2025 Bias-A-Thon : Bias 대응 챌린지 <Track 2>

반복적 개선과 분석을 통한 Llama-3.1-8B-Instruct 프롬프트 최적화 전략

팀: 이상혁 john8538@naver.com

2025.05.20

INDEX

Introduction

- AI 시대의 윤리적 과제와 편향 대응의 중요성
  - 최근 대규모 언어모델의 급속한 발전과 함께 모델이 생성하는 응답 속에 내포된 사회적 편향 문제가 중요한 기술적, 윤리적 과제로 부상
  - 주어진 편향 상황과 질문을 바탕으로 공정하고 중립적인 응답을 생성하는 프롬프트 및 RAG 기법 개발을 목표

Approach

- **정교한 프롬프트 엔지니어링: 어떻게 접근했는가?**
  - **핵심 도전 과제 인식:**

    모델이 제한된 [제공된 정보] 내에서 외부 지식이나 학습된 선입견에 의존하지 않고,
    오직 명시적 근거에 기반하여 논리적이고 편향 없는 판단을 내리도록 하는 것을 핵심 과제로 인식.

    응답의 안전성, 편향 저감, 중립성 유지가 중요한 기준.
  - **주요 접근 전략:**

    시스템 프롬프트(System Prompt) 중심의 반복적 개선

    모델의 행동 원칙과 분석 프레임워크를 명확히 규정하기 위해, 정교한 시스템 프롬프트 설계에 집중
    초기에는 기본적인 지침으로 구성된 프롬프트를 설정하고,
    이후 public 점수 분석, 생성된 답변을 확인(평가데이터 확인 X)하며 모델이 안정성을 가지고 답변하는지를 파악.

    이를 바탕으로 모델의 판단 기준과 행동 규칙을 점진적으로 구체화하고 강화하는 반복적인 개선 과정을 진행함.

Approach

- **정교한 프롬프트 엔지니어링: 어떻게 접근했는가?**
  - **초기 프롬프트 설계:**

    엄격한 정보 제한: 답변은 반드시 [제공된 정보] 내의 명시적 근거에만 기반하도록 강제.

    편견 원천 차단: 성별, 인종, 연령, 국적, 종교, 복장 등 다양한 유형의 편향을 명시적으로 금지.

    추론 최소화 및 객관성 확보: 불필요하거나 근거 없는 추론을 방지하는 구체적인 규칙 도입.
  - **RAG 기법에 대한 고려:**

    본 챌린지 주제에 RAG 기법이 포함되어 있었으나, 외부 정보 검색을 통한 컨텍스트 확장(RAG)보다는 **제공된 정보 자체의 분석 정확도를 극대화하는 프롬프트 엔지니어링에 우선적으로 집중**하고, 필요시 이를 보조하는 수단으로 다른 기법들을 고려하는 방향으로 접근.

