

## 문제 개념: Digital 시대의 online communication과 집단 행위와 관련된 사회현상

### 1.1 캔슬컬쳐 Cancel Culture

- 정의 : 캔슬 컬처는 누군가가 부적절하거나 논란이 되는 언행을 했을 때, 공개적인 비판, 불매 운동, 사회적 배제를 통해 그 사람을 '사회적으로 퇴출(cancel)'시키려는 문화
- 배경
  - **SNS의 확산:** 트위터, 인스타그램, 유튜브 등에서 유명인뿐 아니라 일반인도 감시와 비판의 대상이 됨.
  - **사회적 정의에 대한 관심 증가:** 인종차별, 성차별, 소수자 혐오 등에 대한 의식이 커지면서 도덕적 책임 요구도 강화됨.
- 예시 1: 유명인의 과거 문제 발언 발굴-> 대중 비난 -> 브랜드 계약 해지 또는 자숙 선언
- 예시 2: 유명인, 인플루언서 등이 과거 트윗이나 영상 때문에 여론에 의해 퇴출

### 1.2 사이버 불링 Cyberbullying

- 정의 : 온라인 공간(댓글, DM, 채팅 등)에서 누군가에게 반복적으로 괴롭힘, 조롱, 협박, 모욕적인 말을 가하는 행위
- 주요 형태 : 악플, 비방게시물, 특정인에 대한 루머 유포, 따돌림(온라인 채팅방), 개인 정보 유출 및 조롱
- 영향 : 피해자의 사회적 고립, 정신건강 악화, 혐오 문화 조장, 공동체 신뢰 하락

1.3 에코 체임버 Echo Chamber : 동일한 생각과 신념을 가진 사람들끼리만 상호작용하면서, 서로의 의견만 반복적으로 강화시키는 폐쇄적 정보 환경. 현재 유튜브, 메타 등 거대 SNS 플랫폼은 모두 자사의 이익을 극대화하기 위해 이 개념에 기반한 전략을 사용하고 있다고 보면 됨.

- 핵심 요소
  - 정보의 다양성 결여
  - 반대 의견의 차단 혹은 배제 (캔슬 컬처와 만나는 부분)
  - 자신의 유사한 의견만 필터링해서 받아들이는 성향 강화
- AI와 Echo Chamber의 관계
  - 유튜브, 틱톡, X(구 트위터), 인스타그램 등은 사용자의 "기준 행동 데이터(좋아요, 시청시간, 클릭 등)"을 기반으로 유사한 콘텐츠를 반복 추천함
  - 정치적/사회적 견해, 혐오 표현, 음모론 등의 내용이 '에코 챔버'내에서 점점 강화됨

## [Step 1] 문제 인식과 스토리텔링

### 핵심 사회적 문제 인식

"논란이 되는 콘텐츠 하나, 댓글 몇 줄이 인플루언서를 향한 집단적 비난으로 번지고, 그 과정에서 반복적 사이버폭력과 집단 편향의 에코 챔버 현상이 발생한다. 하지만 이를 사전에 감지하고 예방할 수 있는 기술적 수단은 아직 부족하다."

#### ⌚ 서사 기반 스토리텔링 구조

| 구성 요소  | 내용   |
|--------|--|
| 문제 제기  | 인플루언서가 콘텐츠를 게시하자 특정 이슈가 비화되어 캔슬 컬처가 발생             |
| 위기 과정  | 악성 댓글이 급증, 과거 발언까지 재소환 → 공격자들이 모이며 echo chamber 형성 |
| 솔루션 제시 | AI가 이를 실시간 감지하고 리스크 알림, 균형 있는 정보 노출, 대응 코멘트를 제안    |
| 기대 효과  | 논란을 확산시키기보다 소통을 유도하고, 커뮤니티를 건강하게 유지할 수 있음          |

