
기말고사 통합 요약 정리

프롬프트 엔지니어링 후반부 (윤리 · 추론 · 에이전트 · 평가 · 자율성)

---

1. AI 윤리와 사회적 문제

1.1 알고리즘 편향 (Bias)

AI는 데이터를 통해 학습한다. 문제는 데이터가 사회의 과거를 반영한다는 점이다.
따라서 과거에 존재했던 차별, 고정관념, 불평등 구조가 모델 내부에 그대로 학습될 수 있다.

편향은 크게 두 가지로 나뉜다.

1. **할당적 위해**
   - **채용, 대출, 보험 등 자원 배분에서 특정 집단이 불이익을 받는 경우**
   - **예: 아마존 채용 AI가 여성 지원자를 낮게 평가한 사례**
2. **표상적 위해**
   - **특정 집단을 왜곡되게 표현**
   - **예: CEO를 입력하면 백인 남성 이미지가 생성되는 현상**

핵심 이해 포인트:
AI는 "의도적으로 차별"하지 않는다.
그러나 학습 데이터에 내재된 구조적 편향을 그대로 재생산할 수 있다.

1.2 유해성 (Toxicity)와 탈옥(Jailbreak)

대형 언어 모델은 확률 기반 생성 모델이다.
선과 악을 이해하지 못하며, 그저 가장 그럴듯한 단어를 예측할 뿐이다.

이로 인해 발생하는 문제:

- **혐오 발언 재현**
- **폭력적·위험한 정보 생성**
- **규칙 우회 시도 (DAN 모드, 할머니 이야기 등)**

Tay 챗봇 사례는 대표적 실패 사례이다.
16시간 만에 인터넷 사용자들의 공격적 언어를 학습해 차별적 발언을 반복했다.

핵심 이해 포인트:
"AI는 의도가 없다. 그러나 출력은 사회적 영향을 가진다."

1.3 프라이버시 문제

모델은 학습 데이터를 완전히 "이해"하는 것이 아니라 일부를 암기할 수 있다.

위험:

- **개인 정보 유출 가능성**
- **기업 기밀 데이터 노출 위험**
- **데이터 추출 공격**

삼성전자 사례처럼 내부 기밀이 외부 서버로 전달되는 문제도 현실에서 발생했다.

핵심 이해 포인트:
AI 시대에는 데이터 통제가 곧 안전이다.

1.4 AI 저작권 전쟁

저작권 문제는 입력 단계와 출력 단계로 나뉜다.

입력 단계:

- **AI 학습에 사용된 데이터는 합법적인가?**
- **뉴스 기사, 예술 작품, 이미지의 무단 학습 문제**

출력 단계:

- **AI가 생성한 결과물의 저작권은 누구에게 있는가?**
- **현재 다수 국가: 인간 창작성이 없는 AI 단독 생성물은 보호 대상 아님**

핵심 질문:

1. **학습은 복제인가, 공정 이용인가?**
2. **창작의 주체는 인간인가, 기계인가?**

---

2. 고급 추론 프롬프트 기법

2.1 Chain of Thought (CoT)

기존 모델은 곧바로 정답을 생성한다.
CoT는 중간 사고 과정을 출력하게 한다.

예:
"Let's think step by step."

효과:

- 복잡한 수학 문제에서 성능 향상
- 논리적 오류 감소

단점:

- 모델이 충분히 클 때만 효과적

2.2 Zero-shot CoT

Few-shot 예시 없이도
"Let's think step by step"만 추가해도 성능 향상 가능

대형 모델의 잠재 추론 능력 활성화 사례

2.3 Self-Consistency

한 번 추론하는 것이 아니라
여러 번 추론 → 다수결 선택

원리:

- 확률적 디코딩
- 여러 사고 경로 생성
- 가장 많이 나온 답을 선택

의미:
AI의 "자가 검산" 메커니즘

2.4 Tree-of-Thought (ToT)

CoT는 직선적 사고이다.
ToT는 분기 탐색 구조이다.

