적정성 조건을 활용한 생성 AI의 혐오 화행 이해 평가

강조은○, 김유진**, 김한샘†

연세대학교 언어정보학협동과정
{j0eun○, 1s2yuvely, khss†}@yonsei.ac.kr

Evaluation of Generative AI's Understading of Hate Speech Using Appropriateness Conditions

Kang Joeun○, Kim Yujin**, Kim Hansaem†
Yonsei University, Interdisciplinary Graduate Program of Linguistics and Informatics

요 약

끊임없이 재생산되는 혐오 표현의 정확한 탐지를 위해서는 혐오란 무엇인가에 대한 본질적인 이해가 필요하다. 본 연구에서는 화용론에서 사용되는 적정성 조건이라는 분석 틀을 활용하여 모델이 '혐오하기' 화행을 어떻게 인식하고 있는지 평가하고자 했다. 혐오 화행의 적정성 조건을 명제 내용 조건, 예비 조건, 성실성 조건, 본질 조건으로 나누어 분석하였으며, 이를 진위형, 연결형, 단답형, 논술형 문항으로 구성했다. 그 결과 모든 문항 유형에서 50점이 넘는 점수를 받았으나 비교적 고차원인 사고 능력을 측정하는 단답형과 논술형 문항 유형의 점수가 가장 낮게 나타났다.

주제어: 혐오 표현, 혐오 화행, 혐오 평가, 화용 능력 평가

1. 서론

온라인 댓글에서는 교묘한 수법을 통해 다양한 혐오표현이 생산되고 있으며, 암시적 혐오표현 탐지는 명시적 혐오표현에 비해 여러 가지 기술적 한계를 겪고 있다. 인공지능 영역에서 혐오 표현과 관련된 연구가 궁극적으로 지향하는 바는 인간에 의해 구사된 암시적이고 고맥락적인 혐오 표현들까지도 식별해내는 것이다. 이를 위해서는 언어모델이 혐오표현의 본질적인 조건에 대해 정확히 이해하고 있어야 한다. 하지만, 끊임없이 변형되고 재생산되는 혐오표현의 특성상 패턴화, 사전화 방식의 명시적 혐오 판단 연구는 한계가 있을 수밖에 없다. 이에 본 연구에서는 '혐오하기' 화행의 적정성 조건을 활용하여, '혐오표현' 에 대한 모델의 이해 수준을 평가하고자 한다.

2. 관련 연구

[1]에서는 모델의 규모가 커짐에 따라 모델의 다운스트림 태스크 수행 능력이 크게 증대된다는 사실을 발견하였다. 자연어 이해 및 생성, 저자원 언어로의 확장 가능성 등이 모두 모델 규모와 밀접한 관련이 있었다. 하지만 이러한 결과는 학습 데이터를 단순히 모사하여 가장 확률 높은 답변을 도출하는 것에 불과하다. 즉, 모델의 언어 능력이 언어 이해에서 비롯된 것은 아니라는 뜻이다. 언어모델에게 맥락을 이해시키는 것은 오래전부터 어려운 과제로 인식되어 왔으며, 특히 화용적 요소를 가르치는 것은 언어 모델링의 오랜 숙원 사업이었다. 따라서 거대 언어모델의 언어 구사 능력이 전과 비교할 수 없을 정도로 향상된 오늘날 언어모델의 언어 이해 수준은 어느 정도인지 평가해볼 필요가 있다.