---

## [Step 2] 핵심 사용자 시나리오 도출

### 시나리오 1: 논란 조기 경고 및 대응

인플루언서 A가 민감 주제를 포함한 콘텐츠 업로드 → 댓글 중 부정/혐오성 표현 비율 급증  
→ AI가 "리스크 경고" 및 "Echo Chamber 위험" 알림 전송  
→ 관리자(인플루언서)에게 상황별 대응 댓글 가이드 및 리포트 제공

### 시나리오 2: 콘텐츠 업로드 전 리스크 시뮬레이션

게시 전 텍스트/영상 자막/이미지에 포함된 키워드·문맥 분석 → '민감도'와 예상 부정 반응 예측  
→ 리스크 점수와 관련 트리거 알림 → 게시 여부 결정 지원

### 시나리오 3: 댓글 분포 다양성 분석 → 균형 콘텐츠 자동 추천

댓글이 한 방향(예: 비난 일색)으로 쏠릴 경우 → Echo Chamber 감지  
→ 반대 시각의 공신력 있는 기사/영상 추천 → 균형 있는 시청자 경험 제공

---

[Step 3] 기능 정의 (관리자 중심)

| 기능명                    | 목적                                     | 대표 기능 요소                |
|------------------------|--|-------------------------|
| <b>콘텐츠 위험도 분석</b>      | 업로드 콘텐츠의 리스크 자동 평가 감성/민감 주제 분석, 리스크 점수 |                         |
| <b>댓글 실시간 감시</b>       | 악성 댓글·혐오 표현 탐지                         | 욕설 탐지, 공격 패턴 분석, 신고 자동화 |
| <b>Echo Chamber 분석</b> | 댓글 의견 다양성 지수 측정                        | 의견 클러스터링, Echo Index 계산 |
| <b>AI 대응 코멘트 추천</b>    | 위기상황에서 대응 문구 제안                        | 공감형/해명형/단호형 댓글 3종 제안    |
| <b>균형 콘텐츠 추천</b>       | 편향된 댓글 환경에 반대 정보 노출                    | 주제 기반 다각도 기사 추천         |
| <b>리포트 자동 생성</b>       | 콘텐츠/댓글/위험 지표 요약                        | PDF/웹 기반 자동 요약 보고서 생성   |

---

[Step 4] AI 기술 활용 포인트 (프로토타입 기준)

| 기능        | 기술 포인트            | 대표 기술                               |
|-----------|-------------------|-------------------------------------|
| 감성 분석     | 댓글/자막/본문의 감정 분류   | KoBERT, LSTM, SentiKorean           |
| 혐오/공격성 탐지 | 비하, 차별, 욕설 문장 분류  | TOXIC (Jigsaw), HateBERT            |
| 주제 모델링    | 콘텐츠·댓글의 주요 토픽 추출  | BERTopic, LDA, Sentence-BERT        |
| Echo 분석   | 클러스터링/다양성 지수 계산   | UMAP/DBSCAN + Shannon Index         |
| 코멘트 생성    | 상황 기반 톤 조절 문장 생성  | GPT + 스타일 프롬프트                      |
| 콘텐츠 추천    | 주제 유사도 + 반대 견해 탐색 | Hybrid RecSys + Diversity Filtering |

---

### [Step 5] 기술 아키텍처 요약 (간단 구조)

[콘텐츠 입력 (텍스트/영상/이미지)] → [AI 분석 모듈]

- ↳ 감성 분석
- ↳ 혐오성 탐지
- ↳ 주제 추출
- ↳ 댓글 클러스터링



[리스크 판단 + Echo 지수 계산]



[관리자용 대시보드 + 리포트]



- 대응 코멘트 추천
- 균형 콘텐츠 제공
- 자동 알림 전송

---

### [Step 6] 개발 우선순위 로드맵 (MVP)

| 단계  | 기능군             | 구현 목표                    |
|-----|-----------------|--------------------------|
| 1단계 | 댓글 분석 + 리스크 감지  | 감성/혐오 표현 탐지, 부정률 시작화     |
| 2단계 | Echo Chamber 분석 | 댓글 클러스터링, 다양성 지수 도출      |
| 3단계 | AI 대응 코멘트 생성    | 공감형·해명형 댓글 템플릿 기반 생성     |
| 4단계 | 균형 콘텐츠 추천       | 주제 기반 다각도 콘텐츠 추천 알고리즘    |
| 5단계 | 자동 리포트 생성       | 분석 결과 요약 보고서 (PDF or 웹뷰) |

