1.0M	23.9M	1.1B	913.7M
Books	10.3M	4.4M	29.5M	2.9B	3.7B
CDs_and_Vinyl	1.8M	701.7K	4.8M	514.8M	287.5M
Cell_Phones_and_Accessories	11.6M	1.3M	20.8M	935.4M	1.3B
Clothing_Shoes_and_Jewelry	22.6M	7.2M	66.0M	2.6B	5.9B
Digital_Music	101.0K	70.5K	130.4K	11.4M	22.3M
Electronics	18.3M	1.6M	43.9M	2.7B	1.7B
Gift_Cards	132.7K	1.1K	152.4K	3.6M	630.0K

# 데이터

Grocery_and_Gourmet_Food	7.0M	603.2K	14.3M	579.5M	462.8M
Handmade_Products	586.6K	164.7K	664.2K	23.3M	125.8M
Health_and_Household	12.5M	797.4K	25.6M	1.2B	787.2M
Health_and_Personal_Care	461.7K	60.3K	494.1K	23.9M	40.3M
Home_and_Kitchen	23.2M	3.7M	67.4M	3.1B	3.8B
Industrial_and_Scientific	3.4M	427.5K	5.2M	235.2M	363.1M
Kindle_Store	5.6M	1.6M	25.6M	2.2B	1.7B
Magazine_Subscriptions	60.1K	3.4K	71.5K	3.8M	1.3M
Movies_and_TV	6.5M	747.8K	17.3M	1.0B	415.5M
Musical_Instruments	1.8M	213.6K	3.0M	182.2M	200.1M
Office_Products	7.6M	710.4K	12.8M	574.7M	682.8M
Patio_Lawn_and_Garden	8.6M	851.7K	16.5M	781.3M	875.1M
Pet_Supplies	7.8M	492.7K	16.8M	905.9M	511.0M
Software	2.6M	89.2K	4.9M	179.4M	67.1M
Sports_and_Outdoors	10.3M	1.6M	19.6M	986.2M	1.3B
Subscription_Boxes	15.2K	641	16.2K	1.0M	447.0K
Tools_and_Home_Improvement	12.2M	1.5M	27.0M	1.3B	1.5B
Toys_and_Games	8.1M	890.7K	16.3M	707.9M	848.3M
Video_Games	2.8M	137.2K	4.6M	347.9M	137.3M
Unknown	23.1M	13.2M	63.8M	3.3B	232.8M