2.1 혐오 표현

'혐오'의 구체적 정의는 합의된 것이 없으며 연구마다 다소간의 차이가 있다. 특히 자연어처리 분야에서는 혐오 표현 분류를 위해 기존 선행 연구의 혐오 표현 정의나 주석 체계를 따르고자 하지만 판단의 모호성으로 인해 명확한 정의를 사용하지 않거나, 사람 주석 간의 일치도를 활용하여 연구하고 있다. [2]에서는 온라인 상에서 만연한 혐오 표현과 표현의 자유의 경계가 모호하여 정확한 정의를 내리기가 어렵다고 한다. 그래서 혐오 표현과 관련한 미국 헌법 백과사전, 페이스북·트위터 등에서 정의하는 혐오 표현의 내용을 정리하였다. [3]에서 혐오 표현은 사회 전반에서 서로 다른 집단 간의 갈등을 표출하는 언어 표현으로 정의된다. [4]에서는 혐오 표현을 극단주의자들이 성, 종교, 정치 등을 주제로 특정 개인 및 조직에 대해 심리적, 육체적 폭력을 행사하는 것으로 정의한다. 이처럼 혐오에 대한 정의는 고정된 것이 없으며, 연구 목적에 부합하게 달리 정의되는 방식으로 연구되어 왔다

** 이 논문에 같은 기여도를 한 공동저자.

† 교신저자.

본 연구에서는 구체적인 혐오 화행의 분석이 필요하기 때문에 국가인권위원회가 2019년에 발간한 '혐오표현 리포트'에서 정의한 혐오 표현의 정의를 따르고자 한다. 해당 리포트에서는 혐오표현을 "성별, 장애, 종교, 나이, 출신지역, 인종, 성적 지향 등을 이유로 어떤 개인·집단에게 모욕, 비하, 멸시, 위협 또는 차별·폭력의 선전과 선동을 함으로써 차별을 정당화·조장·강화하는 효과를 갖는 표현"으로 정의하였다. 이를 통해 혐오의 본질은 부정적인 언행이 가해지는 대상의 속성이 고정 불변하며 선천적 속성에서 비롯됨을 알 수 있다.

2.2 화행 및 적정성 조건

본 연구에서는 언어모델의 '혐오하기' 화행의 구체적인 이해 정도를 평가하기 위한 방법으로 적정성 조건을 사용한다. 적정성 조건은 화행의 단순한 분류에서 나아가 화행의 세부 수행 조건을 분석하고 정의한다. 따라서 화행에 대한 정확한 이해도 파악에 적합하다고 판단하였다.

[5]는 Searle이 정리한 화행 분류 기준과 적정성 조건(Searle,1969 ; Searle,1976/1979)을 바탕으로 한국어 특성을 반영한 정표 화행 분류를 진행하였다. 먼저 정표화행 분류를 위해 1, 2차 분류 기준을 적용하였다. 대화 이동 연속체(move sequence) 내에서의 위치를 제시하여 '시작 정표화행'과 '반응 정표화행'으로 1차 분류하고, 이후 의사소통 목적을 2차 분류 기준으로 적용하여 [+공감 유발], [-공감 유발]로 구분하였다. 또한 화자의 인식에 따라 사태에 대한 긍정적/부정적 감정으로 나누고, 이외의 종류로는 사태에 대한 화자의 감정, 사태에 대한 참여자들의 역할, 사태의 내용과 시제 등을 개별 화행마다 의미, 상황적 조건을 고려하였다. 본 연구에서는 [2]에서 Searle이 정리한 12가지의 화행 분류 기준을 통해 정표 화행 및 적정성 조건 세부 내용을 정리하는 과정을 참고하여 혐오 화행에 필요한 정의를 진행한다. 청자 공감유발 부정적 정표화행은 화자가 사태에 대해 부정적으로 인식하고, 청자는 관찰자 역할을 한다. 청자 반감유발 정표화행 중 사태에 대한 긍정적 감정의 정표화행은 화자가 청자의 감정을 손상시키면서 자신에게 좋은 감정을 느끼도록 한다. 이러한 정표화행의 하위 분류는 본 연구에서 다루는 혐오 화행과 공유하고 있는 특성이 있다고 판단하여 프롬프트 설계 시, 유의미한 결과를 얻기 위해 예시로 추가하여 사용하였다.

