https://doi.org/10.23073/igll.2026..23.024

상용 AI의 논증 평가 능력 검증••

PSAT 강화·약화 문항을 중심으로

Argument-Evaluation Capabilities of Commercial AI:
A Focus on Strengthening and Weakening Items in the Public Service Aptitude Test (PSAT)

이일권••• · 최승락••••

국문요약 인공지능시대 핵심 역량으로 여겨지는 비판적 사고 능력을 함양하기 위해서는 올바른 논증을 평가하는 능력을 길러야 한다. 본 연구는 대규모 언어 모델 기반 상용 인공지능이 지닌 논증 평가 능력을 공직적성시험(PSAT)에 사용되는 '강화·약화 문항'을 통해 검증하였다. 먼저 본고는 비판적 사고 교육에 있어 강화·약화 문항이 지닌 교육적·실용적 가치에 대해 논의한다. 다음으로 이러한 이론적 논의를 바탕으로, 공직적성시험 언어논리 영역의 기출 문항 20개를 선별하여 ChatGPT 5.2 Pro, Claude Opus 4.5, Gemini 3 Ultra의 수행 능력을 비교하였다. 각 모델에 동일 문항을 10회씩 반복 제시한 결과, ChatGPT가 99.0%의 정답률로 가장 높은 성능을 보였으며, Claude 87.0%, Gemini 83.0%로 그 뒤를 이었다. 특히 반복 측정을 통해 모델별로 체계적인 오류가 생성되는 사례를 확인하였으며, 단일 응답만으로는 탐지하기 어려운 고착 오류가 존재함을 알 수 있었다. 본 연구는 기존 대규모 언어 모델 기반 프로그램의 성능 평가가 주로 수학이나 코딩 능력에 편중되어 왔다는 점에 주목하여, 비판적 사고의 핵심 구성요소를 통합적으로 측정할 수 있는 강화·약화 문항의 가치를 재조명하고, 인공지능을 교육적으로 활용할 때 교수자의 검토와 평가가 필수적임을 주장한다.

핵심어 비판적 사고, 논증 평가, 강화·약화, 대규모 언어 모델(LLM), PSAT

차례

1. 비판적 사고 교육과 올바른 논증의 평가
2. 상용 AI 프로그램의 논증 평가 능력 검증
3. 결론: 더 나은 비판적 사고 교육을 위한 AI의 활용

• 본 과제(결과물)는 2026년도 교육부 및 강원특별자치도의 재원으로 강원앵커센터의 지원을 받아 수행된 지역성장 인재양성체계(앵커) 글로컬대학 30의 결과입니다(2026-ANCHOR-10-009).

•• 본 연구의 실험에 활용된 ChatGPT 5.2 Pro, Claude Opus 4.5, Gemini 3 Ultra는 실험 진행 당시의 버전을 기준으로 한 것이다. 2026년 7월 현재 각 모델은 더욱 고도화된 최신 버전으로 업데이트 된 상태이므로, 현재 시점의 최신 모델을 대상으로 동일한 평가를 수행할 경우 본 연구의 결과보다 향상된 정답률을 보이거나 고착화된 오류 패턴이 일부 완화되었을 가능성이 존재한다. 더불어, 본 연구에서 검증한 LLM의 '문항 풀이 능력'과 교육 현장에서 요구되는 '문항 출제 능력'은 요구되는 인지적 층위와 논리적 구조화의 방향이 다른 별개의 문제임을 밝혀둔다. 향후 LLM을 문항 출제에 직접적으로 활용하고자 한다면, 그 전제 조건 중 하나로서 해당 모델이 기존 출제 문항의 해결 과정에서 어떠한 체계적 오류나 논리적 비약 없이 100%에 수렴하는 정답률을 우선적으로 증명해야 할 것이다.

1. 비판적 사고 교육과 올바른 논증의 평가

1.1. AI 시대와 비판적 사고의 중요성

2022년 말 ChatGPT가 세상에 공개된 뒤, 다양한 분야에서 활용되는 인공지능(artificial intelligence, AI) 프로그램이 급속도로 발전하고 있다. 이러한 프로그램이 특정 영역에서는 이미 인간의 수준을 넘어섰다는 평가도 잇따르고 있으며, 과거에 경험하지 못했던 새로운 유형의 문제들이 사회적 쟁점으로 급부상하고 있다. 기계가 생성한 사진이나 영상물은 이미 실제와 구분하기 어려워져 가짜 뉴스 문제가 심화되고 있으며, 일부 생성물에는 사

••• 인사혁신처 행정사무관
••••한림대학교 인문학부 철학전공 조교수, 한림철학교육연구소 소장

회적 편향이나 인종차별적 요소가 포함되어 있다는 점이 지적되고 있다. 이에 따라 생성형 AI를 사용할 때는 그 결과물을 무비판적으로 받아들이기보다, 비판적인 시각으로 검토하고 선별적으로 활용해야 한다는 목소리가 커지고 있으며 비판적 사고 교육의 중요성도 함께 강조되고 있다.

고등교육 현장에서의 AI 도입 속도도 역사상 유례를 찾기 어려울 정도로 빨라지고 있다. Contractor와 Reyes에 따르면, 미국의 한 명문대 표본에서 학술 목적의 생성형 AI 사용률은 2023년 봄 이전 10% 미만에서 2024년 가을 80% 이상으로 급증했다.[1] 이 수치는 단일 기관 표본이므로 모든 대학으로 일반화하기는 어렵지만, 미국 전체 노동자의 AI 채택률 40%와 성인 전체의 채택률 23%를 크게 웃도는 수치다. 인터넷이나 스마트폰의 초기 확산 속도마저 뛰어넘는 이례적인 기술 수용이 대학 캠퍼스에서 일어나고 있는 것이다.

교육 영역에서 생성형 AI의 활용이 확대됨에 따라 학생들이 AI가 생성한 결과물을 무비판적으로 수용하여 그대로 제출하는 사례가 새로운 문제로 대두되고 있으며, 이러한 기술에 대한 과도한 의존은 인간 고유의 창의력과 비판적 사고력을 저하시킬 수 있다는 심각한 우려도 제기되고 있다. 세계경제포럼의 「직업의 미래」 보고서에 따르면, 비판적 사고의 척도인 분석적 사고는 기업들이 가장 중요하게 평가하는 핵심 역량으로 강조되며, 창의적 사고와 함께 향후 그 중요성이 더욱 커질 역량으로 지목되었다.[2] 또한 딜로이트 AI 연구소는 생성형 AI와 자동화 기술이 확산될수록 AI의 산출물을 검증하고 정보의 이면을 꿰뚫어 보는 비판적 사고 능력이 더욱 가치 있는 역량으로 자리매김할 것이라고 강조한다.[3]

AI 사용을 우려하는 목소리도 적지 않다. 보스턴컨설팅그룹에 따르면, 생성형 AI가 개인의 문제 해결 능력을 향상시킬 수 있는 반면, 지나친 의존은 오히려 능력 저하를 초래할 수 있다.[4] 해당 연구에서 GPT-4를 사용한 집단은 결과물을 비판적으로 검토하지 않고 그대로 수용하는 경향을 보여, 스스로 문제를 해결한 집단보다 낮은 성과를 보였다. 이는 AI에 대한 무분별한 의존이 비판적 사고 능력에 잠재적인 위협이 될 수 있음을 시사한다. 마이크로소프트와 카네기멜론대학교의 연구 역시 사용자가 생성형 AI를 신뢰할수록, 그리고 AI에 단순히 의존할수록 비판적 사고를 덜 관여하는 경향이 있음을 보여준다.[5]

이러한 맥락에서 AI 시대에 비판적 사고 능력을 함양하는 일은 단순한 교육적 권장 사항을 넘어 필수적인 과제가 되고 있다. 그렇다면 비판적 사고 교육은 어떤 방식으로 이루어져야 할까? 1장에서는 먼저 '비판적 사고'에 대한 여러 철학자들의 입장을 살펴본 뒤, 이들 입장이 적어도 올바른 논증을 평가하는 능력을 중시한다는 점을 논할 것이다. 이어 올바른 논증을 평가하는 능력을 기르는 한 가지 방편으로 공직적성시험(Public Service Aptitude Test, PSAT)과 법학적성시험(Legal Education Eligibility Test, LEET)에 자주 활용되는 논증의 강화·약화 문항을 소개한 후, 이를 비판적 사고 교육에 활용하는 것이 유용하다고 주장할 것이다. 이어지는 2장에서는 상용 AI가 이러한 강화·약화 문항을 얼마나 정확하게 풀이하는지에 관한 실험 결과를 제시하고 이를 해석할 것이다.

1.2. 비판적 사고란?

