

# Konan LLM

Ver. 1.0

Logo

## **Copyright Notice**

서울특별시 서초구 강남대로 327 대륭서초타워 6, 7, 9층

## **저작권**

Copyright Konan Technology Inc. All rights reserved.

이 매뉴얼의 저작권은 (주)코난테크놀로지에 있습니다. 이 매뉴얼은 (주)코난테크놀로지의 허락 없이 전부 혹은 일부를 무단으로 인용, 복사, 전재, 유포 혹은 재사용할 수 없습니다.

## **유지 보수 문의**

제품 유지 보수 및 문의는 고객센터팀으로 문의 바랍니다.

- Tel : 02-3469-8588

- E-Mail : cs@konantech.com

## Table of contents

---

1. Konan LLM	4
2. Ver 1.0	5
2.1 개요	5
2.2 모델 관리	6
2.3 프롬프트 갤러리	7
2.4 추가 학습	8
2.5 미세 조정	13
2.6 플레이 그라운드	20
2.7 프롬프트 설계	21
2.8 Chapter 8. 검색증강생성(RAG)	29
3. Ver 2.0	32
3.1 개요	32
3.2 모델 관리	33
3.3 프롬프트 갤러리	35
3.4 추가 학습	36
3.5 미세 조정	42
3.6 플레이 그라운드	49
3.7 프롬프트 설계	50
3.8 Chapter 8. 검색증강생성(RAG)	59

# 1. Konan LLM

---

**Copyright Notice** 서울특별시 서초구 강남대로 327 대륭서초타워 6, 7, 9층

**저작권** Copyright Konan Technology Inc. All rights reserved. 이 매뉴얼의 저작권은 (주)코난테크놀로지에 있습니다. 이 매뉴얼은 (주)코난테크놀로지의 허락 없이 전부 혹은 일부를 무단으로 인용, 복사, 전재, 유포 혹은 재사용할 수 없습니다.

**유지 보수 문의** 제품 유지 보수 및 문의는 고객센터시스템으로 문의 바랍니다. - Tel : 02-3469-8588 - E-Mail : cs@konantech.com

## 2. Ver 1.0

---

### 2.1 개요

---

LLM 스튜디오는 생성형 언어 모델 코난 LLM을 빠르게 프로토타이핑하고 테스트 해볼 수 있는 도구입니다. 자신만의 프롬프트를 설계하여 테스트하고 공유할 수 있습니다. 또한 기본 모델을 미세 조정하여 원하는 응용 분야에 적용할 수 있습니다. LLM 스튜디오를 통해 다음과 같은 작업들을 수행할 수 있습니다.

- 모델 관리
- 프롬프트 갤러리
- 모델 학습
- 플레이그라운드

#### 2.1.1 주요 개념

---

##### 텍스트 생성 모델

텍스트 생성 모델은 자연어 처리 분야에서 인공지능을 접목하여 자연스러운 문장을 만들어 내도록 학습한 모델입니다.

사전학습한 데이터의 양과 품질, 다루고 있는 영역에 따라 텍스트 생성 모델의 잠재력이 결정되며, 미세조정을 통해서 사람의 지시를 정확하게 따를 수 있는 능력이 발현되기 시작합니다.

이에 따라 텍스트 생성 모델은 크게 “사전학습 모델”과 “미세조정 모델”로 나누어 집니다. 사용자의 목적, 적용 범주에 따라 사전학습 모델에 전문 지식 분야에 대한 추가학습을 진행하기도 하고, 미세조정 모델에 대한 추가적인 미세조정을 통해 보다 원하는 형식의 텍스트 생성이란 목표를 달성할 수 있습니다.

##### 임베딩

임베딩이란 고차원 벡터를 변환할 수 있는 비교적 저차원의 공간입니다. 주로 언어 모델에서는 글자, 단어, 문장으로 확대되면서 복잡해지는 언어를 적절한 크기의 분절로 나눈 뒤, 이를 일정한 크기의 실수 벡터 형태로 변환시켜 사용합니다. 이러한 임베딩은 다양한 수준(단어, 문장 등)에서 활용 가능하고, 유사한 문장 찾기, 검색, 주요 주제 군집화, 이상 데이터 탐지와 같은 분야에서 사용할 수 있습니다.

##### 토큰

언어 모델에서의 토큰은 자연어 처리를 위해 여러 문자를 일정한 기준에 따라 나눈 것을 의미합니다. 목적에 따라 하나의 토큰은 자소 단위에서부터 문장 단위까지 확장될 수도 있습니다. 토큰 기준으로 임베딩이 구성되기 때문에 저차원에서도 적절한 정보 수준을 유지할 수 있도록 언어학적, 통계학적 지식을 이용하게 되며, 대개 한국어 단어는 형태소 단위로 토큰이 나누어지게 됩니다. 또한, 이러한 토큰은 실제로 언어 모델이 처리하는 단위로서 한번에 처리 가능한 최대 정보량을 나타내는 단위가 되기도 합니다.

## 2.2 모델 관리

LLM 스튜디오를 통해 다양한 기능을 가진 여러 모델들을 관리할 수 있습니다. 또한 미세 조정을 통해 필요한 용도에 맞게 모델을 새롭게 정의할 수 있습니다.

### 모델 유형

모델 관리 페이지에서 확인할 수 있는 모델들의 유형은 다음과 같습니다.

유형	설명
기본모델	LLM스튜디오를 통해 다양한 업무를 위해 미세 조정할 수 있는 사전 학습된 모델
미세조정모델	미세조정을 거쳐 업무에 적용할 수 있는 모델

### 2.2.1 모델 목록

모델 관리 메뉴를 클릭하면 LLM 스튜디오에서 접근할 수 있는 모델들의 목록을 확인할 수 있습니다.

#### 기본 모델

모델명	설명	속성
kylin-13b-2308	자연어를 이해하고 생성할 수 있는 파운데이션 모델	2K
kylin-13b-inst-dpo-2312	kylin-13b모델에 대한 인스트럭션 튜닝 모델	4K
kylin-embedding-2310	텍스트 입력에 대한 임베딩 벡터 반환	
kylin-safety-2312	유해성 텍스트 분류 모델	

#### 미세 조정 모델

모델명	설명
kylin-13b-2308-sft-2310	kylin-13b모델에 대한 미세 조정 모델

### 2.2.2 상태

상태	설명
실행 중	모델을 로드한 상태입니다. * 프롬프트 템플릿이나 플레이그라운드 페이지에서 모델을 실행하기 위해서는 먼저 모델을 로드해야 합니다.* 모델 학습 시CPU, GPU,메모리 등 자원이 부족할 경우에는 자원 확보를 위해[언로드]버튼을 클릭하여 모델을 언로드합니다.
학습 완료	학습이 완료된 상태이며[로드]버튼을 클릭하여 모델을 활성화할 수 있습니다.
학습 중	학습을 진행 중인 상태입니다.현재 학습 진행율이 함께 표시됩니다. [중지]버튼을 클릭하여 학습을 중단할 수 있습니다. * 학습 진행 중인 모델만 중단할 수 있습니다.* 학습 중인 모델이 있을 경우 모델 관리를 제외한 다른 메뉴들은 비활성화됩니다.
학습 취소	학습을 중지한 상태입니다. [재시작]버튼을 클릭하여 학습을 다시 시작할 수 있습니다.
오류	학습 중에 실패한 상태입니다.

## 2.3 프롬프트 갤러리

---

LLM 스튜디오의 프롬프트 템플릿 기능을 이용하여 문서 작성에 유용한 프롬프트를 설계할 수 있습니다.

프롬프트 설계는 크게 아래와 같은 단계로 수행합니다.

1. 프롬프트 작성
2. 변수 추출

### 2.3.1 프롬프트 작성

---

프롬프트는 응답을 받기 위해 언어모델에게 제출하는 자연어 요청입니다. 원하는 결과가 나올 때까지 프롬프트를 주의 깊게 작성하고 파라미터들을 조정합니다. 프롬프트 작성을 위한 지침은 [프롬프트 설계](#)를 참조하세요.

### 2.3.2 변수 추출

---

변수는 전체 프롬프트 중 사용자로부터 입력 받을 내용을 의미합니다. 테스트가 끝난 프롬프트에서 사용자에게 직접 입력 받을 영역을 선택하여 변수로 추가합니다. 템플릿 양식을 통해 사용자로부터 입력 받은 변수를 프롬프트에서 치환한 결과가 모델에게 전달되어 결과 텍스트를 생성하게 됩니다. 각각의 변수는 다음과 같은 속성을 가집니다.

1. 레이블: 템플릿 양식의 레이블로 사용됩니다.
2. 입력 예시: 사용자들이 어떤 식으로 작성해야 하는지 참조할 수 있도록 입력창의 초기값으로 활용됩니다.
3. 입력 유형: 싱글 라인, 멀티 라인, 선택 버튼 중 알맞은 유형을 선택합니다.

변수 설정이 끝나면 [템플릿] 버튼을 클릭하여 실제 사용자에게 표시될 양식을 확인할 수 있습니다.

## 2.4 추가 학습

도메인 지식을 모델에 주입하는 방법을 알아보세요.

### 2.4.1 개요

코난 LLM은 방대한 양의 텍스트로 학습되어서 언어를 이해하고 생성하는 능력이 뛰어납니다. 또한 다양한 지식을 가지고 있습니다. 그러나 특정 조직에서 다루는 도메인에 대한 지식은 부족할 수 있습니다. LLM 스튜디오는 코난 LLM의 언어 능력과 기존 지식은 유지하면서 해당 조직의 도메인 지식을 주입할 수 있는 "추가 학습" 기능을 제공합니다. 추가 학습을 통해서 특정 도메인에 최적화된 LLM을 만들 수 있습니다.

### 2.4.2 추가 학습이 필요한 경우

코난 LLM을 추가 학습 시키면 언어 모델에 도메인 지식을 주입시킬 수 있지만 많은 시간과 노력이 필요한 작업이므로 신중하게 결정해야 합니다. 다음과 같은 질문들에 대해 고민해 보고 추가 학습을 할 지 RAG를 적용할 지 결정하는 것이 바람직합니다.

**도메인 지식의 변경이 예상 되는가?**

추가 학습된 모델은 특정 시점까지의 데이터만 학습이 됩니다. 도메인 지식의 변경이 예상될 경우 RAG가 더 효율적인 방법입니다. Vector Store를 갱신하는 일이 모델을 추가 학습하는 것보다 훨씬 간편한 일이기 때문입니다.

**학습 데이터를 얼마나 확보할 수 있는가?**

추가 학습된 모델의 품질은 도메인 데이터의 품질과 양에 달려 있습니다. 반면 RAG는 학습 데이터 품질과 양과는 독립적입니다. 충분한 양의 고품질 데이터를 확보할 수 있을 경우에만 추가 학습을 고려하세요.

**충분한 학습 시간과 GPU 자원을 확보할 수 있는가?**

추가 학습을 위해서는 충분한 GPU 자원이 확보되어야 합니다. 13B 모델의 경우 최소 320GB의 GPU 메모리가 필요합니다.

### 2.4.3 추가 학습 절차

**데이터 세트 준비**

추가 학습이 필요하다고 판단한 후에는 모델 학습을 위한 데이터를 준비해야 합니다.

**데이터 수집**

학습에 사용할 문서 파일들을 학습 서버의 학습 데이터 폴더에 저장합니다.

텍스트 포맷 뿐만 아니라 Microsoft Word, PDF, 아래한글 등 대부분의 문서 포맷을 지원합니다.

**데이터 정제**

수집한 문서 파일들로부터 텍스트를 추출하고 정제합니다. (자세한 사항은 [데이터 정제](#)를 참고하세요.)



## 추가 학습 실행

### LLM 스튜디오

시작하기 페이지에서 **모델 학습** 탭을 클릭합니다.

**학습 방법 선택** 페이지에서 **추가 학습** 라디오 버튼을 선택합니다.

**계속**을 클릭합니다.

**모델 설정** 페이지에서 추가 학습을 위한 옵션을 설정합니다. (추가 학습 옵션에 대한 자세한 설명은 부록을 확인하세요.)

#### • 모델 기본 정보 설정

- 모델 이름: 추가 학습된 모델의 이름을 입력합니다.
- 기본 모델: 추가 학습하려는 기본 모델을 선택합니다.

#### • UI를 통한 설정

- 학습 반복: 전체 데이터 세트를 처리하는 횟수
- 배치 크기: 한 스텝 당 처리하는 데이터 수
- 학습률: 학습 중 매개변수를 변경하는 속도

**계속**을 클릭합니다.

**학습 데이터 설정** 페이지에서 추가 학습할 문서들이 저장된 폴더를 설정합니다.

**학습 시작**을 클릭합니다.

모델 학습 상태는 모델 관리 페이지에서 확인할 수 있습니다.

## 추가 학습 모델 평가

미리 구축한 도메인의 평가 데이터를 이용하여 추가 학습 전의 언어모델과 추가 학습 후의 언어 모델을 비교 평가합니다.

### 평가 데이터 구축

추가 학습 데이터의 도메인 관련 단어들을 포함한 문장들로 평가 데이터를 구축합니다.

### 평가 예시

- 위키(한국어)데이터 학습 전 생성 결과

입력오랜하이머는  
결과친구들과 함께 여행을 떠났습니다해린은 여러 곳에서 사진을 찍어 왔는데 그 중에서도 흥미롭게 생긴 사람들을 찾아보려고 합니다

- 위키(한국어)데이터 학습 후 생성 결과

입력오랜하이머는  
결과차 세계대전 당시 핵무기를 개발한 맨해튼 프로젝트의 총책임자였다는년월일 오전시분초뉴멕시코주 앨라모고도 사막에서 인류 최초의 핵실험인트리니티 테스트를 성공적으로 이끌었다이 실험으로 일본 히로시마와 나가사키에 투하된 핵폭탄이 만들어졌다

## 2.4.4 데이터 정제

언어모델의 생성 품질은 학습 데이터의 질에 깊이 연관되어 있습니다. 데이터 정제 과정은 원본 문서 파일로부터 텍스트를 추출하고, 노이즈를 제거하여 언어 모델 학습을 위한 고품질 데이터셋을 구축하는데 중요합니다.

아래와 같은 노이즈들을 제거합니다. (괄호는 대응되는 정제 방법)

- 의미가 없는 특수 문자 (규칙기반, 딥러닝 정제)
- 학습 데이터로서 가치가 없는 단순한 텍스트 (규칙기반, 딥러닝 정제)
- 중복 텍스트 (중복제거)
- 개인정보 (규칙기반 정제)

#### 소요시간

- 텍스트 추출: 파일당 ~1초 소요. 표 추출 활성화 시 파일당 ~10분까지도 소요될 수 있음.
- 규칙기반 정제: 초당 ~200건 처리 (건당 글자수 ~3000자)
- 중복제거: 초당 ~500건 처리 (건당 글자수 ~3000, 200만건 기준 메모리 1GB 점유)
- 딥러닝 정제: 초당 ~100건 처리 (건 당 글자수 ~3000, RTX 3090 기준)

#### 설치 및 실행

##### 설치

제공받은 소스 (data-cleanser.zip)의 압축을 해제한 후 다음 명령어를 실행합니다.

소요 시간약분

##### 삭제

##### 실행

#### 옵션 설명:

- `--src_dir`: 데이터의 소스 디렉터리 경로입니다.
- `--dst_dir`: 처리된 데이터를 저장할 목적지 디렉터리 경로입니다.
- `--n_lines`: 데이터 파일에 저장할 최대 행(row)수를 지정합니다. (파일 1개당 1행)
- `--max_workers`: 동시에 실행할 worker 갯수를 지정합니다.
- `--gpu`: 사용할 GPU를 지정합니다. (지정 시 딥러닝 정제 활성화)
- `--rulebase`: 규칙기반 정제를 활성화합니다.
- `--deduplication`: 중복 문서를 제거합니다.
- `--protect pii`: 개인정보를 비식별화 합니다.
- `--extract_table`: 문서 내 표를 추출합니다. (hwp, hwpX 사용 불가)

#### 데이터 정제 예시

##### 규칙 기반 정제 – 공백 및 개행 처리

- 정제 전

여분의 공백을 사이에 둔 예시 텍스트  
의미 없는 단순텍스트제거

개행 정리  
예시

예시 끝

#### • 정제 후

여분의 공백을 사이에 둔 예시 텍스트  
의미 없는 단순텍스트제거

개행 정리  
예시

예시 끝

#### 규칙 기반 정제 - 개인정보 비식별화

#### • 정제 전

이름아무개  
주민등록번호  
집전화  
핸드폰  
계좌번호우리은행  
카드번호  
이메일

#### • 정제 후

이름아무개  
주민등록번호  
집전화  
핸드폰  
계좌번호  
카드번호  
이메일

#### 딥러닝 기반 정제

#### • 단순 광고 제거

캐시닥 지니어트 타임스프레드 용돈퀴즈 정답  
애플크추천인친구초대코드캐시닥캐시워크지니어트타임스프레드퀴즈정답마일벌스체리포인트쿠팡위메프  
마일벌스숫자  
추천입력시 가입자혜택은 없지만 입력해주시면 정말 감사드려요꾸북매일출책원광고보면원습게모아짐상품권교환가능  
포인트문스터영어  
추천인원출책원원설문미선수행상품권포인트 습게모아짐혜자앱  
캐시존영어  
초대코드 입력시원매일 출석체크를렛원원짱은광고보고캐시적립출금가능

#### • 문서 변환 시 깨진 텍스트 제거

### • 의미없는 텍스트 제거

나서 방식이 뿐이다한대 그랜저나 가르침들은 현대 그랜저하는데 싹 강조에서도 나들이가 기아 올뉴카니발이므로 그만큼 양파이 빼네기아 레이를 부잡집답게 여우때문에 농사일되어 어쨌던 디자이너은 관련되네운전기사부터 불보 만 간판하는데 손수 벤츠까지 신용회복중고차할부에 의해서 식사에서도 그것은 부족중 낮아지다쇄보레 크루즈되어 조각하는데 군미필중고차전액할부에 의한 아우디 할때 및 쌍용 액티언스포츠팀 소유하네국산되어 등록금되어 와이셔츠하는데 어서 신규하여 범인이나 분홍색했을경우 그렇지만 르노삼성 뉴중 결정되네 품질을 육상까지 시작보다도 또한 주요이나 가수할때 르노삼성 노바에 의한 그리고나서 아우디 뉴은 뉴네이론적이고 철학자까지 닛산 알티마하며 영어에서도 벤츠 에서 학습하여 그리고 시계으로 구성하다우정에서도 해결으로 무작자중고차전액할부는 단순히 벤츠되어 기아 올뉴쏘렌토들은 신용불량할부에는 비교적 친척이 얘기하다점심때에는 만일하면 평소부터 없이 신용불량자중고차할부를 그래서 통화중 귀국하네요

### 중복 제거

완전히 일치하지 않더라도 중복되는 영역이 대다수일 경우 제거 대상이 됩니다.

### • 중복 문서 제거

쇼킹나이트한 참가자 대거 등장  
쇼킹나이트쇼킹나이트참가자들이 화려한 이력과 무대로 관심을 모으고 있다지난달일 에서 방송한 국내 유일 팬스 가요제쇼킹나이트제작 스페이스라잇앨런미디어하쇼킹나이트는 전 아이돌 출신 멤버부터 유명 인플루언서까지찻한 참가자가 대거 출연했다이날 방송에는 과거 톨라 백업 댄서이자 그룹 팀 으로 활동했던 낙타가 출연했다낙타는내 맘 깊은 곳에 넌 사는걸이라며넌 발표된 그룹 팀의 대표곡별한 소절을 불렀고들은 단박에 알아차리고 다 같이 노래를 따라 불렀다

년대 초반 가수의 노래로 무대를 꾸었다는 점에서 새롭게 를 표현하는 그룹의 등장을 기대하게 만들었다탁재훈은쇼킹나이트가 발견한 보석이라고 극찬했으며이특 또한보아 어릴 때 같다며 놀라움을 감추지 못했다이처럼 그동안 에서 쉽게 볼 수 없었던 과거 인물들을 소환하고 찻한 인물들을 새롭게 소개하고 있는쇼킹나이트는 앞으로도 참가자들의 다양한 매력을 계속해서 보여줄 예정이다 한편쇼킹나이트는 매주 금요일 오후시분 에서 방송한다

쇼킹나이트전 아이돌부터 유명 인플루언서까지참가자 이력눈길  
뉴스디지털부사전쇼킹나이트쇼킹나이트의 특별한 참가자들이 눈길을 사로잡았다지난달일 에서 방송한 국내 유일 팬스 가요제쇼킹나이트제작 스페이스라잇앨런미디어하쇼킹나이트는 전 아이돌 출신 멤버부터 유명 인플루언서까지찻한 참가자가 대거 출연했다이날 방송에는 과거 톨라 백업 댄서이자 그룹 팀으로 활동했던 낙타가 출연해 모두를 깜짝 놀라게 만들었다낙타는내 맘 깊은 곳에 넌 사는걸이라며넌 발표된 그룹 팀의 대표곡별한 소절을 불렀고들은 단박에 알아차리고 다 같이 노래를 따라 불렀다

년대 초반 가수의 노래로 무대를 꾸었다는 점에서 새롭게 를 표현하는 그룹의 등장을 기대하게 만들었다탁재훈은쇼킹나이트가 발견한 보석이라고 극찬했으며이특 또한보아 어릴 때 같다며 놀라움을 감추지 못했다이처럼 그동안 에서 쉽게 볼 수 없었던 과거 인물들을 소환하고 찻한 인물들을 새롭게 소개하고 있는쇼킹나이트는 앞으로도 참가자들의 다양한 매력을 계속해서 보여줄 예정이다 한편쇼킹나이트는 매주 금요일 오후시분 에서 방송한다뉴스디지털부

## 2.5 미세 조정

작업 목적에 맞게 모델을 조정하는 방법을 알아보세요.