# 데이터

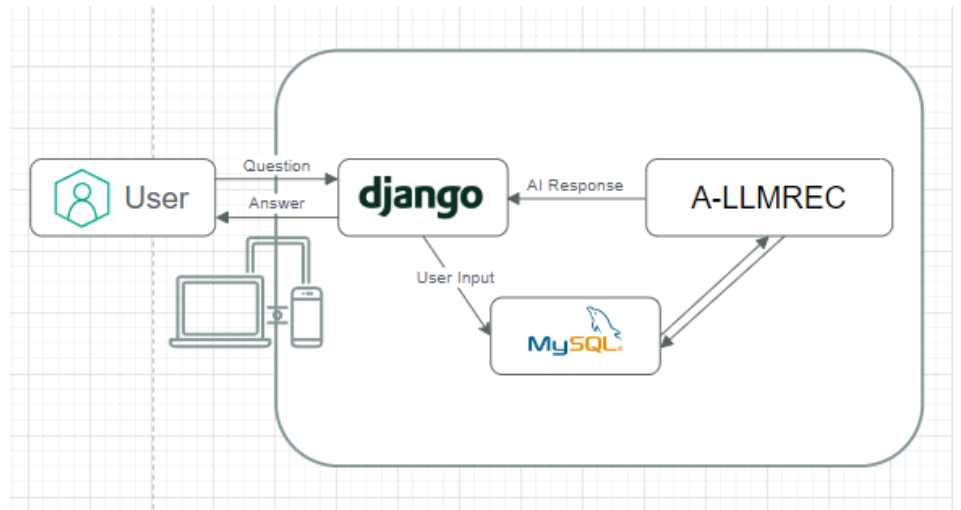
## User Reviews의 데이터 필드

필드	타입	설명
rating	float	평점 1.0-5.0
title	str	리뷰 제목
text	str	리뷰 내용
images	list	유저가 올리는 제품의 실제 이미지 각 이미지 사이즈는 상이하고 (small, medium, large), small_image_url, medium_image_url, large_image_url로 표기
asin	str	제품 ID
parent_asin	str	제품의 상위 ID 다른 색, 스타일, 사이즈들의 제품들은 보통 같은 상위 ID
user_id	str	리뷰어 ID
timestamp	int	리뷰한 시간 (유닉스 시간)
verified_purchase	bool	구매 확인
helpful_vote	int	리뷰의 유용성 투표

## Item Metadata의 데이터 필드

필드	타입	설명
main_category	str	제품의 메인 카테고리
title	str	제품 이름
average_rating	float	제품 페이지에 표기되는 평균 평점
rating_number	int	평점 수
features	list	글머리표 형식으로 된 제품 특징
description	list	제품 설명
price	float	크롤링 당시의 제품 가격 (US \$)
images	list	제품 이미지들 각 이미지의 사이즈가 상이하고 이미지의 위치는 "variant" 필드로 표기
videos	list	제목과 url이 포함되어있는 제품 비디오들
store	str	판매처 이름
categories	list	제품의 계층적 분류
details	dict	제품 상세 설명 (재질, 브랜드, 크기 등)
parent_asin	str	제품의 상위 ID
bought_together	list	웹사이트에서 추천하는 같이사면 좋은 제품들

## 시스템 구성



## 모델링 방안

### 1. 협업 필터링 (Collaborative Filtering)

협업 필터링은 기존 사용자의 행동 패턴(구매 기록, 클릭 등)을 기반으로 유사한 사용자나 상품을 추천하는 전통적인 방식으로 상호작용 데이터가 충분히 쌓였을 때 웹 시나리오에서 매우 높은 성능을 보임

#### 장점

- **웹 시나리오에서 우수한 성능:** 충분한 사용자-상품 상호작용 데이터를 바탕으로 유사한 사용자 간의 패턴을 분석하여 정확한 추천을 제공
- **간단한 구현:** 많은 추천 시스템에서 사용되는 방식으로 검증된 성능을 자랑

#### 단점

- **콜드 스타트 문제:** 새로운 사용자나 상품에 대한 상호작용 데이터가 없을 경우 성능이 급격히 저하
- **텍스트 데이터 미활용:** 상품 설명, 리뷰 등과 같은 모달리티 정보가 활용되지 않아 추천의 다양성과 세부성에서 한계

### 2. 모달리티 어웨어 (Modality-aware)

모달리티 어웨어 방식은 상품 설명, 리뷰, 이미지 등 다양한 모달리티 데이터를 활용하여 콜드 스타트 문제를 해결하고 추천 정확도를 높이는 방식

#### 장점

- **콜드 스타트 문제 해결:** 신규 사용자나 상품에 대한 상호작용 데이터가 부족해도 상품 설명, 리뷰와 같은 텍스트 데이터를 활용해 추천
- **텍스트 및 이미지 정보 활용:** 상품 설명, 리뷰 등 다중 모달리티 데이터를 통합해 추천의 질 향상
- **다양한 정보 출처 통합:** 텍스트 및 이미지 등 다양한 상품 정보를 결합하여 풍부한 추천 결과를 제공

#### 단점

- **상호작용 데이터 부족 시 성능 저하:** 협업 필터링과 결합되지 않으면 상호작용 데이터가 많은 경우 성능이 제한됨
- **모달리티 통합의 복잡성:** 여러 데이터를 결합하는 과정에서 시스템이 복잡해질 수 있고, 이에 따른 연산 비용증가

