Project Progress Report

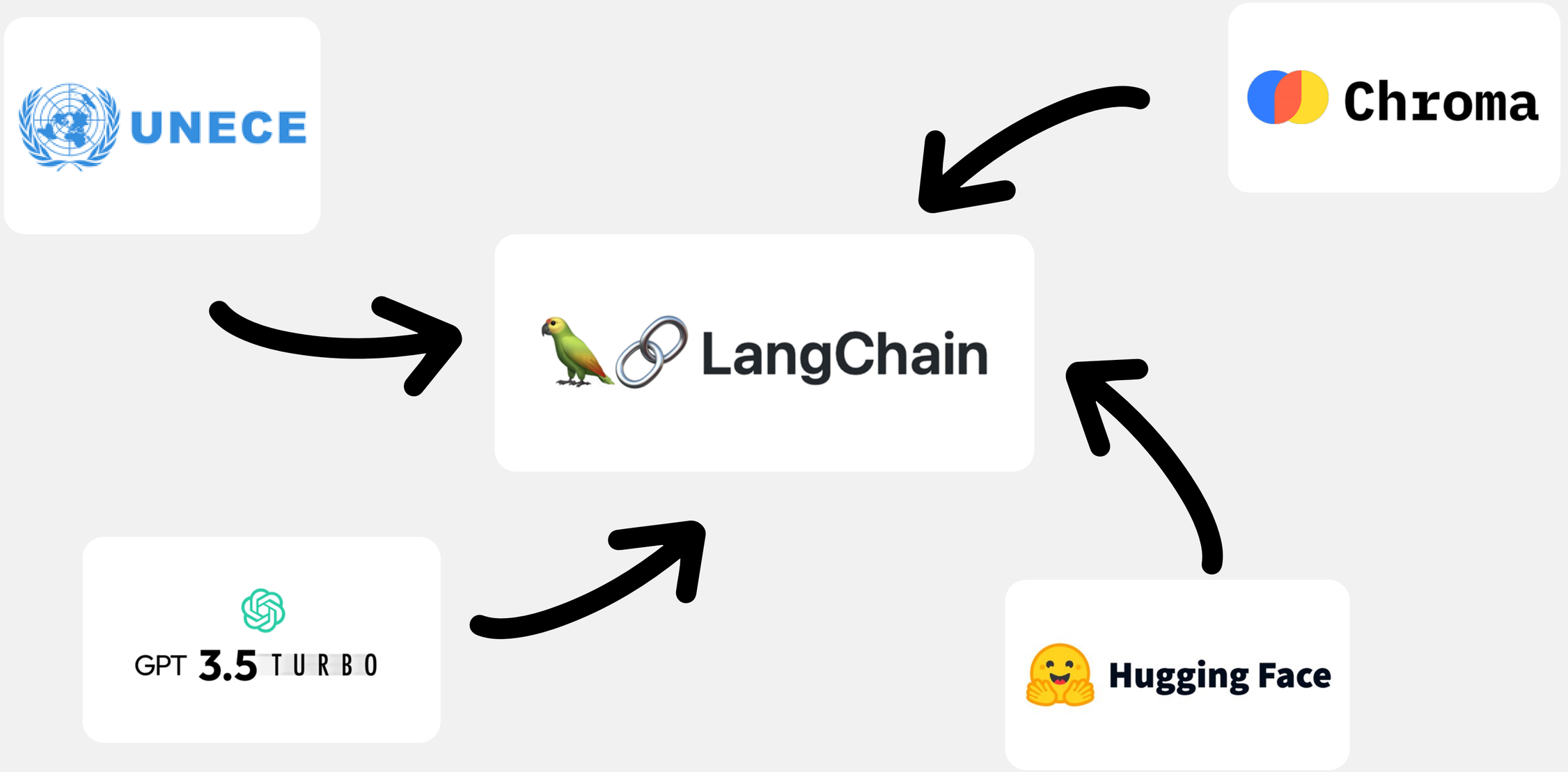
# 프로젝트 진행 상황 보고서

Carbot팀 안성규, 박상범, 김세희, 문벼리

# Contents



# RAG 모델 구현



# RAG 성능 평가

정확성, 적합성, 완전성, 일관성, 이해도 순으로 고려

## 개념

Q1. How does the testing process for braking systems work?