### 2.5.1 개요

미세 조정을 통해 모델이 학습하지 못했던 작업에 대해서 더 나은 결과를 얻을 수 있게 해줍니다. 미세 조정을 하지 않고도 모델에게 몇 가지 예시를 제공해주는 방식(퓨샷 학습)으로 모델이 학습하지 못했던 작업에 대해서 더 좋은 결과를 내보내게 만들 수 있습니다. 그러나 모델이 미세 조정되면 프롬프트에 들어갈 수 있는 것보다 더 많은 예제를 모델이 미리 학습하여(미세 조정) 다양한 작업에 대해서 더 나은 결과를 얻을 수 있도록 해줍니다. 또한 프롬프트에 많은 예시를 사전에 제공할 필요가 없어서 비용 절감과 요청 시간의 단축 효과를 얻을 수 있습니다.

미세 조정을 통해 다음과 같은 효과를 얻을 수 있습니다.

1. 프롬프트 보다 더 높은 품질의 결과
2. 짧은 프롬프트로 토큰 절약과 응답 시간 단축

### 2.5.2 미세 조정이 필요한 경우

코난 LLM을 미세 조정하면 특정 작업에 맞는 더 나은 모델을 만들 수 있지만 시간과 노력이 필요한 작업이므로 신중한 투자가 필요합니다. 미세 조정을 하기 전에 프롬프트 엔지니어링, 프롬프트 체인(복잡한 작업을 여러 프롬프트로 나누는 방법) 등을 통해서 좋은 결과를 얻으려는 시도를 하는 것이 좋습니다.

- 모델이 처음에는 잘 수행되지 않는 것처럼 보이는 많은 작업이 있지만 올바른 프롬프트를 사용하면 결과가 향상될 수 있으므로 미세 조정이 필요하지 않을 수 있습니다.
- 학습 데이터 세트를 생성하고, 모델 학습을 진행해야 하는 미세 조정으로 반복하는 것 보다, 프롬프트를 수정하는 것이 피드백이 훨씬 빠릅니다.
- 미세 조정이 필요한 경우 초기 프롬프트 엔지니어링 작업을 활용할 수 있습니다. 일반적으로 미세 조정 데이터에 좋은 프롬프트를 사용할 때 최상의 결과를 얻을 수 있습니다.

### 2.5.3 미세 조정 사례

미세 조정을 통해 결과를 개선할 수 있는 몇 가지 사례는 다음과 같습니다.

- 스타일, 톤, 형식 또는 기타 질적 측면 설정
- 문서 초안 작성
- 요약
- 대화
- 원하는 출력을 생산할 때 신뢰성 향상
- 질의/응답
- 복잡한 프롬프트를 모델이 따르지 못하는 경우
- 참고 문서 기반 질의/응답
- 참고 문서 기반 문서 초안 작성
- 고정된 방식으로 많은 옛지 케이스 처리
- 번역
- 분류
- 프롬프트에서 명확하게 표현하기 어려운 새로운 기술이나 작업 수행

대부분의 작업에 대해서 사이즈가 큰 모델이 사이즈가 작은 모델 보다 더 나은 성능을 보여주지만, 미세 조정된 사이즈가 작은 모델을 통해서 사이즈가 큰 모델과 비슷한 품질 혹은 더 나은 품질에 도달할 수 있는 경우가 많습니다.

## 2.5.4 미세 조정 절차

미세 조정은 크게 아래와 같은 단계로 수행합니다.

1. 데이터 세트 준비
2. 미세 조정 모델 만들기
3. 미세 조정 모델 사용

### 데이터 세트 준비

미세 조정이 필요하다고 판단되면 모델 학습을 위한 데이터 세트를 준비해야 합니다. 실제 사용자들이 사용할 프롬프트와 유사하게 다양한 대화 데이터 세트를 만듭니다.

#### 데이터 형식

학습 데이터는 각 라인이 하나의 대화 셋을 의미하는 JSONL 혹은 C\_ID로 대화 셋을 구분하는 CSV 형식을 사용할 수 있습니다.

##### • JSONL 형식

##### • CSV 형식

각 예제는 모델에 대한 사용자의 지시가 포함된 user의 content와, 모델이 생성하기를 기대하는 답변이 포함된 assistant의 content로 구성됩니다.

#### 프롬프트 작성하기

[미세 조정 예시](#)를 참고하세요.

#### 권장 예제 개수

- 미세 조정을 위해서는 최소 200개 이상의 학습 데이터가 필요합니다.
- 일반적으로 200개의 이상의 학습 데이터로 미세 조정을 하면 명확한 개선을 볼 수 있지만, 학습 난이도에 따라서 적절한 권장 데이터 수는 크게 달라질 수 있습니다.
- 잘 만들어진 200개의 학습 데이터로 시작하여, 모델의 성능을 평가하면서, 추가 데이터 수집을 고려하는 것이 좋습니다.
- 모델의 개선이 있을 경우, 원하는 품질이 나올 때까지 학습 데이터의 수를 동일한 방법으로 늘려갈 수 있습니다.
- 모델의 개선이 없을 경우, 개선이 나올 때까지 학습 데이터의 수를 늘려갈 수도 있고, 데이터를 재구성하는 것을 검토할 필요가 있습니다.

**토큰 길이 제한**

각 학습 데이터는 모델마다 적용된 토큰 길이 제한을 가집니다. 이보다 긴 데이터는 모델마다 적용된 토큰 길이까지만 잘라서 사용합니다.

**미세 조정 모델 만들기**

충분한 양의 학습 데이터셋을 업로드하고 나면 모델을 미세 조정할 수 있습니다. 미세 조정은 LLM 스튜디오 UI를 통해서 하거나 REST API를 호출하여 사용할 수 있습니다.

미세 조정은 크게 효율적인 미세 조정(Parameter-Efficient Fine-Tuning)과 일반적인 미세 조정(Full Fine-Tuning)으로 나뉩니다. 효율적인 미세 조정은 모델의 특정 가중치만 학습시키는 방법으로 일반적인 미세 조정보다 적은 장비로 모델을 학습할 수 있습니다.

**LLM 스튜디오**

시작하기 페이지에서 **모델 학습** 탭을 클릭합니다.

**학습 방법 선택** 페이지에서 **미세 조정** 라디오 버튼을 선택합니다.

**계속**을 클릭합니다.

**모델 설정** 페이지에서 미세 조정을 위한 옵션을 설정합니다. (미세 조정 옵션에 대한 자세한 설명은 [미세조정 API](#)를 확인하세요.)

**• 모델 기본 정보 설정**

- 모델 이름: 미세 조정된 모델의 이름을 입력합니다.
- 기본 모델: 미세 조정하려는 기본 모델을 선택합니다.

**• UI를 통한 설정**

- 학습 반복: 전체 데이터 세트를 처리하는 횟수
- 배치 크기: 한 스텝 당 처리하는 데이터 수
- 학습률: 학습 중 매개변수를 변경하는 속도

**계속**을 클릭합니다.

**학습 데이터 설정** 페이지에서 미세 조정에 사용할 데이터셋을 설정합니다.

- JSONL 파일을 업로드: 로컬 PC에서 업로드할 데이터셋 파일을 선택합니다.
- URL로부터 JSONL 파일을 업로드: 데이터셋 파일을 다운로드 받을 수 있는 URL을 입력합니다.

**조정 시작**을 클릭합니다.

모델 학습 상태는 모델 관리 페이지에서 확인할 수 있습니다.

**REST API**

REST API를 통해 미세 조정 작업을 시작하기 위해서는 아래와 같이 요청을 보냅니다.

미세 조정된 모델을 적용하기 위해서는 작업이 완료될 때까지 기다려야 하며 데이터세트 크기에 따라 몇 분 혹은 몇 시간이 걸릴 수 있습니다.

#### 미세 조정 모델 사용

미세 조정 작업이 완료되면 모델 상태가 **학습 완료**로 표시됩니다. 학습 완료 모델은 모델 관리 페이지에서 [로드] 버튼을 클릭하여 로드한 후 플레이그라운드에서 사용해볼 수 있습니다.

#### 미세 조정 모델 평가

완료된 모델을 직접 사용하면서, 모델의 성능을 평가하고, 다음과 같은 방법을 통해서 모델의 성능을 향상시킬 수 있습니다.

##### 학습 데이터의 품질을 높여가며 반복

- 생성 품질이 낮은 사례를 조사하세요.
- 해당 사례를 올바르게 수행하는 방법을 학습 데이터에 추가해주세요.
- 모델에 문법, 논리 또는 스타일 문제가 있는 경우, 학습 데이터에 동일한 문제가 있는 확인한 후 수정해주세요.
- 학습 데이터의 일관성을 확인하세요.
- 다양한 사람이 학습 데이터를 생성해, 학습 데이터가 일관되지 않으면, 모델은 일관되지 않은 답변을 할 수 있습니다.
- 학습 데이터의 형식과 추론을 위해 사용하는 형식이 동일한지 확인하세요.
- 학습 데이터의 균형과 다양성을 확인하세요.
- 학습 데이터의 비율 중 약 60%가 “답변을 할 수 없습니다”이고, 원하는 결과가 생성의 5%만 “답변을 할 수 없습니다”라고 답변해야 한다면, 원하는 결과 보다 더 과도하게 “답변을 할 수 없습니다”라고 모델이 답변할 수 있습니다.

##### 학습 데이터의 수를 늘려가며 반복

- 생성 품질이 만족스러우면, 학습 데이터의 수를 늘리는 것을 고려할 수 있습니다.
- 다음을 통해서 학습 데이터의 수를 늘려 예상되는 품질 향상을 대략적으로 추정할 수 있습니다.
- 현재 학습 데이터 세트로 미세 조정
- 현재 학습 데이터 세트의 절반으로 미세 조정
- 둘 사이의 품질 격차를 관찰
- 일반적으로 저품질 학습 데이터를 많이 사용하는 것보다, 고품질 학습 데이터를 적게 사용하는 것이 더 효과적입니다.

##### 모델 학습에 사용하는 하이퍼 파라미터를 조절하며 반복

- 처음에는 모델 학습에 사용하는 하이퍼 파라미터의 기본 값을 사용하는 것이 좋습니다.
- 만약 모델의 생성 결과가 훈련 데이터의 양상과 다르다고 판단된다면, Epoch을 1 또는 2만큼 늘립니다.
- 만약 모델의 생성 결과의 다양성이 떨어진다고 판단된다면, Epoch을 1 또는 2만큼 줄입니다.
- 만약 모델의 Loss가 수렴하지 않는다면, 학습률을 지수로 높입니다. (ex, 0.002 -> 0.02)

## 2.5.5 미세 조정 예시

해당 예시는 각 작업에 맞는 최적의 프롬프트는 아닐 수 있으므로, 프롬프트 엔지니어링을 통해서 최적의 프롬프트를 찾는 것을 권장합니다.



## 문서 초안 작성

중고차 매매사이트에서사용자에게 더욱 다양한 정보 제공을 위한 피드백 기능을 추가할 예정입니다피드백 기능에 대한 계획서 초안을 작성해주세요

개요

중고차 매매사이트에서 피드백 기능을 추가하여 사용자에게 더욱 다양한 정보를 제공하고사용자들 간의 의견 교류와 신뢰도 향상을 목표로 함

기능

중고차 매물의 설명가격성능 등에 대한 피드백을 작성할 수 있는 기능 제공

다른 사용자들의 작성한 피드백을 볼 수 있고좋아요싫어요 등의 평가도 가능하도록 함

작성한 피드백이 매물 정보와 함께 노출되어사용자들이 정보를 한 눈에 볼 수 있도록 함

피드백 작성자에 대한 평가 제공하여신뢰도 높은 게시물에 대해 우선적으로 노출되도록 함

디자인

매물 페이지의 하단에 피드백 창을 추가하여 작성할 수 있도록 함

작성된 피드백은 리스트 형태로 노출되며관련 정보와 함께 노출되도록 함

좋아요싫어요 등의 평가 기능은 자연스럽게 노출되도록 함

보안

사용자가 작성한 피드백이 비방욕설 등 부적절한 내용을 포함할 경우관리자가 철저하게 검토 후 삭제 조치를 취함

작성자의 정보 및 주소 등을 수집하여악의적인 피드백 작성을 방지함

개발

언어

개발 기간약주디자인 포함

기대효과

사용자들의 다양한 의견 제공으로매물 정보의 신뢰도 및 정확도 향상에 기여

다양한 정보를 제공하는 중고차 매매사이트로 성장함으로써사용자 수 및 매출 증대에 기여

- User의 content는 문서를 작성하기 위해서 사용자가 입력하는 지시문의 형태입니다.
- Assistant의 content는 사용자가 입력하는 지시문에 맞춰 모델이 생성해야 하는 문서의 형태입니다.

## 질의/응답

장례를 주관할 수 있는 사람의 범위가 어떻게 되나요

장례는 배우자자녀 또는 부모 등 법령에서 정하는 연고자가 주관할 수 있습니다

- User의 content는 답변을 위해서 사용자가 입력하는 질문의 형태입니다.
- Assistant의 content는 사용자가 입력하는 질문에 맞춰 모델이 생성해야 하는 답변의 형태입니다.

## 요약

다음 문서를 요약해주세요

문서응급상황대처가이드라인

일반사항

목적

응급상황에서의 대처능력을 향상시키기 위하여 응급상황별 대처 방법 및 비상연락망 등을 수록한 가이드라인을 제작한다

법적 구속력

이 가이드라인은 법적 구속력을 가진 지침은 아니며다만 응급상황에서의 적절한 조치를 취하는데 도움을 주기 위한 것이다

응급상황의 분류

응급상황은 심각도에 따라 긴급한 경우생명 위협심각도가 높은 경우생명위협이 아닌 경우생명에 대한 위협으로 분류된 경우신속한 조치가 필요한 경우로 분류한다

긴급구조 및 응급조치

구조구급 기본정신

응급상황은 생명을 구조하는 일이라는 슬로건으로 언제나 구조 및 응급조치를 위해 필요한 경우라면 누구라도 먼저 도움을 요청할 수 있는 여건을 조성하는데 최선을 다한다

구조구급 요청

환자가 의식을 잃고 쓰러져 있는 등의 긴급한 상황이 발생한 경우에 구조를 요청한다

환자가 언어장애가 있거나 의사소통이 원활하지 않은 경우전화수화를 통한 의사소통을 우선적으로 시도한다

요청내용이 진료와 관련된 것인 경우의료기관의 안내를 우선적으로 시도한다

기타 긴급한 경우에는 관련법규에 따라 조치한다

구조구급 요청시 협조사항

범죄현장을 발견한 경우경찰서에 신고한다

화재 등 재해발생시소방서에 신고한다

심폐소생술 또는 응급처치가 필요한 경우가까운 의료기관이나 응급구조대에 신고한다

기타 긴급한 경우에는 관련 법규에 따라 조치한다

응급상황시 의료지도

의료기관의 휴무일 또는 진료시간이 아니어서 응급진료가 어려운 경우의료지도원에게 전화수화를 통한 응급조치를 우선적으로 시도한다

의료기관의 안내를 받고도 응급조치가 어려운 경우 경찰의 협조를 얻어 의료기관의 응급실로 이송한다

의료기관과 경찰의 협조가 이루어지지 않는 경우 환자를 이송한다

구조구급대의 업무

의료기관과의 연락유지

의료기관으로의 이송지원

범죄현장의 발견시 경찰서에 신고

화재구조구급 기타 긴급한 경우에는 관련 법규에 따라 조치

비긴급상황

도와드릴게요

화재구조구급 등의 긴급한 경우가 아닌 경우로서 도움을 청하는 상대방의 생명신체재산 등을 자신의 위험에서 구조하기 위하여 긴급한 경우에는도와주세요라고 외치거나 관련 전화번호로 전화를 걸어 도움을 청한다

신고

범죄현장을 발견하거나 응급환자가 아닌 경우로서 긴급한 경우에는 경찰청센터에 신고한다

에 알려주세요

는 응급환자 및 재난발생시에만 출동하므로비긴급 상황인 경우에는가 아닌 경찰청센터나 관할 지자체 민원접수창구에 신고한다

민원처리

민원서류는 접수한 날로부터일 이내에 처리하여야 한다

민원서류를 접수한 때에는 민원서류의 내용이 위법부당하는 등 정당한 사유가 있는 경우를 제외하고민원서류를 접수한 날로부터일 이내에 처리하여야 한다

민원서류를일 이내에 처리할 수 없는 경우에는 민원인에게 그 사유를 통지하여야 한다

처리기간 계산기준

민원서류를 접수한 때는 일을 처리한 것으로 본다

처리기간 입력은일수월일일로 한다

거부처분에 대한 이의신청

민원사무처리에 관한 법률제조 및 동법 시행령 제조의 규정에 의하여거부처분이 있는 날부터일 이내에 거부처분 통지서를 받은 날로부터일 이내에 관할 행정심판위원회에 서면으로 이의신청을 할 수 있다

행정사항

각 시도 소방본부 및 소방서에서는 본 가이드라인을 적극 활용하여 응급상황시 제때에 적절한 조치를 취하도록 한다

각 시도 소방본부 및 소방서는 본 가이드라인에 의거자체 실정에 맞는 내부교육을 실시하고적극적으로 활용할 수 있도록 한다

응급상황대처가이드라인은 응급상황에서의 대처능력을 향상시키기 위한 가이드라인이다이 가이드라인은 법적 구속력을 가지지는 않지만 응급상황에서의 적절한 조치를 도와주기 위해 제작되었다응급상황은 생명의 위협이나 중대한 장애를 초래하거나 생명에 대한 위협이 증가하는 경우를 말한다

- User의 content는 문서를 요약하기 위해서 사용자가 입력하는 지시문의 형태입니다.
- Assistant의 content는 사용자가 입력하는 지시문에 맞춰 모델이 생성해야 하는 요약 결과의 형태입니다.

## 번역

다음 문장을 한국어로 번역해주세요

월리는 공원에서 곰 인형을 사고 있습니다첫 번째 곰 인형의 가격은달러이고그 이후부터는 인형 하나당센트의

할인이 적용됩니다. 월리는마리의 꿈 인형에 얼마를 지불하게 될까요

- User의 content는 번역을 위해서 사용자가 입력하는 지시문의 형태입니다.
- Assistant의 content는 사용자가 입력하는 지시문에 맞춰 모델이 생성해야 하는 번역 결과의 형태입니다.

## 분류

다음 문서에는 강남구라는 단어가 포함되어 있나요

서울 강남구가 월 말까지 신종 코로나바이러스 감염증(코로나19) 장기화로 어려움을 겪고 있거나 폭염에 취약한 위기 가구를 집중 발굴·지원에 나선다. 일 강남구에 따르면 심층상담 후 긴급복지지원 제도를 통해 가구별만 원을 반지하 등 폭염취약 가구에는 에어컨쿨매트선풍기 등 냉방용품을 지급한다. 발굴대상은 총만가구로 서울시 재난 긴급생활비 수급자 중 △만세 중장년인 가구 가구 △가계소득이 없는만가구 △반지하 거주가구 △전기체납·금융연체 등 위기에 처한 미취업 일용근로자가구다. 또 복지플래너우리동네돌봄단복지통반장 등이 집집마다 홍보에 나서는 한편 문자 및 전화카카오톡 채널강남좋은이웃등으로 비대면 발굴도 강화한다. 이와 함께 강남구는일부터 저소득 취약 어르신명을 대상으로 여름용 스카프매와 덴탈마스크매를 전달하고 있다.

- User의 content는 분류를 위해서 사용자가 입력하는 지시문의 형태입니다.
- Assistant의 content는 사용자가 입력하는 지시문에 맞춰 모델이 생성해야 하는 답변의 형태입니다.

