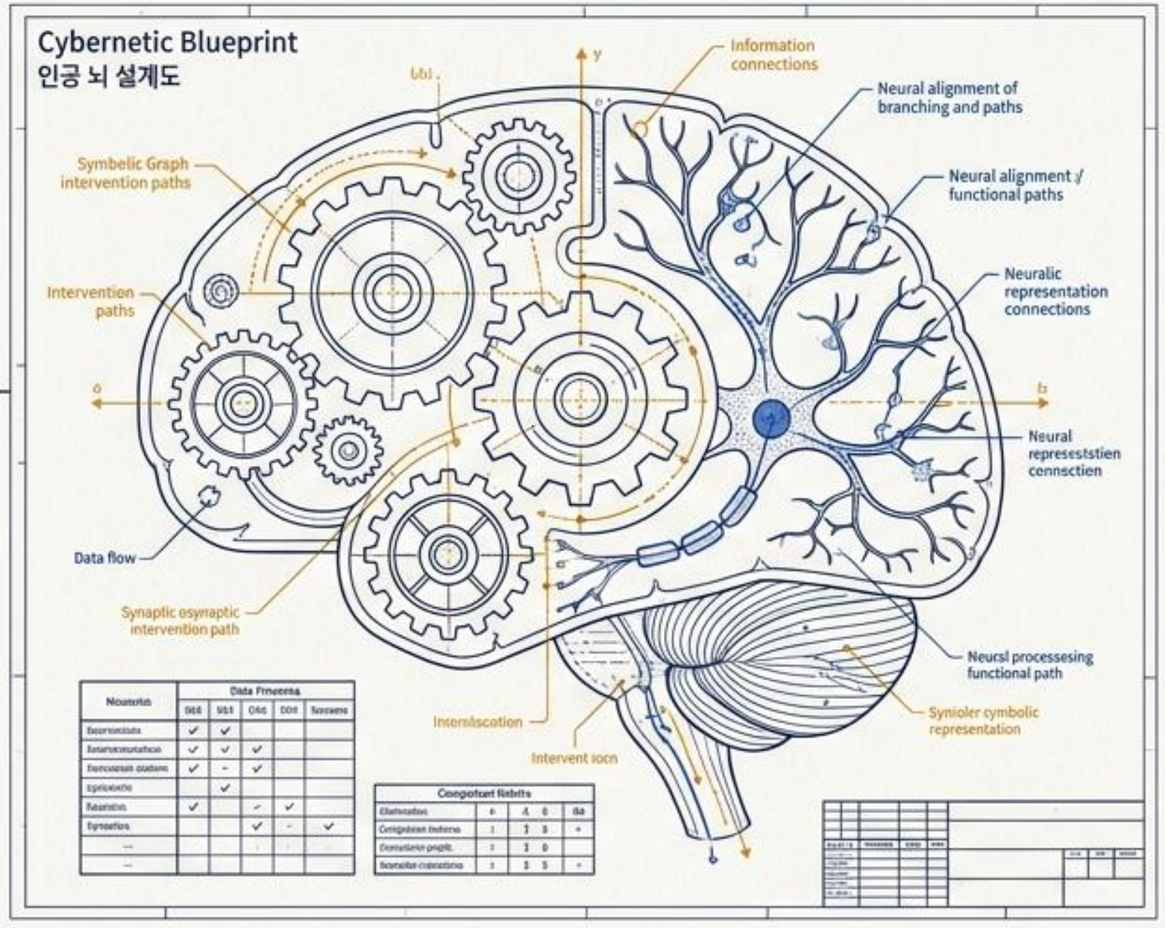
가르치는 기계의 뚜껑을 열다: 교육의 '바깥'을 넘어 '내부 구조' 해부

Part I. 문제 설정과 패러다임 비판



- 교육제도와 조직이라는 외부적 논의를 넘어, 가르치는 기계(AI 튜터)의 내부 구조를 직접 조명
- 거대언어모델(LLM)과 정오답 스칼라 점수라는 껍데기를 파괴하는 새로운 접근 필요
- 뉴로심볼릭(Neuro-Symbolic) 시뮬레이션 학습자 모델의 설계와 검증 과제 제시

시모어 페퍼트의 질문:
'아이가 컴퓨터를 프로그래밍하는가, 컴퓨터가 아이를 프로그래밍하는가?'

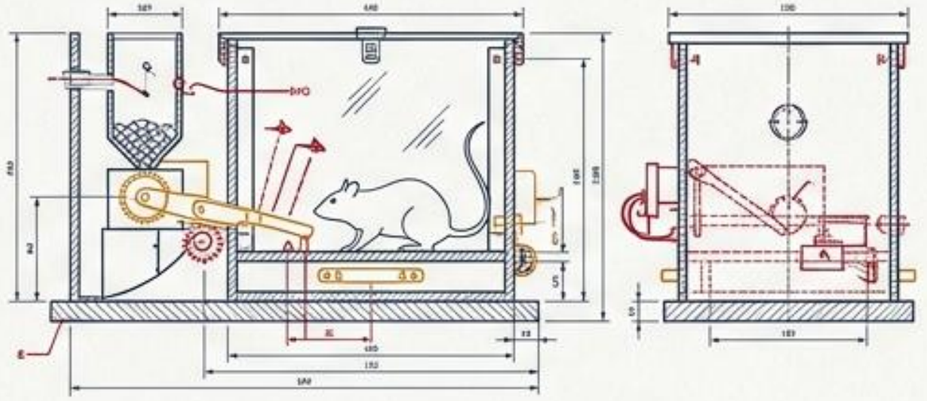
발표자: 이수화 (서울대학교 빅데이터혁신융합대학 연구교수 / 법무법인 디엘지 AI센터장)

Gemini Notebook

지금의 LLM 에듀테크는 1953년 스키너 티칭머신의 '화려한 귀환'일 뿐입니다

Part I. 문제 설정과 패러다임 비판 - 현상 비판

1953 기계식 스키너 박스 도면



=

2026 스마트폰 LLM 체팅 UI



[교안/콘텐츠] → [API 래퍼] → 
[하류에 고정된 학생(수신/채점/감시 대상)]