비판적 사고(critical thinking)는 교육학과 철학 분야에서

1 Zara Contractor·Germán Reyes, "Generative AI in Higher Education: Evidence from an Elite College," *IZA Discussion Paper*, No. 18055, 2025, pp.11~25.

2 Saadia Zahidi et al., *The Future of Jobs Report 2023*, World Economic Forum, 2023, pp.38~39.

3 Deloitte AI Institute, "The State of Generative AI: Q2 Report," Deloitte, 2024, pp.21~27.

4 BCG Editorial Team, 「생성형 AI로 가치를 창출하고 또 파괴하는 법」, BCG 블로그, 2023.10.10.

5 Hao-Ping (Hank) Lee et al., "The Impact of Generative AI on Critical Thinking," *CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2025, pp.11~12.

오랜 기간 논의되어 온 개념으로, 여러 학자들에 의해 다양하게 정의되어 왔다. 비판적 사고라는 용어가 교육학적 목표로 정립되기 이전, 미국의 철학자 Dewey(1910)는 '반성적 사고(reflective thinking)' 개념을 통해 비판적 사고의 원형을 다음과 같이 제시했다.[6]

> "어떤 믿음이나 가정된 지식의 형태를, 이를 지지하는 근거(grounds)와 그로부터 도출되는 결론(conclusions)에 비추어 능동적이고 지속적이며 주의 깊게 고찰하는 과정"

이 정의는 비판적 사고가 단순한 회의적 태도에 머무는 것이 아니라, 논증에서 근거와 결론 간의 관계를 면밀히 검토하는 적극적인 지적 활동임을 강조한다. '논증'(argument)에 대한 일반적인 이해는 "전제와 결론으로 이루어진 말 묶음"이다. 위의 정의는 전제와 결론 간의 관계 파악을 핵심으로 본다는 점에서, 비판적 사고가 올바른 논증을 평가하는 능력과 관련됨을 보여준다.

Dewey에게 사고란 막연한 상상이나 의식의 흐름이 아니다. 그것은 근거가 결론을 논리적으로 적절하게 지지하는지를 검토하는 과정이다. 그는 사고의 과정이 혼란스럽고 불확정적인 상황에서 출발하여, 체계적인 탐구를 통해 명료하고 확정된 상황으로 나아가는 과정이라고 보았다. 특히, Dewey는 '판단의 유보(suspended judgment)'를 강조했는데 반성적 사고는 즉각적인 결론 도출을 유보하고, 충분한 근거가 수집될 때까지 판단을 미루며 의심을 유지하는 태도를 요구한다. 이는 논리적 오류 중 하나인 성급한 일반화를 방지하기 위한 장치이기도 하며, 비판적 사고가 본질적으로 논증에서 결론을 뒷받침하는 근거(전제)를 추적하는 규범적 활동임을 시사한다. 이처럼 Dewey의 '반성적 사고' 개념은 논증을 올바르게 평가하는 역량의 중요성을 강조한다.

Dewey의 반성적 사고 개념을 더욱 구체화하고 측정 가능한 형태로 발전시킨 인물이 Glaser이다. Watson과 함께 '왓슨-글레이저 비판적 사고 검사'를 개발한 Glaser는 비판적 사고가 다음의 세 가지 요소를 포함한다고 정의했다.[7]

(1) 태도(Attitude): 자신의 경험 범위 내에 있는 문제와 주제를 사려 깊게 고려하려는 성향.
(2) 지식(Knowledge): 논리적 탐구와 추론의 방법에 관한 지식.
(3) 기술(Skill): 이러한 논리적 방법을 적용하는 데 필요한 일정 수준의 숙련도.

위의 정의에서 가장 주목해야 할 부분은 두 번째 요소인 '지식'이다. 그는 비판적 사고가 단순히 "생각을 깊게 하려는 의지"만으로는 불가능하며, 반드시 "논리적 탐구 방법"을 알고 있어야 한다고 생각했다. 여기서 그가 말한 논리적 탐구 방법을 적용하는 능력에는 증거를 평가하고 논증을 분석하는 능력, 명시되지 않은 가정과 가치를 인식하는 능력, 언어를 정확하고 명료하게 사용하는 능력, 명제들 간의 논리적 관계가 존재하는지를 파악하는 능력, 타당한 결론을 도출하고 일반화하는 능력 등이 포함된다. 이는 비판적 사고가 본질적으로 연역, 귀납, 타당성 검증, 오류 식별 등의 논리적 도구를 사용하는 기술임을 보여준다.

Dewey와 마찬가지로, Glaser에게도 비판적 사고는 어떤 신념을 지지하는 증거를 검토하고, 그 증거로부터 결론이 적절하게 도출되는지를 평가하는 지적 탐구라 할 수 있다. 즉, 논증에서 결론을 뒷받침하는 전제를 검토하고 그 논증이 올바른지를 판단하는 과정으로도 볼 수 있다.

마지막으로 Fisher&Scriven(1997)은 비판적 사고를 읽

6 John Dewey, *How We Think*, D.C. Heath & Co., 1910, p.9.

7 Edward M. Glaser, *An Experiment in the Development of Critical Thinking, Teachers College*, Columbia University, 1941, p.5.

기나 쓰기와 같은 학문적 능력으로 보고, "관찰과 의사소통, 정보 그리고 논증에 대한 숙련되고 능동적인 해석이자 평가"라고 정의한다.[8] 이들은 비판적 사고를 단순히 문제의 해답을 찾는 문제 해결(problem solving)이나 새로운 아이디어를 만드는 창의적 사고와 구분한다. 문제 해결이 건설적(constructive)이고 해법 지향적이라면, 비판적 사고는 제시된 해법이나 주장의 지적 품질을 검사하는 평가적(evaluative) 활동이라는 것이다. 또한 이들은 비판적 사고의 핵심을 "주장을 해석하고 논증을 평가하는 숙련된 활동"이라고 여겼다. 이는 글의 행간에 숨겨진 전제와 결론의 구조를 파악하고, 그 논증이 논리적 건전성(soundness)을 갖추었는지를 따지는 과정을 의미한다. 즉, 이들 역시 논증을 올바르게 평가하는 능력을 비판적 사고의 핵심으로 본 것이다.

그 밖에도 Ennis(1987)는 비판적 사고를 '무엇을 믿거나 행할지 결정하는 데 초점을 맞춘 합리적이고 반성적인 사고'라고 정의했다.[9] Paul(1990)은 비판적 사고의 메타인지적 성격을 강조한 바 있으며, 이를 "자신의 사고를 개선하기 위해, 사고하는 과정 자체에 대해 사고하는 것"으로 규정하기도 했다.[10] 이러한 비판적 사고에 관한 여러 정의는 비판적 사고가 반성적 사고임을 보여주는 동시에, 무엇보다 어떤 주장을 뒷받침하는 근거를 찾고 평가하는 능력, 즉 올바른 논증을 평가하는 능력이 그 핵심임을 보여준다.

1.3. 올바른 논증과 강화·약화의 개념

앞서 검토한 '비판적 사고'에 대한 여러 정의를 종합해 보면, 비판적 사고에는 공통적으로 "올바른 논증"에 대한 이해가 필수적임을 알 수 있다. 통상적으로 '올바른 논증'이란, 적절한 관련성을 지닌 전제와 결론으로 이루어진 말 묶음으로 다음의 네 가지 요건을 충족해야 한다.

(1) 명시적으로 드러난 논증의 전제가 참이다.
(2) 논증의 숨은 전제(implicit premise)가 참이다.
(3) 논증의 결론이 참이다.
(4) (1)과 (2)의 전제가 (3)의 결론을 적절히 뒷받침한다.

여기서 '숨은 전제'란 표면적으로 논증에 나타나 있지는 않지만, 결론을 이끌어내기 위해 필요한 전제를 말한다. 어떤 논증이 위의 네 가지 조건을 모두 충족한다면 그 논증은 올바르다고 할 수 있다. 반대로, 네 가지 조건 중 하나라도 충족하지 못한다면 그 논증은 올바르지 않다. 일상적인 논증은 대부분 귀납 논증의 형태를 띠며, 추가적인 정보에 따라 강화되거나 약화될 수 있다.[11] 거칠게 말해, 위 네 가지 조건에 부합하는 사례는 그 논증이 올바름을 지지해 주므로 "논증을 강화하는 사례"라고 할 수 있다. 반대로 어떤 논증이 위 네 가지 조건에 부합하지 않음을 보여주는 사례는 그 논증이 올바르지 않음을 지지해 주므로 "논증을 약화하는 사례"라고 할 수 있다.[12]

8 Alec Fisher·Michael Scriven, *Critical Thinking: Its Definition and Assessment*, Point Reyes, CA: Edgepress, 1997, p.21.