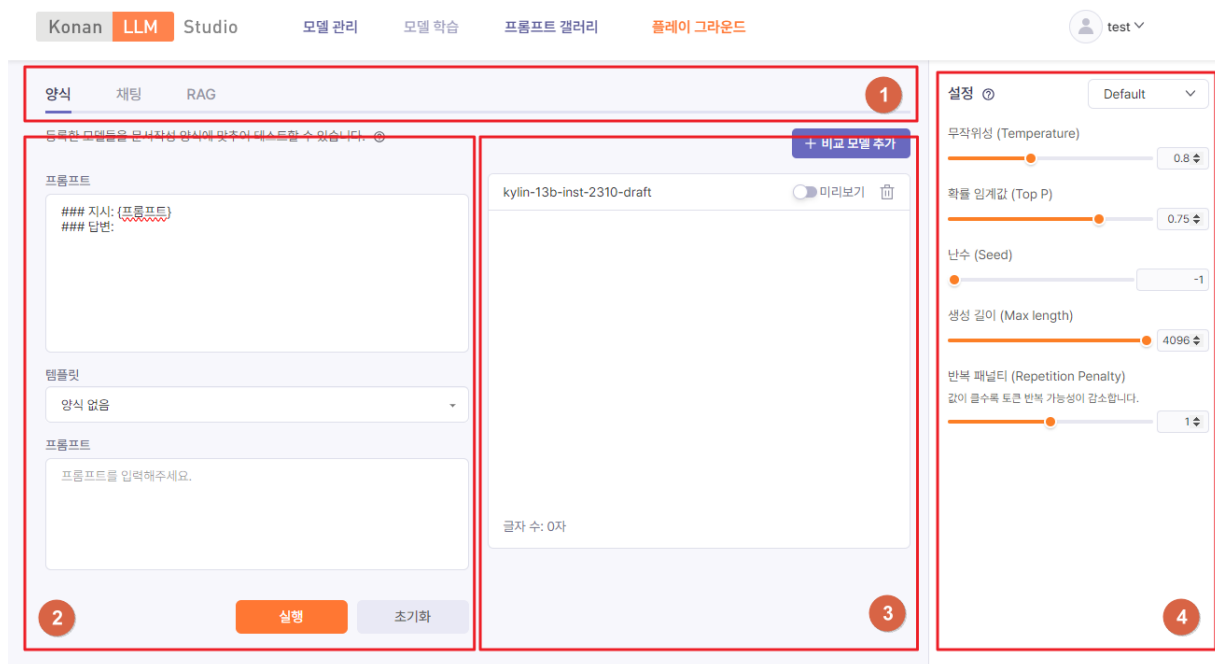
## 2.6 플레이 그라운드

플레이 그라운드를 통해 LLM 관련 지식이나 코딩 없이도 언어 모델을 테스트하고 비교해볼 수 있습니다. 파라미터를 설정하고 프롬프트를 입력하여 텍스트를 생성할 수 있는 공간입니다.

### 2.6.1 페이지 구성

플레이 그라운드 페이지는 아래와 같은 영역으로 구성됩니다.

Figure 6.1. 플레이 그라운드 화면



1. 모드
2. 프롬프트 입력 모드를 선택합니다.
  - 양식: 자유형식 및 템플릿 기반 프롬프트 입력
  - 채팅: 채팅 형식 프롬프트 입력
  - RAG: RAG 쿼리 프롬프트 입력
3. 프롬프트
4. 언어모델에게 전달할 프롬프트를 입력합니다.
5. [실행] 버튼을 클릭하여 텍스트 생성을 시작합니다.
6. 생성결과
7. 언어모델이 생성한 결과 텍스트를 보여줍니다. [+비교모델추가] 버튼을 클릭하면 결과를 생성할 언어모델을 추가하고 각 모델 별 생성 결과를 비교해 볼 수 있습니다.
8. 파라미터
9. 언어모델에 요청 시 전달되는 설정 값입니다. 자세한 설명은 [Completions API](#)를 참조하세요.

## 2.7 프롬프트 설계

프롬프트 설계는 언어 모델에서 원하는 응답을 유도하는 프롬프트를 만드는 프로세스입니다. 이 페이지에서는 프롬프트 설계를 시작하기 위한 기본 개념, 전략, 권장사항 및 예제들을 소개합니다.

### 2.7.1 프롬프트 구조

#### 프롬프트 정의

프롬프트란, 인공지능 모델과의 상호 작용을 위한 입력 문장입니다. 프롬프트에는 질문, 지침, 문맥 정보, 예제 및 모델이 완료하거나 계속하기 위한 부분 입력 등이 포함될 수 있습니다. 효과적인 프롬프트는 명확한 컨텍스트와 구성을 기반으로 모델에게 원하는 결과물을 생성하도록 유도하여 생성된 콘텐츠가 의도한 목적에 부합하는데 중요한 역할을 합니다.

#### 프롬프트 구성 요소

프롬프트에는 다음 유형의 콘텐츠가 하나 이상 포함될 수 있습니다.

- 입력(필수)
- 문맥(선택)
- 예시(선택)

#### 입력(INPUT)

모델에게 응답을 요구하는 것으로, 모델이 답변해야 할 질문, 모델이 수행할 태스크, 모델이 작업을 수행할 대상 등을 포함합니다.

#### 문맥(CONTEXT)

추가 정보를 제공하거나 모델 동작 방법 등을 지정합니다.

#### 예시(EXAMPLES)

프롬프트에 예제를 포함하면 모델이 응답 방법에 대한 패턴과 관계를 식별하고 이를 응답 구성에 적용합니다. 예제가 없는 프롬프트를 zero-shot 프롬프트, 예제가 적은 프롬프트를 few-shot 프롬프트라고 합니다.

#### 프롬프트 구성 예시

이메일의 내용이 광고인지 아닌지를 분류해주세요  
입력  
이메일의 여초내용 또는 특정 키워드를 고려하여 이메일이 광고 범주에 속하는지 문맥  
결정하세요  
예시예시  
이메일 내용당신의 건강을 위한 최고의 선택오가닉 슈퍼푸드를 이제 집에서 편리하게 즐겨보세요첫 구매 고객에게는 특별한 선물을 드립니다지금 바로 확인해보세요  
답변광고입니다  
이메일 내용안녕하세요다음 주 프로젝트 미팅에 대한 일정을 결정하기 위하여 가능한 시간을 알려주시면 감사하겠습니다  
답변광고가 아닙니다  
이메일 내용안녕하세요매뉴얼 작업에 답변 작업 부탁드립니다입력

### 2.7.2 프롬프트 설계 전략

#### 파라미터 설정

- Temperature
- 출력 텍스트의 무작위성을 조절합니다. Temperature 값이 높을수록 창의적이고 다양한 결과가, Temperature 값이 낮을수록 결정적인 결과가 출력됩니다. 가령 Temperature 값이 0이면 가장 확률이 높은 토큰이 결정적으로 선택됩니다.

- Top P
  - Temperature와 마찬가지로 출력 텍스트의 무작위성을 조절합니다. 가능한 모든 토큰을 고려하는 대신, 누적 확률이 특정 임계값(Top P) 이하인 토큰 집합 내에서 무작위적으로 다음 토큰을 선택합니다. 가령 Top P가 0.1이라면 다음 토큰들 중 확률이 높은 순서대로 더하여 누적 확률이 10% 이내인 토큰들 중 무작위로 선택합니다.
- Min-Max Length
  - 생성되는 텍스트의 최소 및 최대 길이를 지정합니다.
- Repetition Penalty
  - 생성되는 텍스트에서 반복되는 결과를 통제하기 위해 반복되는 토큰에는 감점 요소를 부여합니다.

사실을 기반으로 정확성이 요구되는 태스크의 파라미터 설정

- 적용 태스크
  - 보고서 생성, 질의응답, 사실 기반 요약 등
- Temperature
  - 0.2 ~ 0.7 정도의 값이 권장되며, 모델이 높은 확률로 예측하는 단어를 선호하게 하여 출력 텍스트가 더 일관되고 정확해집니다.
- Top P
  - 0.2 ~ 0.7 정도의 값이 권장되며, 단어 선택의 다양성을 줄여서 높은 확률을 가진 단어가 선택됩니다.

일반적으로 temperature와 top\_p 중 하나만 조절하고 다른 하나는 기본값을 유지하는 것이 좋습니다. 두 파라미터를 동시에 낮추면 너무 단조롭거나 예상 가능한 출력이 될 수 있습니다.

창의성이 요구되는 태스크의 파라미터 설정

- 적용 태스크
  - 시 작성, 소설 작성, 광고 문구 생성
- Temperature
  - 0.7 ~ 1.0 정도의 값이 권장되며, 모델이 다양한 단어와 구문을 탐색할 수 있게 하여 창의적인 태스크에 도움이 됩니다.
- Top P
  - 0.7 ~ 0.9 정도의 값이 권장되며, 단어 선택에 다양성을 더해서 예측 가능한 텍스트보다 창의적인 텍스트 생성에 도움이 됩니다.

창의성이 중요한 경우 temperature와 top\_p 모두 높게 설정할 수 있지만, 너무 높은 값은 텍스트가 무작위로 느껴질 수 있으므로 적절한 균형이 필요합니다. 보통 두 파라미터 중 하나만 조절하는 것이 좋습니다.

## 효과적인 프롬프트 작성 전략

### 역할 지정

모델에게 특정 태스크에 대한 역할이나 직업 등을 지정하면, 해당 역할에 맞게 결과물의 일관성이나 품질을 높일 수 있습니다. 예를 들어 고객의 문의에 응대하는 역할을 수행해달라고 하거나, 초등학교 교사라는 직업을 부여하여 학생들 눈높이에 맞춰 답변해달라고 요구할 수도 있습니다.

당신은 의류 쇼핑을 관련 고객 지원 상담원 역할을 수행하여 제품 문제에 대한 고객의 문의에 응답해야 합니다  
고객이 티셔츠의 사이즈를 교환하려면 어떻게 해야 하나요

### 용도 및 타겟 지정

모델에게 태스크에 대한 명확하고 구체적인 용도 및 타겟을 문맥으로 제공하여, 모델의 출력을 원하는 작업의 방향으로 유도하고, 관련 없거나 잘못된 응답을 방지하게끔 합니다.

고객에게 보낼 이메일을 작성해주세요요보다는글로벌 확장 전략에 대한 미팅 후 잠재 고객에게 보낼 후속 이메일을 작성해주세요요라고 구체적인 표현과 문맥을 제공하는 것이 좋습니다

### 원하는 작업의 성공적인 실행 예시 제공

모델에게 실제로 원하는 작업을 수행하도록 요청하기 전에 작업의 예시를 제공(few-shot)하여 모델의 출력을 원하는 방향으로 유도할 수 있습니다.

문장의 감정을 긍정부정중립으로 분류해주세요  
예시  
너가 쓴 안경 정말 멋지다 긍정  
그 음식은 정말 별로였어 부정  
어제 콘서트에 다녀왔어 중립  
  
콘서트가 매우 재미있었어

### 원하는 문서 서식의 유형이나 양식을 명시하여 요청

문서 서식의 유형이나 양식을 명시하면 모델은 그에 맞는 구조와 양식을 지켜서 출력을 생성할 수 있습니다. 이는 모델이 더 구조화되고 목적에 부합하는 문서를 제작하는 데 도움이 됩니다.

주제에 대한 문서를 작성해주세요라는 요청보다는 주제에 대한 전략 보고서 사업 기획서 정책 제안 보고서를 작성해주세요라는 식으로 문서 서식의 유형을 명시하는 것이 좋습니다  
주제에 대한 정책 제안 보고서를 작성해주세요 추진 배경 근거 예산 정책 제안 추진 계획 실행 일정 성과 측정 방법 등을 포함해서 작성해주세요와 같이 문서 양식을 함께 제공해주면 좋습니다

### 원하는 출력 포맷을 명시하여 요청

텍스트 외에도 표나 목록, JSON 등 다양한 형식으로 정보를 요청할 수 있습니다. 이를 통해 더 구조화된 정보를 얻을 수 있습니다. 특히 구조화된 포맷은 모델에게 출력 형태의 가이드라인을 제공하여 모델은 더 정확하고 일관된 정보를 출력할 수 있습니다.

표 형식으로 구조화된 사업 계획을 출력하고 싶은 경우 사업 계획의 주요 단계를 나열해주세요 대신 사업 계획의 주요 단계를 표 형식으로 표현해주세요라고 명시하는 것이 좋습니다

### 마크다운, 구분 기호, 강조 기호 활용하여 프롬프트 구조화

모델이 마크다운이나 기호를 이해하므로 마크다운, 구분 기호, 강조 기호를 사용하여 프롬프트를 구조화하고 중요한 부분을 강조함으로써, 모델이 더 명확하고 정확한 답변을 제공하도록 유도합니다.

또는또는과 같은 구분 기호를 사용하여 지침이나 문맥이 시작되거나 끝나는 곳을 표시하면 모델이 좀 더 프롬프트의 내용을 구조적으로 이해할 수 있습니다 예를 들어 문맥으로 입력할 문서에 대하여 문서 시작과 문서 종료로 감싸줄 수 있습니다  
한편 강조하고 싶은 부분이 있을 경우와 같은 강조 기호를 사용할 수도 있습니다

### 단어, 어미 등 문구를 수정해가면서 입력

프롬프트에 사용된 단어나 어미, 존댓말 여부에 따라 다른 결과를 출력하기도 하므로, 같은 의미의 프롬프트라도 문구와 표현을 약간씩 수정해가면서 원하는 결과물을 도출하도록 유도할 수 있습니다.

초코 케이크 레시피를 알려 주세요라는 입력의 결과가 좋지 않은 경우 초코 케이크를 만드는 방법을 설명해 주세요라는 표현으로 바꾸거나 초코 케이크를 만드는 방법을 설명해줘 또는 초코 케이크를 만드는 방법을 단계별로 설명해줘라는 식으로 약간씩 수정해가면서 더 나은 결과를 출력할 때까지 다양한 시도를 하는 것이 좋습니다

### 각종 매직 프롬프트, 매직 워드 활용

최신 프롬프트 엔지니어링 관련 논문들이나 사용 사례들을 통하여 파악한 동향에 따르면, 특정 표현이나 매직 워드를 통하여 모델의 반응을 유도하고 결과물의 품질을 개선할 수 있습니다.

차근차근 생각해보자	CoT Prompting 기법에서 파생된 Zero-Shot CoT Prompting의 경우엔 질문에 대한 ‘중간 추론 단계’의 예시를 직접 제공해주지 않더라도 ‘Let’s think step by step’와 같은 프롬프트를 추가해주는 것만으로도 모델 출력이 개선될 수 있음을 제시합니다.
심호흡을 하고 차근차근 생각해보자	구글 딥마인드에서 발표한 Large Language Models as Optimizers (2023.09) 논문의 Optimization by PROMpting (OPRO) 방법으로 최적화한 프롬프트 중 ‘Take a deep breath and work on this problem step-by-step’라는 프롬프트를 추가했을 때 PaLM2에서 GSM8K 정답률이 34%에서 80.2%로 개선된 결과가 도출되었습니다.
감정적 자극	MS에서 발표한 Large Language Models Understand and Can be Enhanced by Emotional Stimuli (2023.07) 논문에 따르면 모델에게 감정적 자극(자기 모니터링, 긍정적인 동기부여와 노력 격려, 도전을 성장의 기회로 받아들이도록 유도 등)의 프롬프트를 추가했을 때 성능이 향상되었다는 결과가 제시되었습니다.

### 장황한 표현 피하기

장황한 표현보다는 간결하고 구체적인 지시를 제공하여 모델이 핵심적인 내용에 집중하도록 합니다. 긴 프롬프트가 반드시 좋은 결과를 보장하지 않으며 오히려 중요한 요소를 간과·누락할 위험이 있습니다. 또한 예시를 사용할 때에도 무조건 많다고 좋은 것이 아니며 모델이 다루어야 할 여러 케이스에 대한 패턴을 파악할 수 있는 적절한 예시들을 선택하여 프롬프트에 추가하는 것이 좋습니다.

지나치게 많은 요구사항을 한꺼번에 입력하기보다는 나눠서 여러번에 걸쳐 입력하는 것이 좋습니다. 또한 시에 비슷한 패턴의 예시를 여러 개 넣어주는 것보다는 다른 패턴의 예시를 집약적으로 구성하는 것이 좋습니다. 적절한 예시 개수는 작업에 따라 다르지만 일반적으로 난이도가 평이한 작업의 경우 예시개만으로도 잘 작동합니다.

### 부정문(하지 말아야 할 일) 대신 긍정문(해야 할 일)을 지시

언어 모델은 뭔가를 하지 말라는 부정문보다 해야 할 일을 명확히 지시하는 긍정문에 더 잘 반응하는 편입니다. 따라서 원하지 않는 것을 언급하는 대신 모델이 무엇을 하길 원하는지에 집중하여 지시하면 더 좋은 출력 결과를 얻을 수 있습니다.

길게 설명하지 마세요보다는 짧게 설명하세요와 같은 명확한 지시가 더 좋은 결과를 출력할 수 있습니다.

### 복잡한 작업에 대하여 세부적으로 필요한 단계를 나눠서 요청

복잡한 작업을 수행할 때 각각의 단계를 세부적으로 나눠서 모델에게 요청하면, 모델은 각 단계를 이해하고 순차적으로 수행하면서 오류를 최소화하고 원하는 결과물에 가깝게 출력할 수 있게 됩니다.

문서 시작  
반려동물을 키우는 것은 많은 책임을 수반하며 동물의 건강을 관리하는 것은 이러한 책임 중 중요한 부분입니다. 특히 영양, 운동, 정기적인 건강 검진, 예방접종 그리고 정서적 안정성은 반려동물의 건강한 삶을 위해 필수적인 요소들입니다. 균형 잡힌 식단은 반려동물의 체중을 관리하고 전반적인 건강을 증진시키는 데 도움을 줍니다. 정기적인 운동은 반려동물의 체력을 강화하고 스트레스를 줄이는 데 중요합니다. 한편 동물 병원에서의 정기적인 건강 검진은 잠재적인 건강 문제를 조기에 발견하고 예방하는 데 필수적이며 다양한 질병으로부터 반려동물을 보호하기 위해서는 정기적인 예방접종도 필요합니다. 또한 반려동물의 정서적 안정성도 건강 관리의 중요한 부분입니다. 사랑과 관심, 안정된 환경은 반려동물이 스트레스를 덜 받고 행복하게 생활하는 데 도움을 줍니다.  
문서 종료  
위 문서에서 중요한 키워드들을 뽑아주세요  
각 키워드에 대한 설명을 작성하세요  
키워드와 설명을 보고서 형식으로 정리하여 출력하세요

### 복잡한 작업을 하위 질문으로 나누어 질문하고 연결

복잡한 작업을 하위 질문으로 나눠서 이전 하위 질문의 답변을 다음 하위 질문의 입력에 사용하는 방식으로 Prompt Chaining을 하면, 길고 복잡한 프롬프트 대신 간결하고 명확한 프롬프트를 여러개 전달하여 복잡한



작업을 수행하는 것이 더 효과적일 수 있습니다. 이를 통해 여러 문제 해결을 한꺼번에 수행할 때보다 각 부분적인 문제 해결에 집중할 수 있습니다.

## 1차 프롬프트

주요 마케팅 방법가지를 선정하여 각각에 대하여 설명해 주세요

## 2차 프롬프트

문서 시작  
 1차 프롬프트 결과를 문맥 정보로 입력  
 문서 종료  
 위 문서를 참고하여주요 마케팅 방법을 활용한새로 개봉되는 영화의 홍보 전략보고서를 작성해주세요

### 트러블슈팅

#### 반복 문제 해결

모델이 반복적인 결과를 출력하는 경우, 'repetition penalty' 파라미터를 조정하는 것이 유용합니다. 이 파라미터의 값을 높일수록, 이전에 생성된 토큰과 동일한 토큰이 생성될 때 페널티를 높게 부여하므로 모델이 동일한 단어나 구를 반복하는 것을 억제합니다. 적절한 repetition penalty 값은 상황에 따라 다르지만, 이 값을 너무 높인다면 토큰 다양성이 제한되기도 하므로 유의하면서 조절해야 합니다.

#### 표를 출력하지 못하는 문제 해결

프롬프트에 '표 출력'을 명시적으로 요청하면 대체로 표를 출력하는 편이지만 처음 시도에서 원하는 결과가 나오지 않았더라도 문구를 수정하거나 파라미터 설정을 바꿔가면서 계속 시도해야 합니다. 예를 들어 표에 대한 표현도 '표', '도표', '테이블', '표 구조', '표 형식', '표 형태' 등 다양하게 바꿔보고 표를 출력해달라는 요청도 '표 작성', '표 출력', '표로 표현', '표로 제시', '표로 정리', '표로 변환', '표로 구성' 등 다양한 어휘로 바꿔서 시도해볼 수 있습니다. 그리고 프롬프트 내에서 표를 출력해달라는 요청의 위치를 중간에 넣거나 맨 뒤에 위치하게끔 하거나 괄호 안에 넣거나 하는 식으로 다양하게 시도해볼 필요가 있습니다. 뿐만 아니라 표에 포함되어야 할 내용이나 카테고리 등을 구체적으로 제시하여 원하는 출력 형식을 모델이 이해할 수 있도록 유도하는 것이 좋습니다.