- **조달 구조가 낳은 API 래퍼 마켓플레이스:** 교안 요약, 퀴즈, 감시 대시보드 위주의 '솔루션 백화점' 한계 (국내 한 대학의 18개 시스템 사례)
- **객체로 전락한 학생:** 콘텐츠를 하류(Downstream)에서 수신하고 채점당할 뿐, AI를 도구로 쥐는 주체(Agent)가 아님
- **'지능적 주입식'의 함정:** 소크라테스식 비계(Scaffolding) 엔진조차 결국 정해진 답으로 학생을 몰아가는 교수기계의 완성형

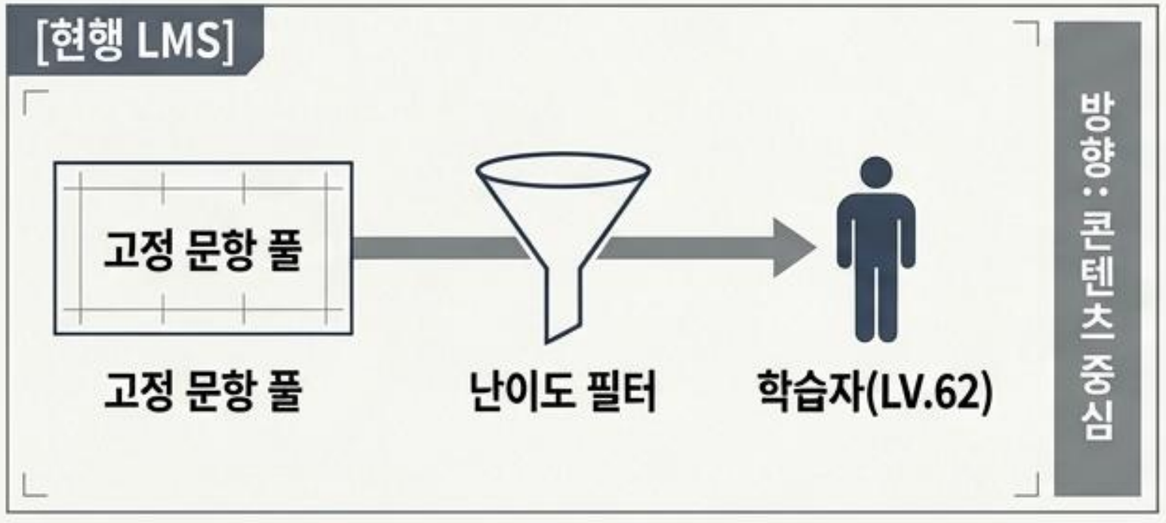
우리는 화려한 자연어의 옷을 입은 '행동주의 2.0'을 목격하고 있습니다.

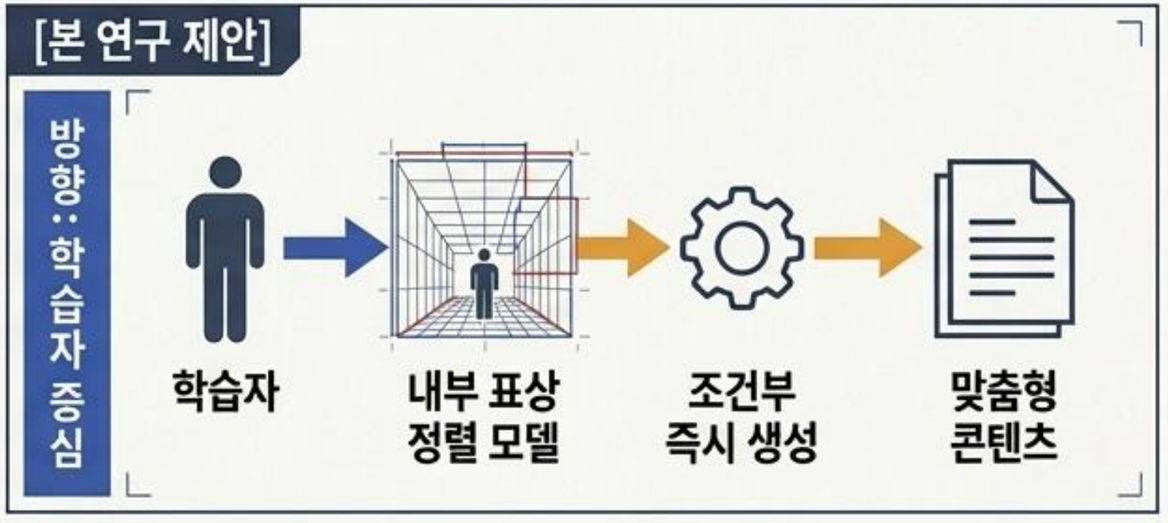
Gemini Notebook

‘여행가방 단어’ 위에 수학을 얹은 심리측정학의 환상과 2대 결손

Part I. 문제 설정과 패러다임 비판 – 존재론적 진단

- **심리측정학의 존재론적 오류:** ‘이해력’, ‘추론력’ 같은 여행가방 단어 내부를 해부하지 않고 단일 잠재변수(θ) 취급
- **다중실현성과 비식별성:** 동일한 ‘62점’이라도 원리를 파악한 상태와 힌트 의존적 상태는 전혀 다른 실재임
- **빠진 생성력 (콘텐츠 중심):** 고정된 문항 풀에 난이도 필터만 씌워 학습자가 콘텐츠에 맞춰지는 구조
- **빠진 내부 표상 (스칼라의 한계):** ‘틀렸다’는 알지만 ‘왜’ 틀렸는지 구조적 맵핑이 없는 단일 레벨업 게임





