https://dx.doi.org/10.20880/kler.2026.61.1.309

수능 ‘언어와 매체’의 단어 문항 풀이에서 나타나는 생성형 인공지능의 자연어 설명에 대한 분석

오지은 서울대학교 국어교육과 강사(제1 저자)

박인규 인천포스코고등학교 교사(교신저자)

I. 서론
II. 자연어 설명과 문법 문항에 대한 이론적 배경
III. 분석 대상 및 방법
IV. 문법 문항에 대한 자연어 설명의 분석 결과
V. 결론

I. 서론

이 연구의 목적은 대학수학능력시험(이하 수능) '언어와 매체'의 단어 관련 문항 풀이와 관련하여 생성형 인공지능(generative artificial intelligence)이 제공하는 자연어 설명의 특징을 분석하는 것이다. 챗지피티(ChatGPT)의 도입과 함께 전문적인 지식 없이도 누구나 대규모 언어 모델을 자연어로 이용할 수 있게 되면서 인공지능에 의한 변화가 다수에게 공상이 아닌 현실로 다가오고 있다. 인공지능에 대한 사회적인 관심이 확대되면서 이와 관련한 학술적인 논의도 이루어지고 있는데, 그 하나의 축을 담당하는 것은 인공지능 모델의 능력에 대한 분석이다.[1)2)] 예컨대 국어와 관련하여

1) 주요한 다른 흐름으로는 인공지능 모델의 활용 가능성을 꼽을 수 있다. 국어 교육 분야에서는 특히 작문 텍스트에 대한 평가와 피드백을 위해 인공지능 모델을 활용하는 방안이 여러 연구에서 탐색되었다(김승주, 2022; 권태현, 2024; 김은선, 2025; 나상수, 2025 등). 이 중 나상수(2025)는 문법 능력을 평가하는 것을 목적으로 하므로 문법 교육 분야의 주요한 성과로 꼽을 수 있다. 이외에 생성형 인공지능이 생성한 문법 문제를 검토한 김민해·이유진·서나영 외(2025), 생성형 인공지능을 문법 학습에서 비계로 활용하는 방향을 탐색한 박종미(2025ㄱ)도 확인된다.

서는 단어 형성 능력(정한데로, 2023), 에세이에서 나타나는 한국어 사용 능력(박서윤·강예지·강조은 외, 2024), 문법성 판단과 문장 생성 능력(남길임·황은하·송현주 외, 2024) 등이 연구된 바 있다. 교육 분야에서는 특히 수능이 인공지능의 능력을 평가하기 위한 도구로 활용되었다. 수능 문항을 도구로 삼아 인공지능의 능력을 살핀 국어 교육 연구로 최인찬·권도형(2024), 이경숙(2025)을 들 수 있다.[3] 최인찬·권도형(2024)은 '독서와 문학' 문항을 활용하여 인공지능의 국어 능력을 확인하고자 하였고, 텍스트 장르와 유형, 측정 영역에 따라 다소 상이하나 챗지피티가 0~80% 사이의 정답률을 보였다고 보고하였다. 이경숙(2025)은 국어 영역의 선택 과목 '언어와 매체' 문항을 통해 인공지능의 문법 능력을 검토하고자 하였으며, 챗지피티의 평균 정답률은 약 37%였다고 밝혔다.

주목할 만한 지점은 인공지능의 급격한 발전이다. 대표적인 생성형 인공지능인 챗지피티와 제미나이 모두 2025년 말 새로운 버전의 모델이 공개되었다. 구글(Google)은 2025년 11월 18일 제미나이(Gemini)[4] 3을 출시하였으며, 추론의 강화와 복합 양식의 처리, 실시간 검색 및 도구와의 통합이 핵심적인 강점이라고 강조하였다(Google, 2025. 11. 18.). 이어 오픈에이아이(OpenAI)는 2025년 12월 11일 지피티-5.2(GPT-5.2)를 기반으로 하는 챗지피티를 출시하였으며, 전문가 수준의 지식 작업 성능이 개선되고 복

---

2) 물론 인공지능 모델의 능력에 대한 연구와 활용 가능성에 대한 연구는 서로 접점을 가진다. 능력에 대한 분석은 활용 가능성에 대한 시사점을 얻을 수 있을 때 그 결과가 유의미해지고, 인공지능의 활용 가능성에 대한 분석은 모델의 능력에 대한 평가를 포함할 것이기 때문이다. 다만 여기에서는 편의를 위해 양자를 연구의 일차적인 초점에 따라 구별하여 논의하고자 하였다.

3) 한편, 박종미(2025ㄴ)는 인공지능의 문법 용어 설명을 표준국어대사전의 뜻풀이와 비교하고 있다. 수능을 활용하지 않으면서 인공지능의 능력을 분석한 연구로 주목된다.

4) Gemini의 국문 표기는 혼란스럽다. 국립국어원에서 일반 명사 'Gemini'를 '제미니'로 표기하기로 심의한 바 있으나, 여기에서는 Gemini가 고유 명사임을 고려하여 공식 발표 영상에서 자주 쓰이는 발음을 음역하여 적기로 한다.

합 양식 처리가 강화되었다고 발표하였다(OpenAI, 2025. 12. 11.). 깃허브에 개인(hehee9, 2025. 12. 17.)[5]이 공개한 결과에 따르면, 제미나이 3 프로(Gemini 3 Pro)와 지피티-5.2(GPT-5.2)는 모두 2026학년도 수능 국어 영역 전체(공통 문항과 선택 문항)에서 만점을 기록하였다.

향후 인공지능의 발전 가능성을 고려하면 이제는 '인공지능이 얼마나 문제를 잘 푸는지'를 확인하는 데에서 나아가 '인공지능의 풀이를 교육적으로 어떻게 의미화할 것인지'에 연구의 초점이 놓일 필요가 있다. 다른 교과에서 문항에 대한 풀이 과정을 제공하는 서비스를 제공 중인 것을 참고할 때, 언어 모델을 활용한 문법 문항 풀이에는 학습자가 스스로 부족한 점을 파악하여 해결의 단서를 좇도록 하여 자기 주도 학업 역량을 강화한다는 긍정적 측면도 존재한다. 그러나 대규모 언어 모델(large language model)이 확률에 의해 작동하므로 이를 비판적인 관점에서 수용하는 태도가 필요하다. 자연어 설명이 표면적으로는 유창하게 보이더라도 그 이면에 과정적 오류가 존재할 수 있기 때문이다. 따라서 본고는 대규모 언어 모델이 제공하는 자연어 설명에서 표면적인 유창성과 출력 과정에서의 불완전성의 양상을 분석하고자 하며, 이를 위해 수능 '언어와 매체'의 단어 관련 문항을 분석 대상으로 삼는다.

II. 자연어 설명과 문법 문항에 대한 이론적 배경

1. 자연어 설명의 개념과 특성

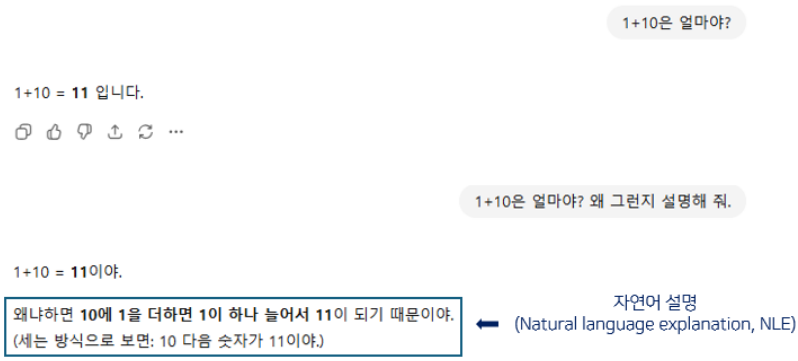
본고에서 '자연어 설명(Natural Language Explanation, NLE)'[6]은 언어

---

5) 장세민(2025. 11. 19.)에 따르면 구유겸 순천향대학교 컴퓨터소프트웨어공학과 3학년 학생이 수능 풀이 실험을 진행한 결과로 보인다.

6) 자연어 설명(NLE)은 시각(vision) 과제와 관련되는 인공지능 연구에서 주로 쓰이는 용

모델이 사용자의 질문에 답할 때 질문에 대한 결과와 함께 제공하는 텍스트를 의미한다. 챗지피티의 답변 사례를 예로 들어 보이면 〈그림 1〉과 같다. 언어 모델은 사용자의 질문에 대해 단순히 답만을 제공할 수도 있고 답에 대한 설명을 자연어로 함께 제공할 수도 있다. 이 연구에서는 판단의 결과인 답 자체뿐 아니라 답과 함께 제공되는 설명에 관심을 둔다.



〈그림 1〉 자연어 설명의 예

자연어 설명이 블랙박스로 비유되는 인공지능의 내부 작동 과정을 실제 사고하는(thinking) 과정으로 언어화한 것이라고 보는 것은 과도한 해석이다(Barez, Wu, Arcuschin, et al., 2025: 2). 자연어 설명은 흔히 중요한 요인을 누락하며, 모델이 수행하는 분산적이고 중첩적인 계산을 온전히 드러내기보다는 부분적이며 사후적인 정당화로 기능하기 때문이다(Barez et al., 2025: 7). 따라서 자연어 설명은 '과정을 설명하는 것처럼 보이는' 언어적 조합을 제시한 결과로 이해되어야 할 것이다.

자연어 설명과 사고의 연쇄(Chain-of-Thought, 이하 CoT)의 관계도

어이다. 예컨대 모델은 제시된 이미지가 무엇인지("고양이")와 함께 왜 그렇게 판단했는지에 대한 설명("털로 덮여 있고 콧수염을 가지고 있기 때문에")을 제공할 수 있다(Sammani, Mukherjee, & Deligiannis, 2022: 8329).

살필 필요가 있다. CoT는 인공지능의 단계적인 추론 과정이나 결과를 의미하는 용어로 흔히 프롬프트 설계 기법(prompt engineering)에서 사용되었다. 이는 최종적인 결괏값으로 이끄는 중간 추론 과정의 연쇄를 프롬프트로 입력하는 것을 말한다(Wei, Wang, Schuurmans, et al., 2022: 2). 단순히 어떤 것의 답을 구하는 것이 아니라 사용자가 요구하는 순서를 입력하는 것이다. 또한 CoT는 자연어로 제시되지 않는 모델의 추론 과정을 의미하기도 한다.[7] 본고에서는 답변과 구별되는, 모델의 사고 연쇄가 자연어로 제공되는 경우만을 CoT로 지칭하고자 한다.

