

2. “비판적으로 보라”는 왜 안 먹히는가

연재 <에이전틱 사고 · 온톨로지> 제2회 — 1부 · 왜 이 문제인가

빈칸으로 돌아온 활동지

활동지에 이런 칸을 만들어 둔 적이 있습니다. ‘AI의 답을 비판적으로 검토하고, 문제점을 적어 보시오.’ 제 스스로 꽤 괜찮은 설계라고 생각했습니다.

학생들의 답안을 건어 보니 세 종류의 답이 나왔습니다. 빈칸이 가장 많았고, 그다음에 ‘자신이 생각하기에 특별한 문제점이 없다’였습니다. 나머지는 ‘AI가 제공한 설명이 조금 어렵다’, ‘예시가 부족하거나 수준에 맞지 않는다’와 같은 말이었습니다. **학습 내용에 대한 평가가 아니라 AI가 얼마나 친절한 지에 대한 평가였습니다.** 학생들이 제 지시를 안 따른 게 아닙니다. **어쩌면 저는 학생들에게 할 수 없는 일을 시켰던 것일지도 모릅니다.** 지난 회에서 접었다고 말씀드린 진단이 여기서 한 번 더 무너집니다.

지침들을 나란히 놓고 보면

그동안 나온 AI 활용 지침들을 모아서 순서대로 늘어놓아 본 적이 있습니다.

프롬프트를 구체적으로 쓸 것. 답변을 그대로 믿지 말 것. 환각을 의심할 것. 출처를 확인할 것. 여러 모델에 물어 비교할 것. 최종 판단은 사람이 할 것.

하나하나 다 옳은 말이라고 생각합니다. 저도 학생들에게 늘상 이 말을 했습니다. 그런데 늘어놓고 보니 공통점이 하나 보였습니다. 첫 항목을 빼면 **전부 AI가 답을 내놓은 다음에 작동합니다.** 즉 사람이 아주 수동적으로 물리게 되는 상황입니다. 의심하라는 것도, 확인하라는 것도, 비교하라는 것도 AI가 만들어낸 결과물이 화면에 떠 있어야 성립하는 말입니다. 지침 전체가 사후에 몰려 있습니다.

이게 왜 문제인지는 한 문장으로 정리됩니다. 검토는 검토할 구조가 이미 있을 때만 가능합니다. 비판할 자리가 마련되어 있지 않은데 비판하라고 하면, 학생이 할 수 있는 일은 표면을 보고 인상을 적는 것뿐입니다. 제가 걷은 활동지가 정확히 그랬습니다.

AI 리터러시 논의는 이 문제를 알고 있습니다

공정하게 말하면, 이 분야의 연구들이 앞서 말한 상황의 문제점을 모르고 있는 것은 아닙니다. Long과 Magerko(2020)는 흩어져 있던 논의를 모아 AI 리터러시를 'AI 기술을 비판적으로 평가하고, AI와 효과적으로 소통하고 협업하며, 온라인과 가정과 직장에서 AI를 도구로 사용할 수 있게 하는 역량의 집합'으로 정의하고 열일곱 가지 역량을 제시했습니다. Ng과 동료들(2021)은 이를 알기·활용하기·평가하기·윤리로 다시 묶었습니다. 유네스코(2023)의 생성형 AI 지침도 큰 틀에서 비슷한 구조를 갖습니다. 다만 UNESCO는 윤리와 인간성을 더욱 강조합니다.

문제는 AI리터러시의 목록 자체가 아닙니다. 목록에 들어 있는 '평가하기'가 교실에서 실제로 발동하려면 무엇이 먼저 있어야 하는지가 잘 다뤄지지 않았다는 점입니다. 그러니까 현재 우리가 사용하는 AI역량은 마치 퀘스트를 깨는 것과 같은 과업 목록 리스트로만 되어 있습니다. 즉 역량 목록은 무엇을 할 수 있어야 하는지를 알려 주지만, 그 능력이 어떤 조건에서 커지는지는 알려 주지 않습니다.

비판적 사고는 옮겨 다니는 기술이 아닙니다

여기서 인지과학 쪽의 오래된 결론 하나를 가져와야겠습니다. Willingham(2007)은 비판적 사고를 자전거 타기 같은 기술로 보는 통념을 정면으로 반박했습니다. 자전거는 한번 배우면 어느 자전거든 탈 수 있지만, 비판적 사고는 그렇게 작동하지 않는다는 것입니다. 분야의 전환에 따른 전이도 잘 되지 않으며 한 번 비판적 사고를 했다고 해서 이것이 지속되지도 않습니다. 그는 비판적 사고가 영역 지식과 분리된 별개의 기술이 아니라 그 영역의 지식과 얽혀 있으며, 배경 지식이 없는 영역에서는 같은 사람이 같은 수준으로 사고하지 못한다고 정리했습니다.