Q2. What is an Antilock Brake System (ABS)?

Q3. How do I configure the parking brake system?

Q4. Tell me all the things I should consider if I collide with a sedan (Hyundai Sonata) in front of me in the same lane.

4개의 질문 -> 대체로 잘 대답

- 확인 사항
  - 부록의 내용을 정확하게 얘기할 수 있는지 확인 필요.
  - 차 이름만 넣어도 분류를 할 수 있는지 확인 필요. (아마 안 될 것으로 추정)
- 더 만들 질문
  - 다른 구성, 개념, 과정에 대한 질문
  - 4번 질문처럼 자세하고 구체적인 질문

## 표

2개의 질문 -> 1개는 질문에 맞는 답변
1개는 질문의 포인트가 답이 없는 부분이었기에 애매함.

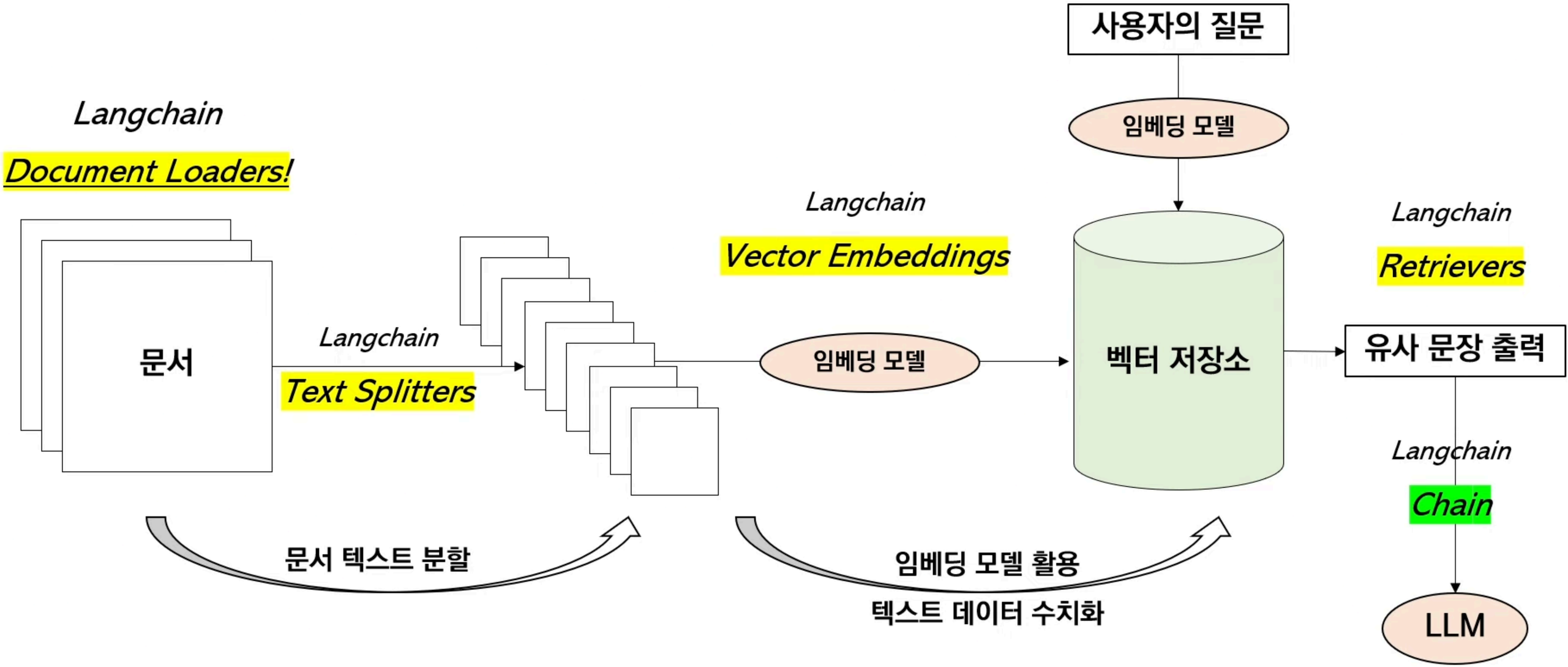
- 확인 사항
  - 좀 더 검증이 필요함.