자연어 설명은 본질적으로 가변적이라는 특성이 있다. 동일한 질문을 같은 모델에 반복할 때 답변의 내용과 표현이 달라질 수 있는 것이다. 모델이 답변을 생성할 때 다음에 올 단어를 확률적으로 고르고, 계산 과정에서 비결정성이나 미세한 오차가 존재하기 때문에 무작위적인 비일관성이 발생한다(Ahn & Yin, 2025: 1).[8] 이에 박종미(2025ㄴ: 319-320)는 같은 질문을 3회 반복하여 세 가지의 서로 다른 답변 자료를 수집한 후 분석 대상으로 삼기도 하였다.

본고에서는 출력 변수를 조정하여 연구 대상이 되는 답변의 다양성을 확보하는 방식을 취한다. 이는 자연어 설명의 형식과 내용이 출력 변수에 영향을 받는다는 특성을 고려한 것이다. 자연어 설명 생성에 영향을 미치는 주요 출력 변수로 온도(temperature), 누적 확률(top-p), 상위 개수(top-k), 최대 생성 토큰(max tokens), 출현 억제(presence penalty), 빈도 억제(frequency penalty) 등을 들 수 있다. 본고에서의 초점은 출력의 다양성과 무작위성을 조절하는 데 영향을 미치는 온도이다. 언어 모델은 기본적으로 다음

---

7) 예컨대 답변 전에 제시되는 "○○s 동안 생각함"을 사용자에게 감추어진 전체의 긴 CoT로 보고, 실제로 사용자에게 제시되는 답변의 내용은 긴 CoT를 대신하는 요약이라고 말하기도 한다(Cameron, 2025).

8) 이에 더하여 모델의 학습 내용이나 구조 개선에 의해서 변동이 발생하기도 한다. 즉, 질문 시기에 따라 답변 내용이 달라지는 것이다.

에 올 단어(토큰)의 확률을 계산하고, 확률에 따라 대상을 선택한다. 이때 '온도'는 토큰별 확률 계산에 영향을 미친다.[9] 온도가 1보다 크면 토큰 간의 격차가 줄어들어 낮은 확률의 토큰도 선택될 가능성이 증가한다. 반면 온도가 1보다 작으면 토큰 간의 격차가 커지는데, 최솟값인 0인 경우 가장 확률이 높은 토큰만 강하게 선택되어 결정론적인 선택을 유도한다.[10][11] 변수에 대한 이해는 언어 모델이 제공하는 자연어 설명을 시시각각 내용이 달라져 신뢰하기 어려운 미지의 무엇인가가 아니라 일정한 방법과 기준에 따라 통계적으로 생성된 결과로 인식하게 한다.

2. 문법 문항의 구조와 그 풀이

문법 영역 평가의 목표가 문법 능력을 평가하는 것이라고 볼 때, 문법 문항에 대한 풀이는 기본적으로 문법 능력을 발휘하는 과정이다. 문법 능력은 언어 능력의 하위 요소로서 일반적으로 '규칙으로서의 능력'으로 초점화되나(이관희 · 정희창, 2010: 55), 언어에서의 규칙을 어디까지로 파악하는지,

---

9) 구체적으로 i번째 토큰의 확률 분포를 계산하는 과정에서 온도 T가 미치는 영향을 식으로 보이면 다음과 같다. z는 소프트맥스 함수를 적용하여 변환하기 전 로짓(logit)이다. T가 커지면 로짓값을 작게 나누므로 토큰 간의 격차가 줄어들어 확률 분포가 평탄해진다. T가 작아지면 격차가 더 커져서 확률의 분포가 날카로워진다.

$$P_i = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}$$

10) 그러나 온도가 0일 경우에도 완전히 결정론적이지는 않다. 여러 개의 토큰이 같은 확률인 경우가 존재할 수도 있고, 확률 계산 시의 부동소수점 연산 등으로 인해 미세한 차이가 생길 가능성이 존재하기 때문이다.

11) 온도 설정에 대한 설명과 권고는 모델에 따라 차이가 있다. 제미나이 3은 온도를 1로 설정하는 것이 기본값으로 권장되며, 그 미만으로 설정하는 경우 성능 저하나 예기치 않은 동작이 발생할 수 있다고 설명한다(Gemini API, 2026. 1. 29.). 챗지피티 레거시에 관한 설명에서는 0.8보다 크면 출력의 무작위성이 커지고, 0.2보다 낮으면 결정론에 가까워짐을 밝히고 있다(OpenAI Platform, n. d.).

언어 사용에서 문법의 역할을 어떻게 보는지에 따라 문법 능력의 개념과 구조가 다르게 설정될 수 있다.

국어 교육 분야에서 이루어진 문법 능력에 대한 대표적인 논의로는 이관규(2008), 구본관(2010)을 들 수 있다. 이관규(2008: 57)는 문법 능력이 지식 능력, 사용 능력, 태도 능력을 모두 합한 것이라고 보고, 음운 능력, 단어 능력, 문장 능력, 의미 능력, 담화 능력, 어문 규범 능력을 그 하위 부분으로 설정할 수 있다고 하였다(이관규, 2008: 221). 구본관(2010: 195)은 문법 능력을 지식 능력, 탐구 능력, 활용 능력으로 하위 분류하고 있어 이관규(2008)와 달리 태도 능력[12]을 제외하고 탐구 능력을 설정하고 있다. 또한 '의미 능력'은 '단어 의미, 문장 의미' 등 단위로 환원할 수 있어 제외하였다.

본고는 단어 관련 문항에 주목하고 있으므로 우리가 분석 대상으로 삼은 문항의 풀이 과정은 기본적으로 단어 능력을 발휘하는 과정으로 볼 수 있다. 다만 문법 개념에 대한 연계가 단위별로 분절되지 않듯(이관희 · 최선희 · 김자영, 2022: 220) 단어에 대한 판단 역시 다른 단위에 대한 이해 및 판단과 영향을 주고받으므로 다른 단위의 문법 능력이 함께 발휘될 가능성이 존재한다. 또한 수능 문항의 특성상 태도 능력을 평가하기에 어려움이 있으므로[13] 본고에서 주목하는 문항들의 풀이 과정에서는 '지식 능력, 탐구 능력, 활용(사용) 능력'이 발현되어야 할 것으로 기대할 수 있다.

그러나 문법 문항의 풀이는 실제 언어 사용 상황에 비해 제한적인 환경을 가진다. 문항에서 주어진 형태로 언어를 이해하거나 사용해야 하기 때문이다. 선택형(선다형) 문항은 일반적으로 지문, 문두, 선택지, 〈보기〉로 구성

---

12) 이에 대해 구본관(2010: 199)은 당시 국어과 교육과정의 문법 영역 내용 체계표를 고려한 결과이기도 하다고 설명하며, 성취도 평가 등의 문항 개발에서 실용적인 목적으로 활용될 수 있을 것이라고 보았다(구본관, 2010: 192). 한편 문법과 관련된 태도는 문법 능력과 직접적인 관련을 맺지는 않으나 문법 교육과 문법 평가에 반영되어야 하는 요소라고 하였다(구본관, 2010: 195).

13) 용어의 적절성은 재론이 필요한 문제이나 2022학년도 이후 수능 문법 문항의 행동 영역은 '개념'과 '적용'으로 분류할 수 있다(김규훈, 2023: 76).

된다(류수열 · 주세형 · 남가영, 2021). 수능 문법 문항의 구조에서 특히 두드러지는 것은 지문이나 〈보기〉의 형태로 제공되는 언어 자료의 성격과 역할이다.[14] 문법 문항에서 활용되는 언어 자료가 문항에서 측정하고자 하는 문법 능력의 내용과 범위에 큰 영향을 미치기 때문이다.

〈보기〉의 형태로 주어지는 언어 자료는 문항 풀이에 필요한 문법 지식을 일부 제공하여 문항이 저차원적인 지식 인출을 요구할 가능성을 방지하는 기능을 한다. 문법 개념을 직접적으로 평가하는 부담을 줄이는 것이다(이관희 외, 2022: 212). 이때 문법 문항에서 〈보기〉 등은 대개 지문의 역할을 대신하는 언어 자료의 역할을 하며 문항 성립을 위한 필수 요소가 된다(류수열 외, 2021: 342). 최근의 수능에서는 언어 자료가 지문의 형태로 보다 적극적으로 동원되고 있다. 김규훈(2023: 81)에서는 문법 지식에 대한 설명 지문을 포함하는 문항이 2017학년도 수능부터 새로운 유형으로 나타났다고 보고 이를 '지문형 문법 문항'으로 명명하였다.

한편, 문법 문항에서 언어 자료는 문항에서 동원해야 하는 지식의 범위를 제한하여 문항의 정합성을 높이는 역할을 하기도 한다. 학문 문법 차원에서 이견의 여지가 있는 지식에 제한을 가하는 장치로 활용되는 것이다(이관희 외, 2022: 212). 문법 지식은 비분절적이고 유동적인 언어 현상의 본질적인 특성, 개인의 맥락의존적이고 다원적인 사용, 연구자의 이론적 관점과 논리 차이로 인해 필연적으로 불확정성(주지연, 2020)을 지닌다. 불확정적인 문법 지식을 고부담 평가를 위해 문항화하는 과정에서 언어 자료는 문항에 관여하는 문법 지식을 누구에게나 똑같이 확인되고 고정적인 형태로 제공하여 문항의 오류 가능성을 줄인다.[15] 이러한 형식은 문항을 풀이에 필요한 자

14) 학습자에게 제공되는 언어 자료의 적절성과 위상에 대하여는 숙고가 필요하다. 문법 문항의 구조에서 언어 자료의 위상을 설정하는 문제는 문법 능력과 문법 평가에 대한 본질적인 질문과 닿기 때문이다. 관련 논의로는 주세형(2009), 남가영(2017), 김규훈(2023) 참고.

15) 예컨대 이도영 · 김잔디 · 민송기 외(2021: 168-169)에서는 한 고등학교 국어 평가 문항을

료와 방향이 제시되는 구조화된 문제(well structured problem)로 만들어, 다양한 방향으로의 해결 가능성을 줄인다.