이 결론을 우리 상황에 옮기면 이렇게 됩니다. 학생이 이차함수 문제에 대한 AI의 풀이를 비판적으로 검토하지 못한 것은 비판적 사고력이 부족해서가 아닙니다. 이차함수에 대해 대조할 만한 것을 명시적으로 갖고 있지 않아서입니다. 비판적 사고를 따로 훈련시키는 방식으로는 이 자리를 메울 수 없습니다. 지난 회에서 본 Lee와 동료들(2025)의 결과가 여기에 정확히 겹칩니다. 지식노동자 319명을 조사했더니, 비판적 사고를 실제로 발동하는지를 예측한 것은 과제에 대한 자기 확신이었습니다. 아는 것이 있다고 느낄 때 따져 보고, AI를 더 믿을수록 덜 따져 봅니다.

게다가 사후 점검은 처음부터 불리한 싸움입니다

설령 학생에게 검토할 구조가 어느 정도 있다 해도, 답이 먼저 나온 뒤에 검토하는 순서 자체가 불리합니다. 인간-자동화 연구가 오래전부터 이 점을 보고했습니다. Parasuraman과 Riley(1997)는 사람이 자동화를 대하는 방식을 사용·오용·불용·남용으로 나누어 정리하면서, 자동화를 지나치게 신뢰해 감시를 소홀히 하는 오용이 흔하게 나타난다고 했습니다. Skitka와 동료들(1999)은 이를 자동화 편향이라는 이름으로 실험했습니다. **자동화된 보조 장치가 있는 조건에서 사람들은 장치가 알려 주지 않은 사건을 놓치는 누락 오류와, 장치가 잘못 지시한 것을 그대로 따르는 실행 오류를 함께 더 많이 냈습니다.**

조종석에서 나온 이야기지만 교실과 멀지 않습니다. 화면에 그럴듯한 답이 이미 떠 있는 상태에서 '한번 의심해 보라'고 하는 것은, 사람이 가장 약해지는 조건에서 가장 어려운 일을 요구하는 셈입니다.

그러면 탐지로 우회하면 되지 않나

이 대목에서 자연스럽게 나오는 생각이 있습니다. 학생이 검토를 못 한다면, 교사가 AI를 썼는지 아닌지라도 가려내면 되지 않겠는가. 국어 교사 박준일 선생님이 이 길을 상세히 검토한 뒤 접었습니다. 박준일(2026)은 탐지에 세 가지 문제가 있다고 정리합니다. 검사하는 AI와 피하는 AI의 끝없는 술래잡기, 탐지기 자체의 낮은 성능, 그리고 위축 효과입니다.

두 번째 항목의 근거가 되는 연구를 직접 확인해 보았습니다. Perkins와 동료들(2024)은 주요 탐지 도구들을 실측했는데, AI가 쓴 글을 AI라고 맞힌 비율이 평균 39.5퍼센트였고, 철자를 흘뜨리거나 문장 길이를 바꾸는 간단한 조작을 가하자 17.4퍼센트까지 떨어졌습니다. 저자들의 결론은 이 도구들을 학업윤리 위반 판정 목적으로 쓰는 것을 권장할 수 없다는 것이었습니다.

세 번째 항목이 교사에게는 더 아픕니다. **탐지기는 반듯한 글을 의심합니다. 문법이 정확하고 구조가 잘 잡힌 글, 그러니까 성실하게 배워서 잘 쓰게 된 학생의 글이 가장 먼저 의심을 받습니다.** 그런데 이런 글이 사람이 쓸 수도 있습니다. 저도 제가 쓴 글을 AI가 썼다고 의심 받았던 기억이 납니다. 워싱턴주 교육청(2024)의 K-12 지침도 탐지 소프트웨어가 현재 신뢰할 만하지 않으며 표절이나 부정행위 판정의 유일한 근거로 써서는 안 된다고 못 박았습니다.

박준일(2026)의 표현을 그대로 옮기면, 검사하는 AI와 피하는 AI의 군비 경쟁에 교사가 심판으로 뛰어드는 순간 교실은 수업이 아니라 수사가 됩니다. 저는 이 문장에 동의합니다. 다만 제가 관심을 두는 지점은 조금 다른 쪽입니다. 탐지도 사후에 작동하는 장치라는 점입니다. 결국 같은 자리에 걸려 있습니다.

검증을 잘하는 사람들은 무엇을 하는가

여기서 방향을 한 번 틀어 보겠습니다. 검토가 안 되는 이유만 늘어놓았으니, 실제로 검토를 잘하는 사람들이 무엇을 하는지 보는 편이 낫겠습니다. Wineburg와 McGrew(2019)의 연구가 이 물음을 정면으로 다룹니다. 이들은 인터넷을 오래 써 온 세 집단에게 온라인 정보의 신뢰성을 평가하게 했습니다. 역사학 박사 10명, 전문 팩트체커 10명, 스탠퍼드 학부생 25명이었습니다.