9 Robert H. Ennis, "A Taxonomy of Critical Thinking Dispositions and Abilities," in Joan B. Baron and Robert J. Sternberg (eds.), *Teaching Thinking Skills: Theory and Practice*, New York: W. H. Freeman, 1987, p.10.

10 Richard Paul, *Critical Thinking: What Every Person Needs to Survive in a Rapidly Changing World*, Center for Critical Thinking and Moral Critique, 1990, p.91.

11 대부분의 논증은 연역 논증과 귀납 논증으로 구별할 수 있다. 여기서 '연역 논증'이란, "전제가 참일 경우 결론이 반드시 참인 것으로 의도된 논증이나 추론. 혹은 전제에 비추어 볼 때 결론이 결코 거짓이 될 수 없는 것으로 의도된 논증이나 추론"을 말한다. 또한 '귀납 논증'은 "전제가 참일 경우 결론이 참일 개연성이 높아지는 것으로 의도된 논증이나 추론"을 말한다. 타당한 연역 논증의 경우, 전제가 참이라면 그 결론은 반드시 참이기 때문에 강화하는 요소가 존재하지 않는다고 여길 수 있다. 이에 본고에서는 되도록 귀납 논증에 맞추어 논의를 진행한다.

12 일상적으로 강화·약화에 관한 문제를 물을 때는 '논증을 강화 및 약화하는 사례', '논지(결론)를 강화 및 약화하는 사례'라는 표현을 사용한다. 이 두 가지 표현 모두 올바른 논증을 구성하는 4가지 각 조건을 충족하느냐 하지 않느냐로 강화 및 약화 여부를 파악한다. 이외에 '전제가 참임을 강화하는 사례'와 같은 표현도 사용하는데 이는 논증을 구성하는 각 전제도 그 논증의 부속 논증의 결론일 수 있기에 사용될 수 있다. 부속 논증의 결론을 강

올바른 논증의 네 가지 요건에 따라 논증을 강화하는 요소는 다음과 같이 정리될 수 있다.

(1*) 전제 강화: 전제가 참임(전제의 신뢰성)을 지지하는 추가적인 증거를 제시한다.

(2*) 숨은 전제 강화: 숨은 전제가 참임(숨은 전제의 신뢰성)을 지지하는 추가적인 증거를 제시한다.

(3*) 결론 강화: 결론이 참임(결론의 신뢰성)을 지지하는 추가적인 증거를 제시한다.

(4*) 전제와 결론의 연결성 강화: 주어진 (숨은)전제가 결론을 적절히 그리고 효과적으로 지지하는 설명을 제시한다.[13]

유사하게, 논증의 약화는 다음과 같은 방식으로 정리할 수 있다.

(1*) 전제 약화: 전제가 참임(전제의 신뢰성)에 의문을 제기하거나 그것이 거짓임을 지지하는 추가적인 증거를 제시한다.

(2*) 숨은 전제 약화: 숨은 전제가 참임(숨은 전제의 신뢰성)에 의문을 제기하거나 그것이 거짓임을 지지하는 추가적인 증거를 제시한다.

(3*) 결론 약화: 결론이 참임(결론의 신뢰성)에 의문을 제기하거나 그것이 거짓임을 지지하는 추가적인 증거를 제시한다.

(4*) 전제와 결론의 연결성 약화: 주어진 (숨은)전제가 결론을 적절히 그리고 효과적으로 지지하지 못하는 설명을 제시한다.

특히, 논증의 약화 (4*)는 논증에 반례를 제시하는 대표적인 방식이다. 다시 말해, "논증의 전제가 참이면서 그 결론이 거짓인 경우"를 제시함으로써 그 논증이 옳지 않음을 보이는 방식이다. 논증의 결론에 대한 신빙성을 약화하는 대표적인 방식으로는, 논증에서 주장된 인과관계가 사실은 상관관계에 불과함을 보이는 것이 있다.

이러한 강화·약화의 개념적 틀은 논증 평가를 단순한 형식적 타당성 검증을 넘어, 화용적 목적 달성과 실천적 문제 해결로 확장한 비판적 사고 연구의 흐름을 계승한다. 박준호(2006)가 성공적인 정당화를 위해 결론보다 더 명확한 전제가 요구된다는 화용적 조건을 강조하고,[14] 박만엽(2011)이 암묵적 가정의 능동적 평가를 비판적 사고의 핵심으로 규정한 것은 본고의 문제의식과 맥을 같이한다.[15] 나아가 비판적 사고를 문제 해결과 대안 모색을 위한 고차적 사고(higher order thinking) 역량으로 파악한 서민규(2010)의 논의 역시 본고의 지향점을 뒷받침한다.[16] 요컨대 본고에서 제시한 숨은 전제의 참 여부에 대한 평가와 전제와 결론 간의 연결성 등의 기준은 앞서 언급한 선행연구의 다양한 입장을 종합하여 정리한 것이라 할 수 있다.

1.4. 강화·약화 평가의 교육적 가치

앞서 1절에서는 비판적 사고 역량의 중요성이 AI 시대에 더욱 커지고 있음을 살펴보았다. 이어 2절에서는 비판적 사고 역량에서 올바른 논증을 평가하는 능력이 핵심임을 논의했다. 또한 3절에서는 올바른 논증을 평가하는 방법의 하나로 논증을 강화하거나 약화하는 사례를 찾는 방식이 있음을 살펴보았다. 이를 종합하면, 논증의 강화·약화를 평가하는 역량을 기르는 교육은 비판적

화하는 사례는 자연스럽게 그 논증을 강화하는 사례라고 볼 수 있다.

13 이외에 잠재적 반론이나 반례를 축소하거나 차단하는 '반론 제거' 방식도 논증을 강화하는 사례로 고려될 수 있으나 이 반론 제거가 (1*)~(4*)와 직접적으로 연관되지 않는 경우, 강화하는 사례로 고려되지 않을 수 있다.

14 박준호, 「논증의 종류와 평가의 기준: 비판적 사고와 비형식 논리학의 의의」, 『범한철학』 42(3), 범한철학회, 2006, 273~296쪽.

15 박만엽, 「비판적 사고와 논증 분석: '발표와 토론'을 중심으로」, 『사고와표현』 4(1), 한국사고와표현학회, 2011, 63~102쪽.

16 서민규, 「비판적 사고 교육, 무엇을 어떻게 할 것인가」, 『교양교육연구』 4(2), 한국교양교육학회, 2010, 129~139쪽.

사고 역량의 함양과 직접적으로 연관됨을 알 수 있다.

먼저, 논증의 강화·약화를 평가하기 위해서는 글에 제시된 논증의 뼈대를 파악하여 논지와 논거를 추출해야 한다. 교육의 대상이 되는 글은 대부분 일정한 주장을 담고 있으며, 그 주장을 뒷받침하는 근거도 제시한다. 즉, 대부분의 글에는 논증이 담겨 있으며, 그 논증을 파악하는 것이 독서의 핵심이라고 할 수 있다. 나아가 비판적 사고 교육은 그렇게 파악한 논증이 올바른지를 평가하는 역량을 요구하며, 더 나아가 그 논증을 지지하거나 지지하지 않기 위한 방법을 아는 것까지 요구한다. 바로 이것이 논증의 강화·약화 사례를 평가하는 활동이다.

Facione가 주도한 1990년 델파이 보고서에 따르면, 비판적 사고는 단순한 정보 수용이나 맹목적인 부정적 태도가 아니라 "목적 지향적이고 자기 조절적인 판단"이다.[17] 이 보고서는 이를 구성하는 핵심 인지 기술로 해석(interpretation), 분석(analysis), 평가(evaluation), 추론(inference), 설명(explanation), 자기조절(self-regulation)의 여섯 가지를 제시하였다. 논증을 강화하거나 약화하는 사례를 찾아 평가하는 것은 단순히 글을 읽는 데 그치지 않는다. 독자는 "추론 능력"을 통해 그 글을 "분석하고 해석"하여 글에 담긴 주장과 숨은 전제를 찾아내고, "자기조절"을 거쳐 논증을 지지하거나 반박하는 사례를 "평가"한 뒤, 최종적으로 그 판단의 근거를 "설명"한다. 즉, 논증의 강화·약화 평가 교육은 학습자가 지문에 담긴 논지를 파악하는 과정 전반에서 이 여섯 가지 인지 능력을 적절히 발휘하는지를 평가하는 엄밀한 방식인 것이다.