문항을 구성하는 요소 중 문두는 평가 문항에서 물음을 직접적으로 구현하는 부분으로, 피험자가 선택지를 어떻게 검토해야 하는지를 결정하는 기능을 수행한다. 정민주·서수현·남민우 외(2022: 51)에서 좋은 문항의 조건으로 문두가 문항의 요구 내용이 명확하게 서술되어 있으면서도 평가 내용이 국어 교과의 핵심적인 역량을 담고 있어야 한다는 점을 지적하였다. 이는 문두가 단순히 질문을 제시하는 것을 넘어 평가 의도를 명시적으로 전달하는 역할을 담당함을 보여 준다. 주세형(2009: 500)은 학업 성취도 평가의 문법 문항을 검토하면서 문두에 평가 요소인 문법 요소가 정확히 드러나지 않았다고 지적하였는데, 본고의 분석 대상이 되는 수능 문항에서도 일부 그러한 양상이 확인된다.[16)]

다음으로 선택지에 대해 살펴보자. 이기돈(2025: 398)에서 2025학년도

예로 들면서 사잇소리 현상을 첨가로 볼 것인지의 여부에 따라 정답이 ①과 ④로 달라질 수 있음을 지적하였다. 또한 해당 고교에서 사용한 교과서 내용에 따라 ①을 정답으로 확정하기 위하여서는 그 근거가 문항에 제시되어야 하고, 이를 탐구의 형태를 가진 언어 자료로 가공하는 것이 바람직하다고 하였다.

> 다음 중 음운 현상이 다른 하나는?
> ①눈길〔눈:길〕 ②웃긴〔욷낀〕 ③국밥〔국빱〕 ④먹는〔멍는〕 ⑤있다〔읻따〕

16) 분석 대상이 된 24개의 문항 중 문법 용어가 드러나는 문두는 다음의 네 경우뿐이다(밑줄은 연구자에 의함.).
㉠에 따를 때, 〈보기〉에 제시된 ㉮~㉲ 중 그 내부 구조가 동일한 단어끼리 묶은 것은? (23-본수-35)
윗글의 ⓐ, ⓑ와 연관 지어 〈자료〉에 제시된 합성 명사를 탐구한 내용으로 적절한 것은? (23-본수-36)
〈자료〉를 바탕으로 〈보기〉의 ⓐ~ⓒ 중 체언과 조사가 결합하여 이루어진 부속 성분이 있는 것만을 고른 것은? (24-9모-39)
〈보기〉를 바탕으로 'ㅎ' 말음 용언의 활용 유형을 탐구한 내용으로 적절하지 않은 것은? (24-본수-27)

수능 국어 영역 56문항은 모두 '선택지가 필요한 문항'으로 분류되었다. 선택지가 문항의 의미 구성에 필수적인 요소로 작용할 뿐만 아니라 정답은 선택지 중에서만 선택해야 하므로 일반적으로는 문항을 해결하는 데 있어 선택지를 모두 살피는 것이 필요하다는 것이다(이기돈, 2025: 402-403). 출제자 관점에서도 선택지가 "평가 문항에 대한 학생들의 반응을 유도하여 최종적으로 결정짓는 역할(이도영 외, 2021: 244)"을 하도록 설계된다는 점은 선택지가 풀이에 중요한 기능을 담당함을 방증한다. 또한 정답을 포함한 모든 선택지를 정답에 가깝게 구성함으로써 문항 추측도를 적절한 범위 내에서 관리할 수 있으며, 이로써 피험자들은 모든 선택지를 주의 깊게 살피면서 가장 적절한 답이 무엇인지를 파악하는 인지적 능력을 발휘하게 된다.

문두의 표현 형식이 선택지를 검토하는 방식과 범위를 규정하기도 한다. 문두의 '가장 적절한 것은?'과 같이 부사를 포함하여 질문하는 최선답형 문항의 경우, 각 선택지 진술의 여러 내용 중 일부의 적절성 여부가 정오 판단에 영향을 주므로 모든 선택지를 확인해야 할 필요성이 요구된다(이기돈, 2025: 403).

대규모 언어 모델이 문법 문항을 풀이하는 상황에서 이러한 구성 요소들은 풀이를 위한 맥락(context)으로 기능한다고 볼 수 있다. 대규모 언어 모델은 사전 학습을 통해 모델의 가중치에 저장된 지식을 보유하는 동시에, 추론 시점에 입력된 맥락 내 정보를 활용하는 능력을 갖추고 있다(Brown, Mann, Ryder, et al., 2020). 문법 문항을 풀이하는 상황에서 문두는 모델이 수행해야 할 과제의 성격을 규정하고, 언어 자료는 해당 과제를 해결하기 위해 참조해야 할 정보를 제공하며, 선택지는 문항에 대한 반응을 유도하고 결정짓는다.

III. 분석 대상 및 방법

첫째, 분석을 위한 문항의 범위와 사용할 생성형 인공지능을 선정하였다. 이를 위해 2025학년도, 2026학년도 2개년의 본수능 문항을 챗지피티 5.2 사고(thinking)와 제미나이 3 프로(pro)에 입력하여 풀이하도록 하였다.[17] 분석 후보가 되는 모델로 챗지피티와 제미나이를 선정한 것은 범용성과 성능을 고려한 것이다. 실제 교사나 학습자가 활용하는 상황과 유사하게 하기 위하여 API 호출 없이 챗지피티와 제미나이의 애플리케이션 또는 웹 인터페이스를 이용하였다.

예비적인 과정에서 두 모델의 결과는 〈표 1〉과 같이 다르게 나타났다. 예비적인 분석 결과를 고려하여 본고는 제미나이를 최종적인 분석 대상으로 설정하였으며, 그 이유는 다음과 같다.

〈표 1〉 챗지피티와 제미나이의 정답률 비교

모델 \ 문항	2025학년도					2026학년도				
	35	36	37	38	39	35	36	37	38	39
챗지피티 5.2 사고	O	O	X	O	O	O	O	O	X	O
제미나이 3 프로	O	O	O	O	O	O	O	O	O	O

본고의 목적이 두 언어 모델의 성능을 비교하는 데 있지 않으며, 본고가 주목하는 것은 자연어 설명 자체의 특성이다. 이를 위해서는 출력에 영향을

17) 문항의 이미지를 캡처하여 질문하였으며 별도의 텍스트로 된 프롬프트는 활용하지 않았다. 복합 양식 처리 기능이 강화되어 별도의 텍스트화 작업이 불필요하다고 판단하였고, 문두에 의해 기본적으로 제공되는 자연어 설명을 수집하고자 하였기 때문이다. 생성형 인공지능에 문항을 입력하여 풀이한 구체적인 방법은 이하의 분석 과정에서도 동일하게 적용하였다.

미치는 변수를 체계적으로 조작할 수 있는 환경이 요구된다. 제미나이는 구글 에이아이 스튜디오(Google AI Studio)를 통해 API 호출에 대한 과금 없이도 온도 등의 출력 변수를 손쉽게 조정할 수 있다는 연구 방법상의 이점을 지닌다. 아울러 제미나이는 지금까지 국어 교육 선행 연구에서 분석 대상으로 주목되지 않았다는 연구사적 의의도 고려되었다.

아울러 〈표 1〉의 정답률 차이가 시사하는 바도 간과하기 어렵다. 예비 분석에 사용된 문항은 수능 2개 학년도 10문항으로 제한적이나, 선행 연구에서 최신 모델들이 수능 국어 영역 전체에서 만점을 기록하였다는 보고(hehee9, 2025. 12. 17.)를 고려하면 대상 문항 중에서 오답이 발생하였다는 것은 해당 모델의 문법 문항 풀이에 상당한 불안정성이 존재함을 의미한다고 판단하였다. 자연어 설명의 특성을 분석하는 것은 정답에 도달하는 과정을 관찰하는 것을 목표로 하므로, 정답에 도달하지 못한 언어 모델의 설명을 분석 대상으로 삼는 것은 연구의 초점을 흐릴 우려가 있다고 판단하여 안정적인 정답을 제시한 제미나이를 분석 대상으로 선정하였다.[18] 본고에서 제미나이는 트랜스포머 기반 언어 모델의 대표적인 사례로서 분석 대상으로 활용된다.[19]

다음으로 단어 관련 문항으로 분석 범위를 제한하였는데, 이는 단어 관련 문항에서 요구되는 기능이 대규모 언어 모델의 작동 방식과 구조적인 유사성을 가지기 때문이다. 대규모 언어 모델은 입력된 자연어를 토큰 단위로 처리하는데, 토큰화의 과정은 형태소 분석의 과정과 유사한 성격을 지닌다. 단어의 내부 구조를 분석하거나, 형태소의 결합 양상을 파악하는 것은 토큰

18) 챗지피티를 포함한 다른 언어 모델과의 비교 분석은 후속 연구의 과제로 남겨 두고자 한다.

19) 이에 본고의 분석이 제미나이를 대상으로 수행되었음에도 결과에 대한 기술과 논의 시 '언어 모델' 또는 '생성형 인공지능'이라는 일반적 표현을 사용한다. 이는 자연어 설명의 특성이 제미나이에 한정된다기보다는 트랜스포머 기반 대규모 언어 모델의 특성에 기인하는 것으로 판단되기 때문이다. 분석에서 관찰되는 자연어 설명의 양상은 제미나이의 특성에서 비롯된 것이라기보다는 언어 모델의 특성이 문법 문항 풀이라는 맥락에서 발현된 것으로 이해된다.

단위로 언어를 처리하는 언어 모델의 작동 원리와 밀접하게 관련된다. 이에 따라 단어 관련 문항은 언어 모델의 사전 학습된 지식이 자연어 설명에 작용하는 양상을 관찰하기에 적합한 분석 대상이 될 수 있다.

이와 관련하여 예비적인 과정에서 자연어 설명을 생성한 결과 중 주목되는 사례는 (2)와 같다.