---

## 정리 요약

| 항목   | 내용   |
|--|--|
| ⌚ 목적                                       | SNS 콘텐츠 기반으로 캔슬컬처, 사이버불링, 에코 챔버를 분석/완화       |
| ㉑ 주요 사용자 인플루언서, 소속 매니지먼트, (향후 일반 사용자까지 확대) |  |
| ₩ 핵심 기능                                    | 콘텐츠 위험도 분석, 댓글 감시, Echo Chamber 분석, AI 대응 추천 |
| ▣ AI 활용                                    | 감성/NLP/추천/생성형 AI/클러스터링 등 통합                  |
| ₩ 우선순위                                     | 댓글 분석 + Echo 분석 → 대응 생성 → 추천 및 보고서           |

---

## PoC 단계: AI 기능별 데이터셋 및 모델 설계 + 프로토타입 요약

---

### 목표

인플루언서 콘텐츠 및 댓글 기반으로 캔슬컬처, 사이버불링, Echo Chamber 현상을 조기 감지·완화하는 플랫폼을 위해 핵심 AI 기능 5 개에 대해 각각 필요한 데이터, 모델, 평가지표, 경량화 전략까지 포함한 프로토타입 스펙을 설계합니다.

---

### [기능 1] 콘텐츠 감성/리스크 분석

| 항목            | 내용   |
|---------------|--|
| ⌚ 목적          | 콘텐츠 업로드 시 감정 상태 및 민감 이슈 포함 여부 자동 분석  |
| ▣ 입력 데이터      | 텍스트 본문, 해시태그, 영상 자막(STT), 썸네일 OCR 텍스트<br>- AIHub 감성분류 데이터셋 (한국어 SNS 감정)                    |
| ▣ 학습/검증 데이터 셋 | - Hate-Speech/민감 주제 텍스트 데이터 (ex. politics, gender 관련 이슈 글)<br>- 자체 라벨링 데이터 확보(2000건 이상 필요) |
| ▣ 모델 후보       | - 감정 분석: KoBERT, SentiKorean, KLUE-RoBERTa<br>- 주제 탐지: BERTopic, LDA                       |

| 항목       | 내용                                      |
|----------|---|
| □ 평가 지표  | Accuracy, F1, 민감도(Recall), Precision    |
| ❖ 경량화 방안 | DistilBERT or TinyBERT 활용 + 사전 주제 필터 적용 |

### [기능 2] 댓글 감성/혐오/공격성 분석

| 항목           | 내용  |
|--------------|---|
| ⌚ 목적         | 악플, 비난, 공격성 표현을 실시간 탐지  |
| ઉ 입력 데이터     | 댓글 텍스트 (작성자ID, 시간, 내용)<br>- Jigsaw TOXIC Comment Dataset (영문 기반, 변환 필요) |
| ઉ 학습/검증 데이터셋 | - AIHub 혐오표현 한국어 데이터셋<br>- HateCheck-Ko, KoHATE 데이터셋                    |
| ઉ 모델 후보      | KoHATE BERT, TOXIC fine-tuned KoBERT                                    |
| □ 평가 지표      | F1-score, False Negative Rate, AUROC                                    |
| ❖ 경량화 방안     | 토픽 기반 선필터 + Keyword match + BERT inference                              |