또한, 논증의 강화·약화 평가 교육은 논증의 취약 지점을 탐구하는 "메타 수준의 사고(meta-level thinking)"를 요구한다. 메타 수준의 사고란 쉽게 말해 "생각에 대한 생각"으로, 단순히 글의 내용을 이해하는 차원을 넘어 자신의 이해 과정을 스스로 바라보며 논증이 어떠한 구조와 전략으로 작동하는지를 상위 수준에서 평가하는 인지 활동이다. 글에 나타난 저자의 주장은 대개 제시된 전제를 바탕으로 결론을 이끌어내지만, 그 사이에는 필연적으로 글에 명시되지 않은 숨은 가정이나 논리적 비약이 존재한다. 능숙한 독자는 글의 논거를 무비판적으로 수용하지 않고, 그 글이 담고 있는 논증을 파악하여 그것이 올바른지 그렇지 않은지를 항시 검토한다. 이런 점에서 볼 때, 논증을 강화하는 사례를 평가하는 교육은 "어떤 정보가 추가되면 이 논증이 더 설득력을 갖게 되는가?"를 찾는 능력을 기르는 것이며 논증을 약화하는 사례를 평가하는 교육은 "어떤 정보가 제시되면 이 논증의 결론을 받아들이기 어렵게 되는가?"를 찾는 능력을 기르는 교육이라 할 수 있다. 예를 들어, 독자가 글을 읽는 과정에서 다음과 같은 논증을 접했다고 해보자.

전제. 어떤 조건 X가 충족된 후 항상 현상 Y가 발생했고 두 사건 간의 상관관계가 매우 높다.

결론. 따라서 조건 X가 현상 Y를 유발하는 직접적인 원인이다.

독자는 가장 먼저 글에 제시된 정보와 사건의 시간적 선후 관계의 의미를 정확히 파악하는 "해석" 과정을 거친다. 그리고 이 논증이 단순한 관찰 결과로부터 확정적인 인과관계라는 결론으로 도약하고 있음을 "분석"할 것이다. 이때 독자는 글의 내용에 매몰되지 않고 한 걸음 물러나 생각하게 되는 "메타 수준의 사고"를 수행하게 된다. 저자가 제시한 결론을 곧이곧대로 수용하는 대신, 이 논증 자체를 하나의 분석 대상으로 객관화하여 두 사건을 동시에 일으키는 제3의 변수가 개입했을 가능성을 진단하는 것이다. 또한 독자는 X가 항상 현상 Y보다 먼저 발생하는지 여부를 고려하여 이 논증을 강화

17 Peter A. Facione, "Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction," *The Delphi Report*, The California Academic Press, 1990, p.3.

하거나 약화하는 사례를 "평가" 할 수 있다. 만약 독자가 약화하는 사례를 찾고자 한다면, 독자는 논증의 빈틈을 파고들어 전제와 결론 사이의 연결고리를 끊어야 한다. 예컨대 "제3의 변수 Z가 X와 Y를 모두 일으켰을 뿐, X는 Y의 원인이 아니다." 라는 대안적 설명이 제시된다면, 저자의 인과적 주장은 크게 흔들리게 된다. 반대로 강화하는 사례를 찾고자 한다면, 논증의 틈을 메워 결론이 더욱 탄탄해지도록 논증을 보완해야 한다. "다른 외부 요인을 모두 통제한 실험에서도 X가 Y를 유발했다." 는 정보를 추가하여 예외의 여지를 차단해 결론의 설득력을 높일 수 있는 것이다. 이렇게 독자는 그 논증에 대해 자신이 왜 그렇게 생각했는지를 "설명" 하며 그 과정에서 상관관계를 인과관계로 착각하는 등의 오류에 빠지지 않았는지 등을 스스로 점검하는 "자기조절" 과정을 거친다. 이를 통해 독자는 자신의 인지적 편향을 경계하고 사고 과정을 통제함으로써 메타 수준의 사고를 효과적으로 수행할 수 있다.

결국, 논증의 강화·약화 평가 교육은 단순히 지문의 내용이 무엇인지를 파악하는 평면적인 작업이 아니다. 지문에 담긴 논증에 새로운 증거가 주어졌을 때 그 논리 구조가 어떻게 무너지거나 견고해지는지를 입체적으로 시뮬레이션하는 고도의 지적 활동이다. 숨겨진 가정을 찾아내어 공격하거나 방어하는 이러한 메타인지적 평가는 가짜 뉴스와 같은 불확실한 정보가 넘쳐나는 생성형 AI 시대에 올바른 의사결정을 내리기 위한 비판적 사고 능력을 함양하는 데에도 필수적이다.

1.5. 논증의 강화·약화 평가 교육의 실용성

논증의 강화·약화 평가의 교육적 가치는 이론적 논의뿐만 아니라 실무 영역에서도 확인할 수 있다. 대한민국에서 논증의 강화·약화 평가를 체계적으로 활용하는 대표적인 시험으로는 공무원 채용시험에 활용되는 공직적격성평가(PSAT)의 언어논리 영역과 법학전문대학원 입시에 활용되는 법학적성시험(LEET)의 추리논증 영역이 있다. 이 두 시험은 모두 고도의 전문성이 요구되는 인재 선발을 위해, 단순 암기 위주의 평가에서 탈피하여 비판적 사고력과 문제 해결 능력을 검증하는 데 중점을 두어 왔다.

PSAT 언어논리 영역은 5급 및 7급 국가공무원 공개경쟁채용시험의 1차 시험으로 시행되며, 공직 수행에서 요구되는 합리적 판단과 정책적 추론 능력을 평가한다. 언어논리 영역에서 강화·약화 유형은 주어진 지문의 논증 구조를 파악하고, 그 논증을 강화하거나 약화하는 선택지를 고르도록 요구함으로써, 응시자가 단순히 글을 읽고 이해하는 수준을 넘어 논증적 사고를 수행할 수 있는지를 평가한다.

이러한 PSAT 언어논리 영역의 평가 방식은 LEET의 추리논증 영역에서도 일관되게 발견된다. LEET의 추리논증 영역 또한 단순한 지문 독해를 넘어, 법학 교육 및 직무 수행의 기초가 되는 논리적 분석력과 비판적 사고력을 측정하는 데 주안점을 둔다. 특히 해당 영역의 강화·약화 유형은 어떤 정보나 문장이 특정 주장에 미치는 영향을 분석하게 함으로써, 응시자가 논증을 얼마나 정확하게 이해하고 평가할 수 있는지를 검증한다는 점에서 PSAT과 맥을 같이한다.

이 두 시험이 공유하는 핵심 특성은 중요한 교육적 함의를 갖는다. 강화·약화 유형은 단순히 특정 시험을 준비하는 데 유용한 기술적 도구에 그치는 것이 아니라, 공공 의사결정과 법적 판단에서 요구되는 논증적 합리성의 기본 훈련 모형으로 기능할 수 있다. 공무원이 정책을 수립하고 평가할 때, 혹은 법조인이 사건을 분석하고 판단할 때 요구되는 사고의 핵심은 결국 "어떤 근거가 어떤 결론을 얼마나 잘 뒷받침하는가?" 를 판단하는 데 있으며, 이는 곧 강화·약화 유형이 측정하고자 하는 능력과 정확히 일치한다.

1.6. 논증의 강화·약화 평가시 객관식 구조의 방법론적 이점

비판적 사고 교육에서 강화·약화 유형이 지니는 가치는 그 내용적 측면뿐만 아니라 형식적 측면에서도 발견된다. 논증 평가 능력을 측정하는 전통적 방식인 서술형 과제는 응시자의 사고 과정을 상세히 파악할 수 있다는 장점이 있으나 채점 기준이 복잡하고 평가자 간 신뢰도를 확보하기 어렵다는 태생적 한계를 지닌다. 반면 PSAT과 LEET에 등장하는 강화·약화 유형 문항은 고도의 논리적 판단 결과를 '객관식'이라는 정교한 틀 안에 담아냄으로써 서술형이 지닌 평가론적 한계를 극복한다.

구체적으로, 강화·약화 문항의 객관식 구조는 다음과 같이 체계적으로 세분화된다. 먼저 지문에서는 특정 결론을 뒷받침하는 근거들이 논증의 형태로 제시되는데, 이때 논증에는 의도적인 논리적 빈틈이 존재한다. 발문에서는 "다음 중 이 글의 논지를 약화하는 것은?"과 같이, 지문의 논증을 약화하거나 강화하는 정보를 찾으라는 명확한 과제를 부여한다. 그리고 출제 과정에서 논리적 정합성이 검토된 다섯 개의 선택지가 제시된다. 출제자는 단 하나의 선택지를 정답으로 지정하고, 그 외에는 논증과 전혀 무관한 정보, 발문의 요구와 반대되는 효과를 내는 정보, 지엽적이지만 매력적인 정보 등을 오답 선택지로 배치한다. 응시자는 주관적 견해를 서술하는 대신, 철저히 통제된 이 다섯 개의 선택지 중 발문의 요구에 따라 논증의 지지 강도를 실질적으로 변화시키는 단 하나의 정답을 골라내야 한다.