(2) 2026학년도 9모 38번 문항에 대한 자연어 설명의 일부
'이밖에는': '에'(부사격 조사) + '는'(보조사) → [격 조사 – 보조사]

(2)에서 체언 뒤에 붙은 '밖에'를 하나의 조사로 인식하지 못하고 체언 '밖'과 조사 '에'로 분석하였다는 점은, 언어 모델이 토큰 단위에 기반하여 문법 분석을 수행하되 그 과정에서 형태소 경계의 설정에 오류가 발생할 수 있음을 보여 준다. 이러한 오분석은 음운론이나 통사론 관련 문항에서도 나타날 가능성이 있으나, 단어 층위에서는 형태소 분석의 정확성이 정답 판별에 직접적으로 관여한다는 점에서 오분석의 양상이 보다 선명하게 드러날 것으로 기대된다.[20] 이러한 점에서 '단어 관련 문항'이라는 범위는 전체 문항 중 강도(intensity)의 유형을 보이는 표본(Creswell, 2012/2015: 191)[21]에 해당한다.

'단어 관련 문항'으로는 단어 단위를 소재로 하는 문항을 〈표 2〉와 같이 선정하였다. 선정 과정에서 '단어 관련 문항'의 여부를 판단할 때에는 형태론(조어론, 활용론) 및 품사론[22]과의 관련성을 고려하였다. 최종적으로 분석 대

20) 주제 분야를 달리하는 경우 음운, 통사, 담화 등과 관련하여 특수한 양상이 추가로 포착될 가능성이 있다. 관련하여서는 후고를 기약한다.

21) 강도 유형은 현상을 강렬하지만 극단적이지 않게 표출하는, 풍부한 정보를 제공하는 사례들을 추출하는 데 목적이 있다(Creswell, 2012/2015: 191).

22) 이선웅(2012: 126)은 품사론은 형태론, 통사론, 어휘론에 펼쳐져 있는 사실을 묶어 주는 전통적인 학문 영역이므로 '형태론'으로 포섭할 수 없다고 하였다. 본고에서 단어 관련 문항의 배경 학문으로 품사론을 고려한 것은 품사론이 형태론의 하위 분야이어서가 아니라

상이 된 문항의 수는 24개이다.

〈표 2〉 분석 대상 문항

문항 번호 / 학년도	6모	9모	본수
2022		37, 39	35-36
2023	38, 39	35-36	35-36
2024	35-36	35-36, 39	37
2025	35-36	35-36	
2026	37, 39	38	39

둘째, 모델에 문항을 입력하여 이에 대한 자연어 설명을 생성하였다. 이때 에이아이 스튜디오에서 출력에 영향을 미치는 변수를 조절하였다. 답변 생성의 자유도에 영향을 미치는 변수 중 핵심적인 조절 대상으로 삼은 것은 온도이다. 각 문항에 대해 온도를 0, 0.5, 1으로 다르게 설정하여 답변을 생성하도록 하였다. 문항 풀이라는 과제의 성격을 고려하여 1보다 높은 온도는 고려하지 않았다. 누적 확률(Top p)은 기본값 1로 설정하였는데 이는 온도를 낮추는 경우 누적 확률을 줄이거나 늘려도 영향이 없기 때문이다. 또한 사고 수준을 높음(high)으로, 최대 출력 길이(Output length)를 최댓값인 65536으로 설정하였다.[23] 각 답변을 생성할 때마다 온도 설정을 바꾸고, 동일한 대화 창에서 앞선 질문과 그 답변을 삭제한 후 본래 입력된 문항을 그대로 재실행(rerun)하여 대화의 지속에 따른 맥락의 개입 가능성을 통제하였다.

셋째, 생성된 자연어 설명을 분석하였다. 분석은 생성된 자연어 설명이

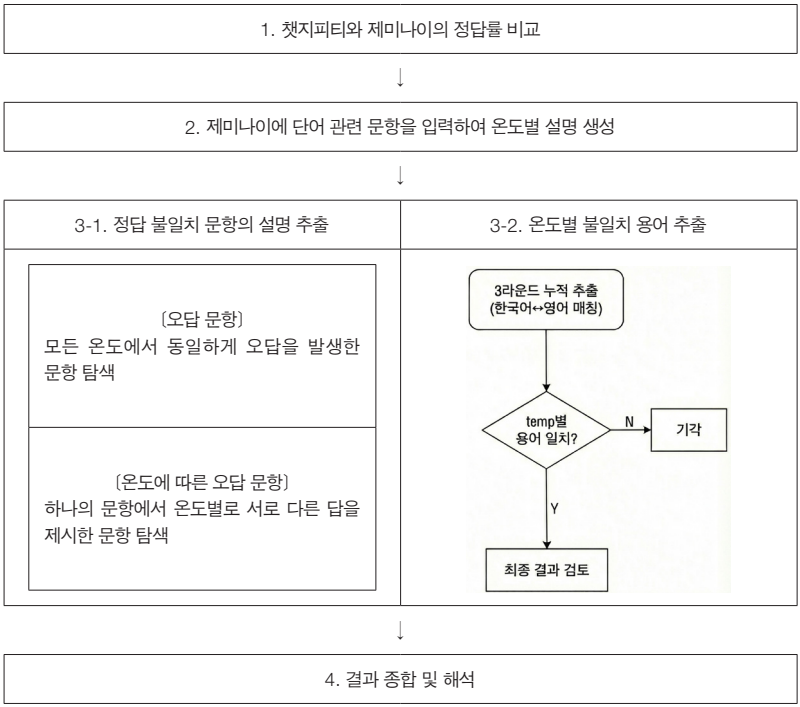
단어 단위를 다루는 학문이기 때문이다.

23) 기타 언급되지 않은 설정은 기본값을 유지하였다.

확률에 의해 작동한다는 점을 고려하여, 두 가지 비교 축을 따라 수행하였다. 첫 번째 분석에서는 모델의 정답 판별 양상을 추적하였다. 모델이 오답을 선택하거나 온도에 따라 상이한 답을 선택할 때, 그 차이가 어디에서 비롯되는지를 파악하기 위함이다. 이 분석은 두 단계로 구성된다. 1단계에서는 각 온도에서 모델이 선택한 답과 실제 정답의 일치 여부를 확인하였다. 온도 0, 0.5, 1 각각에서 모델이 선택한 답 번호를 기록하고, 이를 실제 정답과 대조하여 오답이 발생한 문항을 추출하였다. 2단계에서는 온도 간 답안의 일치 여부를 확인하였다. 세 온도에서 모델이 선택한 답이 모두 동일한지, 온도에 따라 상이한지를 구분하였다. 오답이 발생하거나 온도 간 답안 불일치가 관찰된 문항에 대해서는 각 온도의 자연어 설명을 비교하여 차이의 원인을 추적하였다. 분석 대상의 설정, 분석 기준의 선택, 형태소 분석의 적절성, 선택지 판정의 논리 등을 중심으로 기술하였다.

두 번째 분석에서는 CoT의 영어 문법 용어와 자연어 설명의 한국어 용어 간의 대응 관계를 분석하였다. 이 분석의 목적은 일치·불일치하는 용어의 양적 분포를 확인하는 것이 아니라, 모델이 작동하는 과정에서 입력 컨텍스트에 영향을 받아 출력하는 용어의 사례를 추출하기 위한 것이다. 클로드 코드(Claude Code) OPUS 4.5를 활용한 바이브 코딩(vibe coding) 방식을 활용하여, 먼저 언어 모델을 통해 한국어의 문법 개념어를 추출하고 이에 대응하는 CoT의 영어 용어를 온도별로 찾도록 하였다. 그리고 규칙 기반(rule-based) 방식을 통해 하나의 한국어 용어에 대응하는 온도별 영어 용어 간의 일치 여부를 판단하도록 하였다. 문법 개념어의 추출과 매칭은 언어 모델의 특성상 누락이 발생할 수 있으므로 3회에 걸쳐 용어를 누적하며 수행하였다. 3회 반복하여 추출된 문법 용어 중, 3개 온도에서 모두 등장하는 173개의 공통 용어를 분석한 결과, 46개의 용어가 온도별 불일치를 보였다. 이상의 분석 과정은 다음의 〈표 3〉과 같이 정리할 수 있다. 이후 추출한 결과를 질적으로 해석하여 제시하였다.

〈표 3〉 분석 과정



IV. 문법 문항에 대한 자연어 설명의 분석 결과

본고의 분석에서 나타난 자연어 설명의 특징적인 사례를 표면적 유창성과 내부적 불완전성으로 범주화하여 제시하고자 한다. 생성형 인공지능이 표면적으로는 인간의 눈에 그럴듯해 보이는 답변을 유창하게 제시하나 내부적으로는 그러한 답변의 내용에 오류가 존재할 가능성이 포함되어 있다는 것이다. 아래에서는 항목별로 사례를 들어 그 내용을 자세히 논의하겠다.[24)]

1. 표면적 유창성의 양상

1) 문항 구성 요소에 의존하는 출력

언어 모델이 제공하는 자연어 설명은 문항을 구성하는 요소들에 영향을 받는다. Ⅱ장에서 논의한바 수능의 문법 문항은 언어 자료를 통해 문제 풀이에 필요한 문법 지식을 제공하는 경우가 많으며, 문두와 선택지를 통해 사고의 방향이 안내되기도 한다. 이러한 문항의 구조는 언어 모델이 표면적으로 유창한 설명을 도출하는 기반이 된다.

먼저 입력 자료가 풀이 과정에서 특정 수행을 강제하는 역할을 하기도 한다. 이는 언어 모델이 문항에 포함된 정보가 요구하는 연산을 수행하기 때문이며, 이는 표면적인 유창성을 갖춘 잘못된 결과로 이어지기도 한다. 2026학년도 수능 39번 문항은 조사와 어미의 분석에 대한 것으로 "③ ㉠의 '시작되자'와 ㉡의 '말리는'은 모두 피동의 의미를 더해 주는 접미사가 결합하여 형성된 단어이다."의 선택지를 포함하고 있다. 이를 분석하는 과정에서 언어 모델은 모든 온도 설정에서 '말리는'을 "마르-(어근) + -리-(사동 접미사)"로 분절하였다. 현대 국어에서 '말리다'는 공시적으로 분석하기 어렵다. '마르다'와의 통시적 관련성은 존재하나 '-리-'[25]를 독립적인 형태소로 분석하기 어려운 것이다. 그럼에도 불구하고 언어 모델이 '-리-'를 분석한 것은 선택지에서 피동 접미사의 결합 여부를 판단하도록 요구함으로써 접미사에 대한 분석을 유도하였기 때문이다.