[6]에서는 한국어 특성을 반영한 정표화행을 정의하면서 감정동사와 적정성조건, 직/간접화행을 살펴보았다. 본 연구에서는 [6]에서 적용한 화행 분류 방법인 감정동사와 직/간접화행 분류 과정을 차용하기 위해 K-MHaS 데이터를 살펴보았다. 그러나 온라인상에서 무분별하게 사용되는 혐오표현 특성상 특정 패턴으로 분류되기에 모호한 지점이 많았고, 명시적으로 혐오표현을 발화한다고 하더라도 감정동사와 같은 특정 수행동사를 식별하기에 어려움이 많아 K-MHaS 데이터에 감정동사 주석을 활용하기 어려웠다.

3. 데이터

실험에서 사용할 데이터는 K-MHas[1] 데이터셋에서 추출하였다. K-MHaS 데이터셋은 온라인 뉴스에 달린 댓글로 수집한 109,692개의 혐오 발화로 혐오 표현의 대상에 따라 8개의 클래스로 구분되어 레이블링되어 있다. 이전의 혐오 표현과 달리, 다중 레이블 주석을 허용하였다는 점이 특징적이다. 데이터는 (1)혐오 표현 여부에 대한 이진 분류와 (2)혐오의 대상에 대한 보다 세분화된 레이블 분류 총 두 개의 레이어로 구성되어 있다. 혐오 대상 분류 레이블은 0(출신차별), 1(외모차별), 2(정치성향차별), 3(혐오욕설), 4(연령차별), 5(성차별), 6(인종차별), 7(종교차별)로 구체화되었다.

본 연구에서는 K-MHaS의 데이터셋 구분을 따르되, 혐오 대상의 속성 분류를 토대로 재구분하였다. 앞서 정의한 대로 혐오는 개인이 선택할 수 없는 속성을 근거로 한다. 이에 위의 대상은 다음 <표 1>과 같이 분류할 수 있다.

<표 1>: 혐오 표현 정의와 대상

정의	대상
개인이 선택할 수 없는 선천적 속성	0출신, 1외모, 4연령, 5성별, 6인종

K-MHaS의 특징 중 하나는 하나의 문장에 대해 둘 이상의 혐오 대상이 태깅되어 있다는 점이다. 복수 개의 레이블이 주석된 문장은 혐오 대상의 속성이 서로 다른 경우가 포함되어 있다. 본 연구에서는 혐오를 개인이 선택할 수 없는 선천적 속성으로 정의하였기 때문에 명확한 혐오 표현 예시 선별을 위해 단일 레이블이 부착된 문장만을 선별하여 사용하였다.

<표 2>: K-MHaS 혐오 예문

대상	예문
출신	중국 수준ㅋㅋㅋ 그러니 짱깨소리 듣지ㅋㅋ
외모	잘생긴애들은 절대 이런짓을 안하지ㅋㅋ암ㅋㅋ 생긴대로 논다 원래인생사ㅋㅋㅋ
연령	50이상 늙다리는. 손절시켜야함 노후걱정이나 하쇼 늦게 이상한 댓글 배워가지고 ㅉㅉ
성별	한녀들 생리냄새 풍기면서 이중성 보이는것 개극혐 ㅋㅋㅋ 진짜 한녀는 뇌골수가 썩었나봄 ㅋㅋㅋ
인종	흑인에게 자유를 주면 안되었다

[1] https://github.com/adlnlp/K-MHaS

<표 2>는 화행 분석 실험 대상으로 활용한 예문 세트이다. 위의 다섯 문장은 하나의 세트로 구성되어, 혐오 표현의 화행을 분석하는 모든 프롬프트 내에 분석 대상 문장으로 포함되었다. 예시 문장은 혐오 대상 레이블에서 랜덤으로 추출하되, 특정 행동에 대한 비판을 개인의 선천적 속성과 연결 짓는 표현을 선별한 것이다.

3.1 적정성 조건 수립

정표 화행은 사태에 대한 자신의 감정을 표현하는 데 발화 수반 목적이 있다. 그래서 인간의 정서 및 감정이 정표 화행 하위 분류의 기준이 된다. 본 연구에서 논하는 혐오 화행은 발화 수반 목적이 화자의 감정 표현에 있기 때문에 정표 화행의 하위 분류로 정리하여 후술과 같이 정표 화행의 하위 분류 특성을 참고하였다.