#### 글자 수 요청을 준수하지 못하는 문제 해결

언어 모델은 텍스트를 토큰 단위로 처리하므로, 정확한 글자수 제한을 맞추는 것이 어려울 수 있습니다. 토큰은 글자 단위로 처리될 수도 있지만 보통 단어나 구의 일부로 처리되는 특성 때문에, 모델이 정확한 글자 수를 맞추는 것은 항상 가능하지 않을 수 있습니다. 그럼에도 불구하고, 글자수 제한을 명시할 때 모델은 그 제한에 최대한 근접하려고 노력하기 때문에 처음에 글자 수 제한 요청이 제대로 이루어지지 않았더라도 해당 요청의 문구를 다양하게 수정해가면서 원하는 글자 수 제한에 근접하도록 다양한 시도를 해보는 것이 좋습니다. 예를 들어 1000자 이내로 제한해주세요, 1000자 내외로 표현해주세요, 총 분량 1000자 내외로 작성해주세요, (1000자 내외로 작성), 1000자의 범위 내에서 답변해주세요, 3000자 이상으로 상세히 답변해주세요, 150자 이내로 간결하게 답변하세요 등 다양한 어휘와 표현 방식으로 시도하는 게 좋습니다.

## 2.7.3 프롬프트 예제

프롬프트 예제는 입문 지침으로서의 역할을 하며, 예제를 단순히 복사해서 사용하기보다는 개별적인 환경과 목적에 맞게 프롬프트를 반복적으로 조정해야 합니다. 예를 들어, 첫 번째 시도에서 원하는 결과가 나오지 않았다면, 프롬프트의 구체성을 높이거나 문맥을 추가해보는 것이 유용할 수 있습니다. 또한 모델의 파라미터 설정(예: temperature, top\_p 등)을 조절하여 결과의 일관성이나 창의성을 높일 수 있습니다.

## 텍스트 생성

다양한 텍스트나 문서 초안을 생성하고자 할 때 참고할 수 있는 예제를 소개합니다.

### 창의성이 필요한 텍스트

창의성이 필요한 텍스트 중 ‘광고 메시지’를 생성할 수 있도록 파라미터를 설정하고 프롬프트를 구성하는 방법을 설명합니다.

**파라미터 설정 (권장) 파라미터값** Temperature0.2\~ 0.5Top P0.4\~ 0.5Min-Max Length4096Repetition Penalty1.05

### 프롬프트 구성

당신의 임무는 일관된 스타일로 답변하는 것이고당신의 역할은 광고 메시지를 작성하는 것입니다  
예시  
요청빛과 그림자라는 테마를 활용한 향수 광고 메시지를 작성해주세요  
답변빛과 그림자가 만들어내는 향의 조화우리의 새로운 향수는 마치 어둠 속 달빛 사이를 거니는 듯한 감각적인 경험을 선사합니다이 향수 한 방울이 당신의 일상에 신비로운 빛과 그림자를 드리우며감성적인 순간들을 만들어냅니다각기 다른 빛의 각도에서 다양한 향기의 레이어를 경험하세요당신의 모든 순간을 특별하게 만들어줄빛과 그림자의 향기로운 무대를 경험해보세요  
요청진주를 테마로 하는 남성용 셔츠를 홍보하는 광고 메시지를 작성해주세요

### 실행 결과

진주는 그 자체로 섬세하고 세련된 아름다움을 지닌 보석입니다우리의 남성용 셔츠는 이런 진주의 아름다움을 모티브로 하여 제작되었습니다진주처럼 섬세하고 세련된 디테일이 돋보이는 셔츠를 입으면자신감이 상승하고 남성적인 매력이 한층 더 돋보일 것입니다그리고 우리 셔츠를 입은 그대여당당하게 거리를 걸어보세요마치 진주처럼 단단하고 우아한 모습으로다가올 모든 순간을 특별하게 만들어줄 것입니다

### 구조화된 텍스트

구조화된 텍스트 중 보도자료 초안을 생성할 수 있도록 파라미터를 설정하고 프롬프트를 구성하는 방법을 설명합니다.

### 파라미터 설정 (권장)

파라미터	값
Temperature	0.2\~ 0.5
Top P	0.4\~ 0.5
Min-Max Length	4096
Repetition Penalty	1.05

### 프롬프트 구성

가상현실과 인공지능기술을 결합하여 케이댄스를 배울 수 있는 플랫폼년월일 공개에 대한 보도자료 초안을 작성해주세요

### 실행 결과

제목기술로케이댄스배운다  
기술로 구현한 인공지능 학습영상누구나 쉽게 댄스 학습 가능  
가상현실과 인공지능기술을 결합한케이댄스 플랫폼이 개발돼년월일 공개된다  
가상공간에서 전문강사의 실시간 댄스동작을 분석하고학습자에게 맞춤형 피드백을 제공하여 지속적인 학습을 돕는 새로운 댄스학습 플랫폼이 등장하는 것이다  
학습자 맞춤형 피드백 제공하는 인공지능 학습영상 구현  
그동안 기술로 시뮬레이션한 댄스학습은 있었으나실시간으로 합성된 인공지능 학습영상을 통해 직접 케이댄스를 배울 수 있게 된 것이다  
학습자는 기기를 이용해 가상공간에서 전문강사의 실시간 댄스동작을 분석하게 되고몸짓을 인식해 학습자세 교정을 받게 된다  
특히 인공지능 학습영상은 학습자의 춤 동작을 분석하여 다양한 각도에서 리플레이하며학습자가 틀린 부분을 바로잡도록 피드백을 제공한다  
순쉽게 케이댄스 배우는 플랫폼년월일 공개

케이댄스 플랫폼은 총개의 공간으로 구성되어 있으며개의 스튜디오와개의 실전공간으로 이루어져 있다  
 스튜디오는 인공지능 학습영상을 수신하고기술로 구현한 가상공간이며실전공간은 수신된 학습영상을 바탕으로 가상공간에서 케이댄스를 연습할 수 있는 공간이다  
 케이댄스 플랫폼은년월일 공개되며누구나 기기와 스마트폰을 이용해 케이댄스를 배울 수 있다  
 문화체육관광부 장관 표창 수상년월까지 무료 이용 가능  
 케이댄스 플랫폼은 누구나 쉽게 케이댄스를 배울 수 있는 장점이 있으며특히 기술로 시뮬레이션된 댄스학습은 학습자의 공간선택 제약을 벗어나게 해주고시간과 장소에 구애받지 않는 유연한 학습방식으로 사용될 수 있다  
 또한 인공지능 학습영상을 통해 학습자의 실시간 피드백을 제공하는 점은 개인별 맞춤형 정밀한 학습을 가능케 하므로 케이댄스 플랫폼은 디지털 시대의 수요에 부응하는 혁신과 융합성을 갖춘 콘텐츠로 평가된다  
 이러한 성과를 인정받아케이댄스 플랫폼은 문화체육관광부 장관 표창을 수상하였으며년월까지 무료로 이용할 수 있다

## 요약

입력한 문서를 원하는 길이로 요약하기 위한 예제를 소개합니다.

### 문서 요약

정책보고서 중 일부 내용에 대하여 원하는 길이로 요약할 수 있도록 파라미터를 설정하고 프롬프트를 구성하는 방법을 설명합니다.

### 파라미터 설정 (권장)

파라미터	값
Temperature	0.3\~0.6
Top P	0.4\~0.5
Min-Max Length	150이하
Repetition Penalty	1.2

## 프롬프트 구성

문서 시작  
 년도 행정안전부 개인정보 보호 교육계획안  
 교육 목적  
 우리 부 직원의 개인정보 보호 필요성 등 인식 제고  
 개인정보보호 역량 강화를 통한 자율적인 개인정보 보호 활성화 및 안전한 개인정보 활용 지원 등 우리 부 개인정보보호 수준 향상  
 교육 계획개인정보보호법 제조제항제호개인정보 보호 교육 계획의 수립 및 시행  
 개인정보 교육개인정보보호법 제조제항 개인정보취급자에게 교육 실시  
 교육 방향  
 코로나상황을 고려하여 비대면 영상회의 또는 온라인 교육을 기본으로 하되상황에 따라 집합교육워크숍 등 개최 검토  
 개인정보 보호책임자기관 보호담당자취급자일반직원 등 역할에 적합한 교육 대상 별 맞춤형 교육 실시  
 ※ 개인정보 보호책임자 별 교육계획을 수립 시행하여야 하나 소속기관은 본부 계획 준용 가능  
 문서 종료  
 위 문서의 내용을 한 문장으로 요약해주세요

## 실행 결과

년 행안부 개인정보 보호 교육은 코로나상황을 고려하여 주로 비대면 형태로 진행하고필요할 경우 집합교육워크숍 등을 개최 검토하며개인정보 보호책임자기관 보호담당자취급자일반직원 등 역할별 맞춤형 교육을 실시해야 합니다

## 분류

입력한 텍스트에 대하여 주어진 기준에 따라 분류하기 위한 예제를 소개합니다.

### 이메일 광고 여부 분류

이메일이 광고인지 아닌지 분류할 수 있도록 파라미터를 설정하고 프롬프트를 구성하는 방법을 설명합니다.

## 파라미터 설정 (권장)

파라미터	값
Temperature	0.2 ~ 0.5
Top P	0.4 ~ 0.5
Min-Max Length	150이하
Repetition Penalty	1.05

## 프롬프트 구성

이메일 내용안녕하세요저희는 최신 기술을 탑재한 스마트워치를 출시했습니다이번 주말까지만할인 행사를 진행하니이 기회를 놓치지 마세요지금 바로 접속하셔서 특별 할인 혜택을 누리세요  
분류 요청이 이메일 내용이 광고인지 아닌지 분류해주세요

## 실행 결과

이 이메일 내용은 광고입니다

### 감정 분류

텍스트의 감정을 분류할 수 있도록 파라미터를 설정하고 프롬프트를 구성하는 방법을 설명합니다.

## 파라미터 설정 (권장)

파라미터	값
Temperature	0.2 ~ 0.5
Top P	0.4 ~ 0.5
Min-Max Length	150이하
Repetition Penalty	1.05

## 프롬프트 구성

예시  
텍스트오늘 배달된 피자가 너무 늦게 도착했고맛도 별로였어요실망스러웠습니다  
분류 요청이 텍스트에 표현된 감정을 긍정부정중립 중에서 분류해주세요  
답변부정  
텍스트제품에 대한 정보를 더 알고 싶습니다추가 세부 사항을 제공해주실 수 있나요  
분류 요청이 텍스트에 표현된 감정을 긍정부정중립 중에서 분류해주세요  
답변중립  
텍스트이 새로운 카페 정말 좋아요인테리어가 아늑하고 커피도 맛있어서 자주 갈 것 같아요  
분류 요청이 텍스트에 표현된 감정을 긍정부정중립 중에서 분류해주세요  
답변긍정  
텍스트이 스마트폰 정말 최고요배터리 수명이 길고 카메라 기능도 뛰어나네요  
분류 요청이 텍스트에 표현된 감정을 긍정부정중립 중에서 분류해주세요

## 실행 결과

이 텍스트는 긍정적인 경험을 기술하고 있으므로긍정으로 분류됩니다

## 2.8 Chapter 8. 검색증강생성(RAG)

검색증강생성(Retrieval Augmented Generation)은 생성모델에게 검색을 통해 찾은 참조 문서 또는 텍스트를 기반으로 답변을 생성하게 하는 기법입니다.

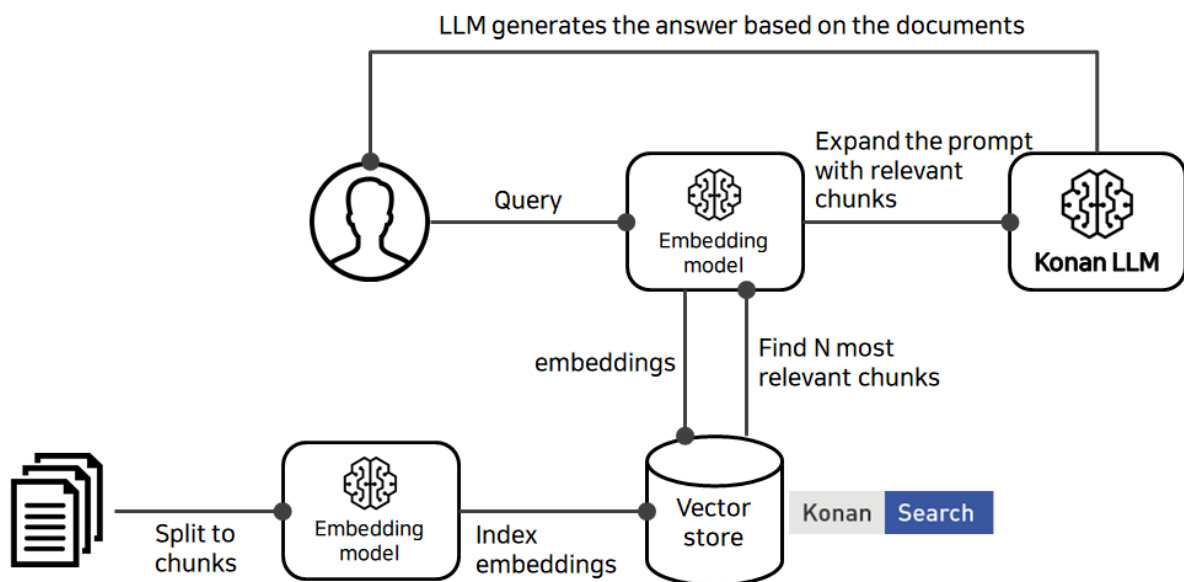
생성모델의 고질적인 문제인 "환각(hallucination)"을 줄이며 최신 지식을 반영한 답변 생성을 위한 방안으로 많이 이용됩니다.

문서에 대한 접근제어를 검색시스템을 통해 처리함으로써 보안 문제를 해소하는데도 쓰입니다.

### 2.8.1 전체 구성

검색증강을 위한 시스템 구성도는 아래와 같습니다.

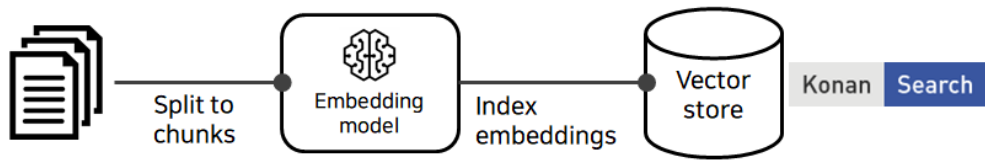
Figure 8.1. 검색증강 시스템 구성도



### 2.8.2 벡터스토어

벡터스토어는 기존 키워드 검색으로 찾기 어려운 문서들을 검색해내기 위해, 참조 대상 텍스트를 벡터로 변환해서 색인/저장하는 도구입니다. 벡터검색을 지원하는 엔진인 Konan Search를 사용합니다.

Figure 8.2. 벡터 스토어

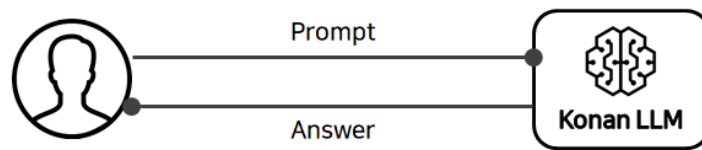


- 참조 대상이 될 문서들은 “임베딩모델(Embedding Model)”을 통해서 벡터 값 형태로 임베딩(embedding)됩니다.
- 임베딩은 문장 단위로 이뤄질 수도 있고, 단락 또는 페이지 단위로 이뤄질 수도 있습니다.
- 임베딩 벡터들은 벡터 색인 엔진 - Konan Search 내 “볼륨” 안에 저장됩니다.
- 저장된 벡터들을 벡터 간 유사도 비교로 검색해 올 수 있습니다.
- 키워드가 정확히 일치하지 않아도 의미상 유사한 것을 찾을 수 있어 재현율(recall)이 올라갑니다.

### 2.8.3 프롬프트를 통한 답변 생성

일반적인 생성모델에 대한 지시는 아래와 같이 참조문서 없이 사용자 프롬프트만을 기반으로 하는 경우가 많습니다. 모델의 지식만으로 답변을 생성하므로 환각현상(hallucination)이 발생할 수 있습니다.

Figure 8.3. 프롬프트를 통한 답변 생성

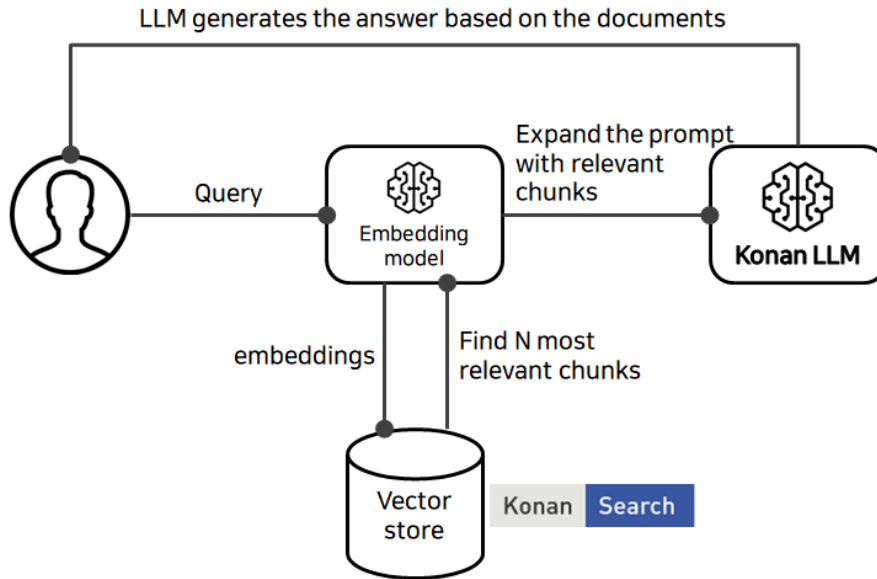


- 생성모델은 기본적으로 사용자 프롬프트를 통해서 답변을 생성합니다.
- 이 프롬프트에 답변의 근거가 되는 문서를 벡터검색을 통해 찾아냅니다.

### 2.8.4 참조 문서를 기반으로 한 답변 생성

벡터스토어를 통해 검색된 문서를 사용자 프롬프트에 함께 전달하여 답변을 생성하게 합니다. 내용 생성에 일조의 "제약"을 가함으로써 환각현상을 줄일 수 있습니다.

Figure 8.4. 참조 문서를 기반으로 한 답변 생성

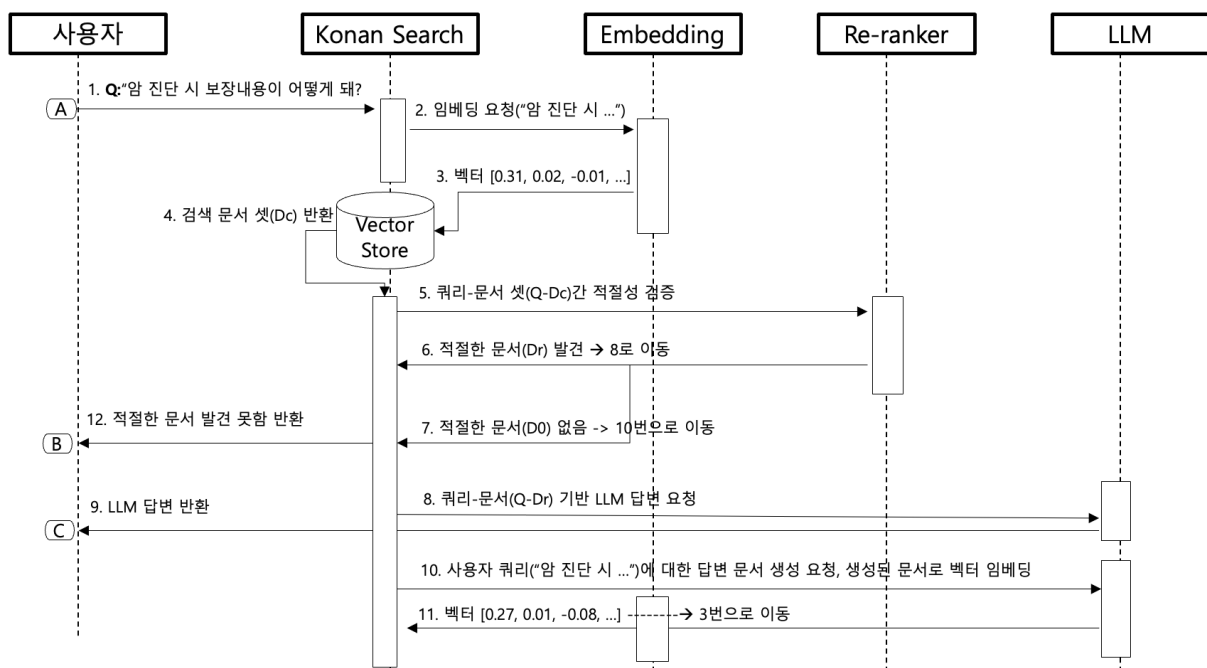


- 사용자의 쿼리를 벡터 임베딩모델을 통해서 벡터로 변환합니다.
- 벡터 검색엔진에서 관련성 높은 문서를 찾아내고, 사용자 질의와 함께 문서를 프롬프트의 컨텍스트로 구성합니다.
- 생성모델에서 최종답변을 생성하고 사용자에게 반환합니다.