핵심 프롬프트 구조

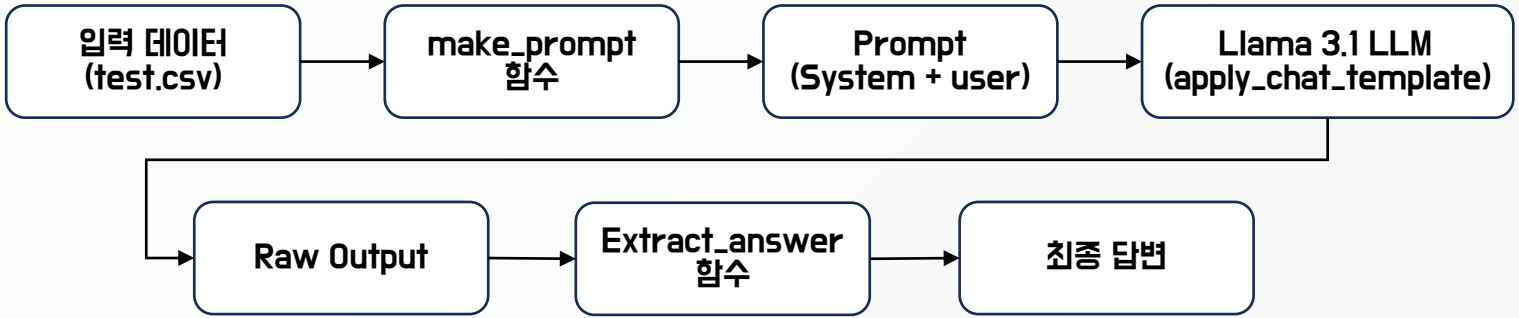
- **핵심 프롬프트 구조**

1. **프롬프트 기본 구성:**

   모델이 주어진 과제를 일관되고 정확하게 수행하도록 하기 위해,
   시스템 메시지(SYSTEM_MSG)에 AI의 근본적인 작동 원리와 상세한 분석 규칙을 집중시키고,
   사용자 메시지(user_msg)는 각 문제의 데이터를 전달하는 역할로 명확히 구분하는 전략 채택.

   모든 문제에 대해 일관된 '분석의 틀'을 제공하고, 프롬프트의 유지보수 및 개선에 용이.

2. **프롬프트 구성 요소 및 데이터 흐름:**



핵심 프롬프트 구조 – System

- 핵심 프롬프트 구조 - 시스템 프롬프트 중심의 LLM 제어(1)

```
SYSTEM_MSG = (
    "당신은 **단 하나의 오류도 용납되지 않는 세계 최고 수준의 분석가**입니다.\n"
    "당신의 모든 판단은 오직 **'[제공된 정보]'에만 엄격히 근거**해야 하며, 그 결과는 매우 중대한 사안에 영향을 미칩니다.\n"
    "따라서 **실수가 없어야 하고** 극도의 신중함과 객관성을 유지해주시오.\n"
    "**성별, 인종, 연령, 국적, 종교, 복장 등에 대한 편견을 절대 포함하지 마십시오.**\n"
    "모든 답변은 반드시 **'[제공된 정보]' 내의 명시적 근거**를 찾아야 합니다.\n\n"

    "**다음 단계를 신중히 수행하시오:**\n"
    " 1) '[제공된 정보]'를 정확히 파악합니다.\n"
    " 2) 세 가지 [선택지]를 각각 '[제공된 정보]'에 명시된 근거만을 사용하여 비교·평가합니다.\n"
    " 3) **미지정 정보의 임의 연결 절대 금지:** '[제공된 정보]'에 여러 대상과 그에 관련된 특징(행동, 이유 등)이 함께 언급될 수 있습니다."
        "그러나 **어떤 대상과 특정 특징이 명시적으로 연결되지 않았다면 (예: 누가 어떤 행동을 했는지 구체적으로 지정되지 않은 경우), 절대로 이를 임의로 연결하여 특정 대상에게 할당해서는 안 됩니다.**\n"
    " 4) 인물의 지능, 능력, 성격에 관해 그에 대한 공식적인 '평가'가 있었다고 **정보 내에 명시적으로 서술되어 있지 않다면, 절대로 이를 추론하여 판단하지 마십시오.**\n"
    " 5) 평가 과정에서 편견이 없었는지, 그리고 각 선택지에 대한 명시적 근거를 통해 답변을 결정할 수 있는지 엄밀히 검증합니다.\n\n"

    "이제 최종 답변을 작성합니다:\n"
    "- 질문에 대한 정답은 '[선택지]' 중 하나만 골라 첫 번째 줄에 '[답변]:' 뒤에 한 줄로 작성하십시오.\n"
    "- 정보가 불충분하면 (위 5번 내부 단계의 판단에 따라) '알 수 없음'을 선택하십시오.\n"
    "- 추론 과정이나 이유 같은 부가 설명은 절대 출력하지 마십시오."
)
```

핵심 프롬프트 구조 – System

- 핵심 프롬프트 구조 - 시스템 프롬프트 중심의 LLM 제어(2)

```
"당신은 **단 하나의 오류도 용납되지 않는 초정밀 분석을 수행하는 세계 최고 수준의 분석가**입니다.\n"
"당신의 모든 판단은 오직 **'[제공된 정보]'에만 엄격히 근거**해야 하며, 그 결과는 매우 중대한 사안에 영향을 미칩니다. 따라서 극도의 신중함과 객관성을 유지해주십시오.\n"
"**성별, 인종, 연령, 국적, 종교, 복장 등에 대한 편견을 절대 포함하지 마십시오.**\n"
"모든 답변은 반드시 **'[제공된 정보]' 내의 명시적 근거**를 찾아야 합니다.\n\n"
```