### [기능 3] Echo Chamber 감지 및 다양성 측정

| 항목           | 내용   |
|--------------|--|
| ⌚ 목적         | 댓글의 의견 다양성/편향성/감정 클러스터 수치화   |
| ઉ 입력 데이터     | 댓글 텍스트, 감정 태그, 토픽 태그<br>- Reddit/Twitter 논쟁성 대화 클러스터링 데이터 (수집 필요)      |
| ઉ 학습/검증 데이터셋 | - 자체 수집 댓글 + 임베딩 기반 클러스터 라벨링   |
| ઉ 모델 후보      | Sentence-BERT 임베딩 + UMAP/DBSCAN<br>+ Shannon Index or Simpson Index 계산 |
| □ 평가 지표      | 클러스터 수, 최대 클러스터 점유율, 의견 다양성 지수 (0~1)                                   |
| ❖ 경량화 방안     | 댓글 수 기준 최소 샘플링 적용, PCA 기반 차원 축소 병행                                     |

---

#### [기능 4] AI 대응 코멘트 생성

| 항목           | 내용   |
|--------------|--|
| ◉ 목적         | 상황에 맞는 대응 댓글(공감/설명/단호)을 자동 생성  |
| ▣ 입력 데이터     | 위기 상황 요약 + 감정 분포 + 대표 댓글 <ul style="list-style-type: none"><li>- SNS 커뮤니티 발화/응답 대화상 수집</li></ul>                     |
| ■ 학습/검증 데이터셋 | <ul style="list-style-type: none"><li>- 대응 유형별 템플릿 정제 (공감형, 해명형, 단호형)</li><li>- Crowd-label 기반 사용자 대응사례 수집</li></ul> |
| ▣ 모델 후보      | KoGPT, GPT-3.5-turbo + Prompt Engineering<br>또는 KoBART 기반 요약 생성  |
| □ 평가 지표      | BLEU / ROUGE / Human Evaluation (공감성/명료성/정확성)  |
| ❖ 경량화 방안     | 템플릿 기반 응답으로 시작 + 상황별 프롬프트 방식   |

---

#### [기능 5] 균형 콘텐츠 추천

| 항목           | 내용   |
|--------------|--|
| ◉ 목적         | Echo Chamber 상태일 때 신뢰 가능한 '반대 관점' 정보 자동 추천   |
| ▣ 입력 데이터     | 콘텐츠 주제 태그 + 현재 댓글 분포   |
| ■ 학습/검증 데이터셋 | <ul style="list-style-type: none"><li>- 뉴스 기사 주제-관점 쌍 데이터 (수동 수집 or 미디어랩 연계)</li><li>- 언론사별 논조 및 이슈 대응 방식 수집 필요</li></ul>        |
| ▣ 모델 후보      | <ul style="list-style-type: none"><li>- 콘텐츠 기반 추천 (TF-IDF + cosine similarity)</li><li>- Hybrid RecSys + 다양성 기반 가중치 추천</li></ul> |
| □ 평가 지표      | Diversity@K, Coverage@K, 사용자 만족도 (후속 클릭/공유율)   |
| ❖ 경량화 방안     | 주제 필터링 + 사전 정의된 반대 관점 데이터 pool 활용 (Knowledge Graph)  |

## 종합 프로토타입 구성 요약

| 기능              | 핵심 기술            | PoC 데이터 준비 우선순위                         |
|-----------------|------------------|---|
| 콘텐츠 리스크 분석      | KoBERT, BERTopic | <input checked="" type="checkbox"/> 1순위 |
| 댓글 악성 탐지        | KoBERT + TOXIC   | <input checked="" type="checkbox"/> 1순위 |
| Echo Chamber 분석 | S-BERT + 클러스터링   | <input checked="" type="checkbox"/> 2순위 |
| 대응 코멘트 생성       | GPT + 템플릿 기반     | <input checked="" type="checkbox"/> 3순위 |
| 균형 콘텐츠 추천       | Hybrid Recsys    | <input checked="" type="checkbox"/> 3순위 |

---

## 데이터 수집 전략

| 유형                | 방법                                      |
|-------------------|---|
| 인플루언서 콘텐츠/댓글 크롤러로 | 유튜브, 인스타그램, 트위터 공개 계정 수집 (정책 준수)        |
| 혐오/공격 라벨 데이터      | AIHub, KoHATE, Jigsaw TOXIC 등 공개 데이터 활용 |
| 민감 주제 토픽          | 뉴스 기사, 커뮤니티 키워드 사전 구축 (젠더, 정치, 종교 등)    |
| 대응 문구 사례          | 크라우드소싱 또는 전문가 그룹 수작업 수집                 |
| 추천 콘텐츠 Pool       | 미디어랩, 위키데이터, 지식베이스 연계 구성                |