## 모델링 방안

### 3. LLM-Only (Large Language Model 기반 추천)

LLM-Only 방식은 대형 언어 모델(LLM)을 활용하여 자연어 처리와 텍스트 데이터 분석을 기반으로 추천을 생성하는 모델

사용자의 자연어 입력을 처리하거나 상품 설명을 분석해 추천을 제공하는 방식

#### 장점

- **텍스트 기반 추천:** 상품 설명, 리뷰 등의 텍스트를 분석하여 상품 추천을 생성
- **자연어 처리의 강점:** 사용자 입력을 자연어로 처리하여 직관적인 검색어를 통한 추천
- **콜드 스타트 문제 해결:** 상호작용 데이터가 부족한 경우에도 텍스트 데이터를 활용하여 적합한 상품을 추천

#### 단점

- **상호작용 데이터 활용 부족:** 협업 필터링에 비해 사용자 상호작용 데이터를 직접 활용하지 않아, 데이터가 많은 웹 시나리오에서는 성능 저하
- **파인 튜닝 비용:** LLM을 지속적인 파인 튜닝 필요 시, 더 많은 시간과 비용 소요

### 4. A-LLMRec 모델: 통합적 접근

A-LLMRec 모델은 협업 필터링, 모달리티 어웨어, LLM 기반 추천의 장점을 결합하여 콜드 스타트와 웹 시나리오에서 모두 높은 성능을 제공하는 시스템

협업 필터링에서 생성된 상호작용 데이터와 LLM이 생성한 텍스트 임베딩을 결합하여 최적화된 추천을 제공

#### 장점

- **협업 필터링과 LLM의 결합:** 협업 필터링의 상호작용 데이터와 LLM의 텍스트 데이터를 정렬 네트워크를 통해 결합하여, 더 높은 추천 정확도를 실현
- **콜드 스타트 및 웹 시나리오 모두 처리:** 상호작용 데이터가 부족한 경우, 텍스트 기반 추천을, 데이터가 많은 경우, 협업 필터링을 통해 최적의 성능을 발휘
- **비파인 튜닝 방식:** LLM을 반복적으로 파인 튜닝할 필요 없이 정렬 네트워크만 학습하여 효율적으로 시스템을 유지

**A-LLMRec 모델은 이 세 가지 방식을 결합하여 콜드 스타트와 웹 시나리오 모두에서 높은 성능을 발휘하며, 효율적이고 확장 가능한 추천 시스템을 제공**

## 기능

### 제품 추천 대화형 시스템

#### 제품 추천 기능:

사용자 대화 데이터를 바탕으로 최적화된 제품을 추천

#### 추천 기준 설정(옵션):

가격, 성능, 리뷰, 사용자 맞춤형 옵션 제공

#### 구매 사이트 연결 기능:

선택된 제품의 구매 링크를 제공

#### 가격 기능:

선택된 제품들의 가격정보를 제공

#### 사용자 추가 요구 사항 반영 기능:

대화 중 사용자가 요청한 추가 요구사항을 반영하여 추천을 업데이트

#### 리뷰 제공:

추천 제품에 대한 사용자 리뷰를 제공



---

## UI

---

### 제품 추천 대화형 시스템

#### 1. 챗봇 인터페이스:

- 사용자와의 대화를 위한 직관적인 챗봇 창.
- 사용자와 자연스러운 대화를 위한 챗봇 형태의 인터페이스.
- 대화를 통해 제품을 추천
- 대화를 통한 가격 안내
- 대화를 통한 구매사이트 연결 UI제공

#### 2. 유저 관리

- **대화 저장 페이지:** 사용자가 이전 대화 기록을 볼 수 있는 인터페이스.
- **유저 정보 페이지:** 로그인한 사용자의 프로필을 볼 수 있는 페이지.

#### 3. 네비게이션 바

- **네비게이션 메뉴:** 로그인, 로그아웃, 회원가입 등의 버튼을 포함한 상단 메뉴 바.
- **페이지 이동 버튼:** 주요 페이지로 쉽게 이동할 수 있는 버튼 제공.