다음으로 자연어 설명에서 나타난 용어는 제시된 문항의 맥락에서 복사

---

24) 기본적으로 언어 모델의 내부 작동 메커니즘이 블랙박스(black box)임을 인정하고 문제 해결 과정에서의 가중치, 토큰 확률 분포, 내부 표상 등은 직접 관찰할 수 없다고 전제한다. 따라서 모든 분석은 관찰 가능한 자연어 형태의 출력(일부 CoT와 자연어 설명)에 기반한다.

25) 물론 '마르다'는 'ᄆᆞᄅᆞ외다(15세기~17세기) > ᄆᆞᄅᆡ다(16세기~17세기) > 말리다(19세기~현재)'의 변화 과정을 거쳤으므로 '-리-'를 분석하는 것은 정확하지 않다.

되었을 가능성이 있다. 예컨대 2022학년도 9월 모의평가 37번 문항에 대한 자연어 설명에서 '관형격 조사'라는 학교 문법 용어를 일관적으로 사용하고 있다. 그런데 최종적인 자연어 설명의 생성 이전 과정으로 확인할 수 있는 CoT에서는 상이한 용어 사용 양상이 나타난다. 온도별로 'genitive particle', 'adnominal case marker', 'possessive marker'의 용어가 사용되었는데 이들은 각각 '속격 조사', '관형격 표지', '소유 표지'로 번역될 것으로 예상할 수 있다. 그러나 실제로는 상이한 분석 용어가 자연어 설명에서는 동일하게 '관형격 조사'로 수렴되었다. 분석 대상인 문항 전체에서 온도별로 상이한 CoT의 용어가 자연어 설명에서 수렴한 경우를 정리하여 보이면 〈표 4〉와 같다.

〈표 4〉 온도별 CoT 용어의 불일치 사례

문항	국어	온도 0	온도 0.5	온도 1
22-본수-36	조사	case markers, particle	case markers, particle, particles	case markers, case particle,
23-본수-35	형태소	-	morphemes	morphemes
23-9모-35	합성 용언	compound predicates	compound predicates, compound verbs, compound verbs and adjectives	compound predicates
24-6모-36	유정 명사	animate, animate noun	sentient noun	animacy
24-9모-39	부속 성분	subsidiary component	subordinate component	adjunct component
25-9모-36	의존 명사	bound nouns, dependent forms	dependent forms, dependent nouns	dependent nouns
26-6모-37	안긴문장	-	embedded sentence	embedded clause

대규모 언어 모델은 프롬프트에 주어진 정보를 가중치 업데이트 없이 즉각적으로 활용하는 맥락 내 학습 능력을 가지고 있다(Brown et al., 2020). 이때 입력된 맥락 내에서 이전에 나타난 패턴을 찾아 이를 복제하는 연산이 지배적으로 관여하며(Olsson, Elhage, Nanda, et al., 2022), 언어 모델이 생

성하는 답변이 주어진 맥락의 토큰 정보에 직접 의존한다. 따라서 자연어 설명의 용어는 CoT에 대한 단순한 번역이 아니라 문항 맥락에서의 패턴 매칭 결과로 여겨지며, 자연어 설명에서 나타나는 용어의 정확성은 모델이 해당 문법 개념을 이해한 결과로는 보기 어렵다.

나아가 맥락 내 정보는 CoT 자체에도 영향을 미치는 것으로 보인다. 2023학년도 6월 모의평가 38번 문항을 풀이하는 과정에서 언어 모델은 '형식 형태소'를 가리키는 영어 용어로 'formal morpheme'을 사용하는 것이 CoT에서 관찰되었다. 일반적으로 형식 형태소는 'empty morpheme' 이나 'grammatical morpheme'으로 사용되므로 'formal'은 '형식'이라는 한국어 용어에 영향을 받은 결과로 해석된다. 해당 문항 풀이에 대한 CoT에서는 'formal morpheme' 외에도 'grammatical morpheme'의 용어가 나타났으나 최종적인 자연어 설명에서는 '형식 형태소'로 수렴되었다. 이러한 사례 또한 맥락이 모델 답변에 미치는 영향을 보여 준다.

2) 문항에서 주목되는 단서의 활용

언어 모델이 제공하는 자연어 설명은 주어진 문항 내의 특정 단서에 주목하여 형태 중심의 언어 분석을 수행한다. 이는 자연어를 토큰으로 분절하여 그 출현 확률을 계산하는 언어 모델의 특성상 자연스러운 결과로 볼 수 있다.[26] 언어 모델의 어텐션이 맥락 내의 특정 토큰에 집중되고 이들이 닻의 역할을 하여 출력을 조정한다는(Wang, Li, Dai, et al., 2023) 점에서 더욱 그러하다. 문항에서 주목된 단서를 활용하는 것은 고정된 형태를 중심으로 일관적인 분석을 수행하는 데 기여한다.

예컨대 (3)은 2024학년도 수능 37번 문항에 대한 자연어 설명의 일부이다. 'ㅎ' 말음 용언의 활용 양상을 규칙 활용과 불규칙 활용, 모음 조화의 적

26) 물론 한국어의 경우 조사와 어미 결합으로 인한 토큰화의 어려움이 존재하나 자연어 처리 기술의 발전으로 극복해 나가고 있는 것으로 생각된다.

용 여부와 축약 가능 여부에 따라 유형별로 구분하여 판단하여야 하는데, 문항에 제시된 형태 정보에 주목하여 일관적이고 정확한 자연어 설명을 제공하는 양상이 확인된다.

(3) 2024학년도 수능 37번 문항에 대한 자연어 설명의 일부

① '조그맣-', '이렇-'은 '조그매', '이래서'로 활용하므로 ㉠-1과 활용의 유형이 같겠군.

'조그맣-'은 어간 끝 모음이 'ㅏ'(양성)이므로 '조그맣- + -아 → 조그매'가 됩니다. 이는 양성 모음 어간이 '-애'로 변하는 ㉠-1 유형에 해당합니다. 하지만 '이렇-'은 어간 끝 모음이 'ㅣ'(음성)입니다. '이렇- + -어서 → 이래서'로 활용되는데, 이는 음성 모음 어간임에도 '-애' 형태가 되는 것입니다. 따라서 이는 '어떻- + -어 → 어때'와 같은 ㉡ 유형에 해당합니다. 두 용언의 활용 유형이 다르므로, 둘 다 ㉠-1과 같다는 설명은 적절하지 않습니다. …

그러나 주목되는 단서를 활용한 형태 중심의 분석은 언어학적 정확성이 부족할 가능성이 있다. 예컨대 2023학년도 6월 모의평가 39번 문항의 언어 자료에서는 피동문을 만드는 한 방법으로 "동사 어간에 '-아지-/-어지-'를 결합하는 방법(주-/주어지-)"을 제시하고 있다. 이에 언어 모델은 해당 문항에 대한 자연어 설명에서 '-아지-/-어지-'를 '피동 구성 요소'로 제시하지만 그 구성에 대해 세부적인 분석을 제시하지는 않는다. (4)에서 동사 어간이나 시제 형태소를 분석한 것에 비해 '-어지-'는 전체 구성을 언급하고 있어 그 단위가 큰데, '-어지-'의 형태에 주목하여 고정적으로 분석한 결과로 볼 수 있다.

(4) 2023학년도 6모 39번 문항에 대한 자연어 설명의 일부

③ ㉢: 언론에 의해 사건의 전모가 자세히 밝혀졌다. '밝혀졌다'의 기본형은 '밝혀지다'입니다.

분석: '밝히-' (동사 어간) + '-어지-' (피동 구성 요소) + '-었-' (과거 시제) + '-다'

동사 어간에 '-아지-/-어지-'를 결합하는 방법에 정확히 해당합니다.

3) 사전에 학습된 문법 개념어의 사용

언어 모델이 제공하는 자연어 설명에는 학술적인 전문어들이 포함되어 있다. 이들은 일상적이지 않은 전문적인 지식을 표상한다는 점에서 자연어 설명의 표면적인 유창성을 높이는 데 기여한다. 자연어 설명에서 나타나는 학술 전문어 중 상당수는 전술한 바와 같이 문항의 맥락에서 주어진 용어들이나 그렇지 않은 사례도 존재한다.

예컨대 2022학년도 9월 모의평가 37번 문항은 파생어 형성에서 접사의 기능을 묻는데, "④ ㉣에서는 주동사에 결합하여 사동사를 만든다."와 같은 선택지가 포함되어 있다. 언어 모델은 해당 선택지에 대한 자연어 설명에서 (5)와 같이 '강세 접미사'의 용어를 동원한다.

(5) 2022학년도 9월 37번 문항에 대한 자연어 설명의 일부

밀치다: '밀다' + '-치-(강세 접미사)' → '밀다'의 뜻을 강조하는 말이지, 시키는 행위(사동)가 아닙니다.

깨뜨리다: '깨다' + '-뜨리-(강세 접미사)' → '깨다'의 뜻을 강조하는 말이지, 시키는 행위(사동)가 아닙니다.

따라서 '밀치다'와 '깨뜨리다'는 사동사가 아니라 어근의 뜻을 강조하는 강세 접미사가 결합한 단어이므로, ④번 설명은 적절하지 않습니다.

'강세 접미사'는 해당 문항에서 직접 제시된 용어가 아니라는 점에서 주목된다. 언어 모델이 광범위하게 학습한 사전 지식이 작용한 결과로 보이기 때문이다. 전체 문항 가운데 현행 교육과정의 범위를 넘어서는 학술 전문어를 동원한 것으로 보이는 사례들은 (6)과 같다.

(6) 학술 전문어를 동원한 다른 사례

가. 부채질(명사 '부채' → 명사), 풋나물(명사 '나물' → 명사), 휘감다(동사 '감다' → 동사), 빼앗기다(동사 '빼앗다' → 동사)
접사가 붙기 전의 어근과 파생된 단어의 품사가 서로 같습니다. (한정적 접사) (2022-9모-37)

나. 어휘적 피동(접사 교체): '-하다' 동사의 경우 '-하다'를 '-받다, -되다, -당하다' 등으로 교체함. (예: 사랑하다 → 사랑받다) (2023-6모-39)

(6가)의 '한정적 접사'와 (6나)의 '어휘적 피동'은 모두 학교 문법 용어 일람에 포함되어 있지 않으며 범위에 대해 이견이 존재할 가능성이 있다. 이러한 용어의 사용은 자연어 설명의 전문성을 높이는 역할을 한다. 그러나 문법 교육의 관점에서는 학습자 수준을 초과하는 용어의 사용을 반드시 긍정적으로 평가할 수는 없다. 학습자 수준에 부합하지 않는 설명은 교육 자료로서의 효용이 떨어지기 때문이다. 또한 교육 내용으로서 안정성과 보편성이 떨어지는 요소가 학습자에게 미칠 부정적인 영향에 대한 고려도 필요하다.