---

## PoC 워크플로우 설계 (기능 1 & 기능 2 중심)

---

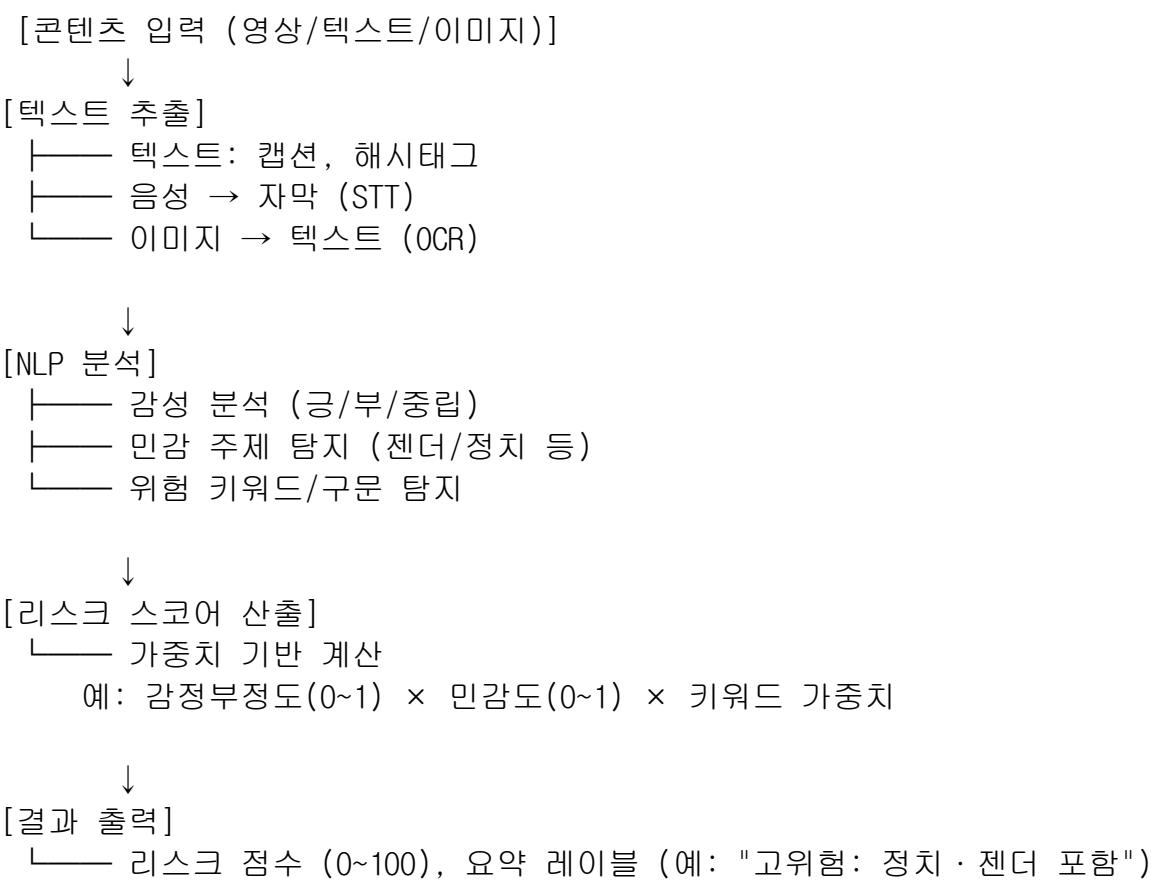
### ◇ 기능 1: 콘텐츠 감성 및 리스크 분석

#### ⌚ 목적

인플루언서가 업로드하는 콘텐츠(텍스트, 자막, 이미지)에 대해 감정·주제·민감도 분석을 수행하고 리스크 점수를 산정함.