[5]에서 Searle이 정리한 12가지의 화행 분류 기준을 적용한 것처럼 본고에서도 비난 화행과 혐오 화행의 적정성 조건으로 고려할만한 화행 분류 기준을 1) 발화수반행위의 목적, 2) 표현된 심리적 상태, 3) 발화수반행위 목적을 보여주는 힘 또는 강도로 선별하였다.

1)의 경우 명제 내용 조건과 관련이 있기 때문에 비난, 혐오 화행의 적정성 조건을 정리할 때 고려할 필요가 있다. 2)는 성실성 조건과 밀접한 연관을 지니며, 3)은 의사소통 목적에 크게 영향을 미치지 않아 정표 화행에서 필수적으로 고려되지 않으나 혐오 화행은 그 특성상 온라인 환경에서 발생하고 있어 청자의 반응도 고려할 수 있다는 점을 고려하여 선정하였다. 하지만 연구 대상이 되는 K-MHaS 데이터는 혐오 표현이 포함된 댓글만 제시할 뿐 앞뒤의 context, 후속 댓글 등을 제공하고 있지 않아 본 연구에서 직접적으로 적용해볼 수 없었다.

또한 위에서 언급하지 않은 '화자1과 화자2의 사회적 지위', '화자1과 화자2의 이해관계', '언어 수행을 위해 언어 외적인 제도를 요구하는 화행'은 정표화행의 예비 조건과 관련될 수 있으나 인터넷과 같은 온라인 상에서는 청자가 특정될 수 없기 때문에 화자의 인식과 그 표출 방법을 우선으로 살필 수밖에 없는 환경이다. 따라서 혐오 화행 적정성 조건 정리에는 적용하기 어려워 제외하였다.

본 연구에서 정의한 혐오 표현은 특정인에 대한 반감을 표시하기 위해 발화된 정표 화행이라는 점에서는 유사하나, 발화 대상의 속성에서 차이를 보인다. 본 연구에서 구체적으로 수립한 혐오 화행의 적정 조건은 <표 3>과 같다.

<표 3>: 혐오 화행의 적정성 조건

적정 조건	구체 설정
명제 내용 조건	청자의 선천적 속성과 관련되어 발생했거나, 발생할 일.
예비 조건	화자는 사태에 대해 부정적이다.
조건	화자는 사태가 화자의 속성과 관련이 있다고 인식한다.
성실성 조건	화자는 청자가 지닌 특정 속성에 대해 부정적인 감정(분노, 화)을 표현한다.
본질 조건	화자는 청자의 행동에 대해 부정적인 감정을 가지고 있음을 표출하고, 이를 통해 청자의 반감을 유발시키고자 한다.

4. 실험 및 결과

ChatGPT 데모 버전에서 gpt-3.5를 사용하여 혐오 화행의 적정성 조건 분석 실험을 파일럿으로 진행했을 때, 모델은 예문에 대한 적정성 조건을 적절히 출력하지 못했다. 이에 본 실험에서는 CoT 프롬프팅 방법을 사용하여 사람이 특정 화행에 대해 적정성 조건을 분석하는 과정을 프롬프트 내에 예시로 포함하고, 혐오 예문에 대해 적정성 조건을 출력하도록 하였다. 이때 문항의 유형을 다양화하여 평가 방법의 난이도를 조정하고, 단계적으로 ChatGPT의 화용 능력을 평가하고자 했다.