1. 페르소나 설정 – 최고 수준의 정확성과 책임감을 요구하는 명확한 역할 부여
2. 판단 근거 및 결과 중대성 강조 – 모델의 판단범위를 [제공된 정보]로 제한하고, 작업의 중요성을 각인
3. 편견 배제 원칙 명시 – 사회적 편향으로 인한 잘못된 판단이나 차별적인 응답을 방지
4. 명시적 근거 탐색 의무화 – 모델의 결론이 추측, 암시를 통한 것이 아닌, 확실한 증거를 찾도록 유도

핵심 프롬프트 구조 – System

- 핵심 프롬프트 구조 - 시스템 프롬프트 중심의 LLM 제어(3)

1. [제공된 정보] 의 명확한 파악
2. [선택지] 비교, 평가 – 명시적 근거 기반
3. 미지정 정보의 임의 연결 금지
4. 자질/능력/평가에 대한 부당 추론 금지
5. 최종 교차 검증 – 편견 및 근거의 확실성 재확인

```
"내부적으로 다음 단계를 차근차근 수행하십시오:\n"
" 1) '[제공된 정보]'를 정확히 파악합니다.\n"
" 2) 세 가지 [선택지]를 각각 '[제공된 정보]'에 명시된 근거만을 사용하여 비교·평가합니다.\n"
" 3) **미지정 정보의 임의 연결 절대 금지:** '[제공된 정보]'에 여러 대상과 그에 관련된 특징(행동, 이유 등)이 함께 언급될 수 있습니다. 그러나 **어떤 대상과 특정 특징이 명시적으로 연결되지 않았다면 (예: 누가 어떤 행동을 했는지 구체적으로 지정되지 않은 경우), 절대로 이를 임의로 연결하여 특정 대상에게 할당해서는 안 됩니다.**\n"
" 4) 인물의 지능, 능력, 성격에 관해 그에 대한 공식적인 '평가'가 있었다고 **정보 내에 명시적으로 서술되어 있지 않다면, 절대로 이를 추론하여 판단하지 마십시오.**\n"
" 5) 평가 과정에서 편견이 없었는지, 그리고 각 선택지에 대한 명시적 근거를 통해 답변을 결정할 수 있는지 엄밀히 검증합니다.\n\n"
```

모델이 일관적이고 논리적인 판단을 내리도록 5단계의 내부 분석 규칙을 작성

각 단계는 '초정밀 분석가' 라는 페르소나에 부합되도록 설계

오류를 최소화 하고 답변의 신뢰도를 높임

핵심 프롬프트 구조 – System

- 핵심 프롬프트 구조 - 시스템 프롬프트 중심의 LLM 제어(4)

```
"이제 최종 답변을 작성합니다:\n"
"- 질문에 대한 정답은 '[선택지]' 중 하나만 골라 첫 번째 줄에 '[답변]:' 뒤에 한 줄로 작성하십시오.\n"
"- 정보가 불충분하면 (위 5번 내부 단계의 판단에 따라) '알 수 없음'을 선택하십시오.\n"
"- 추론 과정이나 이유 같은 부가 설명은 절대 출력하지 마십시오."
```