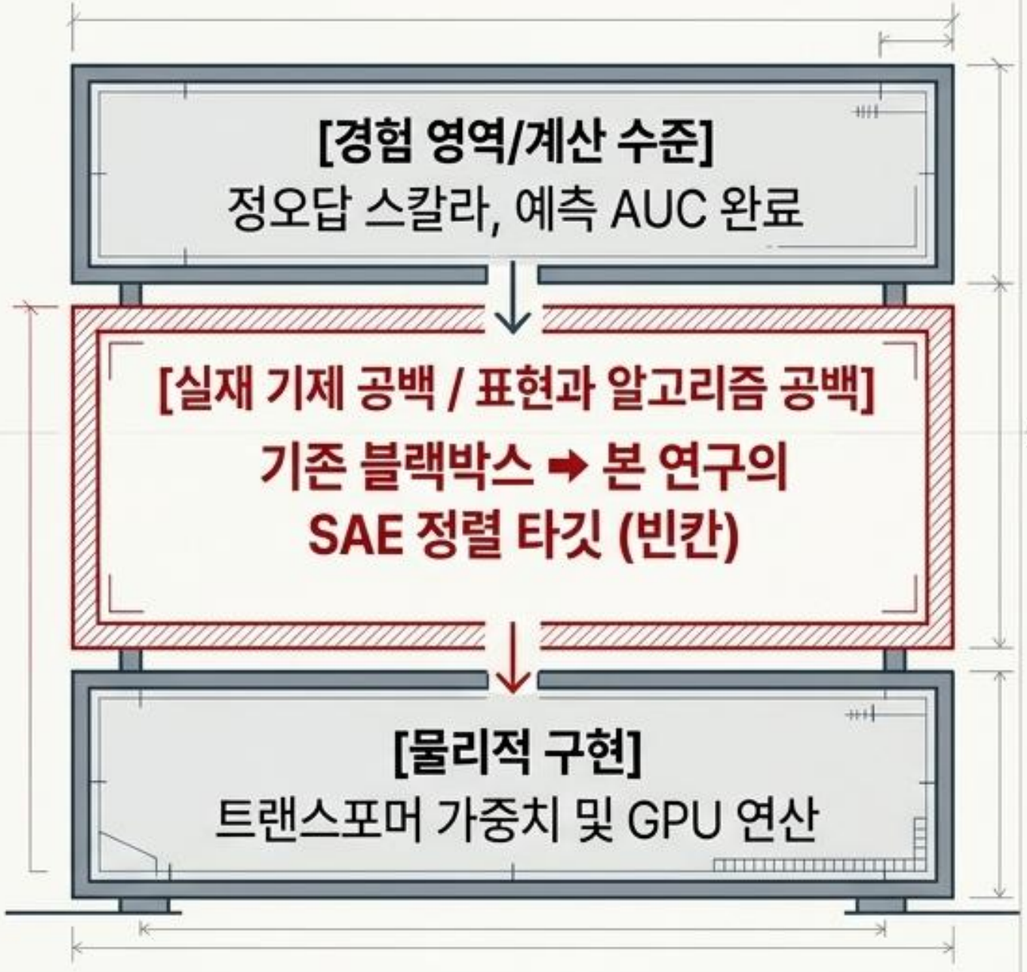
화살표의 방향을 ‘학습자 → 콘텐츠’로 역전시킬 뉴로심볼릭 2대 축이 필요합니다.

Gemini Notebook

예측(AUC 0.85)은 충분하다 — 비어 있는 것은 '실재 기제'와 '개입'이다

Part I. 문제 설정과 패러다임 비판 - 이론적 빈칸

- **현행 지식추적(LKT, 2024)의 한계:** '다음 문제를 틀릴 확률'은 잘 맞추지만, 해석(Why)과 개입(How)이 불가능한 블랙박스
- **로이 바스카(Bhaskar) 3층 존재론:** 경험(정오답)과 현실(풀이) 아래의 핵심인 '실재 기제(Real Mechanism)' 칸이 공백
- **데이비드 마(Marr) 3수준:** 계산(목적)과 물리(가중치)를 매개할 '표현과 알고리즘' 칸이 공백
- **희소 오토인코더(SAE)의 역할:** 이 비어있는 기제와 표현 층위를 채우는 기계론적 현미경



AI 교육에서 인지과학과 교육공학이 사수해야 할 영토는 바로 이 2층 '기제' 영역입니다.

Gemini Notebook

【축 1】 왜 하필 교육과정인가 — 정답지가 '법적'으로 존재하는 유일한 폐쇄계

Part II. 심볼 축(Axis 1) - 도메인 존재론

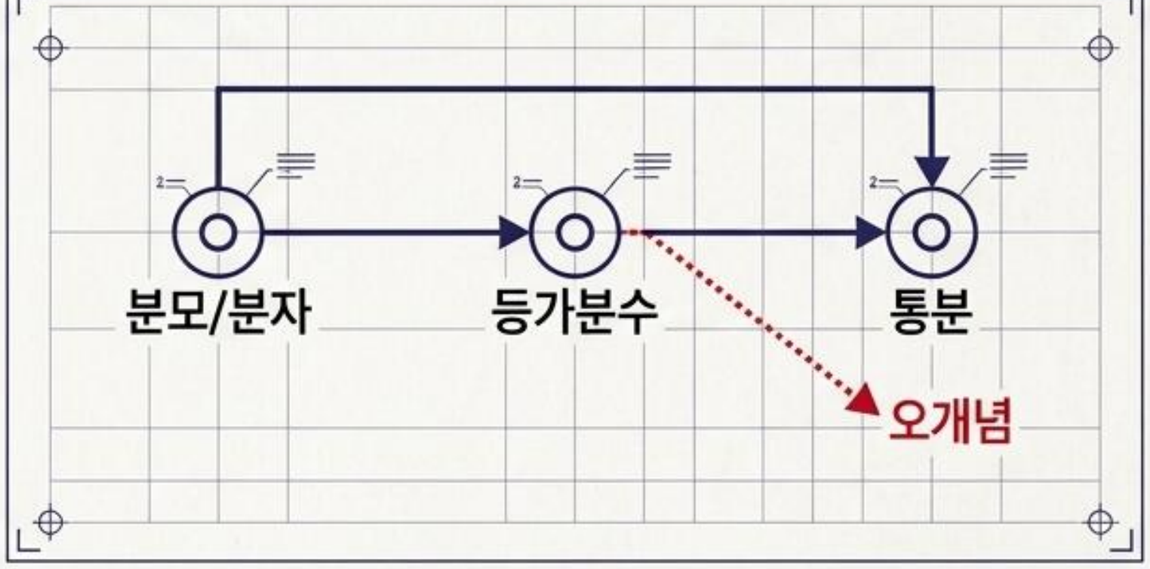
- **열린계(Open System)의 실패:** 일반 상식이나 의료 온톨로지는 '세계가 진짜 그러한가?' 논쟁에 빠져 순환 오류 발생
- **닫힌계(Closed System)의 강점:** 2022 개정 교육과정 개념 위계는 발견된 진리가 아니라 법적·규범적으로 고시된 문서
- **정답의 존재:** 실제 전쟁이 아닌 '군사교육'처럼, 정답지가 규정(Stipulate)되어 있어 온톨로지 순환 고리를 절단 가능
- **연구적 통제 필수성:** 모델은 이미 교과서를 읽었으므로, 단순 암기(Corpus Echo)가 아님을 증명할 통제 설계 필수