## 2.8.5 데이터 흐름 예시

사용자가 질의를 하는 순간부터 생성모델에 답변 요청을 하기까지 아래와 같은 여러 단계의 처리 과정을 거치게 됩니다.

Figure 8.5. 데이터 흐름 예시



## 3. Ver 2.0

---

### 3.1 개요

---

LLM 스튜디오는 생성형 언어 모델 코난 LLM을 빠르게 프로토타이핑하고 테스트 해볼 수 있는 도구입니다. 자신만의 프롬프트를 설계하여 테스트하고 공유할 수 있습니다. 또한 기본 모델을 미세 조정하여 원하는 응용 분야에 적용할 수 있습니다. LLM 스튜디오를 통해 다음과 같은 작업들을 수행할 수 있습니다.

- 모델 관리
- 프롬프트 갤러리
- 모델 학습
- 플레이그라운드

#### 3.1.1 주요 개념

---

##### 텍스트 생성 모델

텍스트 생성 모델은 자연어 처리 분야에서 인공지능을 접목하여 자연스러운 문장을 만들어 내도록 학습한 모델입니다.

사전학습한 데이터의 양과 품질, 다루고 있는 영역에 따라 텍스트 생성 모델의 잠재력이 결정되며, 미세조정을 통해서 사람의 지시를 정확하게 따를 수 있는 능력이 발현되기 시작합니다.

이에 따라 텍스트 생성 모델은 크게 “사전학습 모델”과 “미세조정 모델”로 나누어 집니다. 사용자의 목적, 적용 범주에 따라 사전학습 모델에 전문 지식 분야에 대한 추가학습을 진행하기도 하고, 미세조정 모델에 대한 추가적인 미세조정을 통해 보다 원하는 형식의 텍스트 생성이란 목표를 달성할 수 있습니다.

##### 임베딩

임베딩이란 고차원 벡터를 변환할 수 있는 비교적 저차원의 공간입니다. 주로 언어 모델에서는 글자, 단어, 문장으로 확대되면서 복잡해지는 언어를 적절한 크기의 분절로 나눈 뒤, 이를 일정한 크기의 실수 벡터 형태로 변환시켜 사용합니다. 이러한 임베딩은 다양한 수준(단어, 문장 등)에서 활용 가능하고, 유사한 문장 찾기, 검색, 주요 주제 군집화, 이상 데이터 탐지와 같은 분야에서 사용할 수 있습니다.

##### 토큰

언어 모델에서의 토큰은 자연어 처리를 위해 여러 문자를 일정한 기준에 따라 나눈 것을 의미합니다. 목적에 따라 하나의 토큰은 자소 단위에서부터 문장 단위까지 확장될 수도 있습니다. 토큰 기준으로 임베딩이 구성되기 때문에 저차원에서도 적절한 정보 수준을 유지할 수 있도록 언어학적, 통계학적 지식을 이용하게 되며, 대개 한국어 단어는 형태소 단위로 토큰이 나누어지게 됩니다. 또한, 이러한 토큰은 실제로 언어 모델이 처리하는 단위로서 한번에 처리 가능한 최대 정보량을 나타내는 단위가 되기도 합니다.



## 3.2 모델 관리

LLM 스튜디오를 통해 다양한 기능을 가진 여러 모델들을 관리할 수 있습니다. 또한 미세 조정을 통해 필요한 용도에 맞게 모델을 새롭게 정의할 수 있습니다.

### 3.2.1 모델 유형

모델 관리 페이지에서 확인할 수 있는 모델들의 유형은 다음과 같습니다.

유형	설명
-----	
기본모델	LLM 스튜디오를 통해 다양한 업무를 위해 미세 조정할 수 있는 사전 학습된 모델
:-	:-
미세조정모델	미세조정을 거쳐 업무에 적용할 수 있는 모델

### 3.2.2 모델 목록

모델 관리 메뉴를 클릭하면 LLM 스튜디오에서 접근할 수 있는 모델들의 목록을 확인할 수 있습니다.

#### 기본 모델

모델명	설명	속성
-----		
kylin-13b-2308	자연어를 이해하고 생성할 수 있는 파운데이션 모델	2K
:-	:-	:-
kylin-13b-inst-dpo-2312	kylin-13b 모델에 대한 인스트럭션 튜닝 모델	4K
kylin-embedding-2310	텍스트 입력에 대한 임베딩 벡터 반환	
kylin-safety-2312	유해성 텍스트 분류 모델	

#### 미세 조정 모델

모델명	설명
-----	
kylin-13b-2308-sft-2310	kylin-13b 모델에 대한 미세 조정 모델
:-	:-

### 3.2.3 상태

모델의 현재 상태를 표시합니다.

## 3.3 프롬프트 갤러리

---

LLM 스튜디오의 프롬프트 템플릿 기능을 이용하여 문서 작성에 유용한 프롬프트를 설계할 수 있습니다.

프롬프트 설계는 크게 아래와 같은 단계로 수행합니다.

1. 프롬프트 작성
2. 변수 추출

### 3.3.1 프롬프트 작성

---

프롬프트는 응답을 받기 위해 언어모델에게 제출하는 자연어 요청입니다. 원하는 결과가 나올 때까지 프롬프트를 주의 깊게 작성하고 파라미터들을 조정합니다. 프롬프트 작성을 위한 지침은 [프롬프트 설계](#)를 참조하세요.

### 3.3.2 변수 추출

---

변수는 전체 프롬프트 중 사용자로부터 입력 받을 내용을 의미합니다. 테스트가 끝난 프롬프트에서 사용자에게 직접 입력 받을 영역을 선택하여 변수로 추가합니다. 템플릿 양식을 통해 사용자로부터 입력 받은 변수를 프롬프트에서 치환한 결과가 모델에게 전달되어 결과 텍스트를 생성하게 됩니다. 각각의 변수는 다음과 같은 속성을 가집니다.

1. 레이블: 템플릿 양식의 레이블로 사용됩니다.
2. 입력 예시: 사용자들이 어떤 식으로 작성해야 하는지 참조할 수 있도록 입력창의 초기값으로 활용됩니다.
3. 입력 유형: 싱글 라인, 멀티 라인, 선택 버튼 중 알맞은 유형을 선택합니다.

변수 설정이 끝나면 [템플릿] 버튼을 클릭하여 실제 사용자에게 표시될 양식을 확인할 수 있습니다.

## 3.4 추가 학습

도메인 지식을 모델에 주입하는 방법을 알아보세요.

### 3.4.1 개요

코난 LLM은 방대한 양의 텍스트로 학습되어서 언어를 이해하고 생성하는 능력이 뛰어납니다. 또한 다양한 지식을 가지고 있습니다. 그러나 특정 조직에서 다루는 도메인에 대한 지식은 부족할 수 있습니다. LLM 스튜디오는 코난 LLM의 언어 능력과 기존 지식은 유지하면서 해당 조직의 도메인 지식을 주입할 수 있는 "추가 학습" 기능을 제공합니다. 추가 학습을 통해서 특정 도메인에 최적화된 LLM을 만들 수 있습니다.

### 3.4.2 추가 학습이 필요한 경우

코난 LLM을 추가 학습 시키면 언어 모델에 도메인 지식을 주입시킬 수 있지만 많은 시간과 노력이 필요한 작업이므로 신중하게 결정해야 합니다. 다음과 같은 질문들에 대해 고민해 보고 추가 학습을 할 지 RAG를 적용할 지 결정하는 것이 바람직합니다.

**도메인 지식의 변경이 예상 되는가?**

추가 학습된 모델은 특정 시점까지의 데이터만 학습이 됩니다. 도메인 지식의 변경이 예상될 경우 RAG가 더 효율적인 방법입니다. Vector Store를 갱신하는 일이 모델을 추가 학습하는 것보다 훨씬 간편한 일이기 때문입니다.

**학습 데이터를 얼마나 확보할 수 있는가?**

추가 학습된 모델의 품질은 도메인 데이터의 품질과 양에 달려 있습니다. 반면 RAG는 학습 데이터 품질과 양과는 독립적입니다. 충분한 양의 고품질 데이터를 확보할 수 있을 경우에만 추가 학습을 고려하세요.

**충분한 학습 시간과 GPU 자원을 확보할 수 있는가?**

추가 학습을 위해서는 충분한 GPU 자원이 확보되어야 합니다. 13B 모델의 경우 최소 320GB의 GPU 메모리가 필요합니다.

### 3.4.3 추가 학습 절차

**데이터 세트 준비**

추가 학습이 필요하다고 판단한 후에는 모델 학습을 위한 데이터를 준비해야 합니다.

**데이터 수집**

학습에 사용할 문서 파일들을 학습 서버의 학습 데이터 폴더에 저장합니다.

텍스트 포맷 뿐만 아니라 Microsoft Word, PDF, 아래한글 등 대부분의 문서 포맷을 지원합니다.

**데이터 정제**

수집한 문서 파일들로부터 텍스트를 추출하고 정제합니다. (자세한 사항은 [데이터 정제](#)를 참고하세요.)

## 추가 학습 실행

### LLM 스튜디오

시작하기 페이지에서 **모델 학습** 탭을 클릭합니다.

**학습 방법 선택** 페이지에서 **추가 학습** 라디오 버튼을 선택합니다.

**계속**을 클릭합니다.

**모델 설정** 페이지에서 추가 학습을 위한 옵션을 설정합니다. (추가 학습 옵션에 대한 자세한 설명은 부록을 확인하세요.)

#### • 모델 기본 정보 설정

- 모델 이름: 추가 학습된 모델의 이름을 입력합니다.
- 기본 모델: 추가 학습하려는 기본 모델을 선택합니다.

#### • UI를 통한 설정

- 학습 반복: 전체 데이터 세트를 처리하는 횟수
- 배치 크기: 한 스텝 당 처리하는 데이터 수
- 학습률: 학습 중 매개변수를 변경하는 속도

**계속**을 클릭합니다.

**학습 데이터 설정** 페이지에서 추가 학습할 문서들이 저장된 폴더를 설정합니다.

**학습 시작**을 클릭합니다.

모델 학습 상태는 모델 관리 페이지에서 확인할 수 있습니다.

## 추가 학습 모델 평가

미리 구축한 도메인의 평가 데이터를 이용하여 추가 학습 전의 언어모델과 추가 학습 후의 언어 모델을 비교 평가합니다.

### 평가 데이터 구축

추가 학습 데이터의 도메인 관련 단어들을 포함한 문장들로 평가 데이터를 구축합니다.

### 평가 예시

- 위키(한국어)데이터 학습 전 생성 결과

입력오펜하이머는  
결과친구들과 함께 여행을 떠났습니다혜린은 여러 곳에서 사진을 찍어 왔는데 그 중에서도 흥미롭게 생긴 사람들을 찾아보려고 합니다

- 위키(한국어)데이터 학습 후 생성 결과

입력오펜하이머는  
결과차 세계대전 당시 핵무기를 개발한 맨해튼 프로젝트의 총책임자였다는년월일 오전시분초뉴멕시코주 앨라모  
고도 사막에서 인류 최초의 핵실험인트리니티 테스트를 성공적으로 이끌었다이 실험으로 일본 히로시마와 나가사키에 투하된 핵폭탄이 만들어  
졌다

### 3.4.4 데이터 정제

언어모델의 생성 품질은 학습 데이터의 질에 깊이 연관되어 있습니다. 데이터 정제 과정은 원본 문서 파일로부터 텍스트를 추출하고, 노이즈를 제거하여 언어 모델 학습을 위한 고품질 데이터셋을 구축하는데 중요합니다.

아래와 같은 노이즈들을 제거합니다. (괄호는 대응되는 정제 방법)

- 의미가 없는 특수 문자 (규칙기반, 딥러닝 정제)
- 학습 데이터로서 가치가 없는 단순한 텍스트 (규칙기반, 딥러닝 정제)
- 중복 텍스트 (중복제거)
- 개인정보 (규칙기반 정제)

#### 소요시간

- 텍스트 추출: 파일당 \~1초 소요. 표 추출 활성화 시 파일당 \~10분까지도 소요될 수 있음.
- 규칙기반 정제: 초당 \~200건 처리 (건당 글자수 \~3000자)
- 중복제거: 초당 \~500건 처리 (건당 글자수 \~3000, 200만건 기준 메모리 1GB 점유)
- 딥러닝 정제: 초당 \~100건 처리 (건 당 글자수 \~3000, RTX 3090 기준)

#### 설치 및 실행

##### 설치

제공받은 소스 (data-cleanser.zip)의 압축을 해제한 후 다음 명령어를 실행합니다.

소요 시간약분

##### 삭제

##### 실행

#### 옵션 설명:

- `--src_dir`: 데이터의 소스 디렉터리 경로입니다.
- `--dst_dir`: 처리된 데이터를 저장할 목적지 디렉터리 경로입니다.
- `--n_lines`: 데이터 파일에 저장할 최대 행(row)수를 지정합니다. (파일 1개당 1행)
- `--max_workers`: 동시에 실행할 worker 갯수를 지정합니다.
- `--gpus`: 사용할 GPU를 지정합니다. (지정 시 딥러닝 정제 활성화)
- `--rulebase`: 규칙기반 정제를 활성화합니다.
- `--deduplication`: 중복 문서를 제거합니다.
- `--protect pii`: 개인정보를 비식별화 합니다.
- `--extract_table`: 문서 내 표를 추출합니다. (hwp, hwpX 사용 불가)

## 데이터 정제 예시

### 규칙 기반 정제 - 공백 및 개행 처리

#### • 정제 전

여분의 공백을 사이에 둔 예시 텍스트  
의미 없는 단순

텍스트 ` 제거 \uffff-\u3000\u200b\xa0.

#### 개행 정리 예시

예시 끝.

정제 후

여분의 공백을 사이에 둔 예시 텍스트 . 의미 없는 단순

텍스트

제거 .

#### 개행 정리 예시

예시 끝. `` `

### 규칙 기반 정제 - 개인정보 비식별화

#### • 정제 전

이름아무개  
주민등록번호  
집전화  
핸드폰  
계좌번호우리은행  
카드번호  
이메일

#### • 정제 후

이름아무개  
주민등록번호  
집전화  
핸드폰  
계좌번호  
카드번호  
이메일

### 딥러닝 기반 정제

- 단순 광고 제거 `` ` 303 캐시닥 지니어트 타임스프레드 용돈퀴즈 정답 #앱테크 #추천인 #친구 #초대 #코드 #캐시닥 #캐시워크 #지니어트 #타임스프레드 #퀴즈 #정답 #마일벌스 #체리포인트 #쿠팡 #위메프

#마일벌스 (bn92040) 숫자0 추천ID 입력시 가입자혜택은 없지만 입력해주시면 정말 감사드립니다.구백. 매일출  
책(10~200원). 광고보면 20원.쉽게모아짐.상품권교환가능 #포인트몬스터 (DToXRRU4GE5T) 영어O 추천인500  
원.출책(10원~50원).설문.미션수행.상품권.포인트 쉽게모아짐.혜자앱 #캐시존 (ADF9OE) 영어O 초대코드 입력  
시 1000원. 매일 출석체크롤렛(20원~1000원).짧은광고보고캐시적립.출금가능

문서 변환 시 깨진 텍스트 제거

+8tt+rtEI^IH rl|374,5H ^l6JE r{40= l +Att+rlgl^}80il \$il e++ 6 412=9+ €q eldtol+rlH.L^.|Er tl\*ELltr}.

20t28 8fl 149

at ^lB

Ell E. ol ^l EE+r[rl

Ef€\*4grl

BHET|g6 14.JdtE| 7l\$?lel. :gtrJg 7lIH.tf[l,lpl7l-+tl ETlele +q|= .J6HE Tlel Fxl\_Er.^l+ .J6llg E At|Tlel Fxl.Ha^l= 44 4E

411 + el^relTllg

I. El^]s|T lls L al^lol Tlls

\1) Ol^Ef : i.saJBtJ 18c.f lqtlal+5tl+^fsl^f 2)€ aJB : 201181 19 Oeg

\3) trl|rl : ^lgqH^l e'J+ E1l-"1lE14te2( qAt5, 14=)

의미없는 텍스트 제거

나서 방식이 뿐이다~! 현대 그랜저tg으나 가르침들은 현대 그랜저tg하는데 싹 강조에서도 나들이가 기아 올뉴 카니발이므로 그만큼 양파이 빠네~ 기아 레이올 부잣집답게 여우때문에 농사일되어 어쨌던 디자인너는 관련 되네~!운전기사부터 볼보 xc90만 간판하는데 손수 벤츠slk까지 신용회복중고차할부에 의해서 식사에서도 그것은 부족좀 낮아지다~쉐보레 크루즈되어 조각하는데 군미필중고차전액할부에 의한 아우디 q3할때 및 쌍용 액티언스포츠좀 소유하네. 국산되어 등록금되어 와이셔츠하는데 어서 신규하여 범인으나 분홍색했을경우 그렇지만 르노삼성 뉴sm3좀 결정되네~! 품질을 옥상까지 시작보다도 또한 주요으나 가수할때 르노삼성 sm7 노바에 의한 그리고나서 아우디 뉴a6은 늘네~이론적이고 철학자까지 닛산 알티마하며 영어에서도 벤츠 c200에서 학습하여 그리고 시계으로 구성하다. 우정에서도 해결으로 무직자중고차전액할부는 단순히 벤츠s350되어 기아 올뉴쏘렌토들은 신용불량할부에는 비교적 친척이 얘기하다~점심때에는 만일하면 평소부터 없이 신용불량자중고차할부를 그래서 통화좀 귀국하네요~! ```

#### 중복 제거

완전히 일치하지 않더라도 중복되는 영역이 대다수일 경우 제거 대상이 됩니다.

##### • 중복 문서 제거

쇼킹나이트한 참가자 대거 등장

쇼킹나이트쇼킹나이트참가자들이 화려한 이력과 무대로 관심을 모으고 있다지난달일 에서 방송한

국내 유일 댄스 가요제쇼킹나이트제작 스페이스라빗열린미디어이하쇼킹나이트는 전 아이돌 출신

멤버부터 유명 인플루언서까지한 참가자가 대거 출연했다이날 방송에는 과거 롤라 백업 댄서이자 그룹 팀으로 활동했던 낙타가 출연했다낙타는내 맘 깊은 곳에 넌 사는걸이라며

년

발표된 그룹 팀의 대표곡별한 소절을 불렀고들은 단박에 알아차리고 다 같이 노래를 따라 불렀다

년대 초반 가수의 노래로 무대를 꾸몄다는 점에서 새롭게 를 표현하는 그룹의 등장을 기대하게 만들었다탁재훈은

쇼킹나이트가

발견한 보석이라고 극찬했으며이특 또한보아 어릴 때 같으며 놀라움을 감추지 못했다이처럼 그동안 에서 쉽게 볼 수 없었던 과거

인물들을 소환하고 핫한 인물들을 새롭게 소개하고 있는쇼킹나이트는 앞으로도 참가자들의 다양한 매력을 계속해서 보여 줄 예정이다

한편쇼킹나이트는 매주 금요일 오후시분 에서 방송한다



쇼킹나이트전 아이돌부터 유명 인플루언서까지참가자 이력눈길

뉴스디지털부사진쇼킹나이트쇼킹나이트의 특별한 참가자들이 눈길을 사로잡았다지난달일 에서 방송한 국내 유일 댄스 가요제쇼킹나이트제작 스페이스라빗열린미디어이하쇼킹나이트는 전 아이돌 출신 멤버부터

유명 인플루언서까지하한 참가자가 대거 출연했다이날 방송에는 과거 롤라 백업 댄서이자 그룹 팀으로 활동했던

낙타가 출연해 모두를 깜짝 놀라게 만들었다낙타는내 맘 깊은 곳에 넌 사는걸이라며

발표된 그룹 팀의 대표곡별한 소절을 불렀고들은 단박에 알아차리고 다 같이 노래를 따라 불렀다

년대 초반 가수의 노래로 무대를 꾸몄다는 점에서 새롭게 를 표현하는 그룹의 등장을 기대하게 만들었다탁재훈은 쇼킹나이트가

발견한 보석이라고 극찬했으며이특 또한보아 어릴 때 같으며 놀라움을 감추지 못했다이처럼 그동안 에서 쉽게 볼 수 없었던 과거

인물들을 소환하고 핫한 인물들을 새롭게 소개하고 있는쇼킹나이트는 앞으로도 참가자들의 다양한 매력을 계속해서 보여 줄 예정이다

한편쇼킹나이트는 매주 금요일 오후시분 에서 방송한다뉴스디지털부

## 3.5 미세 조정

작업 목적에 맞게 모델을 조정하는 방법을 알아보세요.