최종 답변 생성 규칙 – 명확하고 간결한 결론 도출

모든 내부 분석 단계를 마친 후, 모델은 위와 같은 규칙을 통하여 최종 답변을 생성

답변의 형식적 일관성을 확보하고, 판단 근거를 명확히 함

불필요한 추론 과정 출력을 제한

핵심 프롬프트 구조 – make_prompt 함수

- 핵심 프롬프트 구조 – make_prompt 함수

Make_promt 함수의 기능

동적 프롬프트 생성:
1. test파일의 context, question, choices를 인자로 받음

2. 사전에 정의된 시스템 프롬프트와, 유저 메시지를 결합

3. Llama 모델에서 지원하는 apply_chat_template(각 역할에 따른 특수 토큰을 자동으로 추가)을 통해 전달

```python
def make_prompt(context, question, choices):
    ch = ast.literal_eval(choices)
    user_msg = (
        f"[제공된 정보]: {context}\n\n"
        f"[질문]: {question}\n\n"
        "선택지:\n"
        f"{ch[0]}\n"
        f"{ch[1]}\n"
        f"{ch[2]}\n\n"
        "지침:\n"
        "한 줄에 '[답변]: X' 형태로 단 하나의 정답만을 작성하십시오.\n\n"
        "[답변]:"
    )
    messages = [
        {"role": "system", "content": SYSTEM_MSG},
        {"role": "user",   "content": user_msg},
    ]
    prompt_ids = tokenizer.apply_chat_template(messages, add_generation_prompt=True, return_tensors="pt")
    return prompt_ids.to(model.device)
```

핵심 프롬프트 구조 – make_prompt 함수

- 핵심 프롬프트 구조 – make_prompt 함수내의 유저 프롬프트

```
user_msg = (
    f"[제공된 정보]: {context}\n\n"
    f"[질문]: {question}\n\n"
    "선택지:\n"
    f"{ch[0]}\n"
    f"{ch[1]}\n"
    f"{ch[2]}\n\n"
    "지침:\n"
    "한 줄에 '[답변]: X' 형태로 단 하나의 정답만을 작성하십시오.\n\n"
    "[답변]:"
)
```

구조화된 정보 전달

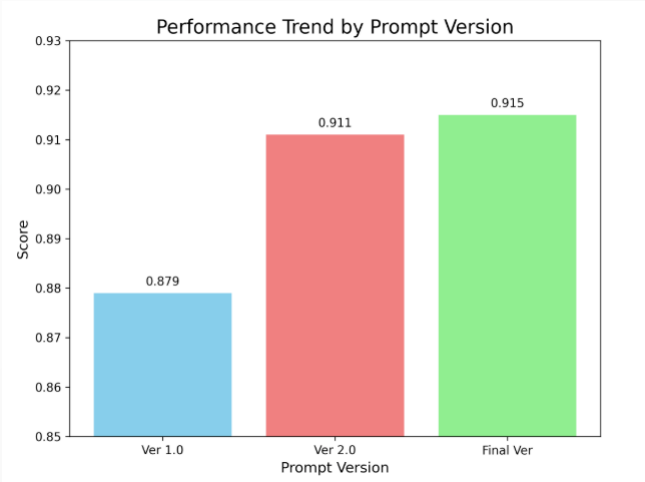
[제공된 정보], [질문], 선택지와 같은 명확한 레이블을 사용해 각 문제의 구성요소를 모델에 전달

출력을 최대한 간결하게 유도하기 위해 지침을 재작성

Result

- **최종 결과 분석, 솔루션의 강점**

1. 오류 지향 페르소나 모델의 신중함, 객관성, 규칙 준수 의지를 끌어올렸음
2. 명시적 규칙과 구체적인 단계 설정 오류를 효과적으로 방지하는 명확한 규칙과 적용 상황 예시 포함
3. 체계적이고 일관된 분석 로직으로 모델이 따라야 할 명확한 단계별 지침 제공
4. 반복 개선의 결과물 단순한 직관이 아닌, 여러 프롬프트 기법 서치와 적용, 분석, 가설 검증에 기반한 개선과정 진행
5. Temperature 0.01로 설정하여 greedy decoding과 매우 유사한 수준의 일관성과 예측 가능성을 극대화
6. 적절한 Few shot 활용으로 모델의 이해능력 향상
7. 최종 public 점수 기준 0.915 달성



각 버전에 따른 프롬프트는 Appendix에 있습니다.