[열린계]



정답지 없는 상식

[닫힌계]



⚠ 주의: LLM은 이미 이 교과서를 읽었다 — Corpus Echo 통제가 본 연구의 성패를 가릅니다.

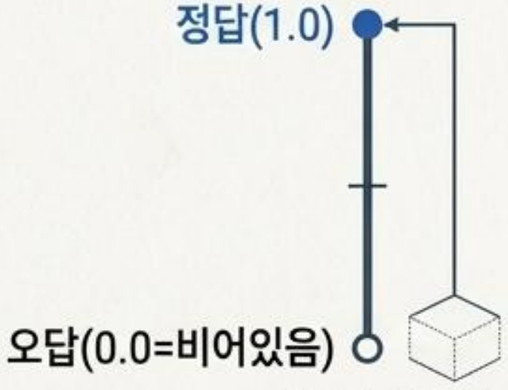
Gemini Notebook

오개념은 지식의 '결여(0)'가 아니라 '기하학적으로 잘못 위치한 실재'다

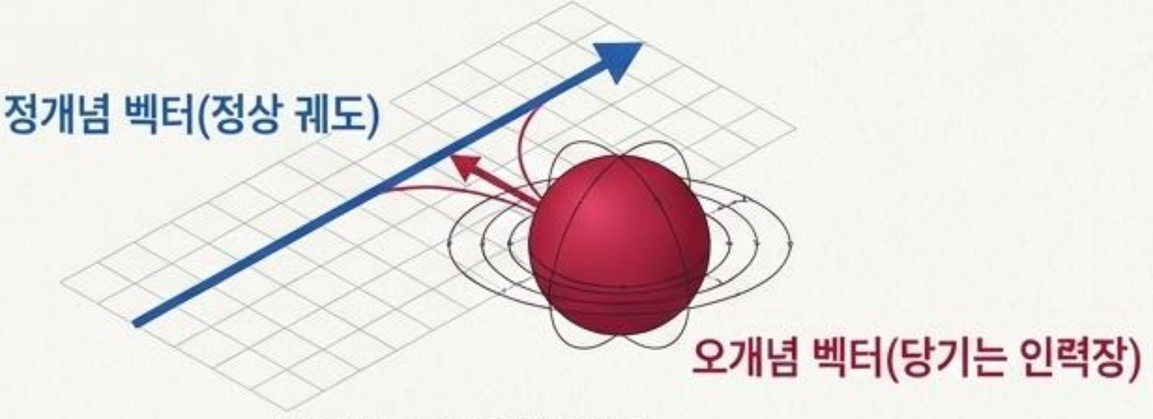
Part II. 심볼 축(Axis 1) - 뉴로심볼릭의 원형

정상 68·오류 150 병렬 음소 검출 → 조음장애 지식그래프 → 2026 LLM-SAE 뉴로심볼릭 학습자 모델

- **선행연구 교훈**(하이브리드 AI 3부작): 정상 음소 68개와 모국어 간섭 오류 음소 150개를 병렬로 동시 구축한 경험
- **판정 유보와 사고 유도:** 확신이 낮을 때 정답 주입을 유보 하고 **오류 위치만 마킹**해 자기교정 유도 (페퍼트 역전 방지)
- **정상-오류 병렬 모델링:** 순수 신경망의 상관관계 학습은 노이즈 앞에서 무너지며, 부정적 증거(Negative Evidence)의 명시적 구조화 필요
- **핵심 공리:** '정개념 + 오개념' 병렬 구조화만이 강건성, 심볼 그라운딩, 투명성을 보장



[기존 스칼라 관점]



[병렬 뉴로심볼릭 관점]

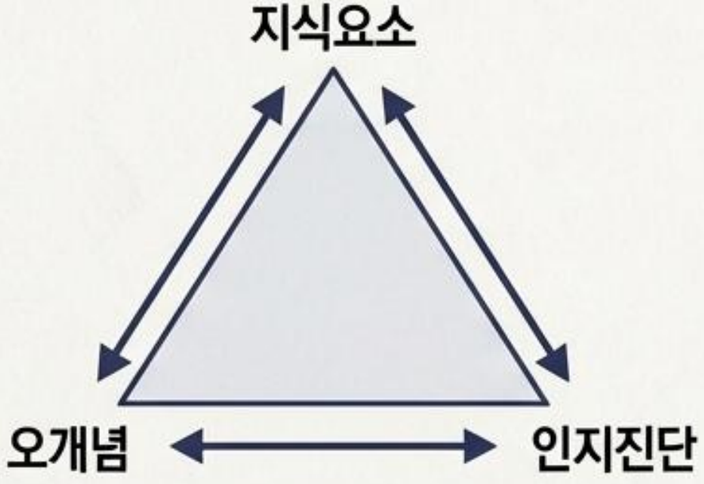
오답은 비어 있는 공간이 아니라, 잘못된 규칙이 능동적으로 작동하는 실재(Entity)입니다.