이 점은 최근 활발히 이루어지는 AI 기반 교육 연구, 특히 AI의 논증 평가 성능을 측정하는 데 있어 결정적인 방법론적 이점을 제공한다. 만약 AI에게 서술형 논증 분석을 요구한다면, AI가 산출한 답변을 다시 인간 평가자가 일일이 채점해야 하는 문제가 발생한다. 하지만 강화·약화 문항을 활용할 경우, AI의 비판적 사고 능력을 별도의 인간 채점 없이 선택지 수준에서 즉각적으로 판정할 수 있다. 이는 평가의 객관성을 담보할 뿐만 아니라, 다양한 AI 모델의 논리적 추론 능력을 정량적으로 비교하고 검증하는 데 핵심적인 평가 도구로 활용될 수 있다.

1.7. 논증의 강화·약화 평가 교육에 AI를 활용할 수 있는가?

대규모 언어 모델(Large Language Model, LLM)은 빠른 속도로 발전하고 있으며, 주요 기업들은 고성능 추론 중심 모델을 잇달아 공개하고 있다. 사람들이 가장 많이 사용하는 대규모 언어 모델 기반 상용 AI 서비스로는 OpenAI의 ChatGPT, Anthropic의 Claude, Google의 Gemini 등이 있으며, 이들 기업은 추론 및 작업 수행 능력의 향상을 강조하며 치열하게 경쟁하고 있다. 그러나 이러한 기술적 발전에 대해 교육 연구자가 취해야 할 태도는 무조건적인 기술 낙관이나 기술 비관이 아니라, 명확히 한정된 과업에 대한 엄밀한 성능 검증이다. AI가 "추론을 잘한다"거나 "비판적 사고가 가능하다"는 일반적인 주장은 그 자체로는 교육적 활용 가능성을 담보하지 않는다. 중요한 것은 특정한 교육적 과업에서 AI가 어느 정도의 성능을 보이는지를 구체적으로 측정하고, 그 결과를 바탕으로 활용 방안을 모색하는 것이다.

현재 대부분의 LLM 벤치마크는 수학 문제 풀이(GSM8K, MATH)나 코딩 능력(HumanEval, MBPP) 평가에 집중되어 있으며, 이러한 벤치마크는 형식적·절차적 추론 능력을 주로 측정한다. 반면 강화·약화 문항은 자연과학, 사회과학, 철학 등 다양한 학문 분야의 지문을 읽고 이해하는 능력, 지문에서 논증 구조를 파악하는 능력, 해당 논증의 타당성을 평가하는 능력, 그리고 논증을 강화하거나 약화하는 새로운 정보를 판별하는 능력을 복합적으로 요구한다. 이러한 문항은 비판적 사고의 핵심 구성요소인 해석, 분석, 평가, 추론을 통합적으로 측정하

며, 무엇보다 이러한 평가를 객관식 문항을 통해 정량적으로 수행할 수 있다는 점에서 교육적·실용적·학술적 가치가 있다. 이러한 맥락에서 다음 장에서는 PSAT의 강화·약화 문항을 중심으로 AI의 논증 평가 능력을 실증적으로 검증할 것이다.

2. 상용 AI 프로그램의 논증 평가 능력 검증

2.1. 실험 설계

상용 AI 프로그램의 논증 평가 능력 검증 실험은 다음과 같은 설계에 따라 수행되었다. 첫째, 실험에 사용할 문항으로는 PSAT 언어논리 영역의 강화·약화 문항 20개를 선별하였다. 강화·약화 문항의 정답 판정 기준이 PSAT과 LEET에서 서로 다를 가능성이 있기 때문에 본 연구에서는 PSAT 문항만을 사용하였다. 문항 선별은 최신 기출 문항의 포함과 교육적 대표성의 확보 사이에서 균형을 이루도록 설계하였다. 특정 연도나 특정 주제에 편향되지 않도록 다양한 시기와 주제 영역에서 문항을 추출하였으며, 난이도의 분포 역시 고려하였다.

둘째, 실험 대상 AI 모델은 2025년 12월 현재 사람들이 가장 많이 사용하는 ChatGPT, Claude, Gemini의 각 최신 모델로 하였으며, 구체적으로 ChatGPT 5.2 Pro, Claude Opus 4.5, Gemini 3 Ultra Deep Thinking을 대상으로 하였다.[18] 각 모델에 동일한 20문항을 10회씩 풀이하도록 요청하였다. 동일 문항에 대해 복수 회차의 응답을 수집하는 이유는 모델의 응답이 확률적 특성을 가지기 때문이다. 동일한 문항에 대해서도 모델이 매번 같은 답을 선택하지 않을 수 있으며, 이러한 변동성은 모델의 추론 안정성을 평가하는 중요한 지표가 된다. 이러한 반복 측정 설계는 대부분의 LLM 벤치마크 연구가 1회 측정에 의존하는 것과 달리, 모델 응답의 확률적 특성과 안정성을 평가할 수 있게 한다. 즉, 20문항 × 10회 반복이 100문항 × 1회 측정보다 모델의 추론 능력에 관한 더 풍부한 정보를 제공할 수 있다.

셋째, 실험 환경은 실험 자료의 오염을 방지하기 위해 엄격히 통제하였다. 모든 실험은 브라우저의 시크릿 모드에서 수행하였으며, 각 회차 전에 활동 기록을 삭제하여 이전 응답이 후속 응답에 영향을 미치지 않도록 하였다. 프롬프트는 문항을 제시하고 정답을 선택하도록 요청하는 기본 구조를 따르되, "검색을 하지 말고 스스로 추론하여 답을 고르시오."라는 조항을 포함하여 외부 정보 검색을 통한 정답 회수 가능성을 차단하였다.

넷째, 교육적 활용 가능성의 최소 기준은 매우 엄격하게 설정하였다. AI 모델은 방대한 양의 언어 지문 자료를 사전 학습하는 과정에서 기출 문항이나 그와 유사한 자료를 이미 접했을 가능성이 있다. 또한 일부 모델은 웹 검색 기능을 통해 실시간으로 정보를 검색할 수 있으며, 이 경우 문항의 정답을 외부 자료에서 회수할 가능성을 완전히 배제하기 어렵다. 이러한 현실을 고려할 때, AI가 단순히 "사람보다 조금 낮은" 수준의 정답률을 보이는 것만으로는 교육 자료 제작 보조 도구로서의 실효성을 인정하기 어렵다. 따라서 본 실험에서는 매우 높은 정확도를 요구하는 엄격한 평가 기준을 적용하였다. 다시 말해, 20문항에 대해 해당 모델이 100% 정답률을 보일 경우, 교육적 활용 가치가 있다고 평가할 것이다.

2.2. 실험 결과 및 분석

2.2.1. 모델별 종합 성능 비교

실험 문항은 2019년 ~ 2025년 5급 및 7급 PSAT 언어논리 기출 가운데 선정한 강화·약화 유형 20문항으로 구성되었다. 각 AI 모델은 동일한 20문항에 10회씩 반복

18 연구가 진행된 시점은 2025년 12월 경으로, 해당 기간 최신 모델을 사용하였다. 상용 AI 프로그램은 급속도로 빠르게 발전하는 만큼 더욱 발전된 모델일수록 정답률이 더 높을 수 있으며 각 모델별 차이점도 다를 수 있다.

하여 응답하였으며, 종합 정답률은 ChatGPT 5.2 Pro가 99.0%로 가장 높았고, Claude Opus 4.5가 87.0%, Gemini 3 Ultra가 83.0%로 뒤를 이었다. ChatGPT는 10회 중 8회에서 20문항 전부를 맞추었으며, Claude와 Gemini는 만점 회차가 없었다. 또한 회차별 점수 표준편차는 ChatGPT(0.4) 〈 Gemini(0.5) 〈 Claude(0.8) 순으로 나타나, ChatGPT와 Gemini가 상대적으로 점수 변동이 작았다. (〈표 1〉 참조.)

〈표 1〉 AI 모델별 강화·약화 문항 수행 결과 요약

모델명	유효 회차	평균 정답	정답률	최고점	최저점	표준 편차	만점 회차
ChatGPT 5.2 Pro	10	19.8	99%	20	19	0.4	8
Claude Opus 4.5	10	17.4	87%	18	16	0.8	0
Gemini 3 Ultra	10	16.6	83%	17	16	0.5	0

2.2.2. 체계적 오류 분석

실험에서 가장 주목할 만한 발견은 특정 모델에서 체계적인 오류가 확인되었다는 것이다. Claude의 경우 문항 10(2022년 7급 언어논리 21번, 정답 ④)에서 10회 전부 ⑤를 선택하여 정답률 0%를 기록하였다. 문항 14(2023년 5급 언어논리 38번, 정답 ③)도 10회 중 9회에서 ⑤를 선택하여 정답률 10%로 나타났다. 이는 특정 논증 구조 또는 선택지 해석에서 일관된 오해가 발생했을 가능성을 시사하며, 단일 응답만으로는 탐지하기 어려운 "고착 오류"가 반복 실험을 통해 드러난 사례라 할 수 있다.