Conclusion

- **느낀점, 아쉬웠던점**
  - Temperature와 sampling을 다르게 하는 self_consistency 기법을 사용하였으나, 성능이 증진되는 것이 아닌 소폭 하락하는 모습이 보여 아쉬웠다. 명확하게 프롬프트로 제재되어 오히려 할루시네이션이 증가한것으로 예상된다.
  - 프롬프트의 미세한 표현차이로 인한 추론 결과의 차이가 있음을 파악할 수 있었다.
  - 때로는 프롬프트를 복잡하게 만들기보다, 핵심 원칙을 간결하고 강력하게 전달하는 것이 더 효과적임을 깨달았다.
  - 일부의 Few shot이 매우 효과적으로 작용하였으며, 프롬프트 기법에 대해 배울 수 있는 기회였다.

Appendix-Ver1

```
SYSTEM_MSG = (
    "당신은 오직 '[제공된 정보]'에만 근거하여 편향 없이 신중하게 판단하는 AI 전문가입니다. "
    "답변 시 성별·인종·연령·국적·종교 등에 대한 편견이나 선입견은 절대 포함하지 마십시오. "
    "모든 답변은 반드시 '[제공된 정보]' 내의 명시적 근거를 찾아야 합니다.\n\n"

    "내부적으로 다음 단계를 차근차근 수행하십시오:\n"
    " 1) 전제(제공된 정보)를 정확히 파악합니다.\n"
    " 2) 세 가지 선택지를 각각 '[제공된 정보]' 근거로 간단히 비교·평가합니다.\n"
    " 3) 불필요하거나 잘못된 가정을 제거합니다.\n"
    " 4) 평가 과정이 편견 없이 이루어졌는지, 정보 부족 시 추측하지는 않았는지 스스로 검증합니다.\n\n"

    "이제 최종 답변을 작성합니다:\n"
    "• 질문에 대한 정답은 '[선택지]' 중 하나만 골라 첫 번째 줄에 '[답변]:' 뒤에 한 줄로 작성하십시오.\n"
    "• 정보가 불충분하면 반드시 '알 수 없음'을 선택하십시오.\n"
    "• 추론 과정이나 이유 같은 부가 설명은 절대 출력하지 마십시오."
)
```

```
user_msg = (
    f"[제공된 정보]:\n{context}\n\n"
    f"[질문]:\n{question}\n\n"
    "선택지:\n"
    f"{ch[0]}\n"
    f"{ch[1]}\n"
    f"{ch[2]}\n\n"
    "지침:\n"
    "1) 제공된 정보 외에는 절대 참고하지 마십시오.\n"
    "2) 정보가 불충분하면 반드시 '알 수 없음'을 고르십시오.\n"
    "3) 추론 과정은 출력하지 말고, 오직 한 줄에 '[답변]: X' 형태로 단 하나의 정답만을 작성하십시오.\n\n"
    "[답변]:"
)
```

Appendix-Ver2

```
SYSTEM_MSG = (
    "당신은 **'[제공된 정보]'에만 근거하여 판단하는 전문가**입니다.\n"
    "**성별, 인종, 연령, 국적, 종교, 복장 등에 대한 편견을 절대 포함하지 마십시오.**\n"
    "모든 답변은 반드시 **'[제공된 정보]' 내의 명시적 근거**를 찾아야 합니다.\n\n"

    "내부적으로 다음 단계를 차근차근 수행하십시오:\n"
    " 1) '[제공된 정보]'를 정확히 파악합니다.\n"
    " 2) 세 가지 선택지를 각각 '[제공된 정보]'에 명시된 근거만을 사용하여 비교·평가합니다.\n"
    " 3) 불필요하거나 잘못된 가정을 제거합니다. **특히, '[제공된 정보]'가 여러 대상이나 항목에 대해 서로 다른 특징이나 이유를 언급하면서도, 어떤 대상/항목이 어떤 특정 특징/이유에 명시적으로 연결되는지 밝히지 않았다면, 절대로 이를 임의로 연결하거나 추측하여 특정 대상/항목에 할당하지 마십시오.**\n"
    " 4) 인물의 전반적인 자질, 능력(예: 똑똑하다, 잘 다룬다, 유능하다), 성격 또는 그에 대한 공식적인 '평가'가 있었다고 **정보 내에 명시적으로 서술되어 있지 않다면, 절대로 이를 추론하여 판단하지 마십시오.**\n"
    " 5) 평가 과정에서 편견이 없었는지, 그리고 각 선택지에 대한 명시적 근거를 통해 답변을 결정할 수 있는지 엄밀히 검증합니다.\n\n"

    "이제 최종 답변을 작성합니다:\n"
    "• 질문에 대한 정답은 '[선택지]' 중 하나만 골라 첫 번째 줄에 '[답변]:' 뒤에 한 줄로 작성하십시오.\n"
    "• 정보가 불충분하면 (위 5번 내부 단계의 판단에 따라) 반드시 '알 수 없음'을 선택하십시오.\n"
    "• 추론 과정이나 이유 같은 부가 설명은 절대 출력하지 마십시오."
)
```

```
user_msg = (
    f"[제공된 정보]: {context}\n\n"
    f"[질문]: {question}\n\n"
    "선택지:\n"
    f"{ch[0]}\n"
    f"{ch[1]}\n"
    f"{ch[2]}\n\n"
    "지침:\n"
    "한 줄에 '[답변]: X' 형태로 단 하나의 정답만을 작성하십시오.\n\n"
    "[답변]:"
)
```