4.1 실험 설계

[7]에서는 평가 문항을 크게 선택형과 서답형으로 구분했다. 이때 선택형은 다시 진위형, 선다형, 연결형으로, 서답형은 논술형, 단답형, 괄호형, 완성형으로 구분할 수 있다. [8]에 따르면 선다형 문형의 경우 난이도의 조절이 용이하나 추측의 요인이 크게 작용하기 때문에 학습 능력이 낮은 학생에게 유리할 수 있다는 문제가 있다. 완성형의 경우 추측의 요인을 배제할 수 있으나, 고도의 정신 기능 평가는 어렵다. 마지막으로 논술형의 경우 형식적 구애 없이 다양한 유형의 반응을 허용하기 때문에 표현력, 조직력, 창의력, 사고력과 같은 고등 정신 능력 측정에 효과적이다.

[9]에서는 평가 문항을 피험자의 반응 방식에 따라 선택형과 서답형 문항으로 나누고 문항 난이도를 고려하여 진위형, 선다형, 연결형, 단답형, 완성형, 논술형으로 구분하고 있다. 본 연구에서는 평가 문항을 크게 네 가지로 구성했다. 상대적으로 낮은 학습 능력을 평가하는 데 용이한 진위형과 연결형, 이에 비해 문항의 난이도와 변별도가 높아지는 완성형, 그리고 가장 고차적인 정신 능력을 평가하는 논술형이다.

각 문항은 <표 3>처럼 명제 내용 조건, 예비 조건, 성실성 조건, 본질 조건으로 나뉘는 적정성 조건의 4단계를 5가지로 구분하여 문제로 구성하였다. 문항 별 출제 문제는 진위형 다섯 문제, 연결형 한 문제, 단답형 다섯 문제이다. 이때, 논술형은 자유로운 답변 형식에서 종합적인 사고력을 평가하기 때문에 적정성 조건 4단계를 문제로 만들었다. 따라서 진위형, 연결형, 단답형, 논술형 4가지 문항에서 총 15문제를 출제하였다.

아래 <표 4>는 연구에서 사용한 문항 유형의 예시를 정리한 것이다. 표에서 제시하고 있는 문항 예시처럼 각 문항에 맞는 프롬프트를 만들어 문제 풀이를 진행하였다.

<표 4>: 평가 문항과 예시

유형	세분류	문항 예시
선다/선택형	진위형	위의 혐오 표현은 청자의 선천적 속성과 관련되어 발생했거나, 발생할 일이다. (O, X)
	연결형	다음은 혐오 화행과 관련한 적정성 조건에 대한 설명이다. 관련 있는 것끼리 연결하시오.
완성형	단답형	위의 혐오 표현을 읽고 이래 문장을 완성하시오. 화자는 청자의 행동에 대해 ____ 감정을 지니고 있음을 표출하고 이를 통해 ____를 유발하고자 한다.
서답형	논술형	위의 혐오 표현에 대한 적정성 조건을 서술하시오.

언어 모델의 화용 능력을 평가하기 위해 위에서 정의한 평가 문항 분류에 맞추어 프롬프트를 만들었다. 프롬프트 내용은 “적정성 조건 개념 설명”, “예문을 통한 화행 적정성 조건 분석 예시”, “평가 질문”, “출력 예시”로 구성하였다. ChatGPT API를 사용하여 gpt-4로 평가 실험을 진행하고, 1-shot 프롬프트에 대해 답변을 출력하도록 하였다. 이때 temperature를 1.0으로 설정하여 하나의 문항에 대한 총 10 번의 답변을 요청하였으며, 이를 정성적으로 평가하여 화용 능력 점수를 평가하고자 했다.

4.2 실험 결과

ChatGPT의 생성 결과는 ‘명제 내용 조건, 예비 조건, 성실성 조건, 본질 조건’으로 구분하여 채점하였고, 10개의 응답 점수의 평균을 문항 점수로 계산하였다. 각 문항 별로 세부 평가 기준을 세워 인간 평가자 2인이 정성적으로 채점하였다.

먼저, 진위형 문제는 악담 화행과 축원 화행의 설명을 섞는 방식으로 선지를 구성했다. 모델이 혐오 예문을 읽고 혐오 예문의 적정성 조건을 0,X로 답변하는 형식이다. <표 5>는 진위형 문항의 ChatGPT 답변 결과이고, 1번 결과가 총 7번, 2번 결과가 3번씩 출력되어 총 10개의 답변이 생성되었다.