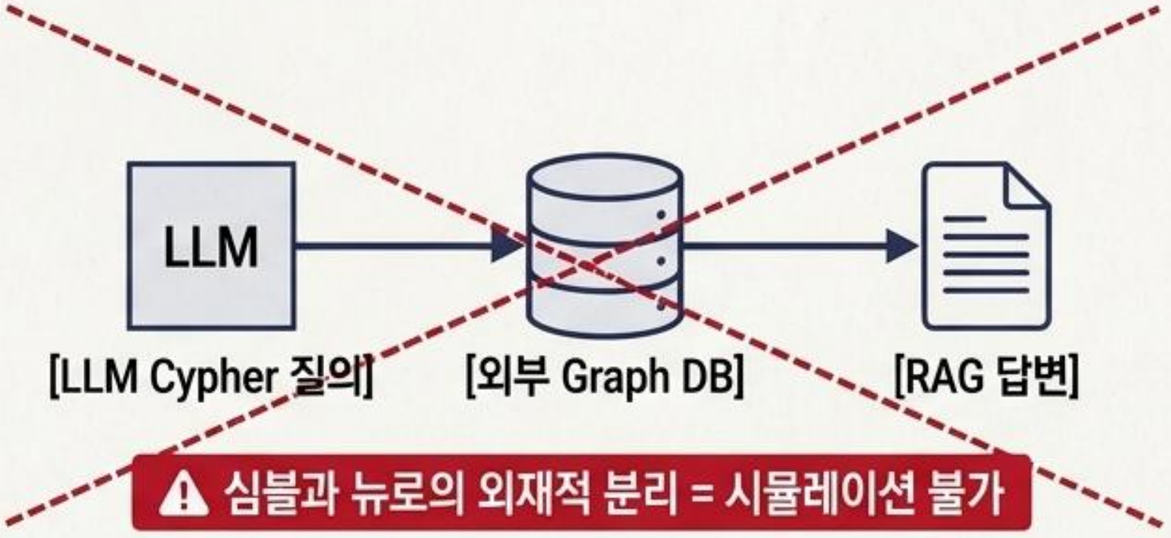
Gemini Notebook

3축 그래프 모델의 성취와 '외재적 결합(External Pipeline)'의 한계

Part II. 심볼 축(Axis 1) - 선행연구 해부

- **최근 국내 한 선행연구의 성취:** 단일 점수 대신 '지식요소-오개념 유형-인지진단 유형'의 3대 이종 그래프 DB로 입체 진단
- **6종 오개념과 12요소 인지진단:** 제한적 사고, 단순인과적 사고 및 알기·적용하기·추론하기 맵핑 (파일럿 F1 0.8391)
- **한계 ① 외재적 파이프라인:** LLM이 Graph DB 밖에서 Cypher 쿼리만 던지는 RAG 껍데기로 쓰여, 내부 추론 시뮬레이션 불가
- **한계 ② 두꺼워진 스칼라:** 학생 노드 속성이 여전히 정답률 숫자(인식하기 0.954)에 머물러 기계론적 조작 불가





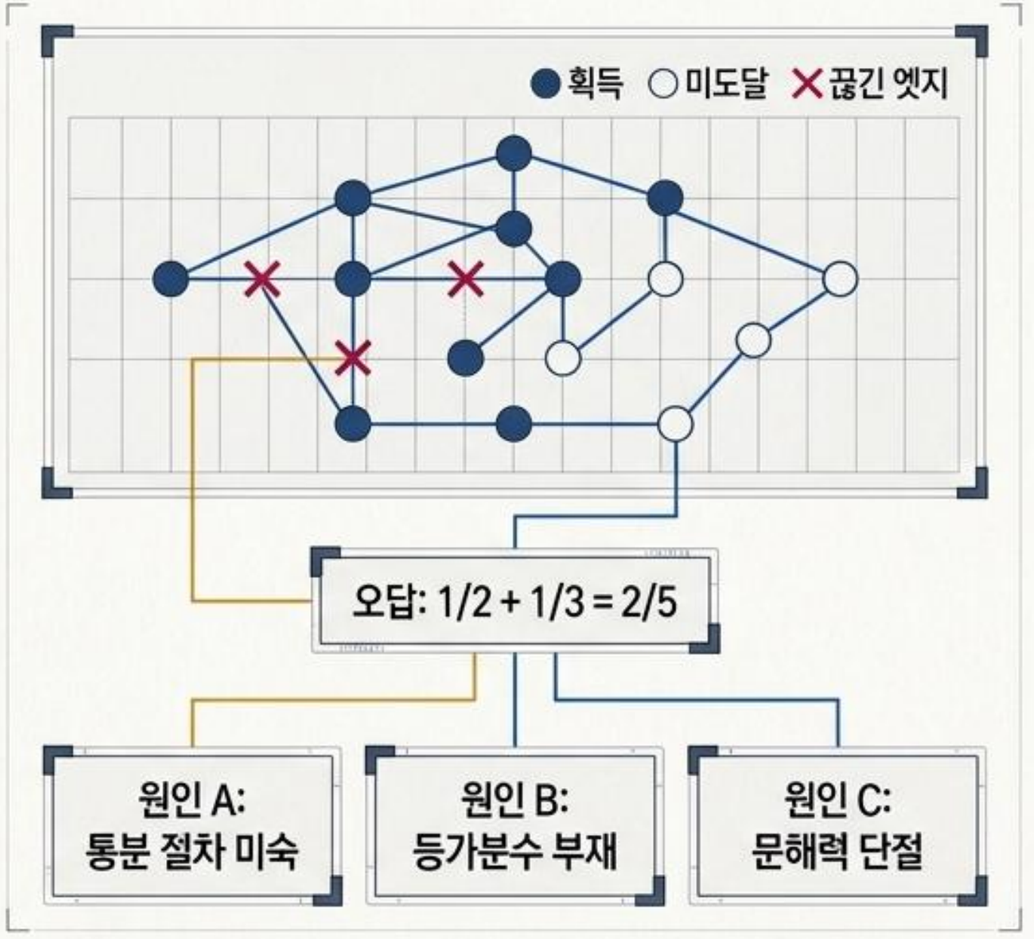
그래프가 LLM 바깥에 머무는 한, '살아있는 뇌'로서의 인공 학습자 조작은 불가능합니다.