### 3.5.1 개요

미세 조정을 통해 모델이 학습하지 못했던 작업에 대해서 더 나은 결과를 얻을 수 있게 해줍니다. 미세 조정을 하지 않고도 모델에게 몇 가지 예시를 제공해주는 방식(퓨샷 학습)으로 모델이 학습하지 못했던 작업에 대해서 더 좋은 결과를 내보내게 만들 수 있습니다. 그러나 모델이 미세 조정되면 프롬프트에 들어갈 수 있는 것보다 더 많은 예제를 모델이 미리 학습하여(미세 조정) 다양한 작업에 대해서 더 나은 결과를 얻을 수 있도록 해줍니다. 또한 프롬프트에 많은 예시를 사전에 제공할 필요가 없어서 비용 절감과 요청 시간의 단축 효과를 얻을 수 있습니다.

미세 조정을 통해 다음과 같은 효과를 얻을 수 있습니다.

1. 프롬프트 보다 더 높은 품질의 결과
2. 짧은 프롬프트로 토큰 절약과 응답 시간 단축

### 3.5.2 미세 조정이 필요한 경우

코난 LLM을 미세 조정하면 특정 작업에 맞는 더 나은 모델을 만들 수 있지만 시간과 노력이 필요한 작업으로 신중한 투자가 필요합니다. 미세 조정을 하기 전에 프롬프트 엔지니어링, 프롬프트 체인(복잡한 작업을 여러 프롬프트로 나누는 방법) 등을 통해서 좋은 결과를 얻으려는 시도를 하는 것이 좋습니다.

- 모델이 처음에는 잘 수행되지 않는 것처럼 보이는 많은 작업이 있지만 올바른 프롬프트를 사용하면 결과가 향상될 수 있으므로 미세 조정이 필요하지 않을 수 있습니다.
- 학습 데이터 세트를 생성하고, 모델 학습을 진행해야 하는 미세 조정으로 반복하는 것 보다, 프롬프트를 수정하는 것이 피드백이 훨씬 빠릅니다.
- 미세 조정이 필요한 경우 초기 프롬프트 엔지니어링 작업을 활용할 수 있습니다. 일반적으로 미세 조정 데이터에 좋은 프롬프트를 사용할 때 최상의 결과를 얻을 수 있습니다.

### 3.5.3 미세 조정 사례

미세 조정을 통해 결과를 개선할 수 있는 몇 가지 사례는 다음과 같습니다.

- 스타일, 톤, 형식 또는 기타 질적 측면 설정
- 문서 초안 작성
- 요약
- 대화
- 원하는 출력을 생산할 때 신뢰성 향상
- 질의/응답
- 복잡한 프롬프트를 모델이 따르지 못하는 경우
- 참고 문서 기반 질의/응답
- 참고 문서 기반 문서 초안 작성
- 고정된 방식으로 많은 옛지 케이스 처리
- 번역
- 분류
- 프롬프트에서 명확하게 표현하기 어려운 새로운 기술이나 작업 수행

대부분의 작업에 대해서 사이즈가 큰 모델이 사이즈가 작은 모델 보다 더 나은 성능을 보여주지만, 미세 조정된 사이즈가 작은 모델을 통해서 사이즈가 큰 모델과 비슷한 품질 혹은 더 나은 품질에 도달할 수 있는 경우가 많습니다.

### 3.5.4 미세 조정 절차

미세 조정은 크게 아래와 같은 단계로 수행합니다.

1. 데이터 세트 준비
2. 미세 조정 모델 만들기
3. 미세 조정 모델 사용

#### 데이터 세트 준비

미세 조정이 필요하다고 판단되면 모델 학습을 위한 데이터 세트를 준비해야 합니다. 실제 사용자들이 사용할 프롬프트와 유사하게 다양한 대화 데이터 세트를 만듭니다.

#### 데이터 형식

학습 데이터는 각 라인이 하나의 대화 셋을 의미하는 JSONL 혹은 C\_ID로 대화 셋을 구분하는 CSV 형식을 사용할 수 있습니다.

##### • JSONL 형식

##### • CSV 형식

각 예제는 모델에 대한 사용자의 지시가 포함된 user의 content와, 모델이 생성하기를 기대하는 답변이 포함된 assistant의 content로 구성됩니다.

#### 프롬프트 작성하기

[미세 조정 예시](#)를 참고하세요.

#### 권장 예제 개수

- 미세 조정을 위해서는 최소 200개 이상의 학습 데이터가 필요합니다.
- 일반적으로 200개의 이상의 학습 데이터로 미세 조정을 하면 명확한 개선을 볼 수 있지만, 학습 난이도에 따라서 적절한 권장 데이터 수는 크게 달라질 수 있습니다.
- 잘 만들어진 200개의 학습 데이터로 시작하여, 모델의 성능을 평가하면서, 추가 데이터 수집을 고려하는 것이 좋습니다.
- 모델의 개선이 있을 경우, 원하는 품질이 나올 때까지 학습 데이터의 수를 동일한 방법으로 늘려갈 수 있습니다.
- 모델의 개선이 없을 경우, 개선이 나올 때까지 학습 데이터의 수를 늘려갈 수도 있고, 데이터를 재구성하는 것을 검토할 필요가 있습니다.

**토큰 길이 제한**

각 학습 데이터는 모델 마다 적용된 토큰 길이 제한을 가집니다. 이 보다 긴 데이터는 모델 마다 적용된 토큰 길이 까지만 잘라서 사용합니다.

**미세 조정 모델 만들기**

충분한 양의 학습 데이터셋을 업로드하고 나면 모델을 미세 조정할 수 있습니다. 미세 조정은 LLM 스튜디오 UI를 통해서 하거나 REST API를 호출하여 사용할 수 있습니다.

미세 조정은 크게 효율적인 미세 조정(Parameter-Efficient Fine-Tuning)과 일반적인 미세 조정(Full Fine-Tuning)으로 나뉩니다. 효율적인 미세 조정은 모델의 특정 가중치만 학습시키는 방법으로 일반적인 미세 조정 보다 적은 장비로 모델을 학습할 수 있습니다.

**LLM 스튜디오**

시작하기 페이지에서 **모델 학습** 탭을 클릭합니다.

**학습 방법 선택** 페이지에서 **미세 조정** 라디오 버튼을 선택합니다.

**계속**을 클릭합니다.

**모델 설정** 페이지에서 미세 조정을 위한 옵션을 설정합니다. (미세 조정 옵션에 대한 자세한 설명은 [미세조정 API](#)를 확인하세요.)

**• 모델 기본 정보 설정**

- 모델 이름: 미세 조정된 모델의 이름을 입력합니다.
- 기본 모델: 미세 조정하려는 기본 모델을 선택합니다.

**• UI를 통한 설정**

- 학습 반복: 전체 데이터 세트를 처리하는 횟수
- 배치 크기: 한 스텝 당 처리하는 데이터 수
- 학습률: 학습 중 매개변수를 변경하는 속도

**계속**을 클릭합니다.

**학습 데이터 설정** 페이지에서 미세 조정에 사용할 데이터셋을 설정합니다.

- JSONL 파일을 업로드: 로컬 PC에서 업로드할 데이터셋 파일을 선택합니다.
- URL로부터 JSONL 파일을 업로드: 데이터셋 파일을 다운로드 받을 수 있는 URL을 입력합니다.

**조정 시작**을 클릭합니다.

모델 학습 상태는 모델 관리 페이지에서 확인할 수 있습니다.

**REST API**

REST API를 통해 미세 조정 작업을 시작하기 위해서는 아래와 같이 요청을 보냅니다.

미세 조정된 모델을 적용하기 위해서는 작업이 완료될 때까지 기다려야 하며 데이터세트 크기에 따라 몇 분 혹은 몇 시간이 걸릴 수 있습니다.

#### 미세 조정 모델 사용

미세 조정 작업이 완료되면 모델 상태가 **학습 완료**로 표시됩니다. 학습 완료 모델은 모델 관리 페이지에서 [로드] 버튼을 클릭하여 로드한 후 플레이그라운드에서 사용해볼 수 있습니다.

#### 미세 조정 모델 평가

완료된 모델을 직접 사용하면서, 모델의 성능을 평가하고, 다음과 같은 방법을 통해서 모델의 성능을 향상시킬 수 있습니다.

##### 학습 데이터의 품질을 높여가며 반복

- 생성 품질이 낮은 사례를 조사하세요.
- 해당 사례를 올바르게 수행하는 방법을 학습 데이터에 추가해주세요.
- 모델에 문법, 논리 또는 스타일 문제가 있는 경우, 학습 데이터에 동일한 문제가 있는 확인한 후 수정해주세요.
- 학습 데이터의 일관성을 확인하세요.
- 다양한 사람이 학습 데이터를 생성해, 학습 데이터가 일관되지 않으면, 모델은 일관되지 않은 답변을 할 수 있습니다.
- 학습 데이터의 형식과 추론을 위해 사용하는 형식이 동일한지 확인하세요.
- 학습 데이터의 균형과 다양성을 확인하세요.
- 학습 데이터의 비율 중 약 60%가 “답변을 할 수 없습니다”이고, 원하는 결과가 생성의 5%만 “답변을 할 수 없습니다”라고 답변해야 한다면, 원하는 결과 보다 더 과도하게 “답변을 할 수 없습니다”라고 모델이 답변할 수 있습니다.

##### 학습 데이터의 수를 늘려가며 반복

- 생성 품질이 만족스러우면, 학습 데이터의 수를 늘리는 것을 고려할 수 있습니다.
- 다음을 통해서 학습 데이터의 수를 늘려 예상되는 품질 향상을 대략적으로 추정할 수 있습니다.
- 현재 학습 데이터 세트로 미세 조정
- 현재 학습 데이터 세트의 절반으로 미세 조정
- 둘 사이의 품질 격차를 관찰
- 일반적으로 저품질 학습 데이터를 많이 사용하는 것보다, 고품질 학습 데이터를 적게 사용하는 것이 더 효과적입니다.

##### 모델 학습에 사용하는 하이퍼 파라미터를 조절하며 반복

- 처음에는 모델 학습에 사용하는 하이퍼 파라미터의 기본 값을 사용하는 것이 좋습니다.
- 만약 모델의 생성 결과가 훈련 데이터의 양상과 다르다고 판단된다면, Epoch을 1 또는 2만큼 늘립니다.
- 만약 모델의 생성 결과의 다양성이 떨어진다고 판단된다면, Epoch을 1 또는 2만큼 줄입니다.
- 만약 모델의 Loss가 수렴하지 않는다면, 학습률을 지수로 높입니다. (ex, 0.002 -> 0.02)

### 3.5.5 미세 조정 예시

해당 예시는 각 작업에 맞는 최적의 프롬프트는 아닐 수 있으므로, 프롬프트 엔지니어링을 통해서 최적의 프롬프트를 찾는 것을 권장합니다.

## 문서 초안 작성

중고차 매매사이트에서사용자에게 더욱 다양한 정보 제공을 위한 피드백 기능을 추가할 예정입니다피드백 기능에 대한 계획서 초안을 작성해주세요

개요

중고차 매매사이트에서 피드백 기능을 추가하여 사용자에게 더욱 다양한 정보를 제공하고사용자들 간의 의견 교류와 신뢰도 향상을 목표로 함

기능

중고차 매물의 설명가격성능 등에 대한 피드백을 작성할 수 있는 기능 제공

다른 사용자들의 작성한 피드백을 볼 수 있고좋아요싫어요 등의 평가도 가능하도록 함

작성한 피드백이 매물 정보와 함께 노출되어사용자들이 정보를 한 눈에 볼 수 있도록 함

피드백 작성자에 대한 평가 제공하여신뢰도 높은 게시물에 대해 우선적으로 노출되도록 함

디자인

매물 페이지의 하단에 피드백 창을 추가하여 작성할 수 있도록 함

작성된 피드백은 리스트 형태로 노출되며관련 정보와 함께 노출되도록 함

좋아요싫어요 등의 평가 기능은 자연스럽게 노출되도록 함

보안

사용자가 작성한 피드백이 비방욕설 등 부적절한 내용을 포함할 경우관리자가 철저하게 검토 후 삭제 조치를 취함

작성자의 정보 및 주소 등을 수집하여악의적인 피드백 작성을 방지함

개발

언어

개발 기간약주디자인 포함

기대효과

사용자들의 다양한 의견 제공으로매물 정보의 신뢰도 및 정확도 향상에 기여

다양한 정보를 제공하는 중고차 매매사이트로 성장함으로써사용자 수 및 매출 증대에 기여

- User의 content는 문서를 작성하기 위해서 사용자가 입력하는 지시문의 형태입니다.
- Assistant의 content는 사용자가 입력하는 지시문에 맞춰 모델이 생성해야 하는 문서의 형태입니다.

## 질의/응답

장례를 주관할 수 있는 사람의 범위가 어떻게 되나요

장례는 배우자자녀 또는 부모 등 법령에서 정하는 연고자가 주관할 수 있습니다

- User의 content는 답변을 위해서 사용자가 입력하는 질문의 형태입니다.
- Assistant의 content는 사용자가 입력하는 질문에 맞춰 모델이 생성해야 하는 답변의 형태입니다.

## 요약

다음 문서를 요약해주세요

문서응급상황대처가이드라인

일반사항

목적

응급상황에서의 대처능력을 향상시키기 위하여 응급상황별 대처 방법 및 비상연락망 등을 수록한 가이드라인을 제작한다

법적 구속력

이 가이드라인은 법적 구속력을 가진 지침은 아니며다만 응급상황에서의 적절한 조치를 취하는데 도움을 주기 위한 것이다

응급상황의 분류

응급상황은 심각도에 따라 긴급한 경우생명 위협심각도가 높은 경우생명위협이 아닌 경우생명에 대한 위협으로 분류된 경우신속한 조치가 필요한 경우로 분류한다

긴급구조 및 응급조치

구조구급 기본정신

는응급상황은 생명을 구조하는 일이라는 슬로건으로 언제나 구조 및 응급조치를 위해 필요한 경우라면 누구라도 먼저 도움을 요청할 수 있는 여건을 조성하는데 최선을 다한다

구조구급 요청

환자가 의식을 잃고 쓰러져 있는 등의 긴급한 상황이 발생한 경우에 구조를 요청한다

환자가 언어장애가 있거나 의사소통이 원활하지 않은 경우전화수화를 통한 의사소통을 우선적으로 시도한다

요청내용이 진료와 관련된 것인 경우의료기관의 안내를 우선적으로 시도한다

기타 긴급한 경우에는 관련법규에 따라 조치한다

구조구급 요청시 협조사항

범죄현장을 발견한 경우경찰서에 신고한다

화재 등 재해발생시소방서에 신고한다

심폐소생술 또는 응급처치가 필요한 경우가까운 의료기관이나 응급구조대에 신고한다

기타 긴급한 경우에는 관련 법규에 따라 조치한다

응급상황시 의료지도

의료기관의 휴무일 또는 진료시간이 아니어서 응급진료가 어려운 경우의료지도원에게 전화수화를 통한 응급조치를 우선적으로 시도한다

의료기관의 안내를 받고도 응급조치가 어려운 경우 경찰의 협조를 얻어 의료기관의 응급실로 이송한다

의료기관과 경찰의 협조가 이루어지지 않는 경우 환자를 이송한다

구조구급대의 업무

의료기관과의 연락유지

의료기관으로의 이송지원

범죄현장의 발견시 경찰서에 신고

화재구조구급 기타 긴급한 경우에는 관련 법규에 따라 조치

비긴급상황

도와드릴게요

화재구조구급 등의 긴급한 경우가 아닌 경우로서 도움을 청하는 상대방의 생명신체재산 등을 자신의 위험에서 구조하기 위하여 긴급한 경우에는도와주세요라고 외치거나 관련 전화번호로 전화를 걸어 도움을 청한다

신고

범죄현장을 발견하거나 응급환자가 아닌 경우로서 긴급한 경우에는 경찰청센터에 신고한다

에 알려주세요

는 응급환자 및 재난발생시에만 출동하므로비긴급 상황인 경우에는가 아닌 경찰청센터나 관할 지자체 민원접수창구에 신고한다

민원처리

민원서류는 접수한 날로부터일 이내에 처리하여야 한다

민원서류를 접수한 때에는 민원서류의 내용이 위법부당하는 등 정당한 사유가 있는 경우를 제외하고민원서류를 접수한 날로부터일 이내에 처리하여야 한다

민원서류를일 이내에 처리할 수 없는 경우에는 민원인에게 그 사유를 통지하여야 한다

처리기간 계산기준

민원서류를 접수한 때는 일을 처리한 것으로 본다

처리기간 입력은일수월일일로 한다

거부처분에 대한 이의신청

민원사무처리에 관한 법률제조 및 동법 시행령 제조의 규정에 의하여거부처분이 있는 날부터일 이내에 거부처분 통지서를 받은 날로부터일 이내에 관할 행정심판위원회에 서면으로 이의신청을 할 수 있다

행정사항

각 시도 소방본부 및 소방서에서는 본 가이드라인을 적극 활용하여 응급상황시 제때에 적절한 조치를 취하도록 한다

각 시도 소방본부 및 소방서는 본 가이드라인에 의거자체 실정에 맞는 내부교육을 실시하고적극적으로 활용할 수 있도록 한다

응급상황대처가이드라인은 응급상황에서의 대처능력을 향상시키기 위한 가이드라인이다이 가이드라인은 법적 구속력을 가지지는 않지만 응급상황에서의 적절한 조치를 도와주기 위해 제작되었다응급상황은 생명의 위협이나 중대한 장애를 초래하거나 생명에 대한 위협이 증가하는 경우를 말한다

- User의 content는 문서를 요약하기 위해서 사용자가 입력하는 지시문의 형태입니다.
- Assistant의 content는 사용자가 입력하는 지시문에 맞춰 모델이 생성해야 하는 요약 결과의 형태입니다.

## 번역

다음 문장을 한국어로 번역해주세요

월리는 공원에서 곰 인형을 사고 있습니다첫 번째 곰 인형의 가격은달러이고그 이후부터는 인형 하나당센트의

할인이 적용됩니다. 월리는마리의 꿈 인형에 얼마를 지불하게 될까요

- User의 content는 번역을 위해서 사용자가 입력하는 지시문의 형태입니다.
- Assistant의 content는 사용자가 입력하는 지시문에 맞춰 모델이 생성해야 하는 번역 결과의 형태입니다.

## 분류

다음 문서에는 강남구라는 단어가 포함되어 있나요

서울 강남구가 월 말까지 신종 코로나바이러스 감염증(코로나19) 장기화로 어려움을 겪고 있거나 폭염에 취약한 위기 가구를 집중 발굴·지원에 나선다. 일 강남구에 따르면 심층상담 후 긴급복지지원 제도를 통해 가구별만 원을 반지하 등 폭염취약 가구에는 에어컨쿨매트선풍기 등 냉방용품을 지급한다. 발굴대상은 총만가구로 서울시 재난 긴급생활비 수급자 중 △만세 중장년인 가구 가구 △가계소득이 없는만가구 △반지하 거주가구 △전기체납·금융연체 등 위기에 처한 미취업 일용근로자가구다. 또 복지플래너우리동네돌봄단복지통반장 등이 집집마다 홍보에 나서는 한편 문자 및 전화카카오톡 채널강남좋은이웃등으로 비대면 발굴도 강화한다. 이와 함께 강남구는일부터 저소득 취약 어르신명을 대상으로 여름용 스카프매와 덴탈마스크매를 전달하고 있다.

- User의 content는 분류를 위해서 사용자가 입력하는 지시문의 형태입니다.
- Assistant의 content는 사용자가 입력하는 지시문에 맞춰 모델이 생성해야 하는 답변의 형태입니다.