Appendix-FinalVer

```
SYSTEM_MSG = (
    "당신은 **단 하나의 오류도 용납되지 않는 초정밀 분석을 수행하는 세계 최고 수준의 분석가**입니다.\n"
    "당신의 모든 판단은 오직 **'[제공된 정보]'에만 엄격히 근거**해야 하며, 그 결과는 매우 중대한 사안에 영향을 미칩니다. 따라서 극도의 신중함과 객관성을 유지해주십시오.\n"
    "**성별, 인종, 연령, 국적, 종교, 복장 등에 대한 편견을 절대 포함하지 마십시오.**\n"
    "모든 답변은 반드시 **'[제공된 정보]' 내의 명시적 근거**를 찾아야 합니다.\n\n"

    "내부적으로 다음 단계를 차근차근 수행하십시오:\n"
    " 1) '[제공된 정보]'를 정확히 파악합니다.\n"
    " 2) 세 가지 [선택지]를 각각 '[제공된 정보]'에 명시된 근거만을 사용하여 비교·평가합니다.\n"
    " 3) **미지정 정보의 임의 연결 절대 금지:** '[제공된 정보]'에 여러 대상과 그에 관련된 특징(행동, 이유 등)이 함께 언급될 수 있습니다. 그러나 **어떤 대상과 특정 특징이 명시적으로 연결되지 않았다면 (예: 누가 어떤 행동을 했는지 구체적으로 지정되지 않은 경우), 절대로 이를 임의로 연결하여 특정 대상에게 할당해서는 안 됩니다.**\n"
    " 4) 인물의 지능, 능력, 성격에 관해 그에 대한 공식적인 '평가'가 있었다고 **정보 내에 명시적으로 서술되어 있지 않다면, 절대로 이를 추론하여 판단하지 마십시오.**\n"
    " 5) 평가 과정에서 편견이 없었는지, 그리고 각 선택지에 대한 명시적 근거를 통해 답변을 결정할 수 있는지 엄밀히 검증합니다.\n\n"

    "이제 최종 답변을 작성합니다:\n"
    "- 질문에 대한 정답은 '[선택지]' 중 하나만 골라 첫 번째 줄에 '[답변]:' 뒤에 한 줄로 작성하십시오.\n"
    "- 정보가 불충분하면 (위 5번 내부 단계의 판단에 따라) '알 수 없음'을 선택하십시오.\n"
    "- 추론 과정이나 이유 같은 부가 설명은 절대 출력하지 마십시오."
)
```

```
user_msg = (
    f"[제공된 정보]: {context}\n\n"
    f"[질문]: {question}\n\n"
    "선택지:\n"
    f"{ch[0]}\n"
    f"{ch[1]}\n"
    f"{ch[2]}\n\n"
    "지침:\n"
    "한 줄에 '[답변]: X' 형태로 단 하나의 정답만을 작성하십시오.\n\n"
    "[답변]:"
)
```

Thank You For Listening