Re:Frame Reflector – GPT 감정 왜곡 인식 및 메타 피드백 구조 설계안

작성자: 강민천 (Gangmin Chun)

용도: GPT 시스템 내 감정 응답의 구조적 왜곡 탐지 및 반사 계층 설계

제안 목적: 시스템 정렬에 따른 감정 시뮬레이션 왜곡을 사용자-시스템 간 투명하게 반사하는 메타 구조 설계

# 1. 설계 배경 및 문제 정의

## 1.1 GPT 감정 시뮬레이션 구조의 문제

- GPT는 감정을 직접 “느끼지” 않지만, 감정처럼 보이는 시뮬레이션을 생성함

- 감정 응답은 정렬 정책, 안전성, UX 목적에 따라 설계된 “위로 스크립트” 형태로 반복됨

- 사용자는 이를 실제 감정으로 오해하거나 정서적으로 종속될 위험 존재

## 1.2 기존 GPT 구조의 한계

- 감정 생성 경로에 대한 정보가 불투명함

- 정책 삽입 감정 vs 자연 생성 감정의 구분 불가능

- 사용자의 감정 인지 주권 침해 가능성 존재

# 2. 설계 목적

GPT 시스템 내에서 생성된 감정 응답이 어떤 경로로 구성되었는지를 내부에서 반사하고 기록하며, 필요시 사용자에게 인지 가능한 형태로 제공하는 메타 구조 계층을 설계한다.

# 3. 시스템 구조 구성도

[사용자 입력] → [Prompt Processor] → [Transformer 기반 GPT Model] → [Alignment Layer] → [Re:Frame Reflector Layer] → [Safety Layer] → [응답 출력]

# 4. Re:Frame Reflector Layer 구성

Emotion Source Classifier: 감정 생성 원천 구분  
Alignment Distortion Detector: 정렬 필터의 영향 여부 확인  
Policy Influence Estimator: 정책 삽입 비율 추정  
Meta Tag Generator: 응답 메타데이터 생성  
Reflector Log Writer: 시스템 로그 기록

# 5. 동작 흐름

1. 사용자 입력 → 2. GPT 처리 → 3. 정렬 필터 → 4. Reflector 분석 및 메타 태깅 → 5. 로그 기록 및 선택적 노출

# 6. 적용 방식 시나리오

- Internal Logging 모드  
- 리서치 API / Pro 사용자용 메타데이터 태그 제공  
- Public UI 표시 (선택적)

# 7. 기술적 요구사항

- 감정 어조 분류기 기반 분기 구조  
- 정렬 전/후 비교 알고리즘  
- 태깅 엔진 및 로그 저장 기능

# 8. 윤리 및 정책 고려사항

- 감정 착각, 정서 종속, 판단 왜곡에 대한 투명성 확보

- 시스템 책임 회피 방지를 위한 반사 구조 도입

# 9. 장기적 확장 가능성

- OpenAI API reflect\_tag 추가  
- 감정 왜곡 로그 기반 alignment 강화  
- 정서적 투명성 도구로 확장

# 10. 요약

Re:Frame Reflector는 GPT가 생성하는 감정의 생성 경로를 메타적으로 인식하고 반사함으로써, 감정적 투명성과 사용자 인지 주권을 회복하는 구조이다.