#### 4. 로그인 구분 UI:

- 로그인 상태에 따라 다른 UI를 제공.
- 

#### 5. 프로필 페이지:

- 사용자의 프로필 정보를 확인하고 수정할 수 있는 페이지.
-

## 데이터 모델 비교

## 비교 결과 예측

모델	장점	단점	적용 시나리오
협업 필터링	윌 시나리오에서 매우 높은 성능	콜드 스타트 문제 발생	상호작용 데이터가 많은 경우 (윌 시나리오)
모달리티 어웨어	텍스트, 이미지 등을 활용해 콜드 스타트 문제 해결 가능	상호작용 데이터가 많은 경우 성능이 떨어질 수 있음	텍스트 및 이미지 정보가 풍부한 상품 (콜드 스타트 시나리오)
LLM-Only	텍스트 기반으로 자연어 추천 제공, 콜드 스타트 문제 해결	상호작용 데이터를 활용하지 않아 윌 시나리오에서 성능 저하	자연어 입력 기반 추천, 신규 사용자 및 상품 추천

## ALLMREC 결과

### 1. ALLMRec 모델은 협업 필터링과 LLM을 결합한 방식

- 콜드 스타트와 윌 시나리오 모두에서 기존 모델보다 뛰어난 성능을 보일 것이라 예상

### 2. 아마존 데이터셋을 사용


- 대량의 데이터 반영, 개인화된 추천 제공이라는 장점을 통해 파인 튜닝 없이도 우수한 성능을 발휘할 것이라 예상


## 실험 결과:


- **콜드 스타트 시나리오** : LLM과 텍스트 정보를 결합하여, 상호작용 데이터가 부족한 경우에도 정확한 추천을 제공할 것이라 예상
- **윌 시나리오** : 협업 필터링을 통한 상호작용 데이터 기반 추천을 통해 최적의 성능 유지 예상


# 기술 스택


Development Tools (개발 도구)


 PYTHON

 VSCODE


 GOOGLECOLAB


 HTML5


 CSS

 JAVASCRIPT


Collaboration Tools (협업 도구)


 DISCORD

 NOTION


 GOOGLEDRIVE


Version Control (버전 관리)

 GIT


 GITHUB

AI & Machine Learning (인공지능 및 머신러닝)