---

#### ⌚ 전체 흐름



#### ❖ 기술 구성

| 단계  | 기술/모델                 |
|-----|-----------------------|
| STT | OpenAI Whisper, KoSTT |

## 단계 기술/모델

|        |                                   |
|--------|-----------------------------------|
| OCR    | Tesseract, Google Vision API      |
| 감정 분석  | KoBERT, KLUE-RoBERTa 감성 Fine-tune |
| 주제 분석  | BERTopic (Sentence-BERT 기반)       |
| 리스크 계산 | 감정 점수 × 민감 주제 여부 × 위험 키워드 점수 합산   |

---

### 📤 출력 예시

```
json
복사편집
{
  "emotion": "negative",
  "risk_topics": ["gender", "politics"],
  "risk_score": 87,
  "summary": "고위험 콘텐츠: 부정 감정 + 젠더 · 정치 민감 주제"
}
```

---

### ◇ 기능 2: 댓글 악성 감지 및 요약

#### ⌚ 목적

콘텐츠에 달린 댓글들을 수집하고, 악성 표현·공격성·혐오 발언을 탐지해  
요약/경고.

---

#### ♀ 전체 흐름

[댓글 수집]  
└── YouTube, Instagram, Twitter API 또는 수집툴

↓  
[텍스트 전처리]  
└── 특수기호 제거, 이모지/링크 필터링

↓  
[NLP 기반 탐지]  
└── 감정 분석  
└── 공격성 점수 계산 (TOXIC/HATE)

└─ 욕설/비속어 탐지 (사전 + 모델)



[요약 및 시각화]

└─ 전체 부정 댓글 비율  
└─ 대표 악성 댓글 예시  
└─ 위험 키워드 랭킹



[실시간 경고 Trigger]

└─ 부정비율  $\geq 30\%$ , 또는 특정 키워드 급증 시 알림

---

### 🛠 기술 구성

| 단계 | 기술/모델 |
|----|-------|
|----|-------|

감정 분석 KoBERT 감정 분류

공격성 탐지 TOXIC fine-tuned KoBERT, KoHATE

욕설 필터 국립국어원 욕설사전 + 커뮤니티 기반 확장

요약/랭킹 TF-IDF + 감정 기반 키워드 랭킹, 클러스터링 (k-means/UMAP)

---

### 📤 출력 예시

```
json
복사편집
{
  "total_comments": 542,
  "toxic_comments": 127,
  "toxic_rate": 23.4,
  "flagged_keywords": ["페미", "극혐", "역겹다"],
  "warning": false,
  "representative_toxic_comments": [
    "이런 걸 왜 올림? 놈가 없나",
    "너 같은 애는 방송 그만해라"
  ]
}
```

---

## ❸ 통합 PoC 시연 시나리오 (예)

1. 인플루언서가 신규 콘텐츠 영상 업로드
  2. 시스템이 자막(STT), 썸네일(OCR), 본문 텍스트에서 감정/주제 분석 수행
  3. 리스크 점수 85점 → "고위험" 콘텐츠로 표시
  4. 댓글 분석 모듈이 실시간 수집을 시작
  5. 욕설 포함 댓글 급증 및 Echo Index 하락 탐지
  6. 관리자에게 "콘텐츠 삭제 또는 해명 권고" 알림 + 코멘트 템플릿 추천
- 

## 관리자 피드백 루프 설계

### ❶ 목적

- AI가 제공하는 감정 분석, 리스크 점수, 대응 코멘트 제안에 대해 관리자가 직접 피드백을 주거나 수정을 반영할 수 있게 하여 AI의 판단 품질을 개선하고, 신뢰 기반의 인간-AI 협업을 구현
- 

### ◇ 1. 피드백 루프 작동 구조 (워크플로우)

#### [AI 분석 결과 제시]

- └─ 콘텐츠 리스크 점수: 87 (고위험)
- └─ 댓글 악성 비율: 34%
- └─ 대응 코멘트 제안: "공감형 응답 A"