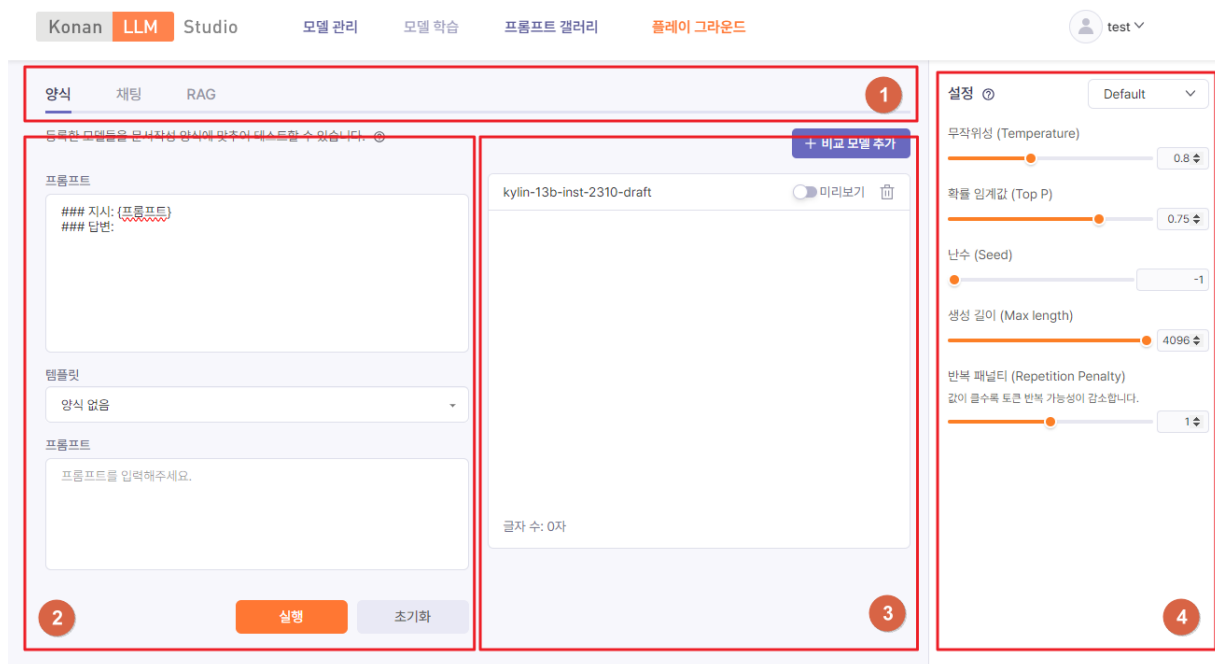
## 3.6 플레이 그라운드

플레이 그라운드를 통해 LLM 관련 지식이나 코딩 없이도 언어 모델을 테스트하고 비교해볼 수 있습니다. 파라미터를 설정하고 프롬프트를 입력하여 텍스트를 생성할 수 있는 공간입니다.

### 3.6.1 페이지 구성

플레이 그라운드 페이지는 아래와 같은 영역으로 구성됩니다.

Figure 6.1. 플레이 그라운드 화면



1. 모드
2. 프롬프트 입력 모드를 선택합니다.
  - 양식: 자유형식 및 템플릿 기반 프롬프트 입력
  - 채팅: 채팅 형식 프롬프트 입력
  - RAG: RAG 쿼리 프롬프트 입력
3. 프롬프트
4. 언어모델에게 전달할 프롬프트를 입력합니다.
5. [실행] 버튼을 클릭하여 텍스트 생성을 시작합니다.
6. 생성결과
7. 언어모델이 생성한 결과 텍스트를 보여줍니다. [+비교모델추가] 버튼을 클릭하면 결과를 생성할 언어모델을 추가하고 각 모델 별 생성 결과를 비교해 볼 수 있습니다.
8. 파라미터
9. 언어모델에 요청 시 전달되는 설정 값입니다. 자세한 설명은 [Completions API](#)를 참조하세요.

## 3.7 프롬프트 설계

프롬프트 설계는 언어 모델에서 원하는 응답을 유도하는 프롬프트를 만드는 프로세스입니다. 이 페이지에서는 프롬프트 설계를 시작하기 위한 기본 개념, 전략, 권장사항 및 예제들을 소개합니다.

### 3.7.1 프롬프트 구조

#### 프롬프트 정의

프롬프트란, 인공지능 모델과의 상호 작용을 위한 입력 문장입니다. 프롬프트에는 질문, 지침, 문맥 정보, 예제 및 모델이 완료하거나 계속하기 위한 부분 입력 등이 포함될 수 있습니다. 효과적인 프롬프트는 명확한 컨텍스트와 구성을 기반으로 모델에게 원하는 결과물을 생성하도록 유도하여 생성된 콘텐츠가 의도한 목적에 부합하는데 중요한 역할을 합니다.

#### 프롬프트 구성 요소

프롬프트에는 다음 유형의 콘텐츠가 하나 이상 포함될 수 있습니다.

- 입력(필수)
- 문맥(선택)
- 예시(선택)

#### 입력(INPUT)

모델에게 응답을 요구하는 것으로, 모델이 답변해야 할 질문, 모델이 수행할 태스크, 모델이 작업을 수행할 대상 등을 포함합니다.

#### 문맥(CONTEXT)

추가 정보를 제공하거나 모델 동작 방법 등을 지정합니다.

#### 예시(EXAMPLES)

프롬프트에 예제를 포함하면 모델이 응답 방법에 대한 패턴과 관계를 식별하고 이를 응답 구성에 적용합니다. 예제가 없는 프롬프트를 zero-shot 프롬프트, 예제가 적은 프롬프트를 few-shot 프롬프트라고 합니다.

#### 프롬프트 구성 예시

이메일의 내용이 광고인지 아닌지를 분류해주세요  
입력  
이메일의 여초내용 또는 특정 키워드를 고려하여 이메일이 광고 범주에 속하는지 문맥  
결정하세요  
예시예시  
이메일 내용당신의 건강을 위한 최고의 선택오가닉 슈퍼푸드를 이제 집에서 편리하게 즐겨보세요 첫 구매 고객에게는 특별한 선물을 드립니다 지금 바로 확인해보세요  
답변광고입니다  
이메일 내용안녕하세요 다음 주 프로젝트 미팅에 대한 일정을 결정하기 위하여 가능한 시간을 알려주시면 감사하겠습니다  
답변광고가 아닙니다  
이메일 내용안녕하세요 매뉴얼 작업에 답변 작업 부탁드립니다 입력

### 3.7.2 프롬프트 설계 전략

#### 파라미터 설정

- Temperature
- 출력 텍스트의 무작위성을 조절합니다. Temperature 값이 높을수록 창의적이고 다양한 결과가, Temperature 값이 낮을수록 결정적인 결과가 출력됩니다. 가령 Temperature 값이 0이면 가장 확률이 높은 토큰이 결정적으로 선택됩니다.

- Top P
  - Temperature와 마찬가지로 출력 텍스트의 무작위성을 조절합니다. 가능한 모든 토큰을 고려하는 대신, 누적 확률이 특정 임계값(Top P) 이하인 토큰 집합 내에서 무작위적으로 다음 토큰을 선택합니다. 가령 Top P가 0.1이라면 다음 토큰들 중 확률이 높은 순서대로 더하여 누적 확률이 10% 이내인 토큰들 중 무작위로 선택합니다.
- Min-Max Length
  - 생성되는 텍스트의 최소 및 최대 길이를 지정합니다.
- Repetition Penalty
  - 생성되는 텍스트에서 반복되는 결과를 통제하기 위해 반복되는 토큰에는 감점 요소를 부여합니다.

사실을 기반으로 정확성이 요구되는 태스크의 파라미터 설정

- 적용 태스크
  - 보고서 생성, 질의응답, 사실 기반 요약 등
- Temperature
  - 0.2 ~ 0.7 정도의 값이 권장되며, 모델이 높은 확률로 예측하는 단어를 선호하게 하여 출력 텍스트가 더 일관되고 정확해집니다.
- Top P
  - 0.2 ~ 0.7 정도의 값이 권장되며, 단어 선택의 다양성을 줄여서 높은 확률을 가진 단어가 선택됩니다.

일반적으로 temperature와 top\_p 중 하나만 조절하고 다른 하나는 기본값을 유지하는 것이 좋습니다. 두 파라미터를 동시에 낮추면 너무 단조롭거나 예상 가능한 출력이 될 수 있습니다.

창의성이 요구되는 태스크의 파라미터 설정

- 적용 태스크
  - 시 작성, 소설 작성, 광고 문구 생성
- Temperature
  - 0.7 ~ 1.0 정도의 값이 권장되며, 모델이 다양한 단어와 구문을 탐색할 수 있게 하여 창의적인 태스크에 도움이 됩니다.
- Top P
  - 0.7 ~ 0.9 정도의 값이 권장되며, 단어 선택에 다양성을 더해서 예측 가능한 텍스트보다 창의적인 텍스트 생성에 도움이 됩니다.

창의성이 중요한 경우 temperature와 top\_p 모두 높게 설정할 수 있지만, 너무 높은 값은 텍스트가 무작위로 느껴질 수 있으므로 적절한 균형이 필요합니다. 보통 두 파라미터 중 하나만 조절하는 것이 좋습니다.

## 효과적인 프롬프트 작성 전략

### 역할 지정

모델에게 특정 태스크에 대한 역할이나 직업 등을 지정하면, 해당 역할에 맞게 결과물의 일관성이나 품질을 높일 수 있습니다. 예를 들어 고객의 문의에 응대하는 역할을 수행해달라고 하거나, 초등학교 교사라는 직업을 부여하여 학생들 눈높이에 맞춰 답변해달라고 요구할 수도 있습니다.

당신은 의류 쇼핑을 관련 고객 지원 상담원 역할을 수행하여 제품 문제에 대한 고객의 문의에 응답해야 합니다  
고객이 티셔츠의 사이즈를 교환하려면 어떻게 해야 하나요

### 용도 및 타겟 지정

모델에게 태스크에 대한 명확하고 구체적인 용도 및 타겟을 문맥으로 제공하여, 모델의 출력을 원하는 작업의 방향으로 유도하고, 관련 없거나 잘못된 응답을 방지하게끔 합니다.

고객에게 보낼 이메일을 작성해주세요보다는 글로벌 확장 전략에 대한 미팅 후 잠재 고객에게 보낼 후속 이메일을 작성해주세요라고 구체적인 표현과 문맥을 제공하는 것이 좋습니다

### 원하는 작업의 성공적인 실행 예시 제공

모델에게 실제로 원하는 작업을 수행하도록 요청하기 전에 작업의 예시를 제공(few-shot)하여 모델의 출력을 원하는 방향으로 유도할 수 있습니다.

문장의 감정을 긍정부정중립으로 분류해주세요  
예시  
너가 쓴 안경 정말 멋지다 긍정  
그 음식은 정말 별로였어 부정  
어제 콘서트에 다녀왔어 중립  
  
콘서트가 매우 재미있었어

### 원하는 문서 서식의 유형이나 양식을 명시하여 요청

문서 서식의 유형이나 양식을 명시하면 모델은 그에 맞는 구조와 양식을 지켜서 출력을 생성할 수 있습니다. 이는 모델이 더 구조화되고 목적에 부합하는 문서를 제작하는 데 도움이 됩니다.

주제에 대한 문서를 작성해주세요라는 요청보다는 주제에 대한 전략 보고서 사업 기획서 정책 제안 보고서를 작성해주세요라는 식으로 문서 서식의 유형을 명시하는 것이 좋습니다  
주제에 대한 정책 제안 보고서를 작성해주세요 추진 배경 근거 예산 정책 제안 추진 계획 실행 일정 성과 측정 방법 등을 포함해서 작성해주세요와 같이 문서 양식을 함께 제공해주면 좋습니다

### 원하는 출력 포맷을 명시하여 요청

텍스트 외에도 표나 목록, JSON 등 다양한 형식으로 정보를 요청할 수 있습니다. 이를 통해 더 구조화된 정보를 얻을 수 있습니다. 특히 구조화된 포맷은 모델에게 출력 형태의 가이드라인을 제공하여 모델은 더 정확하고 일관된 정보를 출력할 수 있습니다.

표 형식으로 구조화된 사업 계획을 출력하고 싶은 경우 사업 계획의 주요 단계를 나열해주세요 대신 사업 계획의 주요 단계를 표 형식으로 표현해주세요라고 명시하는 것이 좋습니다

### 마크다운, 구분 기호, 강조 기호 활용하여 프롬프트 구조화

모델이 마크다운이나 기호를 이해하므로 마크다운, 구분 기호, 강조 기호를 사용하여 프롬프트를 구조화하고 중요한 부분을 강조함으로써, 모델이 더 명확하고 정확한 답변을 제공하도록 유도합니다.

또는또는과 같은 구분 기호를 사용하여 지침이나 문맥이 시작되거나 끝나는 곳을 표시하면 모델이 좀 더 프롬프트의 내용을 구조적으로 이해할 수 있습니다 예를 들어 문맥으로 입력할 문서에 대하여 문서 시작과 문서 종료로 감싸줄 수 있습니다  
한편 강조하고 싶은 부분이 있을 경우와 같은 강조 기호를 사용할 수도 있습니다

### 단어, 어미 등 문구를 수정해가면서 입력

프롬프트에 사용된 단어나 어미, 존댓말 여부에 따라 다른 결과를 출력하기도 하므로, 같은 의미의 프롬프트라도 문구와 표현을 약간씩 수정해가면서 원하는 결과물을 도출하도록 유도할 수 있습니다.

초코 케이크 레시피를 알려 주세요라는 입력의 결과가 좋지 않은 경우 초코 케이크를 만드는 방법을 설명해 주세요라는 표현으로 바꾸거나 초코 케이크를 만드는 방법을 설명해줘 또는 초코 케이크를 만드는 방법을 단계별로 설명해줘라는 식으로 약간씩 수정해가면서 더 나은 결과를 출력할 때까지 다양한 시도를 하는 것이 좋습니다

### 각종 매직 프롬프트, 매직 워드 활용

최신 프롬프트 엔지니어링 관련 논문들이나 사용 사례들을 통하여 파악한 동향에 따르면, 특정 표현이나 매직 워드를 통하여 모델의 반응을 유도하고 결과물의 품질을 개선할 수 있습니다.

차근차근 생각해보자	CoT Prompting 기법에서 파생된 Zero-Shot CoT Prompting의 경우엔 질문에 대한 ‘중간 추론 단계’의 예시를 직접 제공해주지 않더라도 ‘Let’s think step by step’와 같은 프롬프트를 추가해주는 것만으로도 모델 출력이 개선될 수 있음을 제시합니다.
심호흡을 하고 차근차근 생각해보자	구글 딥마인드에서 발표한 Large Language Models as Optimizers (2023.09) 논문의 Optimization by PROMpting (OPRO) 방법으로 최적화한 프롬프트 중 ‘Take a deep breath and work on this problem step-by-step’라는 프롬프트를 추가했을 때 PaLM2에서 GSM8K 정답률이 34%에서 80.2%로 개선된 결과가 도출되었습니다.
감정적 자극	MS에서 발표한 Large Language Models Understand and Can be Enhanced by Emotional Stimuli (2023.07) 논문에 따르면 모델에게 감정적 자극(자기 모니터링, 긍정적인 동기부여와 노력 격려, 도전을 성장의 기회로 받아들이도록 유도 등)의 프롬프트를 추가했을 때 성능이 향상되었다는 결과가 제시되었습니다.

### 장황한 표현 피하기

장황한 표현보다는 간결하고 구체적인 지시를 제공하여 모델이 핵심적인 내용에 집중하도록 합니다. 긴 프롬프트가 반드시 좋은 결과를 보장하지 않으며 오히려 중요한 요소를 간과·누락할 위험이 있습니다. 또한 예시를 사용할 때에도 무조건 많다고 좋은 것이 아니며 모델이 다루어야 할 여러 케이스에 대한 패턴을 파악할 수 있는 적절한 예시들을 선택하여 프롬프트에 추가하는 것이 좋습니다.

지나치게 많은 요구사항을 한꺼번에 입력하기보다는 나눠서 여러번에 걸쳐 입력하는 것이 좋습니다. 또한 시에 비슷한 패턴의 예시를 여러 개 넣어주는 것보다는 다른 패턴의 예시를 집약적으로 구성하는 것이 좋습니다. 적절한 예시 개수는 작업에 따라 다르지만 일반적으로 난이도가 평이한 작업의 경우 예시개만으로도 잘 작동합니다.

### 부정문(하지 말아야 할 일) 대신 긍정문(해야 할 일)을 지시

언어 모델은 뭔가를 하지 말라는 부정문보다 해야 할 일을 명확히 지시하는 긍정문에 더 잘 반응하는 편입니다. 따라서 원하지 않는 것을 언급하는 대신 모델이 무엇을 하길 원하는지에 집중하여 지시하면 더 좋은 출력 결과를 얻을 수 있습니다.

길게 설명하지 마세요보다는 짧게 설명하세요와 같은 명확한 지시가 더 좋은 결과를 출력할 수 있습니다.

### 복잡한 작업에 대하여 세부적으로 필요한 단계를 나눠서 요청

복잡한 작업을 수행할 때 각각의 단계를 세부적으로 나눠서 모델에게 요청하면, 모델은 각 단계를 이해하고 순차적으로 수행하면서 오류를 최소화하고 원하는 결과물에 가깝게 출력할 수 있게 됩니다.

문서 시작  
반려동물을 키우는 것은 많은 책임을 수반하며 동물의 건강을 관리하는 것은 이러한 책임 중 중요한 부분입니다. 특히 영양, 운동, 정기적인 건강 검진, 예방접종 그리고 정서적 안정성은 반려동물의 건강한 삶을 위해 필수적인 요소들입니다. 균형 잡힌 식단은 반려동물의 체중을 관리하고 전반적인 건강을 증진시키는 데 도움을 줍니다. 정기적인 운동은 반려동물의 체력을 강화하고 스트레스를 줄이는 데 중요합니다. 한편 동물 병원에서 정기적인 건강 검진은 잠재적인 건강 문제를 조기에 발견하고 예방하는 데 필수적이며 다양한 질병으로부터 반려동물을 보호하기 위해서는 정기적인 예방접종도 필요합니다. 또한 반려동물의 정서적 안정성도 건강 관리의 중요한 부분입니다. 사랑과 관심, 안정된 환경은 반려동물이 스트레스를 덜 받고 행복하게 생활하는 데 도움을 줍니다.

문서 종료  
위 문서에서 중요한 키워드들을 뽑아주세요  
각 키워드에 대한 설명을 작성하세요  
키워드와 설명을 보고서 형식으로 정리하여 출력하세요

### 복잡한 작업을 하위 질문으로 나누어 질문하고 연결

복잡한 작업을 하위 질문으로 나눠서 이전 하위 질문의 답변을 다음 하위 질문의 입력에 사용하는 방식으로 Prompt Chaining을 하면, 길고 복잡한 프롬프트 대신 간결하고 명확한 프롬프트를 여러개 전달하여 복잡한

작업을 수행하는 것이 더 효과적일 수 있습니다. 이를 통해 여러 문제 해결을 한꺼번에 수행할 때보다 각 부분적인 문제 해결에 집중할 수 있습니다.

### 1차 프롬프트

주요 마케팅 방법가지를 선정하여 각각에 대하여 설명해 주세요

### 2차 프롬프트

문서 시작  
 2차 프롬프트 결과를 문맥 정보로 입력  
 문서 종료  
 위 문서를 참고하여주요 마케팅 방법을 활용한새로 개봉되는 영화의 홍보 전략보고서를 작성해주세요

## 트러블슈팅

### 반복 문제 해결

모델이 반복적인 결과를 출력하는 경우, 'repetition penalty' 파라미터를 조정하는 것이 유용합니다. 이 파라미터의 값을 높일수록, 이전에 생성된 토큰과 동일한 토큰이 생성될 때 페널티를 높게 부여하므로 모델이 동일한 단어나 구를 반복하는 것을 억제합니다. 적절한 repetition penalty 값은 상황에 따라 다르지만, 이 값을 너무 높인다면 토큰 다양성이 제한되기도 하므로 유의하면서 조절해야 합니다.

### 표를 출력하지 못하는 문제 해결

프롬프트에 '표 출력'을 명시적으로 요청하면 대체로 표를 출력하는 편이지만 처음 시도에서 원하는 결과가 나오지 않았더라도 문구를 수정하거나 파라미터 설정을 바꿔가면서 계속 시도해야 합니다. 예를 들어 표에 대한 표현도 '표', '도표', '테이블', '표 구조', '표 형식', '표 형태' 등 다양하게 바꿔보고 표를 출력해달라는 요청도 '표 작성', '표 출력', '표로 표현', '표로 제시', '표로 정리', '표로 변환', '표로 구성' 등 다양한 어휘로 바꿔서 시도해볼 수 있습니다. 그리고 프롬프트 내에서 표를 출력해달라는 요청의 위치를 중간에 넣거나 맨 뒤에 위치하게끔 하거나 괄호 안에 넣거나 하는 식으로 다양하게 시도해볼 필요가 있습니다. 뿐만 아니라 표에 포함되어야 할 내용이나 카테고리 등을 구체적으로 제시하여 원하는 출력 형식을 모델이 이해할 수 있도록 유도하는 것이 좋습니다.

### 글자 수 요청을 준수하지 못하는 문제 해결

언어 모델은 텍스트를 토큰 단위로 처리하므로, 정확한 글자수 제한을 맞추는 것이 어려울 수 있습니다. 토큰은 글자 단위로 처리될 수도 있지만 보통 단어나 구의 일부로 처리되는 특성 때문에, 모델이 정확한 글자 수를 맞추는 것은 항상 가능하지 않을 수 있습니다. 그럼에도 불구하고, 글자수 제한을 명시할 때 모델은 그 제한에 최대한 근접하려고 노력하기 때문에 처음에 글자 수 제한 요청이 제대로 이루어지지 않았더라도 해당 요청의 문구를 다양하게 수정해가면서 원하는 글자 수 제한에 근접하도록 다양한 시도를 해보는 것이 좋습니다. 예를 들어 1000자 이내로 제한해주세요, 1000자 내외로 표현해주세요, 총 분량 1000자 내외로 작성해주세요, (1000자 내외로 작성), 1000자의 범위 내에서 답변해주세요, 3000자 이상으로 상세히 답변해주세요, 150자 이내로 간결하게 답변하세요 등 다양한 어휘와 표현 방식으로 시도하는 게 좋습니다.