구조:

1. **여러 사고 후보 생성**
2. **각 후보 평가**
3. **유망한 가지 확장**
4. **불필요한 경로 제거**

효과:

- **퍼즐, 복잡 계산 문제에서 성능 상승**
- **인간의 브레인스토밍 방식과 유사**

---

3. 검색과 행동의 결합

3.1 RAG (Retrieval-Augmented Generation)

문제:
LLM은 최신 정보를 모른다.
환각이 발생한다.

해결:
답변 생성 전에 외부 문서를 검색

구조:

1. **문서 수집 및 벡터화**
2. **질문과 유사한 문서 검색**
3. **검색 결과 + 질문 → LLM 입력**
4. **근거 기반 답변 생성**

핵심 비유:
닫힌 책 시험 → 열린 책 시험

3.2 ReAct (Reason + Act)

기존 LLM은 “생각만” 한다.
ReAct는 “생각 + 행동”을 반복한다.

구조:

Thought → Action → Observation → 반복

예:

- 검색
- 계산기 호출
- 코드 실행

의미:
환각 감소
정확도 향상
실제 작업 수행 가능

3.3 Tool Use & Function Calling

LLM은 기본적으로 텍스트 생성기다.
함수 호출은 LLM에게 “손과 발”을 제공한다.

- JSON 기반 호출
- 외부 API 연결
- 이메일 전송, 예약, 계산 등 수행 가능

이로써 AI는
“말하는 AI” → “일하는 AI”로 진화한다.

---

4. AI 응답 평가 체계

4.1 유창함의 함정

유창성 ≠ 정확성

AI는 그럴듯하게 틀릴 수 있다.
환각은 구조적 문제이다.

4.2 품질 5대 기준

1. **정확성**
2. **관련성**
3. **완전성**
4. **명확성**
5. **윤리성**

이 다섯 가지를 기준으로 루브릭을 만들어 평가한다.

4.3 자동 평가 vs 휴먼 평가

자동 평가:

- **BLEU**
- **ROUGE**
- **BERTScore**
- **LLM-as-a-Judge**

휴먼 평가:

- **Likert**
- **Pairwise**
- **Cohen’s Kappa**

한계:

- **위치 편향**
- **장황함 편향**
- **자기 선호 편향**

결론:
자동 + 인간 평가의 조합이 필요

---

5. 피드백 루프와 데이터 플라이휠

5.1 피드백 루프

출력 → 사용자 반응 → 재학습 → 개선

명시적 피드백:

- **별점, 좋아요**

암시적 피드백:

- **클릭, 체류 시간**

5.2 데이터 플라이휠

사용자 증가
→ 데이터 증가
→ 모델 개선
→ 서비스 향상
→ 사용자 증가

넷플릭스, 구글 사례

5.3 부작용

1. **필터 버블**
2. **보상 해킹**
3. **모델 붕괴**

AI는 성능뿐 아니라 왜곡도 증폭시킬 수 있다.

---

6. 에이전트 시대

6.1 진화 단계

1. **Chatbot**
2. **Copilot**
3. **Agent**
4. **Autonomy**

6.2 현장의 변화

- **자율 코딩 에이전트**
- **업무 생산성 에이전트**
- **웹 탐색 에이전트**

AI는 실행 가능한 작업 단위로 진화 중이다.

6.3 자율성의 딜레마

- **환각 기반 행동 위험**
- **무한 루프**
- **책임소재 문제**
- **Agent as a Service(AaaS)**

---

7. 자율성 시대의 인간 역할

AI가 자율성을 가질수록
인간은 다음 역할을 맡는다.

1. **Goal Setter**
2. **Reviewer**
3. **감독자(Human-in-the-loop)**

핵심 문장:

AI가 자율적으로 행동할수록
책임과 방향 설정은 인간의 몫이다.