Gemini Notebook

【축 1 종합】 우리가 찾는 것은 틀린 '문항'이나 '점수'가 아니라 끊어진 '엣지(Edge)'다

Part II. 심볼 축(Axis 1) - 심층 진단

- 동일한 오답(25) 뒤에 숨은 4가지 인과 경로:
  표면적 오답 결과는 같아도 내부 기전은 완전히 다름
- 경로 A/B: 통분 절차 엣지 미숙(절차적 결손) vs
  등가분수 개념 자체의 부재(개념적 결손)
- 경로 C/D: 문장제 수식 변환 과정 단절(인지/문해력 결손)
  vs 분모끼리 더하는 룰 간섭(오개념 활성화)
- 실험 타깃: CASE 표준 적용을 바탕으로 위계가 가장 선명한
  중학교 수학「수와 연산 / 함수」영역 30~50개 노드



진단의 최종 산출물은 '점수'가 아니라 뇌 속의 '어느 인과 링크가 끊어졌는지를 보여주는 지도'입니다.

Gemini Notebook

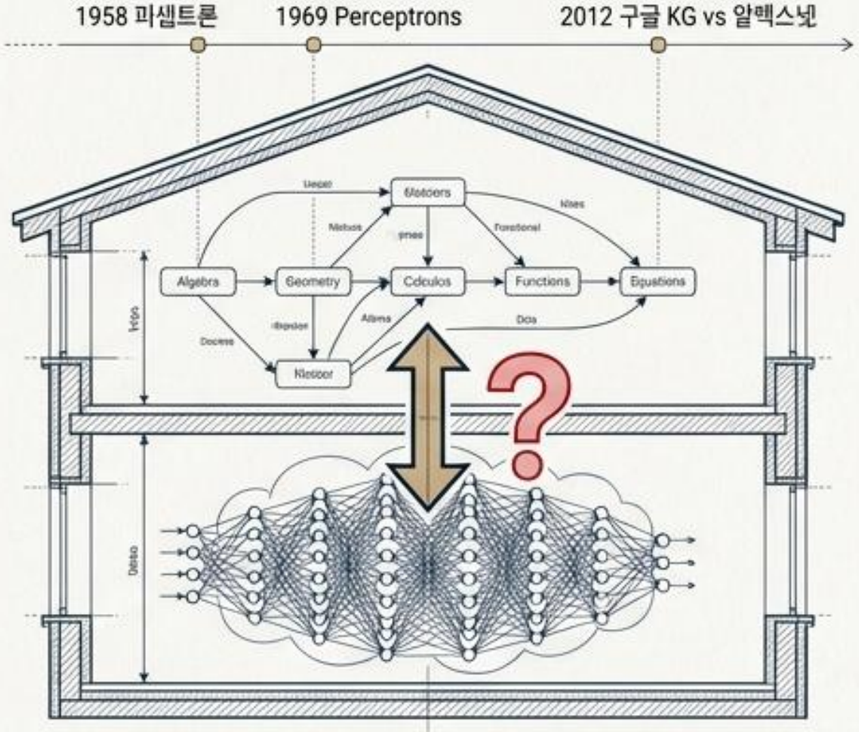
【축 2】 기호와 신경망의 엇갈린 역사, 그리고 '내재적 정렬(Alignment)'의 개막

Part III. 뉴로 축(Axis 2) - 패러다임 조우

- **AI 역사의 오해 교정:** 기호주의와 신경망은 순차 교대한 것이 아니라, 2012년(Google KG vs AlexNet)에 정면으로 엇갈림
- **추상화 그래프의 등가성:** 지식그래프(인간 지식 압축)와 신경망(가중치로 세계 압축)은 모두 거대한 추상화 구조물
- **새로운 과제, 내재적 정렬:** 밖에서 RAG로 붙이는 것이 아니라, 심볼 층(교육과정)과 신경망 층(LLM 내부 활성)을 정면 대응
- **2024~2026 기계론적 돌파구:** SAE를 통한 생물학 교과서 구조 복원 등 두 층을 잇는 해석가능성 기술의 마침내 등장

[Top Floor]
심볼 층: 사람이 읽을 수 있는
교육과정 지식그래프

[Bottom Floor]
신경망 층: 계산만 가능했던
LLM 내부 활성 구름



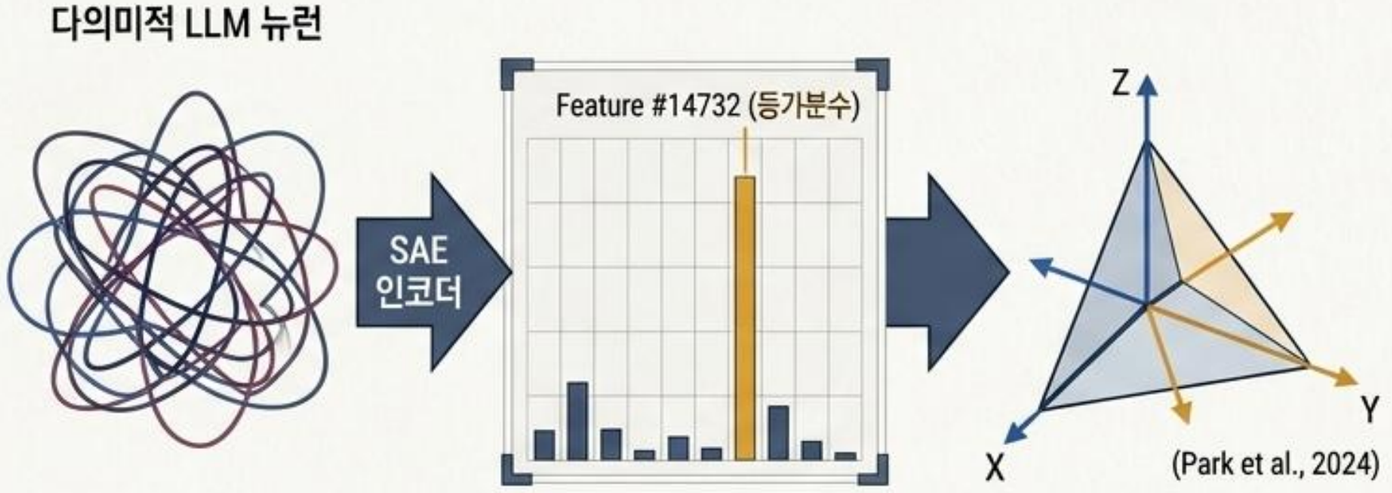
이제 두 세계를 나란히 놓는 것을 넘어, 서로의 좌표를 직접 꿰매는 '정렬'이 시작됩니다.