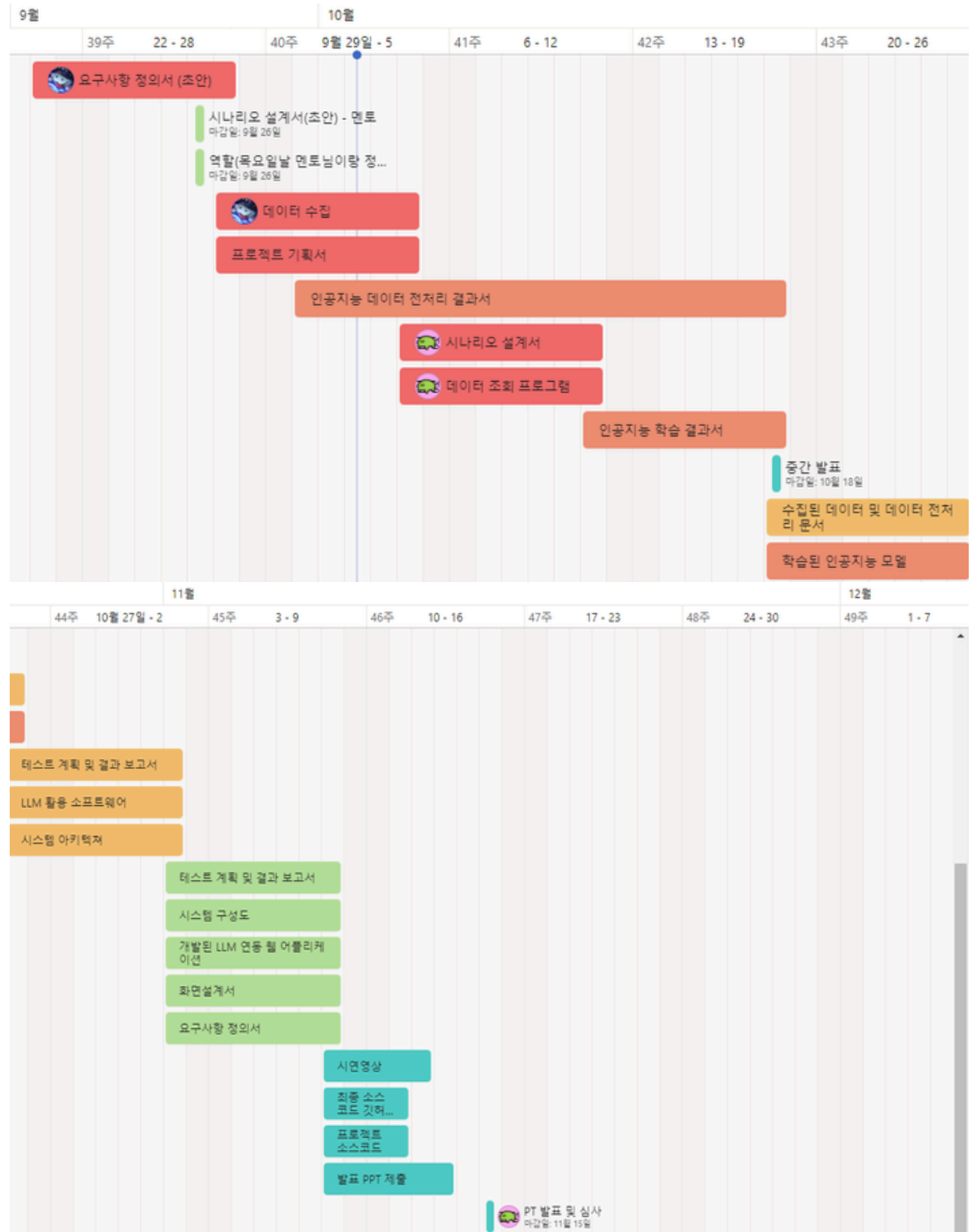
 OPENAI

 LANGCHAIN

Web Development (웹 개발)

 DJANGO

## 일정



## 문서화

### 1. 요구사항 정의서

- 프로젝트의 요구사항을 명확히 정의한 문서.
- 시스템의 주요 기능과 성능 요구사항, 사용자 요구사항에 따른 우선순위 설정.

### 2. 전처리 결과서

- 프로젝트에서 데이터를 본격적으로 활용하기 전에 수행된 데이터 정제 및 전처리 과정에 대한 설명과 결과를 기록

### 3. 인공지능 학습 결과서

- 프로젝트에서 데이터를 본격적으로 활용하기 전에 수행된 데이터 정제 및 전처리 과정에 대한 설명과 결과를 기록
- 모델이 해결하고자 하는 문제를 정의 및 결과

### 2. 기술 문서

- 시스템 아키텍처, 추천 알고리즘 설계, 화면 설계서, 시스템 구성도, 시나리오 설계서
- 예시 문서:
  - 데이터베이스 구조 및 스키마
  - UI/UX 설계 가이드라인

### 3. 사용자 가이드

- 시스템의 사용법 및 주요 기능에 대한 설명서.
- 목표: 일반 사용자가 시스템을 쉽게 이해하고 사용할 수 있도록 지원.

### 4. 테스트 계획서 및 결과 보고서

- 시스템의 기능 및 성능을 검증하기 위한 테스트 절차, 테스트 환경 및 범위 설정. 테스트 결과에 따른 문제점 분석 및 수정 사항 기록.

## 커뮤니케이션

### 도구 :

- Discord , Notion , Asana , Google Drive

### 주기적인 회의: 매주 월/목 (19:00)

- 진행 상황 체크 및 이슈 공유.
- 다음 일정 공유

## R&R

역할	주책임자	부책임자
AI 모델 설계	진정현	서종호
데이터분석 및 전처리	서종호	장정원
웹 프론트	장정원	진정현
웹 백엔드	서종호	사재민
테스트	진정현	사재민
문서	사재민	장정원