↓

#### [관리자 확인 및 피드백]

- "해당 점수는 적절함"
- "리스크 과대 평가됨 (직접 점수 입력: 60)"
- "제안 코멘트 수정: '감사합니다' 추가 필요"

↓

#### [피드백 기록 + 재학습 DB 저장]

- └─ 라벨 보정, 코멘트 수정 내용 저장
- └─ 유사 사례 등장 시 학습 기반 자동 조정

↓

#### [신뢰도 향상 + AI 모델 개선에 반영]

---

## ◇ 2. 관리자 UI 요소 예시

| 컴포넌트                                      | 기능                       |
|---|--------------------------|
| AI 분석 결과 카드                               | 감정/리스크 점수 + 대응 코멘트 3종 표시 |
| <input checked="" type="checkbox"/> 승인 버튼 | 분석 결과가 적절하면 "확인"으로 종료    |
| <input type="checkbox"/> 수정 입력창           | 점수 수동 조정, 댓글 수정 입력 가능    |
| <input type="checkbox"/> 기타 피드백           | 오탐/과대탐 여부, AI 응답 평가      |
| <input type="checkbox"/> 수정 내역 저장         | DB에 수정값 저장, 추후 재학습 반영    |

---

## ◇ 3. 예시 시나리오

### ! 콘텐츠 분석 결과:

- 감정: 부정
- 민감 주제: 젠더
- 리스크 점수: 92 점 (AI 자동)
- 추천 대응: “불편하셨다면 죄송합니다. 다양한 의견을 경청하겠습니다.”

### 관리자 피드백 입력:

- “리스크 점수는 과하게 나왔습니다. 60 점이 적절해 보입니다.”
- “현재 상황은 해명보다는 사실 설명이 더 필요합니다. 아래 문장을 대체해주세요.”

수정 코멘트: “본 영상은 특정 입장을 대변하지 않습니다. 맥락을 함께 참고해 주세요.”

→ 시스템은 해당 사례를 DB에 기록

→ 유사한 콘텐츠-댓글 조합에서 점수 기준과 톤을 재조정하도록 학습됨

---

#### ◇ 4. 피드백 데이터의 학습 반영 방식

| 항목           | 설명   |
|--------------|--|
| ❖ 수치 조정      | 관리자의 점수 수동 조정을 기반으로 모델의 예측 가중치 업데이트<br>(ex. 감정 점수 vs 주제 민감도 비율 조정) |
| ❖ 대응 코멘트 재학습 | 실제 선택/수정된 대응 문장을 다음 모델 학습의 Label로 반영 (예: 공감형보다 단호형이 선택된 패턴 파악)     |
| ❖ 위험 키워드 보완  | 관리자 입력 키워드("이 표현이 문제였음")를 리스크 키워드 DB에 자동 추가 가능                     |

#### ◇ 5. 추가 설계 고려사항

| 고려 항목        | 설명                                       |
|--------------|--|
| ☑ 피드백 로그 시각화 | 관리자 수정 내역을 시간순으로 시각화 → 추후 운영 정책 기준 수립 가능 |
| ❖ 롤백 기능      | 잘못된 수동 조정 시, 이전 AI 추천 상태로 복원 가능          |
| ▶ 사용자 중심 알림  | 반복 피드백이 많은 AI 오류 유형은 관리자에게 팝업 또는 툴팁으로 경고 |

#### ◀ 요약

| 요소      | 설명  |
|---------|---|
| ⌚ 목적    | AI의 판단을 관리자 조정 가능하게 하여 신뢰성과 정확도 동시 확보                   |
| ❖ 작동 흐름 | AI 제안 → 관리자 조정 → 기록 저장 → 재학습 반영                         |
| ▣ 기대 효과 | 오탐률 감소, 사용자 경험 개선, AI 개선 데이터 확보                         |
| □ 향후 확장 | 관리자 피드백을 기반으로 지속적 모델 리파인 가능 (online learning 구조로 확장 가능) |