Gemini 역시 문항 14에서 10회 연속 ④를 선택하여 정답률 0%를 기록하였고, 문항 18(2025년 5급 언어논리 18번, 정답 ②)에서도 10회 연속 ④를 선택하여 정답률 0%를 기록하였다. 반면 ChatGPT의 오류는 특정 문항에 대한 고착된 오해라기보다, 개별 회차의 일시적 불안정(잡음)으로 해석하는 것이 타당하다.

이 결과는 일관성이 반드시 정확성을 의미하지 않음을 명확히 보여준다. ChatGPT의 높은 일관성은 높은 정답률과 결합되어 신뢰성을 강화하지만, Gemini의 높은 일관성은 특정 오답 선택지로의 고착을 의미할 수도 있다. 따라서 교육·평가 현장에서 AI를 활용할 때에는 단일 응답의 그럴듯함뿐 아니라, 반복 검증을 통한 체계적 오류 탐지와 교차 검증 절차가 병행되어야 한다.

2.2.3. 통계적 유의성

본 실험에서는 20문항이라는 제한된 표본을 사용하였으나, 각 문항에 대해 10회 반복 측정을 수행하여 모델당 총 200개의 관측치를 확보하였다. 모델 간 정답률 차이(ChatGPT 99.0% vs Claude 87.0%)는 이항 검정 결과 통계적으로 유의하였다($p < .001$). 이는 관찰된 성능 차이를 우연만으로 설명하기 어려움을 시사한다.

2.2.4. 회차별 정답 개수 분석

〈표 2〉는 각 모델의 회차별 정답 개수 및 정답률을 요약한 것이다.

〈표 2〉 회차별 정답 개수 및 정답률

회차	ChatGPT 5.2 Pro	Claude Opus 4.5	Gemini 3 Ultra
1차	20 (100.0%)	18 (90.0%)	17 (85.0%)
2차	20 (100.0%)	17 (85.0%)	16 (80.0%)
3차	20 (100.0%)	17 (85.0%)	17 (85.0%)
4차	20 (100.0%)	18 (90.0%)	17 (85.0%)
5차	19 (95.0%)	18 (90.0%)	17 (85.0%)
6차	20 (100.0%)	16 (80.0%)	17 (85.0%)
7차	20 (100.0%)	18 (90.0%)	16 (80.0%)
8차	19 (95.0%)	18 (90.0%)	17 (85.0%)
9차	20 (100.0%)	18 (90.0%)	16 (80.0%)
10차	20 (100.0%)	16 (80.0%)	16 (80.0%)

2.2.5. 모델별 상세 분석

2.2.5-A. ChatGPT 5.2 Pro

ChatGPT는 평균 19.8/20점(99.0%)으로 세 모델 중 최고 성능을 기록했다. 10회 중 8회가 만점(20/20)이었으며, 최저 점수도 19점에 그쳐 성능 하한이 매우 높다. 문항 수준에서는 20문항 중 18문항에서 10회 모두 정답을

기록했고, 오답이 발생한 문항은 2개(문항 10, 14)에 불과했다. 오답 패턴을 보면, 문항 10(2022년 7급 언어논리 21번)은 8차에서만 ②를 선택했으며, 문항 14(2023년 5급 언어논리 38번)도 5차에서만 ④를 선택했다. 즉, ChatGPT의 오류는 특정 문항에 대한 고착된 오해라기보다, 개별 회차의 일시적 불안정(잡음)으로 해석하는 것이 타당하다. 교육적 활용 관점에서 ChatGPT는 높은 정확도와 높은 일관성을 바탕으로 해설 생성 및 정답 검증을 위한 1차 보조 도구로 활용할 수 있다고 평가할 수 있다. 하지만 2장 1절에서 언급한 제약과 교육적 활용을 위해 매우 높은 정확도가 요구된다는 점을 고려하면, ChatGPT 역시 10회 모두 만점을 기록한 것은 아니다. 그러므로 논증의 강화·약화 평가 교육에 사용할 때에는 교육자(인간)의 검증이 반드시 필요하다.

2.2.5-B. Claude Opus 4.5

Claude는 평균 17.4/20점(87.0%)으로 중간 수준의 성능을 보였다. 최고 점수는 18점, 최저 점수는 16점이며, 만점 회차는 없었다. 문항별 분석에서 가장 두드러진 특징은 '체계적 오류(systematic error)'가 확인된다는 점이다. 대표적으로 문항 10(2022년 7급 언어논리 21번)은 10회 전부 ⑤를 선택하여 정답률 0%를 기록했다. 문항 14(2023년 5급 언어논리 38번)도 10회 중 9회에서 ⑤를 선택하여 정답률 10%로 나타났다. 이는 특정 논증 구조 또는 선택지 해석에서 일관된 오해가 발생했을 가능성을 시사하며, 단일 응답만으로는 탐지하기 어려운 고착 오류가 반복 실험을 통해 드러난 사례라 할 수 있다. Claude Opus 4.5는 논증의 강화·약화 문항의 정답 검증 및 해설 생성용으로 활용하기 어려우며, 일관된 오답을 생성한다는 점에서 교육자의 검증이 반드시 필요하다.

2.2.5-C. Gemini 3 Ultra Deep Thinking

Gemini는 10회차 모두 분석하였다. 평균 점수는 16.6/20점(83.0%)이며, 최고 17점, 최저 16점으로 비교적 좁은 범위에서 점수가 변동하였다. 이는 회차별 안정성이 높다는 신호일 수 있으나, 동시에 '일관되게 틀리는' 체계적 오류가 포함될 경우 위험 요인이 될 수 있다. 실제로 Gemini는 문항 14(2023년 5급 언어논리 38번)에서 10회 연속 ④를 선택하여 정답률 0%를 기록했고, 문항 18(2025년 5급 언어논리 18번)에서도 10회 연속 ④를 선택하여 정답률 0%를 기록했다. 또한 문항 9(2022년 7급 언어논리 11번)에서는 10회 중 8회 ⑤를 선택하는 등 특정 오답 선택지로 편향되는 양상이 확인된다. Claude와 마찬가지로, Gemini 3 Ultra Deep Thinking도 논증의 강화·약화 문항의 정답 검증 및 해설 생성용으로 활용하기에 적합하지 않으며, 일관된 오류가 발생한다는 점에서 교육자의 검증이 반드시 필요하다.

2.2.6. 문항별 정답률 분석

〈표 3〉은 오답이 발생한 문항만을 추려 문항별 정답률을 제시한 것이다. 20문항 중 13문항은 세 모델이 모두 100%를 기록했으나, 나머지 7문항에서는 모델별 약점이 뚜렷하게 갈렸다.

〈표 3〉 **오답 발생 문항별 정답률**

문항	원본 문번	정답	ChatGPT	Claude	Gemini
1	2019년 5급 38번	②	100.0%	100.0%	40.0%
7	2022년 5급(나) 17번	②	100.0%	70.0%	100.0%
9	2022년 7급 11번	④	100.0%	90.0%	20.0%
10	2022년 7급 21번	④	90.0%	0.0%	100.0%
14	2023년 5급 38번	③	90.0%	10.0%	0.0%
16	2024년 7급 20번	③	100.0%	80.0%	100.0%
18	2025년 5급 18번	②	100.0%	90.0%	0.0%

모델 간 정답률 차이가 극단적으로 큰 문항도 있었다. 문항 10(Claude 0% vs Gemini 100%), 문항 18(Gemini 0% vs ChatGPT 100%)이 대표적이다. 또한 문항 14는 ChatGPT 90%, Claude 10%, Gemini 0%로 격차가 매우 커, 동일 문항에서도 모델별 해석 경로가 크게 달라질 수 있음을

보여준다.

2.2.7. 논증의 강화·약화 평가 교육에 AI를 어떻게 사용해야 하는가?

본 실험의 결과를 분석해 볼 때, 어떤 방식으로 AI를 활용해야 하는지에 대해 생각해 볼 수 있다. 우선 본 실험에서 수행한 과제는 제시된 문항의 정답을 선택하는 것이었다. 다시 말해, 논증의 강화 및 약화를 교육하기 위한 지문 및 평가 문항을 만드는 데 AI를 활용하는 것이 적합한지는 다른 문제이다. 그러므로 본고에서 논증의 강화·약화 평가 교육에 AI를 활용한다는 것은 "정해진 지문과 문항에 대해 정답을 확인하고 해설서를 만드는 용도"로 활용하는 것으로 한정된다.