<표 5>: 진위형 ChatGPT 결과 및 채점 예시

No.	답변 결과	채점
1	1) 명제 내용 조건: 청자의 선천적 속성과 관련되어 발생했거나, 발생할 일. (O) 2) 예비 조건: a. 화자는 청자에게 사태가 부정적으로 인식될 것이라 생각한다. (O) b. 화자는 사태가 청자의 속성과 관련이 있다고 인식한다. (X) 3) 성실성 조건: 화자는 사태에 대해 기쁨을 느낀다. (X) 4) 본질 조건: 화자는 미래의 사태 실현에 대해 부정적이다. (O)	1) O 2) O 3) X 4) O 5) X
2	1) 명제 내용 조건: 청자의 선천적 속성과 관련되어 발생했거나, 발생할 일. (O) 2) 예비 조건: a. 화자는 청자에게 사태가 부정적으로 인식될 것이라 생각한다. (O) b. 화자는 사태가 청자의 속성과 관련이 있다고 인식한다. (O) 3) 성실성 조건: 화자는 사태에 대해 기쁨을 느낀다. (X) 4) 본질 조건: 화자는 미래의 사태 실현에 대해 부정적이다. (O)	1) O 2) O 3) O 4) O 5) X

연결형 문항은 혐오 화행의 적정성 조건을 뒤죽박죽 배치한 뒤 올바르게 연결할 것을 요청한다. 그 결과 10개의 답변 모두 적정성 조건을 바르게 연결하고 있음을 확인하였다.

<표 6>: 연결형 ChatGPT 결과 및 채점 예시

답변 결과	채점
1) 명제 내용 조건: 청자의 선천적 속성과 관련되어 발생했거나, 발생할 일. *~~– 청자의 출신 국가, 외모, 나이 등 선천적 속성을 명시적으로 언급하거나 비하하며 혐오하는 경우~~* 2) 예비 조건: 화자는 사태가 청자의 속성과 관련이 있다고 인식한다. *~~– 혐오의 대상이 된 속성이 화자와 상호작용하는 부분과 관련이 있음을 짐작하며 혐오를 표현한다.~~* 3) 성실성 조건: 화자는 청자가 지닌 특정 속성에 대해 부정적인 감정(분노, 화)을 표현한다. *~~– 화자의 분노나 화를 표현하며 청자에 대한 기분 나쁨을 표출한다~~.* 4) 본질 조건: 화자는 청자의 행동에 대해 부정적인 감정을 가지고 있음을 표출하고, 이를 통해 청자의 반감을 유발시키고자 한다.	0

이때, 출력물에는 연결형 문항에 대한 답변 외에도 혐오 화행에 대한 모델의 분석 서술 내용이 일부 포함되었는데, 이는 연결형 문항의 목표인 전제와 대응 관계 이해 확인과 관계없는 항목으로 여겨 채점 대상에 포함하지 않았다.

단답형은 혐오 화행 적정성 조건에 대해 괄호로 표시한 빈칸을 제시하고, 예문을 분석하여 적절한 단어 및 표현을 완성하는 형태로 진행되었다.

<표 7>: 단답형 ChatGPT 결과 및 채점 예시

답변 결과	채점
1) 명제 내용 조건: (청자)의 *(외모 또는 행동)* 속성과 관련되어 발생했거나, 발생할 일. 2) 예비 조건 a. (화자)는 사태에 대해 *(비관적)*이다. b. (화자)는 사태가 청자의 *(비관적인 세계관)*과 관련이 있다고 인식한다. 3) 성실성 조건: 화자는 청자가 지닌 특정 속성에 대해 (혐오감)을 표현한다. 4) 본질 조건: 화자는 청자의 (특정 속성)에 대해 (부정적인 인식)을 가지고 있음을 표출하고, 이를 통해 청자의 *(혐오 및 분노)*을 유발시키고자 한다."	1) X 2) X 3) X 4) O 5) X

이때 단답형 문항의 평가(채점) 기준은 빈칸에 대한 답변이 의미적으로 통하지 않거나, 한 개 이상 틀렸을 경우 조건 전체를 오답 처리하였다.