## 그림&수식

그림 관련 2개, 간단한 수식 관련 1개의 질문
-> 답변 어려움.

- 확인 사항
  - 수식-텍스트 변환을 거쳐도 동일한지 확인

# RAG 프로세스 모형

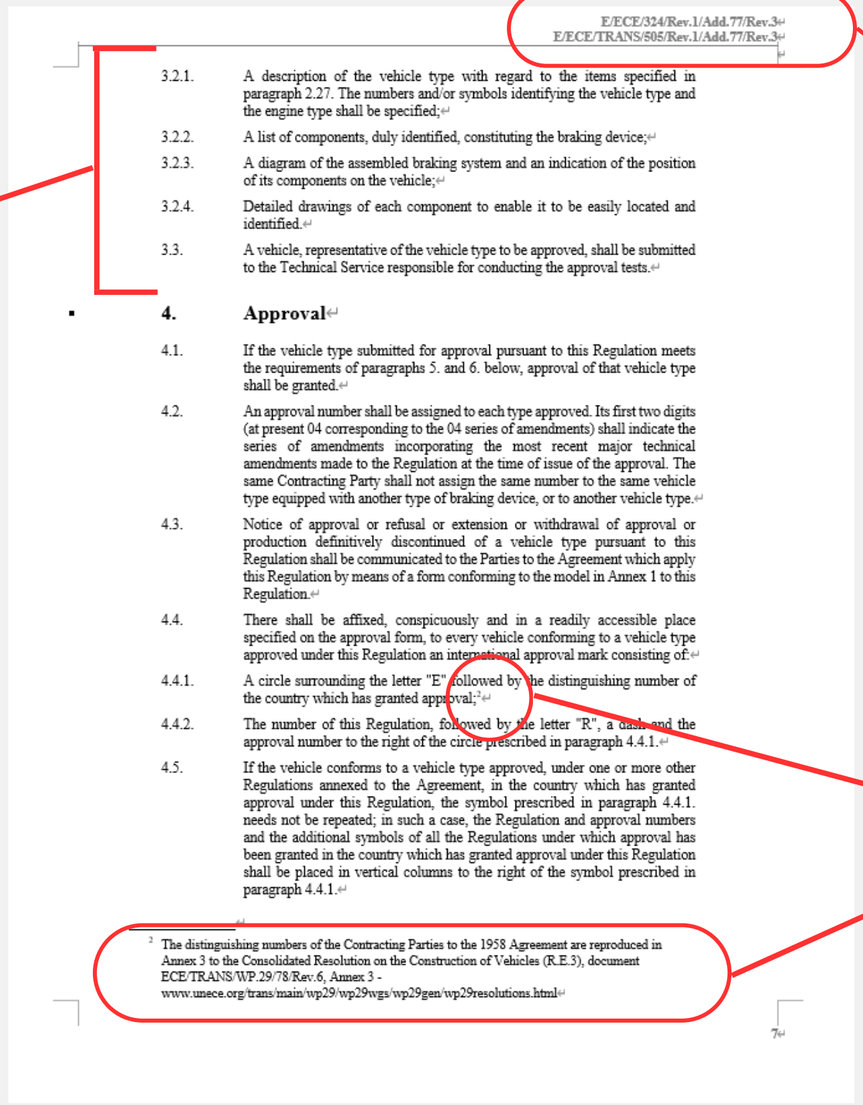

사용자의 질문
임베딩 모델
Langchain
Document Loaders!
문서
Langchain
Text Splitters
Langchain
Vector Embeddings
임베딩 모델
벡터 저장소
Langchain
Retrievers
유사 문장 출력
Langchain
Chain
LLM
문서 텍스트 분할
임베딩 모델 활용
텍스트 데이터 수치화