결과가 뜻밖입니다. 역사학자와 학생들은 웹사이트 안에 머무르면서 아래로 읽어 내려갔습니다. 로고가 공식적으로 보이는지, 도메인 이름이 그럴듯한지, 참고문헌이 달려 있는지를 보았습니다. 그리고 자주 속았습니다. 이런 표식은 만들어 내기 쉽기 때문입니다.

팩트체커들은 달랐습니다. 사이트를 훑고 나서 곧바로 그 사이트를 떠났습니다. 새 탭을 열어 '이 단체가 누구인가'를 밖에서 찾았습니다. 저자들은 이것을 옆으로 읽기라고 불렀습니다. 팩트체커들은 훨씬 적게 읽고 훨씬 빨리, 그리고 더 타당한 판단에 도달했습니다.

이 연구에서 제가 주목한 것은 역사학자들이 실패했다는 대목입니다. 이들은 사료 비판 훈련을 받은 사람들이고, 텍스트를 의심하는 일에서는 전문가입니다. 그런데도 자기 영역 밖의 디지털 정보 앞에서는 학부생과 비슷하게 행동했습니다. Willingham(2007)의 결론이 여기서 다시 확인됩니다. 비판적으로 읽는 능력은 훈련받은 영역 안에서 작동합니다. 그리고 팩트체커들이 잘한 이유도 일반적인 회의심이 아니었습니다. 그들에게는 '이런 이름의 단체는 대개 어디에 속해 있다'는 영역 지식과, 무엇을 어디서 확인할지에 대한 구체적인 절차가 있었습니다.

교실에 옮기면 두 가지가 따라옵니다. 하나, 검토는 대상 안에서 하는 것이 아니라 밖에서 대조하는 것입니다. AI의 답을 아무리 뜯어보아도 답이 나오지 않습니다. **이미 AI에 의해 생성된 답은 검토의 대상이 되지 못합니다.** 둘, 대조하려면 대조할 것이 있어야 합니다. 팩트체커에게는 그것이 축적된 영역 지식이었고, 학생에게는 그것이 아직 없습니다.

그래서 시점을 앞으로 당깁니다

지금까지 나온 것을 모아 보면 이렇습니다. 사후 검토는 검토할 구조가 없으면 작동하지 않고, 검토할 구조는 영역 지식과 분리되지 않으며, 답이 먼저 나온 상태는 사람이 가장 흔들리기 쉬운 조건입니다. 그리고 사후에 붙이는 장치는 검토든 탐지든 같은 한계를 공유합니다. 그렇다면 남는 길은 하나입니다. **문제는 답이 나온 뒤가 아니라 답이 나오기 전에 발생합니다.**

한 가지는 분명히 해 두겠습니다. 앞에 나온 지침들이 틀렸다는 말이 아닙니다. 출처를 확인하고 환각을 의심하는 일은 여전히 해야 합니다. 다만 그 지침들이 작동하려면 앞쪽에

무언가가 놓여 있어야 하는데, 그 앞쪽을 우리가 거의 다루지 않았다는 것입니다. 이 연재는 지침을 대체하려는 게 아니라 지침 앞에 자리를 하나 만들려는 시도입니다.

그러면 그 앞쪽에 무엇을 두어야 하는가. 저는 이것을 “학생이 무엇을 명시적으로 갖고 있어야 하는가”라는 물음으로 바꿔 놓고 붙들었습니다. 그리고 답을 찾다가 뜻밖의 분야에서 개념을 하나 빌려 오게 되었습니다. 다음 회에서는 그 개념, 곧 온톨로지라는 말을 왜 굳이 빌려 오는지를 말씀드리겠습니다. 미리 한 가지만 적어 두면, 철학의 존재론도 아니고 정보학의 형식적 명세도 그대로 쓰지 않습니다. 그렇게 쓰면 곧바로 반박당한다는 것을 알고 있고, 그 반박도 하겠습니다.

참고 문헌

- Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.
- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16.
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253.
- Perkins, M., Roe, J., Vu, B. H., 외 (2024). GenAI detection tools, adversarial techniques and implications for inclusivity in higher education. *International Journal for Educational Integrity*. (arXiv:2403.19148)
- Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5), 991–1006.
- Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. Paris: UNESCO.
- Washington Office of Superintendent of Public Instruction (2024). *Human-centered AI guidance for K–12 public schools*.
- Willingham, D. T. (2007). Critical thinking: Why is it so hard to teach? *American Educator*, 31(2), 8–19.
- Wineburg, S., & McGrew, S. (2019). Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record*, 121(11), 1–40.
- 박준일 (2026). 「AI 탐지에서 삶과 삶의 연결로: 학생의 현명한 AI 활용을 이끄는 국어 수업 설계」, 미간행 원고.

※ 박준일(2026)은 조작 후 탐지 정확도를 22.2퍼센트로 적었습니다. 본문에 쓴 17.4퍼센트는 학술지에 실린 최종본의 수치입니다. 사전 공개본과 최종본 사이에 차이가 있는 것으로 보이며, 인용하실 때는 최종본 쪽을 권합니다.