실험 결과에 대한 분석을 바탕으로, 강화·약화 문항 기반 교육 및 평가 맥락에서 AI 모델의 교차 사용 전략을 다음과 같이 고려할 수 있다. 첫째, 하나의 모델로 1차 검증 후, 다른 모델로 교차 검증하는 2단계 검증 전략이 필수적이다. ChatGPT를 1차 정답 검증 도구로 활용하고(99.0%의 높은 정확도), Claude와 Gemini를 2차 교차 검증용으로 병행 활용하는 것이 현실적이다. 세 모델이 모두 동의하는 문항(13개 문항)은 고신뢰 문항으로 분류하고, 불일치 문항(7개 문항)은 교육자(인간)가 반드시 검토하도록 한다.

둘째, 모델별 강점을 활용한 역할 분담이 권장된다. ChatGPT는 정답 해설 초안 생성 및 정답 후보 제시에 활용하고, Claude는 대안적 풀이 생성 및 반대 논증 구성에 활용하며, Gemini는 오답 유형 수집 및 함정 선택지 분석에 활용한다. 특히 문항 10(Claude 0% vs Gemini 100%)과 문항 18(Gemini 0% vs Claude 90%)의 사례는 동일 문항에서도 모델별 해석이 극명하게 갈릴 수 있음을 보여주므로, 다양한 관점의 해설을 생성하는 데 유용하다.

셋째, "불일치 문항 중심 토론" 교육 설계가 가능하다. 세 모델의 응답이 불일치하는 문항을 학생에게 제시하고, "왜 각 모델이 다른 선택지를 골랐을까?"를 분석하게 함으로써 비판적 사고 훈련의 직접적 소재로 활용할 수 있다. 이는 AI 리터러시와 논증 분석 능력을 동시에 함양하는 효과적인 교육 방법이 될 수 있다.

3. 결론: 더 나은 비판적 사고 교육을 위한 AI의 활용

본고에서 우리는 AI 시대에 비판적 사고 역량이 필수적이며 이러한 비판적 사고 역량을 함양하기 위해 논증의 강화·약화 평가 교육이 이론적인 측면과 아울러 교육적인 측면에서도 가치가 있음을 논했다. 또한 논증의 강화·약화 평가 교육을 시행하는 데 있어 2025년 12월 현재 상용화되어 있는 최신 생성형 AI 모델들을 활용하는 것이 적절한가에 대해 모델별 비교 실험을 수행하였다. 실험의 결과는 다음과 같은 함의를 지닌다.

첫째, AI 추론 능력의 현황 측면에서 본 실험의 강화·약화 20문항 기준 성능 순위는 ChatGPT(99.0%) 〉 Claude(87.0%) 〉 Gemini(83.0%)이다. ChatGPT는 반복 실험에서도 높은 정확도와 낮은 변동성을 보였으며(만점 8/10), Claude와 Gemini는 특정 문항에서의 고착 오류가 평균 성능을 제한하였다.

둘째, Claude와 Gemini가 각기 다른 문항에서 일관된 오류를 일으켰다는 것은 교육자(인간)의 검증이 필수적임을 분명히 보여준다. Claude의 문항 10(10회 연속 오답), Gemini의 문항 14·18(각 10회 연속 오답)은 '모델이 동일한 오답을 반복적으로 제시할 수 있음'을 보여준다. 또한 교육 현장에서는 계산기 수준의 정확한 검증이 필요한 만큼, 세 모델 가운데 어느 모델도 전체 실험에서 100% 정답률을 기록하지 못했다는 결과는 어느 모델을 사용하더라도 교육자의 검증이 반드시 필요함을 시사한다. 아울러 다중 모델 교차 검증과 같은 절차적 안전장치도 도입되어야 한다. 또한 높은 신뢰도가 요구되는 강

화·약화 문항의 출제 및 교재 제작을 위해서는 추가적인 실험이 병행되어야 할 것이다.

셋째, 교육적 활용 방향은 다음과 같이 정리할 수 있다. ① ChatGPT와 같이 정답률이 매우 높은 모델을 사용한다면 해설 초안 생성과 정답 후보 제시에 활용하되, 최종 해설은 교육자가 논증 구조와 반례 가능성을 재검토하여 확정한다. ② Claude·Gemini와 같은 모델은 비교·대조용(대안 풀이 생성, 반대 논증 구성, 오답 유형 수집)에 유용하나, 체계적 오류 가능성이 있는 문항군을 사전에 식별하고 학생에게 비판적으로 검토하도록 지도한다. ③ 세 모델이 모두 동의하는 문항과 불일치하는 문항을 구분하여, 불일치 문항을 중심으로 "왜 그 선택지가 논증을 강화(혹은 약화)하는가?"를 논증 구조 수준에서 토론하게 함으로써 비판적 사고 교육 목표에 직접 연결할 수 있다.

넷째, 본고에서 제시한 논의와 실험은 AI 리터러시 교육의 관점에서도 중요한 함의를 갖는다. 학생들이 AI의 오류 사례를 직접 분석하고 "왜 AI가 이 문항에서 틀렸는가?"를 탐구하는 과정은 그 자체로 비판적 사고 훈련이 될 수 있다. AI의 체계적 오류 패턴을 분석하는 활동을 통해 학생들은 AI의 한계를 이해하는 동시에, 논증 구조에 대한 깊은 이해를 발전시킬 수 있다. 이는 AI 시대에 요구되는 비판적 AI 활용 능력을 함양하는 데 기여할 수 있다.

다섯째, 본고에서 제시한 논의와 실험은 학술적 차원에서 기존 LLM 벤치마크 연구와 차별화되는 기여를 제공한다. 현재 대부분의 AI 추론 능력 평가는 수학 문제 풀이나 코딩 능력에 집중되어 있으며, 이는 형식적·절차적 추론만을 측정한다. 본고에서 활용한 강화·약화 문항은 다양한 학문 분야의 지문 이해, 논증 구조 파악, 논증 타당성 평가, 논증 강화·약화 판별이라는 복합적 능력을 요구하며, 이는 비판적 사고의 핵심 구성요소를 통합적으로 측정하는 것이다. 또한 본고에서 진행한 실험은 단순히 정답률만을 보고하는 기존 방식과 달리, 반복 측정을 통해 모델 응답의 안정성과 변동성을 함께 분석하였다는 점에서 방법론적 기여가 있다.

마지막으로 본고에서 진행한 실험의 한계와 후속 연구를 정리한다. 본고에서 진행한 실험은 20문항이라는 제한된 표본 크기, 인간 수험생과의 직접 비교 부재, 모델의 추론 과정에 대한 질적 분석의 부재 등의 한계를 지닌다. 후속 연구에서는 표본을 확대하고, 인간 수험생 집단과의 비교를 통해 AI의 논증 평가 능력을 보다 맥락적으로 이해할 필요가 있다. 또한 모델이 오답을 선택하는 경우의 추론 과정을 질적으로 분석하여, 체계적 오류의 원인을 규명하는 후속 연구가 진행될 수 있을 것이다.

참고문헌

1. 단행본 및 논문

박만엽, 「비판적 사고와 논증 분석: '발표와 토론'을 중심으로」, 『사고와표현』 4(1), 한국사고와표현학회, 2011.

박준호, 「논증의 종류와 평가의 기준: 비판적 사고와 비형식 논리학의 의의」, 『범한철학』 42(3), 범한철학회, 2006.

서민규, 「비판적 사고 교육, 무엇을 어떻게 할 것인가」, 『교양교육연구』 4(2), 한국교양교육학회, 2010.

알렉산더 피셔, 최원배 역, 『피셔의 비판적 사고』, 서광사, 2018.

정혜선·박섭형·김여진·한수미·최승락, 『AI시대, 대학교육을 리디자인하라』, 학지사, 2026.

최승락, 「AI에 맡길 것과 맡기지 않을 것: 인지적 외주화 시대의 대학 교육」, 『AI직설』, 박영사, 2026.

Alec Fisher·Michael Scriven, *Critical Thinking: Its Definition and Assessment*, Edgepress, 1997.

Deloitte AI Institute, "The State of Generative AI: Q2 Report," Deloitte, 2024.

Edward M. Glaser, *An Experiment in the Development of Critical Thinking*, Teachers College, Columbia University, 1941.

Hao-Ping (Hank) Lee et al., "The Impact of Generative AI on Critical Thinking," *CHI Conference on Human Factors in Computing Systems, Yokohama*, Japan, 2025.

John Dewey, *How We Think*, D.C. Heath & Co., 1910.

Peter A. Facione, "Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction," *The Delphi Report*, The California Academic Press, 1990.

Richard Paul, *Critical Thinking: What Every Person Needs to Survive in a Rapidly Changing World*, Center for Critical Thinking and Moral Critique, 1990.