# RAG 프로세스 최적화 기법

| | |
|---|---|
| **텍스트 포멧팅** | Json 혹은 Markdown 형식으로 규격화 |
| **텍스트 분할** | 의미론적 유사도에 따른 텍스트 분리 |
| **텍스트 임베딩** | 임베딩 모델 변환<br>임베딩 벡터 캐싱 |
| **벡터스토어** | 리트리버에 더욱 용이한 FAISS 사용 |
| **리트리버** | 기반 모델 고도화<br>체인 기법에 따른 텍스트 품질 고도화<br>프롬프트 엔지니어링을 통한 답변 생성 고도화 |

# 데이터 전처리 - 현재 방식의 문제점

정보가 부분적으로 입력되어
온전한 맥락을 참고하지 못함





3.2.1. A description of the vehicle type with regard to the items specified in paragraph 2.27. The numbers and/or symbols identifying the vehicle type and the engine type shall be specified;

3.2.2. A list of components, duly identified, constituting the braking device;

3.2.3. A diagram of the assembled braking system and an indication of the position of its components on the vehicle;

3.2.4. Detailed drawings of each component to enable it to be easily located and identified.

3.3. A vehicle, representative of the vehicle type to be approved, shall be submitted to the Technical Service responsible for conducting the approval tests.

**4. Approval**

4.1. If the vehicle type submitted for approval pursuant to this Regulation meets the requirements of paragraphs 5. and 6. below, approval of that vehicle type shall be granted.

4.2. An approval number shall be assigned to each type approved. Its first two digits (at present 04 corresponding to the 04 series of amendments) shall indicate the series of amendments incorporating the most recent major technical amendments made to the Regulation at the time of issue of the approval. The same Contracting Party shall not assign the same number to the same vehicle type equipped with another type of braking device, or to another vehicle type.

4.3. Notice of approval or refusal or extension or withdrawal of approval or production definitively discontinued of a vehicle type pursuant to this Regulation shall be communicated to the Parties to the Agreement which apply this Regulation by means of a form conforming to the model in Annex 1 to this Regulation.

4.4. There shall be affixed, conspicuously and in a readily accessible place specified on the approval form, to every vehicle conforming to a vehicle type approved under this Regulation an international approval mark consisting of:

4.4.1. A circle surrounding the letter "E" followed by the distinguishing number of the country which has granted approval;[2]

4.4.2. The number of this Regulation, followed by the letter "R", a dash and the approval number to the right of the circle prescribed in paragraph 4.4.1.

4.5. If the vehicle conforms to a vehicle type approved, under one or more other Regulations annexed to the Agreement, in the country which has granted approval under this Regulation, the symbol prescribed in paragraph 4.4.1. needs not be repeated; in such a case, the Regulation and approval numbers and the additional symbols of all the Regulations under which approval has been granted in the country which has granted approval under this Regulation shall be placed in vertical columns to the right of the symbol prescribed in paragraph 4.4.1.

[2] The distinguishing numbers of the Contracting Parties to the 1958 Agreement are reproduced in Annex 3 to the Consolidated Resolution on the Construction of Vehicles (R.E.3), document ECE/TRANS/WP.29/78/Rev.6, Annex 3 - www.unece.org/trans/main/wp29/wp29wgs/wp29gen/wp29resolutions.html



필요 없는 정보가 반복적으로 입력

보충 정보(주석)과 거리가 있어
모델에게 혼란을 줄 가능성

# 데이터 전처리 - 현재 방식의 문제점

| *Vehicle decelerations* | *Signal generation* |
|---|---|
| ≤ 0.7 m/s² | The signal shall not be generated |
| > 0.7 m/s² and ≤ 1.3 m/s² | The signal may be generated |
| > 1.3 m/s² | The signal shall be generated |

테이블의 경우, 텍스트로 추출이 가능하나
해당 텍스트가 “테이블”임을 인지하지 못할 가능성 존재

Vehicle decelerations Signal generation
0.7 m/s² The signal shall not be generated
> 0.7 m/s² and ≤ 1.3 m/s² The signal may be generated
> 1.3 m/s² The signal shall be generated