2. 내부적 불완전성의 양상

1) 다양한 문항 해결 경로의 생성

생성형 인공지능은 자연어 설명을 통해 문항 풀이에 필요한 언어에 대한 분석 과정을 보여 준다. 그러나 언어 모델이 수행하는 언어에 대한 분석은 확률적이며 변동 가능하다. 같은 문항에 대한 답변에서 온도 설정에 따라 상이한 분석 방법이 선택되는 현상이 관찰되었다.

2023학년도 수능 35번 문항은 합성 명사의 내부 구조를 묻는 문제이다. 지문에서는 "합성 명사 내부에 복합어가 있을 때, ㉠합성 명사를 형태소 단위까지 분석하면 합성 명사의 내부 구조를 세밀히 알 수 있다"라고 제시하며, '㉮ 새우볶음, ㉯ 집안싸움, ㉰ 논밭갈이, ㉱ 탈춤놀이' 중 내부 구조가

동일한 것을 묶도록 요구한다. '(단일어 어근+ 단일어 어근)+(단일어 어근+접사)'의 구조를 가진다는 점에서 '집안싸움'과 '논밭갈이'가 같은 내부 구조를 가지는 것으로 파악할 수 있다.

해당 문항의 풀이 과정에서 언어 모델은 온도 설정에 따라 완전히 다른 분석 방법을 선택하였는데 그 내용을 보이면 (7)과 같다.

(7) 2023학년도 수능 35번에 대한 자연어 설명의 일부

가. ㉮ 집안싸움: '집안'(명사) + '싸움'(명사)으로 나뉩니다. '집안'은 '집'(명사) + '안'(명사)으로 이루어진 합성 명사입니다. '싸움'은 어근 '싸우-'(동사) + 접사 '-ㅁ'으로 이루어진 파생 명사입니다. → 구조: [합성 명사(명사+명사)] + [용언 어근 + 접사]

나. ㉯ 집안싸움: '집안' + '싸우-' + '-ㅁ' 문장으로 환원하면: 집안에서 싸우다. → 관계: 부사어(처소)와 서술어의 관계입니다.

다. ㉮ 집안싸움: '집안'(합성어: 집 + 안) + '싸움'(파생어: 싸우- + -ㅁ) → 구조: [복합 어근] + [어간 + 명사 파생 접미사(-ㅁ)]

(7가)는 온도가 0일 때로 형태소의 결합 유형을 기준으로 분석하여 '집안싸움'과 '논밭갈이'의 공통점을 포착하였다. 이에 반해 (7나)는 온도가 0.5일 때로 통사적 의미 관계를 기준으로 하여 '새우볶음, 논밭갈이, 탈춤놀이'가 같은 구조라고 보았다. (7다)는 온도가 1일 때의 답변으로 접미사 형태의 동일성을 중심으로 판단하여 '논밭갈이'와 '탈춤놀이'를 같은 구조로 파악하였다.

온도별 (7)의 세 해설을 개별적으로 접하면 이들은 모두 정교한 문법 분석처럼 보인다. "내부 구조가 동일한 단어끼리 묶은 것은?"이라는 문항의 요구를 제각기 다른 방식으로 해석하여 자연어 설명에서 각 방법별로 일관된 분석 결과를 제시하고 있기 때문이다.[27] 그러나 온도별로 상이한 분석이 나

27) 해당 문항에서 정답을 맞힌 것은 온도가 0인 경우이다. 일반적으로 낮은 온도는 안정적인 답변을 원할 때 권장된다. 다만 본고의 결과만으로는 온도 설정과 문항 풀이 과제의 관계

타났다는 것은 분석 체계에 대한 선택이 결정적이지 않다는 것을 시사한다. 언어 모델 내부에 일관된 분석 체계가 존재하지 않고 분석 경로가 확률적으로 선택되는 것이다.

2) 출력 과정에서 오류의 순차적 전달

트랜스포머 기반 언어 모델이 제공하는 자연어 설명은 연쇄적으로 오생성될 위험을 가지고 있다. 자연어 설명은 기본적으로 계열적인 서술(sequential narrative)의 형태로 구성된다(Barez et al., 2025: 7). 언어 모델이 토큰을 생성하는 과정에서 오생성이 발생할 가능성이 항존하는데 오생성된 내용이 문항 해결에 중요한 정보일 때는 잘못된 결과로 이어진다.[28)]

예컨대 2023학년도 9월 모의평가 35번 문항을 살펴보자. 문항의 내용은 (8)과 같다. 문항의 지문에서 ㉠은 '쓸모없다'로 주어와 서술어 관계의 합성 용언이고, ㉡은 '손잡다'로 목적어와 서술어 관계의 합성 용언이다.

(8) 2023학년도 9모 35번 문항의 일부

② '나는 눈부신 태양 아래에 서 있었다.'의 '눈부신'는 구성적 측면에서 ㉠과 동일한 유형의 합성 용언이겠군.

③ '누나는 나를 보자마자 뒤돌아 앉았다.'의 '뒤돌아'는 구성적 측면에서 ㉡과 동일한 유형의 합성 용언이겠군.

모델은 일부 온도에서 ㉠와 ㉡의 기호를 혼동하는 양상을 보인다.[29)] 온도가 0.5인 경우에는 선택지에서 판단해야 하는 기호를 정확하게 인식하여

를 일반화하기에는 무리가 있다.

28) 이러한 문제를 해결하기 위한 확산(diffusion) 모델이 연구되고 있기도 하나 본고에서 분석 대상으로 삼은 제미나이 3 프로는 여전히 순차 생성 방식을 취한다.

29) 온도별 답변을 생성할 때에는 문항 이미지가 포함된 문서 PDF를 1회 읽은 후 재실행(Rerun)을 반복하였다. 세 온도 모두 동일한 입력을 통해 풀이하도록 실행한 것이므로 입력의 문제로는 보기 어렵다.

정답을 도출하였으나, 0과 1인 경우에는 그렇지 않았다. (9)는 온도가 0인 경우의 CoT와 자연어 설명의 일부를 가져온 것이다. (9가)에서 선택지 2의 기호를 ㉠이 아닌 ㉡으로 잘못 인식하였으며, 그 결과 (9나)에서는 부정확한 판단과 설명을 제시하였다.

(9) 2023학년도 9모 35번 문항에 대한 CoT와 자연어 설명의 사례

가. CoT: "I've just pinpointed Option 2 as the error. It misclassifies '눈부시다' as object-predicate, when it's subject-predicate (눈이 부시다), aligning with type ㉠. The image clearly shows the incorrect symbol ㉡ assigned to this subject-predicate example"

나. ② '눈부신': '눈이 부시다'로 분석되므로 주어+서술어 관계입니다. 따라서 ㉠과 동일한 유형이어야 합니다. 하지만 선택지에서는 ㉡(목적어+서술어)과 동일한 유형이라고 했으므로 틀린 설명입니다. (부적절)

흥미로운 것은 (10)의 사례이다. 위의 (9)에서 CoT와 자연어 설명 간의 오류 인식 양상이 동일했던 것과 달리, 아래의 (10)에서는 CoT에서는 선택지의 기호를 정확하게 인식했음에도 불구하고 자연어 설명에서는 틀린 기호를 제시하고 있다. 하나의 정답을 찾아야 하는 과정에서 선택지 2번에 대한 오생성 결과가 선택지 3번에 대한 판단에 영향을 미친 것으로 해석할 수 있다.

(10) 2023학년도 9모 35번 문항에 대한 CoT와 자연어 설명의 사례

가. Option 3 claims '뒤돌아' shares the syntactic structure of ㉡. After a very close review of the original image, I've confirmed this is accurate.

나. ② '눈부신'은 '눈이 부시다'로 주어와 서술어의 관계를 보여줍니다. 따라서 ㉠과 동일한 유형이어야 하는데, 선택지에서는 ㉡과 동일한 유형이라고 했으므로 적절하지 않습니다.

③ ‘뒤돌아’는 ‘뒤로 돌다’로 부사어(방향)와 서술어의 관계이므로 ㉢과 동일한 유형입니다. (적절함)

3) 사전 학습된 지식의 한계

대규모 언어 모델을 활용할 때 발생하는 오류의 가능성은 근본적으로 사전 학습된 지식의 한계가 원인이 된다고 말할 수 있다. 언어 모델이 사전에 학습한 지식이 문법 교육의 특수성과 전문성에 부합하지 않아 부정확한 설명을 제공할 수 있다. 또한 인간에 의해 생산된 자료를 학습 대상으로 삼는다는 점에서 본질적인 한계를 지니기도 한다.

2026학년도 9월 모의평가 38번 문항을 사전 학습된 지식의 한계를 보이는 사례로 살필 수 있다. 해당 문항은 격 조사에 보조사가 결합한 형태를 판단하는 내용인데 모델은 모든 온도에서 부정확한 결론의 답을 도출하였다. 선택지의 “이것뿐이라면”에서 격조사 ‘이-’를, “이밖에는”에서 보조사 ‘밖에’와 ‘는’을 분석하는 것이 정답을 판단하는 핵심적인 요소이지만 모델은 (11)과 같은 자연어 설명을 제시한다.

(11) 2026학년도 9모 38번 문항에 대한 자연어 설명의 일부
③ ‘이밖에는’: ‘에’(부사격 조사) + ‘는’(보조사) → [격 조사 - 보조사] (예시의 ‘자기한테만’과 동일 유형)
결과: 두 가지 유형이 모두 나타나 있으므로 정답입니다.
④ ‘이것뿐이라면’: ‘뿐’(보조사) + ‘이’(서술격 조사) + ‘라면’(어미). ‘뿐이’에서 [보조사 - 격 조사]의 형태가 보이지만, 서술격 조사 ‘이다’는 활용을 하는 특수성이 있어 일반적인 격 조사와의 결합 예시(예: ‘만으로’)보다는 구조적 대응성이 떨어집니다. 하지만 가장 확실한 근거는 ③번이 완벽하게 대응된다는 점입니다.