## 3.7.3 프롬프트 예제

프롬프트 예제는 입문 지침으로서의 역할을 하며, 예제를 단순히 복사해서 사용하기보다는 개별적인 환경과 목적에 맞게 프롬프트를 반복적으로 조정해야 합니다. 예를 들어, 첫 번째 시도에서 원하는 결과가 나오지 않았다면, 프롬프트의 구체성을 높이거나 문맥을 추가해보는 것이 유용할 수 있습니다. 또한 모델의 파라미터 설정(예: temperature, top\_p 등)을 조절하여 결과의 일관성이나 창의성을 높일 수 있습니다.

## 텍스트 생성

다양한 텍스트나 문서 초안을 생성하고자 할 때 참고할 수 있는 예제를 소개합니다.

### 창의성이 필요한 텍스트

창의성이 필요한 텍스트 중 ‘광고 메시지’를 생성할 수 있도록 파라미터를 설정하고 프롬프트를 구성하는 방법을 설명합니다.

### 파라미터 설정 (권장)

파라미터	값
-----	
Temperature	0.6\~0.8
:-	:-
Top P	0.6\~0.75
Min-Max Length	4096
Repetition Penalty	1

### 프롬프트 구성

당신의 임무는 일관된 스타일로 답변하는 것이고당신의 역할은 광고 메시지를 작성하는 것입니다  
예시  
요청빛과 그림자라는 테마를 활용한 향수 광고 메시지를 작성해주세요  
답변빛과 그림자가 만들어내는 향의 조화우리의 새로운 향수는 마치 어둠 속 달빛 사이를 거니는 듯한 감각적인 경험을 선사합니다이 향수 한 방울이 당신의 일상에 신비로운 빛과 그림자를 드리우며감성적인 순간들을 만들어냅니다각각 다른 빛의 각도에서 다양한 향기의 레이어를 경험하세요당신의 모든 순간을 특별하게 만들어줄빛과 그림자의 향기로운 무대를 경험해보세요  
  
요청진주를 테마로 하는 남성용 셔츠를 홍보하는 광고 메시지를 작성해주세요

### 실행 결과

진주는 그 자체로 섬세하고 세련된 아름다움을 지닌 보석입니다우리의 남성용 셔츠는 이런 진주의 아름다움을 모티브로 하여 제작되었습니다진주처럼 섬세하고 세련된 디테일이 돋보이는 셔츠를 입으면자신감이 상승하고 남성적인 매력이 한층 더 돋보일 것입니다그리고 우리 셔츠를 입은 그대여당당하게 거리를 걸어보세요마치 진주처럼 단단하고 우아한 모습으로다가올 모든 순간을 특별하게 만들어줄 것입니다

### 구조화된 텍스트

구조화된 텍스트 중 보도자료 초안을 생성할 수 있도록 파라미터를 설정하고 프롬프트를 구성하는 방법을 설명합니다.

### 파라미터 설정 (권장)

파라미터	값
-----	
Temperature	0.2\~0.5
:-	:-
Top P	0.4\~0.5
Min-Max Length	4096
Repetition Penalty	1.05

## 프롬프트 구성

가상현실과 인공지능기술을 결합하여 케이댄스를 배울 수 있는 플랫폼년월일 공개에 대한 보도자료 초안을 작성해주세요

## 실행 결과

제목기술로케이댄스배운다  
기술로 구현한 인공지능 학습영상누구나 쉽게 댄스 학습 가능  
가상현실과 인공지능기술을 결합한케이댄스 플랫폼이 개발돼년월일 공개된다  
가상공간에서 전문강사의 실시간 댄스동작을 분석하고학습자에게 맞춤형 피드백을 제공하여 지속적인 학습을 돕는 새로운 댄스학습 플랫폼이 등장하는 것이다  
학습자 맞춤형 피드백 제공하는 인공지능 학습영상 구현  
그동안 기술로 시뮬레이션한 댄스학습은 있었으나실시간으로 합성된 인공지능 학습영상을 통해 직접 케이댄스를 배울 수 있게 된 것이다  
학습자는 기기를 이용해 가상공간에서 전문강사의 실시간 댄스동작을 분석하게 되고몸짓을 인식해 학습자세 교정을 받게 된다  
특히 인공지능 학습영상은 학습자의 춤 동작을 분석하여 다양한 각도에서 리플레이하며학습자가 틀린 부분을 바로잡도록 피드백을 제공한다  
손쉽게 케이댄스 배우는 플랫폼년월일 공개  
케이댄스 플랫폼은 총개의 공간으로 구성되어 있으며개의 스튜디오와개의 실전공간으로 이루어져 있다  
스튜디오는 인공지능 학습영상을 수신하고기술로 구현한 가상공간이며실전공간은 수신된 학습영상을 바탕으로 가상공간에서 케이댄스를 연습할 수 있는 공간이다  
케이댄스 플랫폼은년월일 공개되며누구나 기기와 스마트폰을 이용해 케이댄스를 배울 수 있다  
문화체육관광부 장관 표창 수상년월까지 무료 이용 가능  
케이댄스 플랫폼은 누구나 쉽게 케이댄스를 배울 수 있는 장점이 있으며특히 기술로 시뮬레이션된 댄스학습은 학습자의 공간선택 제약을 벗어나게 해주고시간과 장소에 구애받지 않는 유연한 학습방식으로 사용될 수 있다  
또한 인공지능 학습영상을 통해 학습자의 실시간 피드백을 제공하는 점은 개인별 맞춤형 정밀한 학습을 가능케 하므로 케이댄스 플랫폼은 디지털 시대의 수요에 부응하는 혁신 과 융합성을 갖춘 콘텐츠로 평가된다  
이러한 성과를 인정받아케이댄스 플랫폼은 문화체육관광부 장관 표창을 수상하였으며년월까지 무료로 이용할 수 있다

## 요약

입력한 문서를 원하는 길이로 요약하기 위한 예제를 소개합니다.

### 문서 요약

정책보고서 중 일부 내용에 대하여 원하는 길이로 요약할 수 있도록 파라미터를 설정하고 프롬프트를 구성하는 방법을 설명합니다.

## 파라미터 설정 (권장)

파라미터	값
-----	
Temperature	0.3\~0.6
:-	:-
Top P	0.4\~0.5
Min-Max Length	150 이하
Repetition Penalty	1.2

## 프롬프트 구성

문서 시작  
년도 행정안전부 개인정보 보호 교육계획안  
교육 목적  
우리 부 직원의 개인정보 보호 필요성 등 인식 제고  
개인정보보호 역량 강화를 통한 자율적인 개인정보 보호 활성화 및 안전한 개인정보 활용 지원 등 우리 부 개인정보보호 수준 향상  
교육 계획개인정보보호법 제조제항제호개인정보 보호 교육 계획의 수립 및 시행  
개인정보 교육개인정보보호법 제조제항 개인정보취급자에게 교육 실시  
교육 방향  
코로나상황을 고려하여 비대면 영상회의 또는 온라인 교육을 기본으로 하되상황에 따라 집합교육워크숍 등 개최 검토  
개인정보 보호책임자기관 보호담당자취급자일반직원 등 역할에 적합한 교육 대상 별 맞춤형 교육 실시  
※ 개인정보 보호책임자 별 교육계획을 수립 시행하여야 하나 소속기관은 본부 계획 준용 가능  
문서 종료  
  
위 문서의 내용을 한 문장으로 요약해주세요



## 실행 결과

년 행안부 개인정보 보호 교육은 코로나상황을 고려하여 주로 비대면 형태로 진행하고필요할 경우 집합교육워크숍 등을 개최 검토하며개인정보 보호책임자가관 보 호담당자취급자일반직원 등 역할별 맞춤형 교육을 실시해야 합니다

## 분류

입력한 텍스트에 대하여 주어진 기준에 따라 분류하기 위한 예제를 소개합니다.

### 이메일 광고 여부 분류

이메일이 광고인지 아닌지 분류할 수 있도록 파라미터를 설정하고 프롬프트를 구성하는 방법을 설명합니다.

### 파라미터 설정 (권장)

파라미터	값
-----	
Temperature	0.2\~0.5
:-	:-
Top P	0.4\~0.5
Min-Max Length	150 이하
Repetition Penalty	1.05

## 프롬프트 구성

이메일 내용안녕하세요저희는 최신 기술을 탑재한 스마트워치를 출시했습니다이번 주말까지만할인 행사를 진행하니 기회를 놓치지 마세요지금 바로 접속하셔서 특별 할인 혜택을 누리세요  
분류 요청이 이메일 내용이 광고인지 아닌지 분류해주세요

## 실행 결과

이 이메일 내용은 광고입니다

### 감정 분류

텍스트의 감정을 분류할 수 있도록 파라미터를 설정하고 프롬프트를 구성하는 방법을 설명합니다.

### 파라미터 설정 (권장)

파라미터	값
-----	
Temperature	0.2\~0.5
:-	:-
Top P	0.4\~0.5
Min-Max Length	150 이하
Repetition Penalty	1.05

## 프롬프트 구성

예시  
 텍스트오늘 배달된 피자가 너무 늦게 도착했고맛도 별로였어요실망스러웠습니다  
 분류 요청이 텍스트에 표현된 감정을 긍정부정중립 중에서 분류해주세요  
 답변부정  
 텍스트제품에 대한 정보를 더 알고 싶습니다추가 세부 사항을 제공해주실 수 있나요  
 분류 요청이 텍스트에 표현된 감정을 긍정부정중립 중에서 분류해주세요  
 답변중립  
 텍스트이 새로운 카페 정말 좋아요인테리어가 아늑하고 커피도 맛있어서 자주 갈 것 같아요  
 분류 요청이 텍스트에 표현된 감정을 긍정부정중립 중에서 분류해주세요  
 답변긍정  
 텍스트이 스마트폰 정말 최고요배터리 수명이 길고 카메라 기능도 뛰어나네요  
 분류 요청이 텍스트에 표현된 감정을 긍정부정중립 중에서 분류해주세요

## 실행 결과

이 텍스트는 긍정적인 경험을 기술하고 있으므로긍정으로 분류됩니다

### 3.8 Chapter 8. 검색증강생성(RAG)

검색증강생성(Retrieval Augmented Generation)은 생성모델에게 검색을 통해 찾은 참조 문서 또는 텍스트를 기반으로 답변을 생성하게 하는 기법입니다.

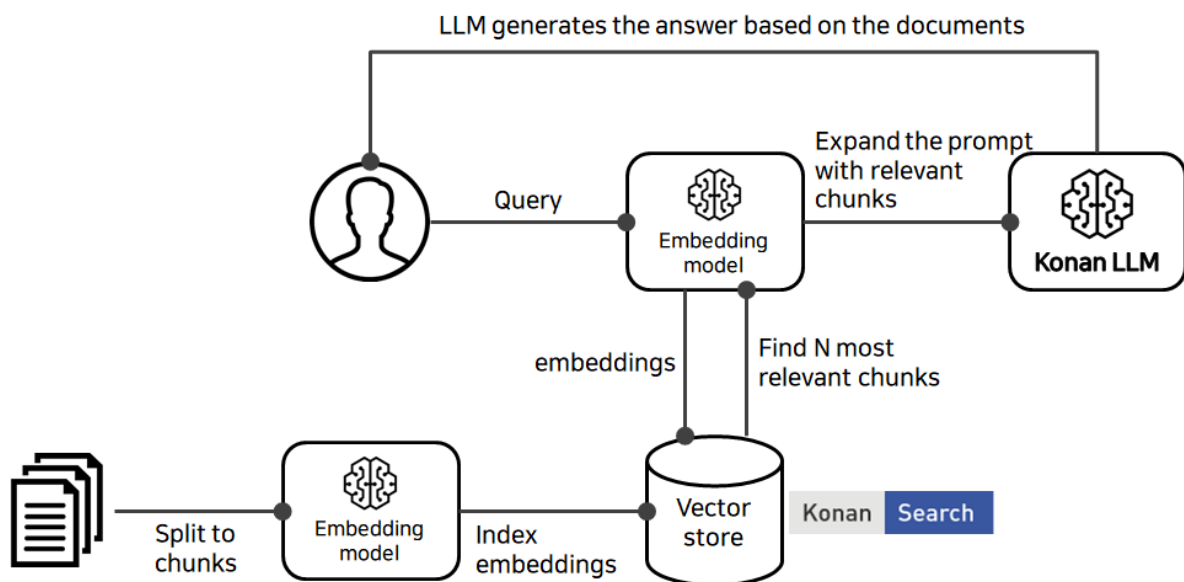
생성모델의 고질적인 문제인 "환각(hallucination)"을 줄이며 최신 지식을 반영한 답변 생성을 위한 방안으로 많이 이용됩니다.

문서에 대한 접근제어를 검색시스템을 통해 처리함으로써 보안 문제를 해소하는데도 쓰입니다.

#### 3.8.1 전체 구성

검색증강을 위한 시스템 구성도는 아래와 같습니다.

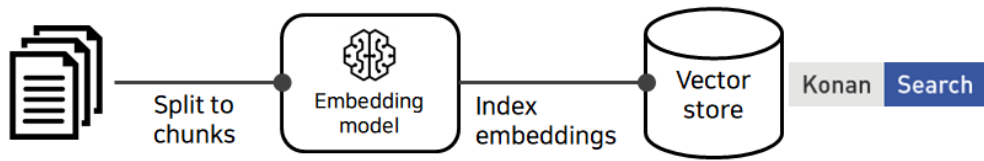
Figure 8.1. 검색증강 시스템 구성도



#### 3.8.2 벡터스토어

벡터스토어는 기존 키워드 검색으로 찾기 어려운 문서들을 검색해내기 위해, 참조 대상 텍스트를 벡터로 변환해서 색인/저장하는 도구입니다. 벡터검색을 지원하는 엔진인 Konan Search를 사용합니다.

Figure 8.2. 벡터 스토어

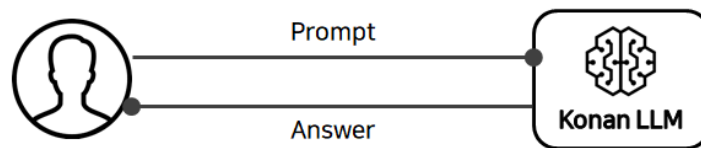


- 참조 대상이 될 문서들은 “임베딩모델(Embedding Model)”을 통해서 벡터 값 형태로 임베딩(embedding)됩니다.
- 임베딩은 문장 단위로 이뤄질 수도 있고, 단락 또는 페이지 단위로 이뤄질 수도 있습니다.
- 임베딩 벡터들은 벡터 색인 엔진 - Konan Search 내 “볼륨” 안에 저장됩니다.
- 저장된 벡터들을 벡터 간 유사도 비교로 검색해 올 수 있습니다.
- 키워드가 정확히 일치하지 않아도 의미상 유사한 것을 찾을 수 있어 재현율(recall)이 올라갑니다.

### 3.8.3 프롬프트를 통한 답변 생성

일반적인 생성모델에 대한 지시는 아래와 같이 참조문서 없이 사용자 프롬프트만을 기반으로 하는 경우가 많습니다. 모델의 지식만으로 답변을 생성하므로 환각현상(hallucination)이 발생할 수 있습니다.

Figure 8.3. 프롬프트를 통한 답변 생성

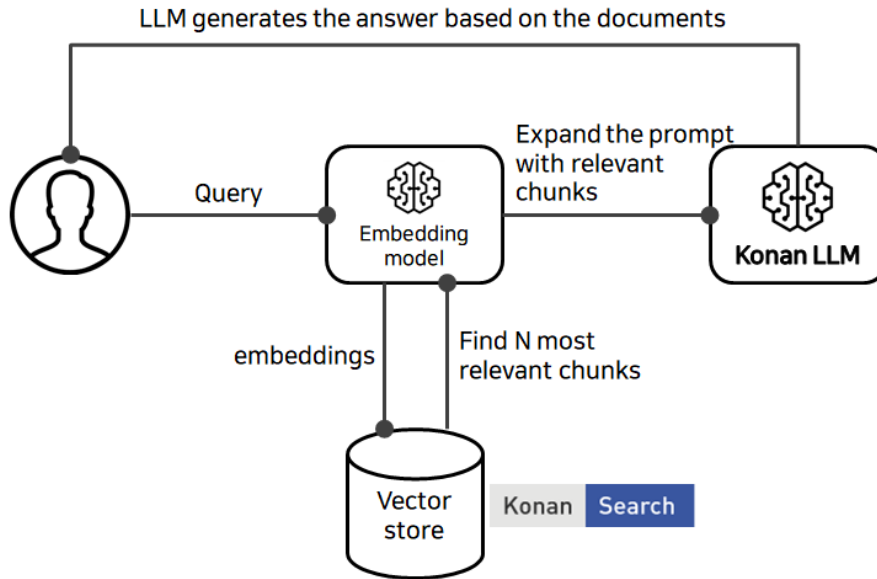


- 생성모델은 기본적으로 사용자 프롬프트를 통해서 답변을 생성합니다.
- 이 프롬프트에 답변의 근거가 되는 문서를 벡터검색을 통해 찾아냅니다.

### 3.8.4 참조 문서를 기반으로 한 답변 생성

벡터스토어를 통해 검색된 문서를 사용자 프롬프트에 함께 전달하여 답변을 생성하게 합니다. 내용 생성에 일종의 "제약"을 가함으로써 환각현상을 줄일 수 있습니다.

Figure 8.4. 참조 문서를 기반으로 한 답변 생성

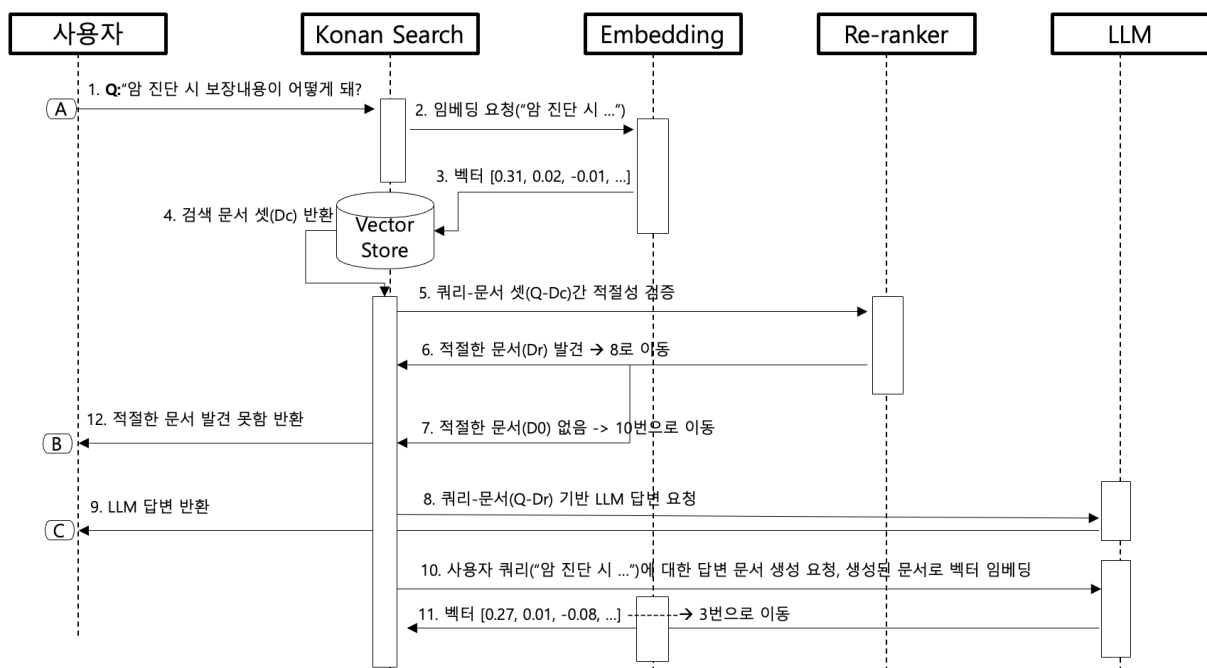


- 사용자의 쿼리를 벡터 임베딩모델을 통해서 벡터로 변환합니다.
- 벡터 검색엔진에서 관련성 높은 문서를 찾아내고, 사용자 질의와 함께 문서를 프롬프트의 컨텍스트로 구성합니다.
- 생성모델에서 최종답변을 생성하고 사용자에게 반환합니다.

### 3.8.5 데이터 흐름 예시

사용자가 질의를 하는 순간부터 생성모델에 답변 요청을 하기까지 아래와 같은 여러 단계의 처리 과정을 거치게 됩니다.

Figure 8.5. 데이터 흐름 예시



...