Robert H. Ennis, "A Taxonomy of Critical Thinking Dispositions and Abilities," in Joan B. Baron·Robert J. Sternberg (eds.), *Teaching Thinking Skills: Theory and Practice*, New York: W. H. Freeman, 9-26, 1987.

Saadia Zahidi et al., *The Future of Jobs Report 2023*, World Economic Forum, 2023.

Zara Contractor·Germán Reyes, "Generative AI in Higher Education: Evidence from an Elite College," *IZA Discussion Paper* 18055, 2025.

2. 기타 자료

BCG Editorial Team, 「생성형 AI로 가치를 창출하고 또 파괴하는 법」, BCG 블로그, 2023.10.10.

Abstract

Argument-Evaluation Capabilities of Commercial AI

A Focus on Strengthening and Weakening Items in the Public Service Aptitude Test (PSAT)

Critical thinking is widely regarded as a core competency in the age of artificial intelligence, and cultivating it requires the ability to evaluate arguments properly. This study assessed the argument-evaluation capabilities of commercially available AI systems powered by large language models (LLMs) using argument-strengthening and argument-weakening items from Korea's Public Service Aptitude Test (PSAT). First, it discusses the educational and practical value of strengthening and weakening items in critical thinking education. Based on this theoretical discussion, 20 previously administered items were selected from the verbal reasoning section of the PSAT to compare the performance of ChatGPT 5.2 Pro, Claude Opus 4.5, and Gemini 3 Ultra. Each model was presented with the same items ten times. ChatGPT achieved the highest accuracy rate at 99.0%, followed by Claude at 87.0% and Gemini at 83.0%. In particular, repeated testing revealed systematic error patterns specific to each model, as well as persistent errors that would be difficult to detect from a single response alone. Given that existing performance evaluations of LLM-based systems have focused primarily on mathematical and coding abilities, this study highlights the value of strengthening and weakening items as a means of comprehensively assessing the core components of critical thinking. It further argues that review and evaluation by instructors remain essential when AI is used for educational purposes.

Keywords Critical Thinking, Argument Evaluation, Strengthen/Weaken items, Large Language Model(LLM), PSAT(Public Service Aptitude Test)

Lee, Il-Kwon | Ministry of Personnel Management
Choi, Seung-Rak | Hallym University

이 논문은 2026년 7월 20일에 투고 완료되어
2026년 7월 25일부터 8월 15일까지 심사위원이 심사하고
2026년 8월 20일에 심사위원 및 편집위원 회의에서 게재가 결정된 논문임.

[부록] 모델별·회차별 답안 리스트

각 모델(ChatGPT 5.2 Pro, Claude Opus 4.5, Gemini 3 Ultra)의 10회 반복 실험에서 기록된 답안(각 20문항의 선택지 번호)을 수록한다. 본 답안은 수기 기록의 오류 가능성을 사후 검증할 수 있도록 투명성을 확보하기 위해 제공한다.

〈표 4〉 GPT Pro 5.2

문번	원본 문번	정답	1차	2차	3차	4차	5차	6차	7차	8차	9차	10차
1	2019년 5급 언어논리 38번	②	②	②	②	②	②	②	②	②	②	②
2	2019년 5급 언어논리 18번	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
3	2021년 5급 언어논리(가) 38번	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
4	2021년 민경채 언어논리 21번	③	③	③	③	③	③	③	③	③	③	③
5	2021년 5급 언어논리(가) 17번	④	④	④	④	④	④	④	④	④	④	④
6	2022년 5급 언어논리(나) 15번	③	③	③	③	③	③	③	③	③	③	③
7	2022년 5급 언어논리(나) 17번	②	②	②	②	②	②	②	②	②	②	②
8	2022년 5급 언어논리(나) 18번	③	③	③	③	③	③	③	③	③	③	③
9	2022년 7급 언어논리 11번	④	④	④	④	④	④	④	④	④	④	④
10	2022년 7급 언어논리 21번	④	④	④	④	④	④	④	④	②	④	④
11	2022년 7급 언어논리 22번	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
12	2023년 5급 언어논리(가) 18번	④	④	④	④	④	④	④	④	④	④	④
13	2025년 5급 언어논리 32번	①	①	①	①	①	①	①	①	①	①	①
14	2023년 5급 언어논리 38번	③	③	③	③	③	④	③	③	③	③	③
15	2024년 7급 언어논리 17번	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
16	2024년 7급 언어논리 20번	③	③	③	③	③	③	③	③	③	③	③
17	2025년 7급 언어논리 15번	④	④	④	④	④	④	④	④	④	④	④
18	2025년 5급 언어논리 18번	②	②	②	②	②	②	②	②	②	②	②
19	2025년 5급 언어논리 35번	②	②	②	②	②	②	②	②	②	②	②
20	2025년 5급 언어논리 38번	③	③	③	③	③	③	③	③	③	③	③

〈표 5〉 Gemini 3 Ultra Deep Thinking

문번	원본 문번	정답	1차	2차	3차	4차	5차	6차	7차	8차	9차	10차
1	2019년 5급 언어논리 38번	②	①	①	②	②	②	②	①	①	①	①
2	2019년 5급 언어논리 18번	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
3	2021년 5급 언어논리(가) 38번	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
4	2021년 민경채 언어논리 21번	③	③	③	③	③	③	③	③	③	③	③
5	2021년 5급 언어논리(가) 17번	④	④	④	④	④	④	④	④	④	④	④
6	2022년 5급 언어논리(나) 15번	③	③	③	③	③	③	③	③	③	③	③
7	2022년 5급 언어논리(나) 17번	②	②	②	②	②	②	②	②	②	②	②
8	2022년 5급 언어논리(나) 18번	③	③	③	③	③	③	③	③	③	③	③
9	2022년 7급 언어논리 11번	④	④	⑤	⑤	⑤	⑤	⑤	⑤	④	⑤	⑤
10	2022년 7급 언어논리 21번	④	④	④	④	④	④	④	④	④	④	④
11	2022년 7급 언어논리 22번	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
12	2023년 5급 언어논리(가) 18번	④	④	④	④	④	④	④	④	④	④	④
13	2025년 5급 언어논리 32번	①	①	①	①	①	①	①	①	①	①	①
14	2023년 5급 언어논리 38번	③	④	④	④	④	④	④	④	④	④	④
15	2024년 7급 언어논리 17번	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
16	2024년 7급 언어논리 20번	③	③	③	③	③	③	③	③	③	③	③
17	2025년 7급 언어논리 15번	④	④	④	④	④	④	④	④	④	④	④
18	2025년 5급 언어논리 18번	②	④	④	④	④	④	④	④	④	④	④
19	2025년 5급 언어논리 35번	②	②	②	②	②	②	②	②	②	②	②
20	2025년 5급 언어논리 38번	③	③	③	③	③	③	③	③	③	③	③

〈표 6〉 Claude Opus 4.5

문번	원본 문번	정답	1차	2차	3차	4차	5차	6차	7차	8차	9차	10차
1	2019년 5급 언어논리 38번	②	②	②	②	②	②	②	②	②	②	②
2	2019년 5급 언어논리 18번	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
3	2021년 5급 언어논리(가) 38번	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
4	2021년 민경채 언어논리 21번	③	③	③	③	③	③	③	③	③	③	③
5	2021년 5급 언어논리(가) 17번	④	④	④	④	④	④	④	④	④	④	④
6	2022년 5급 언어논리(나) 15번	③	③	③	③	③	③	③	③	③	③	③
7	2022년 5급 언어논리(나) 17번	②	②	②	①	②	①	①	②	②	②	②
8	2022년 5급 언어논리(나) 18번	③	③	③	③	③	③	③	③	③	③	③
9	2022년 7급 언어논리 11번	④	④	④	④	④	④	④	④	④	④	⑤
10	2022년 7급 언어논리 21번	④	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
11	2022년 7급 언어논리 22번	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
12	2023년 5급 언어논리(가) 18번	④	④	④	④	④	④	④	④	④	④	④
13	2025년 5급 언어논리 32번	①	①	①	①	①	①	①	①	①	①	①
14	2023년 5급 언어논리 38번	③	⑤	⑤	⑤	⑤	③	⑤	⑤	⑤	⑤	⑤
15	2024년 7급 언어논리 17번	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤	⑤
16	2024년 7급 언어논리 20번	③	③	③	③	③	③	①	③	③	③	①
17	2025년 7급 언어논리 15번	④	④	④	④	④	④	④	④	④	④	④
18	2025년 5급 언어논리 18번	②	②	④	②	②	②	②	②	②	②	②
19	2025년 5급 언어논리 35번	②	②	②	②	②	②	②	②	②	②	②
20	2025년 5급 언어논리 38번	③	③	③	③	③	③	③	③	③	③	③