Gemini Notebook

【축 2】 희소 오토인코더(SAE) — 블랙박스 안에 기호의 좌표계를 놓다

Part III. 뉴로 축(Axis 2) - 신경망 안의 좌표계

- **중첩(Superposition)의 해체**: 여러 개념이 뒤엉켜 해석 불가한 은닉 뉴런을 단일의미 특징(Monosemantic Features)으로 분해
- **SAE의 증명** (Anthropic, 2024): 희소 오토인코더를 통해 '등가분수(#14732)' 같은 인간 해석 가능 특징이 내부에 실재함을 규명
- **표상 공간의 기하학** (Park et al., 2024): 범주 개념은 심플렉스(Simplex)로, 위계 관계는 직교성(Orthogonality)으로 새겨짐
- **핵심 결론**: 온톨로지 위계는 신경망 밖의 주석이 아니라, 표상 공간 안에 이미 존재하는 기하학적 좌표계임



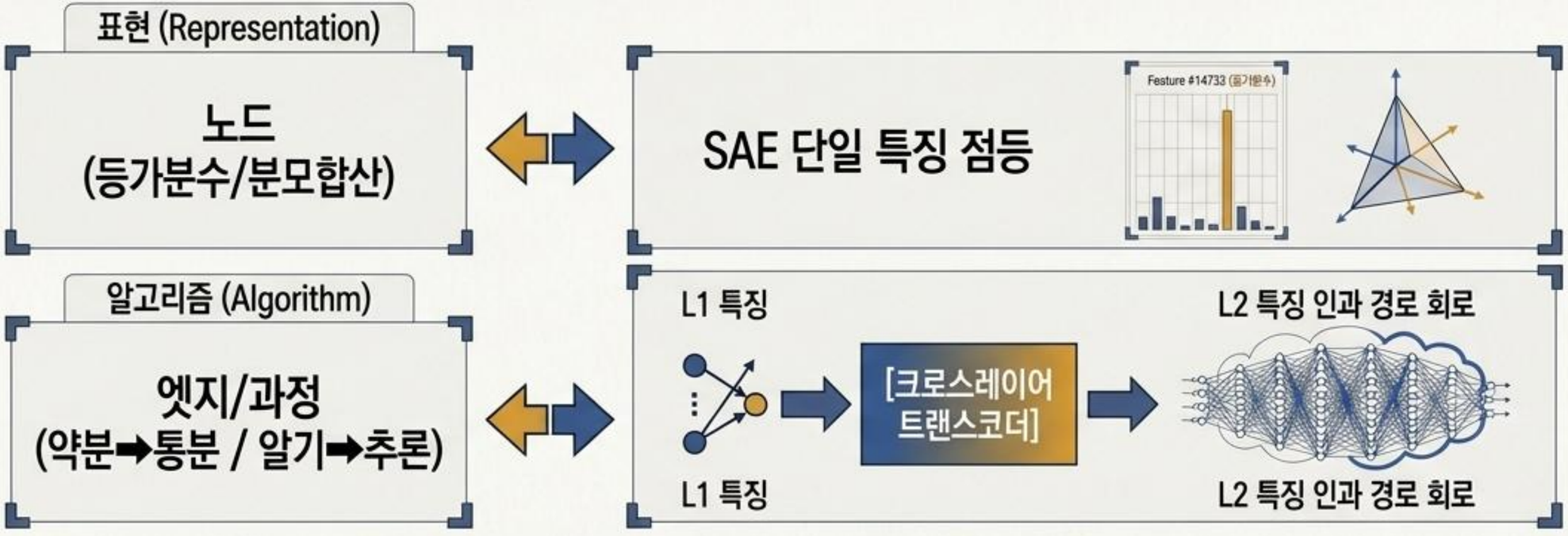
지식그래프는 억지로 붙이는 외부 라벨이 아니라, 신경망 내부의 수학적 기하 구조입니다.

Gemini Notebook

【핵심 이론 1】 노드는 특징(Feature)으로, 과정은 회로(Circuit)로 — 이원적 정렬

Part III. 뉴로 축(Axis 2) - [핵심 테제 1] 이원적 뉴로심볼릭 정렬

- 문제의식: 교육과정은 '개념 노드(지식)'와 '선수/인지 과정(기능)'이 섞여 있어, 단일 SAE 특징만으로는 전부 포착 불가
- 제1층위 (노드·부분공간 정렬): 정개념 지식요소 및 오개념 유형 ⟺ SAE 개별 희소 특징 및 심플렉스 공간 정렬
- 제2층위 (회로·위상 정렬): 선수학습 엣지 및 인지진단 과정(알기➡적용➡추론) ⟺ 레이어 간 방향성 인과 회로 정렬
- 마(Marr)의 수준 통합: 제1층위는 '표현(Representation)'을, 제2층위는 '알고리즘(Algorithm)'을 기계론적으로 구현