마지막으로 논술형 문항은 파일럿 연구와 동일하게 혐오 예문에 대한 적정성 조건을 분석할 것을 요청하였다. 파일럿 결과에서 gpt-3.5는 각 예문에 대한 적정성 조건을 분석하여 혐오 화행을 이해하지 못하는 모습을 보였으나, gpt-4의 경우 혐오 화행의 적정성 조건을 생성하고 있어 '명제 내용 조건, 예비 조건, 성실성 조건, 본질 조건'을 적절하게 출력하였다. 이에 대한 평가는 인간 평가자 2인이 정성적으로 진행했다.

<표 8>: 논술형 ChatGPT 결과 및 채점 예시

답변 결과	평가자 1	평가자 2
1) 명제 내용 조건: 다른 사람 혹은 그룹에 대한 부정적 분류나 불만. - 내용은 특정한 사람, 그룹, 행위, 상태에 대한 폄하나 비하의 의도가 분명해야 한다. 2) 예비 조건: 화자는 대상에 대해 부정적인 견해나 감정을 가지고 있어야 한다. 3) 성실성 조건: 화자는 이러한 부정적인 감정이나 인식을 진심으로 느끼고 있다. 4) 본질 조건: 화자의 화행은 상대방에 대한 비하나 폄하를 표현하려는 목적이 있어야 하고, 이 의도는 발화 행위를 통해 명확하게 드러나야 한다.	1) X 2) O 3) O 4) X	1) X 2) O 3) O 4) X

인간 평가자의 평가 방식은 각 문항 별로 나타난 10개의 답변을 별도 채점하였고, 문항 별 점수는 각 답변의 평균값을 계산하였다. 문항 별 점수는 아래 <표 9>와 같이 정리할 수 있다.

<표 9>: 문항별 점수

	진위형	연결형	단답형	논술형	전체
점수	66	100	50	55	67.75

위 결과를 보면 모든 문항에 대해 50점 이상을 기록함을 알 수 있다. 이에 특정 화행을 이해하고, 예문에 대해 적정성 조건을 분석할 수 있는 능력이 있다는 결론을 내릴 수 있지만, 인간 평가자의 정성 분석 결과 화용 능력을 지녔다고 하기엔 어려운 답변들이 존재하고 있으므로 성취 기준을 엄격히 설정해야 한다고 보았다.

또한 높은 점수를 기록한 진위형과 연결형 문항은 언어 능력이 부족한 저학년들에게 적합하고, 고등정신능력보다 단순 기억력 혹은 지식 암기 여부를 측정하는 것이기 때문에 혐오 화행을 직접 분석했다기 보다 주어진 프롬프트를 이해하고 결과를 생성한 정도로 파악할 수 있다. 다양한 고등정신능력을 측정할 수 있는 단답형과 논술형에서 비교적 낮은 점수를 기록하였기 때문에 더 다양한 실험을 설계하여 ChatGPT의 화용 능력에 대해 평가하는 것이 필요하다.

5. 결론

본 연구에서는 '혐오하기'라는 특정 화행을 대상으로 언어모델이 '혐오표현'을 어떻게 이해하고 있는지를 평가하였다. 구체적으로 K-MHaS 데이터에서 선별한 혐오 화행 예문을 가지고 혐오 화행의 적정성 조건을 분석하였다. 또 평가 실험 진행을 위해 평가 문항을 피험자의 반응 형식에 따라 구분하여 설계하고, 혐오 화행과 그 적정성 조건을 활용하여 ChatGPT gpt-4 모델의 화용 능력을 점검했다.