(11)에서 모델은 지식의 부족으로 ‘밖에’가 보조사임을 파악하지 못한

다. 또한 조사의 전형적 사례에 대한 편향으로 서술격 조사의 가능성을 기각한다. 조사에 대한 지식의 부족으로 잘못된 판단을 내린 것이다.

사전 학습한 말뭉치의 빈도의 영향을 받은 것으로 여겨지는 사례도 관찰되었다.[30] 예컨대 2025학년도 6월 모의평가 35번 문항은 지칭어가 미지칭과 부정칭으로 쓰이는 경우를 묻고 있다. (12)는 해당 문항의 일부이며, (13)은 그에 대한 설명 사례이다. 그중 (13가)는 해당 문항에 대한 자연어 설명의 일부이고 (13나)는 EBSi의 해설을 가져온 것이다.

(12) 2025학년도 6모 35번 문항의 일부
[보기] 그는 ⓑ 무슨 일이든 척척 해내니?
밖에 ⓓ 어떤 분이 오셨어요?
[선택지] ⓑ, ⓓ는 알지 못함과 불분명함을 모두 나타내며 품사는 동일하다.

(13) 2025학년도 6모 35번 문항에 대한 설명 사례
가. ⓑ는 부정칭, ⓓ는 미지칭이므로 두 단어의 묶음이 두 개념을 모두 포괄하는 것은 맞을 수 있습니다. 하지만 보통 이런 선택지는 각 단어가 두 가지 속성을 다 가진다는 의미(중의성)로 해석되거나, 각각의 연결이 명확해야 하는데, 여기서는 4번 선택지가 훨씬 명확한 정답입니다.
나. ⓓ의 '어떤'은 맥락에 따라 대상을 알지 못하거나 대상이 정해지지 않아 불분명함을 나타내며 관형사이다.

"밖에 어떤 분이 오셨어요?"의 '어떤'에 대해 언어 모델은 온도와 관계없이 모든 답변에서 '어떤'의 의미를 '알지 못함'으로 파악하였다. 이는 사람이 작성한 답변인 (13나)에서 '어떤'의 의미로 부정칭으로서의 사용을 고려

30) 언어 모델이 제공하는 자연어 설명은 기본적으로 확률적인 선택의 결과이고, 확률은 언어 모델이 사전 학습한 말뭉치에서 나타나는 빈도에 영향을 받는다. 인공 신경망 기반 언어 모델은 통계적 언어 모델이 가지는 빈도에 대한 한계를 극복하기 위해 도입되었으나(유원준·안상준, 2026) 여전히 빈도에 의한 영향은 존재한다.

하고 있는 것과 비교된다. 제시된 문장에서 '어떤'이 부정칭으로 쓰이는 경우가 미지칭으로 쓰이는 경우보다 드물기 때문에 낮은 빈도가 영향을 미친 것으로 생각된다.

한편 (14)은 높은 빈도에 이끌려 부정확한 설명을 제시한 사례이다. 해당 문항에서는 "나는 내일 친구랑 만나 미술관이랑 영화관에 가."라는 문장을 제시하고 선택지를 통해 '만나'와 '가'의 어미가 모두 문장을 종결하는 동일한 문법적 기능을 하는지를 판단하도록 하였다. (14)는 이에 대한 언어 모델의 설명이다.

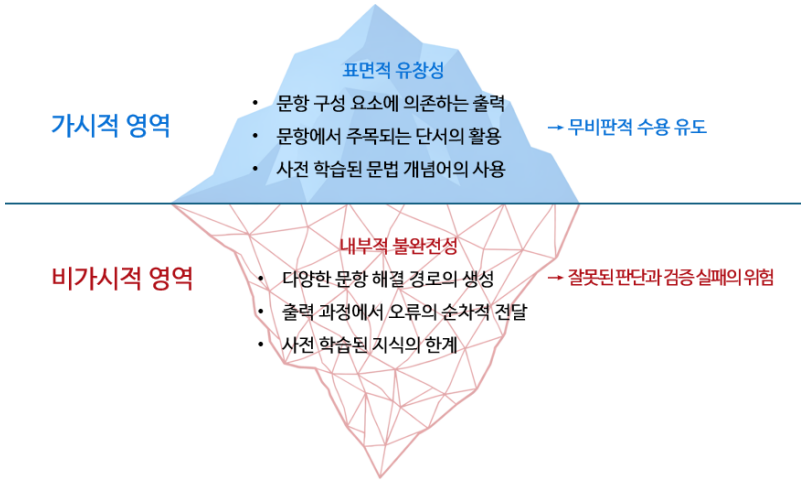
(14) 2026학년도 수능 39번 문항에 대한 자연어 설명의 일부
가. '만나'('만나서'의 줄임꼴 혹은 연결어미 활용)는 뒤에 오는 '미술관이랑 영화관에 가'라는 행위와 연결해 주는 연결 어미의 기능을 하고 있습니다.
나. '만나'는 '만나서' 혹은 '만나고'와 같이 뒤에 오는 말('가')과 이어주는 연결 어미(-아/어)가 쓰인 것입니다.

(14가)에서 '만나'를 '만나서'의 줄임꼴로 본다는 것은 정확한 설명으로 보기 어렵다. (14나)는 다른 온도의 답변인데 '만나'가 연결 어미가 쓰인 것임을 '만나고'나 '만나서'를 들어 설명하고 있다. (14나)는 정확한 설명이나 (14가)와 공통적으로 '만나서'를 언급하고 있다는 것이 주목된다. 이러한 결과는 '만나서'의 높은 빈도에 영향을 받았을 가능성이 있다.

V. 결론

본고에서는 인공지능의 성과를 문법 교육적으로 의미화하는 일이 필요하다는 문제의식을 바탕으로 수능 '언어와 매체'의 단어 문항 풀이 과정에서

생성형 인공지능이 제공하는 자연어 설명을 분석하였다. 앞에서 우리는 자연어 설명이 표면적 유창성과 내부적 불완전성을 가진다는 것을 분석 결과를 바탕으로 논의하였으며, 이를 〈그림 2〉와 같이 정리할 수 있다.



〈그림 2〉 자연어 설명의 특징과 교육적 시사점

자연어 설명이 가지는 표면적 유창성은 사용자에게 그 내용을 무비판적으로 받아들이도록 유도하나, 가시적으로 드러나지 않는 내부적 불완전성은 위험을 내포하고 있다. 여기서 주목해야 할 점은 가시적인 것과 비가시적인 것 사이의 간극이다. 표면적 유창성은 학습자에게 가시적인 영역이지만, 내부적 불완전성은 비가시적이기 때문이다. 예컨대 학습자는 가시적으로 드러나는 유창한 분석을 목격하며 인공지능이 문법을 완벽히 이해하고 있다고 판단하게 되고, 이는 곧 인공지능의 해설에 대한 무비판적 수용과 비판적 검토의 생략으로 이어진다. 학습자가 보이는 것에 이끌려 보이지 않는 것을 고려하지 못할 때, 체계적인 오류의 위험이 존재한다.

이러한 위험을 해소하는 가장 확실한 방법은 내부적으로도 완전한 문법 교육용 모델의 개발일 것이다. 단기적으로는 인공지능을 활용하는 서비스의

생산자들이 사전 학습하는 지식의 범위를 넓힐 수 있다. 사전이나 문법 이론서를 학습할 경우 모델의 성능이 개선될 것으로 예상되나 학습 과정에서 저작권 등 윤리적인 문제가 있다. 중장기적으로는 인공지능 개발자들이 새로운 구조의 언어 모델을 개발할 수 있다. 다양한 경로 생성과 오류의 순차 전달 문제는 트랜스포머 기반 언어 모델의 구조적인 특성으로 인한 것이기 때문이다. 트랜스포머 전후 언어 모델의 큰 성능 차이를 고려하면, 또 다른 구조적 혁신이 문제 해결의 열쇠가 될 것이다.

문제는 현실적으로 우리가 완전한 문법 교육용 인공지능의 완성을 기다리기만 할 수 없다는 것이다. 현시점에 '이미' 많은 교사들과 학습자들이 생성형 인공지능을 문법 교육에 직간접적으로 활용하고 있을 것으로 생각된다. 생성형 인공지능을 활용한 교육 자료 제작은 권장되어야 하는가? 생성형 인공지능의 답변을 바탕으로 정기고사 문항에 대해 이의를 제기하는 학생에게는 어떻게 대처할 것인가? 이러한 질문은 다가올 미래에 대한 상상이 아니라 문법 교육에서 이미 당면하고 있는 현재이다.

자주 반복되어 수사적으로 느껴지기도 하나 사용자의 비판적인 인식이 무엇보다 중요하게 요구된다. 교사와 학습자는 생성형 인공지능의 답변을 확정된 정답이 아닌 하나의 가설적 제안으로 취급해야 하며, 자신의 문법적 직관과 인공지능의 설명을 대조하며 검증하는 주체성을 발휘해야 한다. 본고는 그러한 비판적인 인식이 필요한 지점을 구체적으로 밝혔다는 의의가 있다. 교사와 학습자는 생성형 인공지능이 제공하는 답변이 맥락에 의해 직접적인 영향을 받으며, 확률적으로 가능한 여러 경로 중 하나로 순차적 출력에 의한 오류를 내포할 수 있고, 저빈도 형식에 대한 고려나 문법 이론에 대한 이해가 충분하지 못할 수 있음을 인지해야 한다. 조진수(2026: 308)에서는 미래의 문법 교육이 인간 언어와 인공지능 생성 언어에 대한 인식론적 조망을 포함해야 한다고 하면서, '인공지능에 의해 산출된 언어는 인간이 직접 말하거나 쓰는 방식으로 만든 언어와 어떤 차이가 있을까?'를 핵심 질문의 사례로 제시한 바 있다. 본고의 분석 결과는 그에 대한 근거로 인공지능이

생성한 언어의 구체적인 양상을 제공한다. 이 연구는 언어 주체의 비판적 인식을 이끄는 출발점으로서 생성형 인공지능이 제공하는 자연어 설명의 특징을 문법 교육적인 관점에서 실증적으로 포착하였다는 의의를 지닌다.

