DISCORD

1.3 LLM 생성 과정

모델 개발을 위한 라이프 사이클

데이터 수집 및 준비 > 모델 설계 > 모델 학습 > 평가 및 검증 > 모델 배포 및 유지보수

1. 데이터 수집 및 준비

- **(1) 데이터 수집 :** 양도 많고 종류도 다양한 텍스트 데이터가 필요
  - HTML, PDF, 텍스트파일, 데이터베이스 등으로부터 수집
  - 수집 시 저작권, 개인정보 보호 등 법적인 문제 고려하여 수집
- **(2) 데이터 정제 :** 데이터 품질이 높어야 LLM이 정확하게 학습할 수 있음
  - 중복 제거
  - 노이즈 제거 : 오타, 잘못된 문장 부호, 비정상적인 문자 등
  - 토큰화 : 텍스트를 작은 단위로 나누는 과정

```
test = 'Hello how are you ?'
test.split(' ')
```

```
['Hello', 'how', 'are', 'you', '?']
```

  - 정규화 : 대문자 통일, 어간 추출(stemming) 등을 통해 단어의 기본 형태로 변환

```
print(stemmer.stem('runs'))
print(stemmer.stem('run'))
print(stemmer.stem('running'))
```

```
run
run
run
```

- **(3) 데이터 형식 변경 :** 데이터 형식 일치 ex) 날짜를 'YYYYMMDD' 형식으로 맞추기

2. 모델 설계

- LLM은 주로 **트랜스포머 모델** 기반

- 학습률, 배치크기 등과 같은 하이퍼파라미터 설정 필요

3. 모델 학습

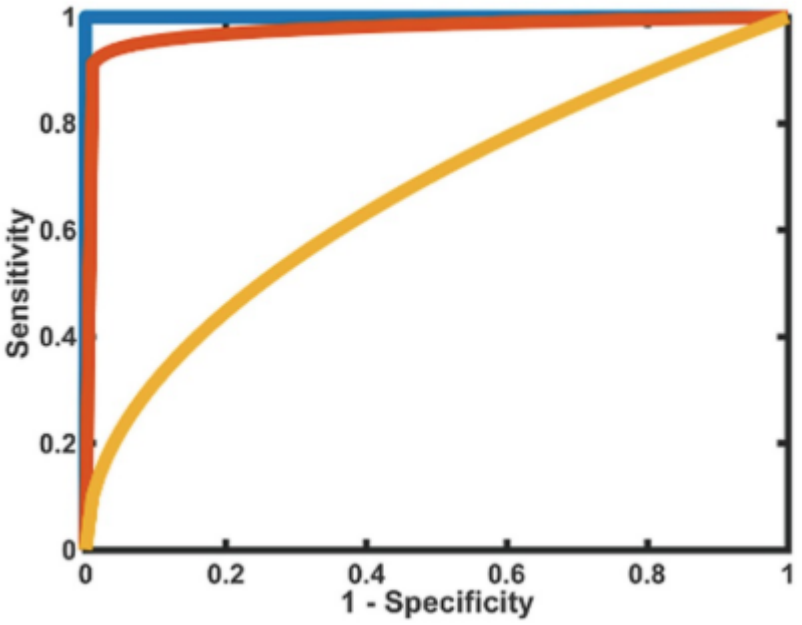
- 모델이 데이터로부터 패턴을 학습하고, 텍스트를 생성하거나 번역하는 등의 작업 수행
- 모델링 : 모델이 데이터로부터 중요한 특징이나 관계를 학습하고, 수학적 구조로 표현하는 과정

4. 평가 검증

- 모델이 얼마나 잘 작동하는지 평가하고 실제 서비스했을 때 어느 정도의 성능을 낼 수 있는지 확인
- 훈련 데이터 : 훈련시킬 때 사용
- 검증 데이터 : 모델 성능 평가하기 위한 용도. 하이퍼파라미터를 조정하고 중간 평가.
- 테스트 데이터 : 모델 성능 평가하기 위한 용도실제로 서비스했을 때 어떤 성능을 보일지 평가.

💡 모델 지표

- 정확도(Accuracy) : 얼마나 정확한가
- 정밀도(precision) : 양성으로 예측된 사례 중 실제 양성인 사례의 비율. ex) 스팸 메일
- 재현율(recall) : 실제 양성 중 양성으로 예측.
- F1 점수 (F1 score): 정밀도와 재현율의 조화평균
- ROC 곡선 및 AUC :
  - ROC Curve : 파란색일 수록 좋은 모델
  - AUC : Curve 안의 면적. 1일수록 좋은 모델



5. 배포 및 유지보수

- 배포 : 서비스(ex. QnA 챗봇)를 사용자가 이용하는 것
- 유지보수 : 문제(오류 등)발생 시 수정

1.4 LLM 생성 후 추가 고려 사항

1. 윤리적 고려 및 보정

- 책임감 있는 AI
    - 인공지능을 설계, 개발, 배포할 때 윤리적, 법적, 사회적 책임을 고려하는 접근 방식
- 공정성 : 성별, 인종, 나이 등에 대한 비편향
- 신뢰성&안정성 : 안전하고, 예측 가능한 위험을 관리하고 예방
- 프라이버시 : 개인정보 보호, 데이터 보안 유지
- 포용성 & 다양성 : 차별없이 혜택을 받을 수 있도록 환경 조성
- 윤리적 사용 : 사회적, 도덕적 기준에 부합하는 방식으로 사용 필요
- 투명성 : 의사 결정 과정이 명확하고 이해 가능해야하며, 사용자는 AI의 작동 방식과 결정 기준을 알 권리가 있음
- 책임성 : 결정에 대한 책임을 명확히 하여 문제 발생 시 적절한 해결책 제시

2. 지속적인 모니터링

- 악의적으로 사용하는 경우를 방지하기 위해서 지속적 검사가 필요
    - AI 스스로 악의적 문구를 탐지
    - 사람이 질문과 답변을 지속적으로 점검