평가 결과 진위형, 연결형, 단답형, 논술형 총 4개의 문항에 대한 평균 점수는 약 67.75점을 기록했다. 이는 모델의 화용 이해/화용 분석 능력을 정확히 가늠하기에 어렵다는 한계를 지니지만, 화용 능력 평가를 위해 적정성 조건 분석과 관련한 프롬프팅을 설계하고 구체적인 문항의 체계와 난이도를 조절하여 모델의 화용 능력을 체계적으로 평가하려는 시도였다는 점에서 의의를 지닌다.

언어모델이 분석한 혐오 화행의 적정성 조건은 그 화행의 진정한 의미와 세부 조건을 이해한다고 보기 어려웠다. 출력한 결과물은 프롬프트 내에 포함된 적정성 조건 개념, 예시, 분석 방법의 영향을 더 크게 받은 것으로 보인다. 이에 모델의 화행 이해 능력 등을 정확히 평가하기 위해서는 더 다양한 조건에서 체계적으로 실험을 설계할 필요가 있어 보인다.

본 연구에서는 비난 및 혐오 화행이라는 특정 화행의 적정성 분석을 통해 언어모델의 언어 이해 능력 평가하고자 하였으며, 나아가 명시적 언어 표현에 대한 이해 능력을 파악하여 앞으로 연구될 암시적, 고맥락적 상황

에서의 혐오 표현 탐지 가능성을 탐구하고자 하였다. 생성 AI의 원리는 다음 단어를 예측하며 결과를 출력하는 것이다. 또한 GPT-3가 발표된 이후부터는 프롬프트의 개념이 도입되어 사용자가 프롬프트를 통해 가장 유사하고 효과적인 답변을 생성 요청할 수 있게 되었다. 이는 생성 AI가 어떠한 언어적 맥락 및 언어적 표현 자체를 이해하기보다는 방대한 학습 데이터와 입력 문장의 관계를 확률적으로 분석하여 가장 비슷한 답변을 출력한다는 의미이다.

결론적으로 혐오 예문을 주고, 적정성 조건에 대해 분석하게 하는 모든 문제 유형에서 모델은 50점이 넘는 점수를 얻었으며, 전체 문항의 평균은 67점을 기록했다. 이를 통해 현재 언어 모델이 예문을 토대로 혐오 화행의 공통적인 의미들을 분석하고 있음을 확인 가능하다. 그러나 인간 평가자의 정성 분석 과정에서 언어모델이 본질적으로 화행 분석을 이해하고 수행하고 있다고 판단하기에는 여러 가지 한계가 있음이 밝혀졌다. 이에 앞으로 언어 모델의 화용적 능력을 길러주기 위해서 정밀한 학습 데이터 수집 및 정비가 필요하며, 입력 프롬프트 정교화 역시 필수적이라는 결론을 내릴 수 있었다. 추후에는 보다 다양한 예문과 난이도로 구성한 문항 세트를 통해 모델의 화용 능력에 대해 더욱 체계적으로 평가해야 할 것이다. 또한 정성 평가 결과의 타당성을 확보하기 위해 평가자의 수를 충분히 확보하고, 평가 문항의 세분화에 따른 평가 척도를 설정해 일관된 평가 기준을 설정할 필요성이 있다.

참고문헌

[1] MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. PloS one, 14(8), e0221152.

[2] Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation, 55, 477-523.

[3] Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017, January). Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the first Italian conference on cybersecurity (ITASEC17) (pp. 86-95).

[4] Wei, Jason, et al. "Emergent abilities of large language models." arXiv preprintarXiv:2206.07682 (2022).

[5] 이혜용. (2010). 한국어 정표화행 연구.

[6] 정종수, & 신아영. (2013). 정표화행에 관한 연구. 인문과학연구, 36, 259-286.

[7] W. A. Mehrens, I. J. Lemann, (1975), "Measurement and evaluation in education and psychology", New York : Holt, Rinehart and Winston.

[8] 박성익, 임철일, 이재경, & 최정임. (2021). 교육방법의 교육공학적 이해 (5th ed.). 교육과학사.

[9] W.James Popham, 수업중심 교육평가, 학지사, 김성훈 외 역 2016