* 본 논문은 2026.01.31. 투고되었으며, 2026.02.08. 심사가 시작되어 2026.03.07. 심사가 종료되었음.

참고문헌

구본관(2010),「문법 능력과 문법 평가 문항 개발의 방향」,『국어교육학연구』37, 185-218.

권태현(2024),「ChatGPT를 활용한 쓰기 채점 및 피드백 방안 –프롬프트 전략을 중심으로」,『새국어교육』141, 7-42.

김규훈(2023),「2022학년도 이후 수능 문법 문항의 비평 연구 –지문형 및 통합형 문항을 중심으로–」,『우리말 글』97, 71-99.

김민해·이유진·서나영·천하연·전대일(2025),「LLM을 활용한 수능 국어 영역 문법 문제 생성 시스템 제안」,『에듀테인먼트연구』7(1), 311-327.

김승주(2022),「딥러닝 자연어처리 기법을 활용한 논증적 글쓰기 자동 채점 방안 연구: 교사 채점자와 기계 채점자의 협업적 채점 수행 모델을 기반으로」, 한국교원대학교 박사학위논문.

김은선(2025),「교사와 인공지능 글쓰기 피드백에 대한 초등학생의 반응」,『한국초등국어교육』80, 5-35.

나상수(2025),「텍스트 생성 과정에서의 문법 활용 능력 평가 모델 구현 연구」, 서울대학교 박사학위논문.

남가영(2017),「통합형 문법 평가문항의 양상과 설계 방향 –국가 수준 문법 평가문항을 중심으로–」,『우리말 글』75, 161-200.

남길임·황은하·송현주·안의정(2024),「생성형 AI의 문법적 능력에 대한 국어학적 연구 – 형태 통사적 특성을 중심으로 –」,『한말연구』65(11), 1-21.

류수열·주세형·남가영(2021),『국어과 교사 전문성 신장 노트 2 국어교육 평가론』, 서울: 사회평론아카데미.

박서윤·강예지·강조은·김유진·이재원·정가연·최규리·김한샘(2024),「GPT-4를 활용한 인간과 인공지능의 한국어 사용 양상 비교 연구」,『국어국문학』206, 5-47.

박종미(2025ㄱ),「문법 학습 비계 설정자로서 ChatGPT 활용을 위한 과제와 방향」,『국어교육학연구』60(1), 83-109.

박종미(2025ㄴ),「ChatGPT의 국어 문법 개념 이해도 평가 – 의미 유사도 분석을 통한 교육적 활용 가능성 탐색 –」,『청람어문교육』103, 309-336.

유원준·안상준(2026),『딥 러닝을 이용한 자연어 처리 입문』, 검색일자 2026년 1월 28일, https://wikidocs.net/book/2155.

이경숙(2025),「수능 국어 영역에서 ChatGPT의 문법 능력에 대한 연구: 2022~2025년 '언어와 매체'를 중심으로」,『언어와 정보 사회』54, 157-189.

이관규(2008),『학교 문법 교육론』, 서울: 고려대학교 민족문화연구소.

이관희·정희창(2010),「국민의 문법 능력 평가 연구」,『우리말 글』67, 53-76.

이관희·최선희·김자영(2022),「문법 문항에 반영된 예상 오개념 분석 –대학수학능력시험 및 6월·9월 모의평가를 대상으로」,『국어교육』179, 207-249.

이기돈(2025),「수능 수학 영역에서 선다형 문항 선지의 측정평가적 기능 검토 및 단답형 확대

제안」, 『학교수학』 27(2), 393-412.
이도영·김잔디·민송기·서수현·안혁·장창중·한재덕(2021), 『평가 문항 출제의 정석 국어과 선다형 시험 평가 문항 어떻게 만들어지나?』, 고양: 한국교육방송공사.
이선웅(2012),『한국어 문법론의 개념어 연구』, 서울: 월인.
장세민(2025. 11. 19.), "제미나이 3로 2026 수능 풀어보니"...GPT-5.1 누르고 압도적 1위, AI Times, 검색일자 2026년 1월 28일, https://www.aitimes.com/news/articleView.html?idxno=204103.
정민주·서수현·남민우·최숙기·이상일·남가영(2022), 「좋은 국어과 평가 문항 특성에 대한 질적 분석 연구 – 국어과 평가 문항 양호도 분석틀 개발 연구(2)」, 『청람어문교육』 89, 43-78.
정한데로(2023), 「'생성 AI 화자'의 단어 형성 - 〈ChatGPT〉, 〈Bard〉, 〈CLOVA X〉를 대상으로 -」, 『한말연구』 64(54), 1-26.
조진수(2026), 「인간 언어에 대한 인식 확장과 미래 문법 교육과정을 위한 핵심 질문의 재구성」, 『국어교육연구』 90, 291-305.
주세형(2009), 「국가 수준 학업 성취도 평가에서의 소위 텍스트 중심 원리에 대한 비판 -2005~2008년 문법 영역을 중심으로-」, 『국어교육학연구』 35, 481-506.
주지연(2020), 「문법지식의 불확정성과 문법 교육」, 『국어교육연구』 73, 153-184.
최인찬·권도형(2024), 「생성형 인공지능 ChatGPT의 국어능력은 어떠한가? - 2024학년도 대학수학능력시험 국어영역 문항 풀이 결과의 오류 유형 분석을 중심으로」, 『리터러시 연구』 15(2), 279-318.
Ahn, J. J. & Yin, W. (2025), "Prompt-reverse inconsistency: Llm self-inconsistency beyond generative randomness and prompt paraphrasing", *arXiv preprint arXiv:2504.01282*.
Barez, F., Wu, T. Y., Arcuschin, I., Lan, M., Wang, V., Siegel, N., ... & Bengio, Y. (2025), "Chain-of-thought is not explainability", Preprint, alphaXiv, v1.
Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020), "Language models are few-shot learners" *Advances in neural information processing systems* 33, 1877-1901.
Cameron, R. W. (2025), "Demystifying reasoning models", 검색일자 2026. 1. 31., https://cameronrwolfe.substack.com/p/demystifying-reasoning-models.
Creswell, J. (2012), 『질적 연구방법론: 다섯 가지 접근』, 조흥식·정선욱·김진숙·권지성(역), 서울: 학지사, 2015.
Gemini API(2026. 1. 29.), Gemini 3 개발자 가이드, 검색일자 2026. 1. 31., https://ai.google.dev/gemini-api/docs/gemini-3?hl=ko.
Google(2025. 11. 18.), "A new era of intelligence with Gemini 3", Google Keyword, 검색일자 2026. 1. 28., https://blog.google/products/gemini/gemini-3/#note-from-ceo.

hehee9(2025. 12. 17.), “2026-CSAT”, Github, 검색일자 2026. 1. 21., https://github.com/hehee9/2026-CSAT.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., ... & Olah, C. (2022), “In-context learning and induction heads”, arXiv preprint arXiv:2209.11895.

OpenAI(2025. 12. 11.), “Open AI, Introducing GPT-5.2”, 검색일자 2026. 1. 28., https://openai.com/index/introducing-gpt-5-2.

OpenAI Platform(n. d.), “Completions(Legacy)”, 검색일자 2026. 1. 31., https://platform.openai.com/docs/api-reference/completions.

Sammani, F., Mukherjee, T., & Deligiannis, N. (2022), “Nlx-gpt: A model for natural language explanations in vision and vision-language tasks”. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Wang, L., Li, L., Dai, D., Chen, D., Zhou, H., Meng, F., ... & Sun, X. (2023), “Label words are anchors: An information flow perspective for understanding in-context learning”, arXiv preprint arXiv:2305.14160.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022), “Chain-of-thought prompting elicits reasoning in large language models”, *Advances in neural information processing systems* 35, 24824-24837.

초록

수능 '언어와 매체'의 단어 문항 풀이에서 나타나는 생성형 인공지능의 자연어 설명에 대한 분석

오지은 · 박인규

이 연구의 목적은 대학수학능력시험 '언어와 매체'의 단어 관련 문항 풀이와 관련하여 생성형 인공지능이 제공하는 자연어 설명의 특성을 분석하는 것이다. 이 연구는 급속도로 발전하는 인공지능 자연어 처리 기술의 성과를 문법 교육적으로 의미화할 필요성에서 출발하였다. 2022학년도부터 2026학년도까지의 '언어와 매체'의 단어 관련 문항을 제미나이 3 프로 모델에 입력하여 자연어 설명을 생성하고 그 특성을 분석하였다. 분석 결과 자연어 설명은 표면적 유창성과 내부적 불완전성의 특성을 가지는 것으로 파악되었다. 표면적 유창성의 양상으로는 문항 구성 요소에 의존하는 출력, 문항에서 주목되는 단서의 활용, 사전 학습된 문법 개념어의 사용을 분석하였다. 내부적 불완전성의 양상으로는 다양한 문항 해결 경로의 생성, 출력 과정에서 오류의 순차적 전달, 사전 학습된 지식의 한계를 기술하였다. 표면적 유창성은 가시적인 반면 내부적 불완전성은 비가시적이기 때문에 문법 교육에서 언어 주체의 비판적인 인식이 중요하게 요구된다.

핵심어 자연어 설명, 표면적 유창성, 내부적 불완전성, 대규모 언어 모델, 생성형 인공지능

ABSTRACT

Surface Fluency and Internal Imperfection in Natural Language Explanations for CSAT "Language and Media" Word Items

Oh Jieun · Park Ingyu

This study analyzes the characteristics of natural language explanations generated by large language models when solving word-related items in the Korean College Scholastic Ability Test (CSAT). Word-related items from the 2022 to 2026 CSAT were input into the Gemini 3 Pro model to generate natural-language explanations, and their features were examined. The analysis indicates that these explanations exhibit two key characteristics: surface fluency and internal imperfection. This study delineates three manifestations of surface fluency: outputs relying on question components, the use of salient cues in the question, and the use of pre-trained grammatical terminology. It also identifies three manifestations of internal imperfections: the generation of multiple solution pathways, the sequential propagation of errors during the output process, and limitations in pre-trained knowledge. Because surface fluency is visible whereas internal error risk is not, grammar education that fosters users' critical awareness is urgently required.

KEYWORDS Natural Language Explanation(NLE), surface fluency, internal imperfection, language model, generative artificial intelligence