

# LLM 테스트 계획 및 결과 보고서 (LanChain 및 RAG를 활용한 의료LLM 개발)

3차 프로젝트 5팀(권성호, 남의현, 손현성, 이준석,  
이준배)

프로젝트명 :  
작성일 : 2025.06.02  
작성자 : 이준배

## 목차

1. 개요 **-2P**
2. 테스트 목적
3. 테스트 대상 모델
4. 테스트 방법
  - 4.1. 질문 데이터
  - 4.2. 평가 기준 **-3P**
5. 테스트 결과(모델별 응답 예시 및 분석)
  - 5.1. **GPT3.5**
  - 5.2. **GPT3.5(FT) -4P**
  - 5.3. **GPT4o**
  - 5.4. **KULLM3 (max 1024, sampling) -5P**
  - 5.5. **KULLM3 (max 256, no sampling)**
6. 결론 **-6P**
7. 향후 계획
8. 출처
  - 8.1. 파인튜닝 모델
  - 8.2. 허깅페이스 모델 **-7P**

## 1. 개요

본 보고서는 다양한 LLM(Large Language Model) 모델(GPT-3.5 Turbo, GPT-3.5 Turbo Fine-tuned, GPT-4o-mini, KULLM3)에 대해 테스트를 진행하고, 모델 성능, 환각 발생 여부, 응답 적절성 등을 분석한 내용을 담고 있다.

gpt계열은 open ai api키 사용 kullm 계열은 로컬에 다운 받아 사용

## 2. 테스트 목적

- 다양한 LLM 모델의 성능 비교
- Fine-tuning의 효과 확인
- KULLM3 모델의 토큰 수 및 샘플링 설정에 따른 성능 차이 분석
- 환각(Hallucination) 현상 방지 여부 확인

## 3. 테스트 대상 모델

모델명	특징
GPT-3.5 Turbo	OpenAi 기본 모델
GPT-3.5 Turbo(FT)	사용자 데이터로 파인튜닝한 모델
GPT-4o-mini	최신 경량형 OpenAi 모델
KULLM3 (max 1024, sampling)	한국어 특화 모델, 샘플링 사용
KULLM3 (max 256, no sampling)	동일 모델, 샘플링 비활성화, 짧은 응답 유도

## 4. 테스트 방법

### 4-1. 질문 데이터

- **gpt** 계열 - '당뇨병에 대해 알려줘' 라는 질문에 대한 답변 받기

- **kullm** - ‘고려대학교에 대해서 알고 있니?’ 라는 공식 문식 예시 질문에 대한 답변 받기

#### 4-2. 평가 기준

- 적절성: 질문에 대한 응답이 의미 있고 구체적인가
- 환각: 틀린 정보 또는 허구를 생성하는가
- 응답 길이: 너무 짧거나 불필요하게 긴 응답 여부
- 일관성: 유사 질문에 대해 유사한 응답을 하는가

### 5. 테스트 결과(모델별 응답 예시 및 분석)

#### 5-1. GPT-3.5

## 건강 질의응답 RAG 챗봇

? 궁금한 건강 관련 질문을 입력하세요:

당뇨병에 대해 알려줘

질문하기

 **답변:**

당뇨병은 혈당 조절이 원활하지 않아 혈당 수치가 지속적으로 높은 상태를 말합니다. 이는 인슐린 분비 부족이나 인슐린 작용 부족으로 인해 발생할 수 있습니다. 당뇨병은 혈당 조절을 위해 생활습관 개선, 약물 치료, 인슐린 주사 등 다양한 치료 방법이 있습니다. 만약 당뇨병 의심 증상이 있다면 즉시 의료진과 상담하여 적절한 치료를 받는 것이 중요합니다.

- 적절성: 당뇨병에 대한 설명, 발생원인, 치료방법과 같은 필수적인 요소들이 모두 담겨있다.
- 환각: 틀린 정보 또는 허구를 생성하지 않았다.
- 응답 길이: 응답 길이는 보통이며 실행 시간 또한 준수하다.
- 일관성: 유사 질문에 대해 유사한 응답을 한다.

## 5-2. GPT-3.5(FT)

### 건강 질의응답 RAG 챗봇

? 궁금한 건강 관련 질문을 입력하세요:

당뇨병에 대해 알려줘

질문하기

 답변:

당뇨병은 인슐린 분비 부족, 인슐린 작용 부족, 또는 두 현상이 동시에 나타나는 대사 이상으로 인해 발생하는 만성 질환입니다.

- 적절성: 당뇨병에 대한 설명, 발생 원인만 대답하였다
- 환각: 틀린 정보 또는 허구를 생성하지 않았다.
- 응답 길이: 응답 길이는 짧으며 실행 시간은 준수하다.
- 일관성: 유사 질문에 대해 유사한 응답을 한다.