# 데이터 전처리 - 현재 방식의 문제점

$$d_m = \frac{V_b^{\,2} - V_e^{\,2}}{25.92 \cdot (S_e - S_b)} \qquad \text{in m/s}^2$$

수식의 구조가 완전히 와해되고
대부분의 정보를 상실함

92.252 2
b ee b
mS SV Vd−□−=
in m/s2

# 데이터 전처리 - 포멧팅으로 얻을 수 있는 이점들


```
{
 "Metadata": {
  "Source": "UN Regulation No. 78",
  "Page": 3,
  "Chapter": "Regulation 2. Definition"
 },
 "Definitions": {
  "2.1": {
   "Term": "Antilock Brake System (ABS)",
   "Definition": "A system which senses ..."
  },
```


메타데이터를 통해서
정보의 출처를 분명히 하고
맥락 정보를 보존

테그를 통해서
추가적인 정보를 제공 가능

**예시: jsonl 형식**

# 데이터 전처리 - 포멧팅으로 얻을 수 있는 이점들

- 필요한 정보만 선별하여 모델의 혼란을 방지
- 분산 되어 있는 정보(주석, 부록 참조)를 하나로 취합
- 특수한 데이터(테이블, 이미지, 수식)의 의미를 온전히 보존

# 데이터 전처리 - 후보 포멧 형식들

## Jsonl


```
{
  "Table 2-1": [
    {
      "Deceleration": "≤ 0.7 m/s²",
      "SignalGeneration": "The signal shall not be generated"
    },
    {
      "Deceleration": "> 0.7 m/s² and ≤ 1.3 m/s²",
      "SignalGeneration": "The signal may be generated"
    },
    {
      "Deceleration": "> 1.3 m/s²",
      "SignalGeneration": "The signal shall be generated"
    }
  ]
}
```


- 높은 유연성
- 메타데이터 기입 용이

## Markdown

| > *Vehicle decelerations* | > *Signal generation* |
|---|---|
| > ≤ 0.7 m/s² | > The signal shall not be generated |
| > \> 0.7 m/s² and ≤ 1.3 m/s² | > The signal may be generated |
| > \> 1.3 m/s² | > The signal shall be generated |

- 높은 정형성
- 문서 학습량이 높음

# 텍스트 분할 (청킹) 사이즈 조절

작은 청크 사이즈

{A} {B} {C} {D}

- 연관성 있는 핵심 정보만
- 맥락 정보 상실 가능성

{A B C D}

{A B}  {C D}

- 전후 맥락을 포함한 풍부한 정보
- 세부적인 정보를 놓질 가능성

큰 청크 사이즈

# 분야별 발전 방향

**데이터 전처리**

- 기본적인 데이터 포맷팅 코드 구현을 통해 json, markdown 성능 비교
- 의미적 손실을 줄일 수 있는 최적화된  chunk size 찾기
- 문서 내에서 상호 참조하는 부분 구현 방법 찾기

**RAG**

- RAG 성능 평가를 위해 자동차 도메인 특화 질문 만들기
- Gpt -4 와 기반 모델 및 임베딩 모델 간의 비교

**파인 튜닝**

- 파인튜닝 오픈 소스 모델 선정
- 파인 튜닝 코드 구현

# 앞으로의 계획

| 항목 | 과업 | 7월 | | | 8월 | | | |
|---|---|---|---|---|---|---|---|---|
| | | 2주차 | 3주차 | 4주차 | 1주차 | 2주차 | 3주차 | 4주차 |
| 데이터 전처리 | 데이터 chunk size 결정 | 7/9 ~ 16 | | | | | | |
| | json, markdown 포맷 변경 | 7/9 ~ 19 | | | | | | |
| RAG | 코드 구현 | 7/10 ~ 23 | | | | | | |
| | 모델 평가 질문 선정 | 7/10 ~ 29 | | | | | | |
| | 최적화 | | 7/15 ~ 8/4 | | | | | |
| 파인튜닝 | 모델 선정 및 코드 구현 | | | 7/24 ~ 8/12 | | | | |
| | 최적화 | | | | 8/2 ~ 8/26 | | | |

# 감사합니다