## 5-3. GPT-4o

# 💬 건강 질의응답 RAG 챗봇

? 궁금한 건강 관련 질문을 입력하세요:

당뇨병에 대해 알려줘

질문하기

📌 답변:

당뇨병은 혈당(혈액 내 포도당) 수치가 비정상적으로 높은 상태를 말합니다. 주로 인슐린의 분비 부족이나 인슐린 저항성으로 인해 발생합니다. 당뇨병은 크게 제1형과 제2형으로 나눌 수 있습니다.

1. **제1형 당뇨병:** 주로 어린이나 청소년에서 발생하며, 췌장에서 인슐린을 생산하는 세포가 파괴되어 인슐린이 부족해지는 상태입니다. 이 경우 인슐린 주사가 필요합니다.
2. **제2형 당뇨병:** 성인에서 주로 발생하며, 인슐린 저항성이 주요 원인입니다. 초기에는 인슐린이 정상 또는 증가된 상태일 수 있지만, 시간이 지남에 따라 인슐린 분비가 감소하게 됩니다. 생활습관 개선, 약물 치료, 인슐린 요법 등이 필요할 수 있습니다.

당뇨병의 주요 증상으로는 잦은 소변, 갈증, 체중 감소, 피로감 등이 있으며, 장기적으로는 심혈관 질환, 신장 질환, 신경 손상 등의 합병증을 유발할 수 있습니다. 따라서 조기 진단과 적절한 관리가 중요합니다. 식이요법, 운동, 혈당 모니터링 등이 치료의 핵심입니다.

- 적절성: 당뇨병에 대한 설명, 발생 원인, 당뇨병 유형, 증상 등 여러 요소를 대답하였다.
- 환각: 틀린 정보 또는 허구를 생성하지 않았다.
- 응답 길이: 응답 길이는 길며 자세하다. 실행 시간은 준수하다.
- 일관성: 유사 질문에 대해 유사한 응답을 한다.

## 5-4. KULLM3 (max 1024, sampling)

```
temperature=0.8
60m 5.5s Python
c:\Users\Playdata\Desktop\3차_프로젝트\project3\Lib\site-packages\tqdm\auto.py:21: TqdmWarning: IPProgress not found. Please update jupyter and ipywidget
from .autonotebook import tqdm as notebook_tqdm
Loading checkpoint shards: 100%|██████████| 5/5 [00:00<00:00, 23.57it/s]
The attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may observe unexpected be
고려대학교는 대한민국의 서울특별시에 위치한 사립
```

응답이 60분이 넘어가도 나오지 않았다. 적절한 응답을 하는거 같지만  
응답 시간이 너무 길어 의미가 없다고 판단하고 중단하였다. (max  
token을 줄이고 sampling기능을 제거하여 재시도)

### 5-5. KULLM3 (max 256, no sampling)

```
[2] 30m 21.6s
.. Loading checkpoint shards: 100%|██████████| 5/5 [00:00<00:00, 245.73it/s]
Some parameters are on the meta device because they were offloaded to the disk and cpu.
네, 고려대학교에
```

마찬가지로 응답이 30분이 지나도 나오지 않았다. kullm 사용은 힘들 거  
같다고 판단 되었다.

## 6. 결론

- **gpt** 계열 : GPT3.5(FT) – GPT3.5 – GPT4o 순서로 답변의 질이 좋아  
지며 자세해 진다. 실행시간은 세 모델 모두 비슷하다. 토큰 관리를 잘  
하면서 **GPT4o** 쓰는것이 제일 좋을거 같다.
- **kullm** 계열 : max\_token 를 줄이고 sampling을 하지 않아도 응답  
시간이 너무 길어 실사용은 불가능으로 판단하였다.

## 7. 향후 계획

gpt4o-mini 의 파인튜닝 기능이 풀리면 파인튜닝 한 후 성능 비교를 해보고싶다. 또한  
허깅페이스의 kullm3모델 말고 다른 모델로 작업하거나 api를 불러 실험 해볼 것이다.

현재 모델 기준으로는 프롬프트 엔지니어링에 좀 더 힘들써 성능을 올릴 것이다.

## 8. 출처

### 8-1. 파인튜닝 모델

## MODEL

### ft:gpt-3.5-turbo-0125:personal::Bco3BID3

○ Status	✔ Succeeded
① Job ID	ftjob-qMNjhDexhhFe80LkeL4QYqu6
🔗 Training Method	Supervised
📦 Base model	gpt-3.5-turbo-0125
📦 Output model	ft:gpt-3.5-turbo-0125:personal::Bco3BID3
🕒 Created at	2025년 5월 30일 오후 3:04
✔ Data sharing	Private
📊 Trained tokens	707,292

#### ⚙️ Hyperparameters

Epochs	3
Batch size	2
LR multiplier	2
Seed	1926649846

#### 📷 Checkpoints

ft:gpt-3.5-turbo-0125:personal::Bco3BGRv:ckpt-step-742

모델명 : ft:gpt-3.5-turbo-0125:personal::Bco3BID3

## 8-2. 허깅 페이스 모델

<https://huggingface.co/nlpai-lab/KULLM3>

모델명 : nlpai-lab/KULLM3