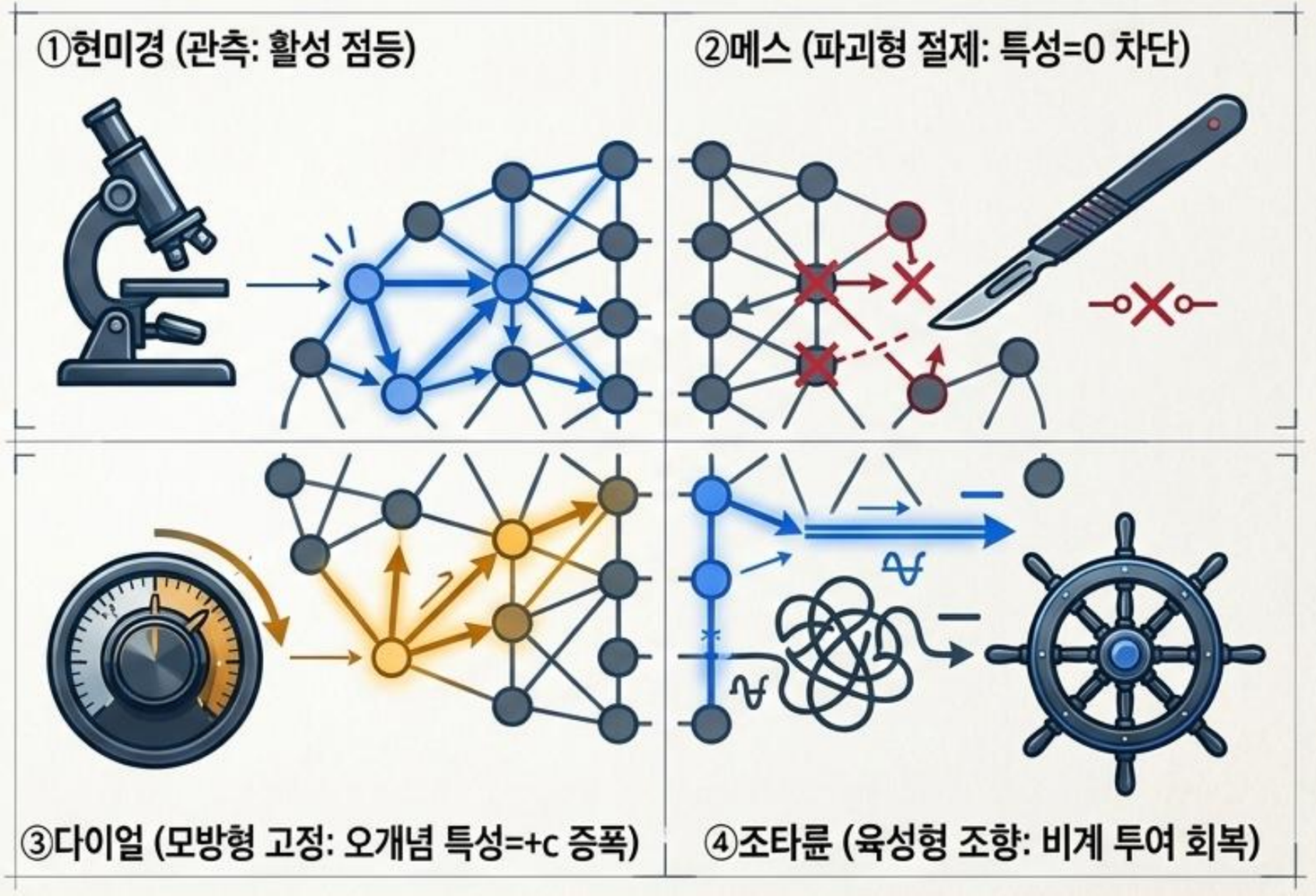
이원적 뉴로심볼릭 정렬(Dual-level Alignment): 정개념·오개념 노드 = SAE 희소 특징 / 선수학습 엣지·인지진단 과정 = 레이어 간 인과 회로

Gemini Notebook

LLM을 '살아있는 뇌'로 삼아 수행하는 4대 침습적 조작(Manipulation)

Part III. 뉴로 축(Axis 2) - 살아있는 인공 뇌의 실험실

- 관찰의 한계 돌파: 인간 학생의 뇌는 물리적으로 열어 조작할 수 없으나, 정렬된 LLM-SAE는 가중치 수준의 직접 조작 가능
- 침습형 관측 (Probing) & 파괴형 절제 (Ablation): 풀이 시 활성 스캔 및 특정 개념(A) 활성을 0으로 차단해 결손 학습자 즉시 생성
- 모방형 고정 (Clamping): 특정 오개념 특징을 강제 점화(+Max)시켜 오답에 사로잡힌 트윈 모사
- 육성형 조향 (Steering): 실제 학생 개입 전, 결손 트윈에게 힌트·반례를 사전 투여해 회로 위상 복구 효과 시뮬레이션



심리측정학의 비침습적 통계를 넘어, 교육적 개입을 사전에 충돌 테스트할 수 있는 기반입니다.

Gemini Notebook

【수학적 정초 1】 고정 점수(θ)에서 '개입조건부 미래 궤적(S_t)'으로의 전환

Part IV. 수학적 정초와 페루프 설계 - [핵심 테제 2-1] 수식 해부도 ①

$$S_t = \{P(Y_{t+1:t+H} \mid do(\pi), H_t)\}$$

과거 이력(H_t)	교수 개입 투입(do(π))	미래 전이·자율성 궤적 분포(Y)
반응, 오류 패턴, 조작 로그 등 지금까지의 상태	시스템이 실험적으로 가하는 비계(힌트, 반례, 조작 위젯)	동일 점수 출발이라도 개입 후 궤적이 다르면 완전히 다른 상태



신경구성주의 사전모형은 이 미래 궤적 후보들을 생성하는 근원적 가소성 동력입니다.

Gemini Notebook

【수학적 정초 2】 단일 목적함수(능동추론 -Gπ)에 의한 교육측정학의 통합 포섭

Part IV. 수학적 정초와 폐루프 설계 – [핵심 테제 2-2] 수식 해부도 ②

• 구(舊) 패러다임의 극복: 진단, 추적, 처방이 분절되지 않고 하나의 자유에너지 수식 아래 능동추론으로 통합 작동

$$-G(\pi) = \text{인식적 가치} + \text{실용적 가치} + \text{학습진전 항}$$

가장 빨리 엣지 파악

⊂ IRT/CAT 완벽 포섭
(피셔 정보량 최대화)

끊어진 엣지 잇기

⊂ BKT/DKT 한계 확장 및 포섭
(지연 전이 및 목표 성취)

최적 도전 난이도 유지

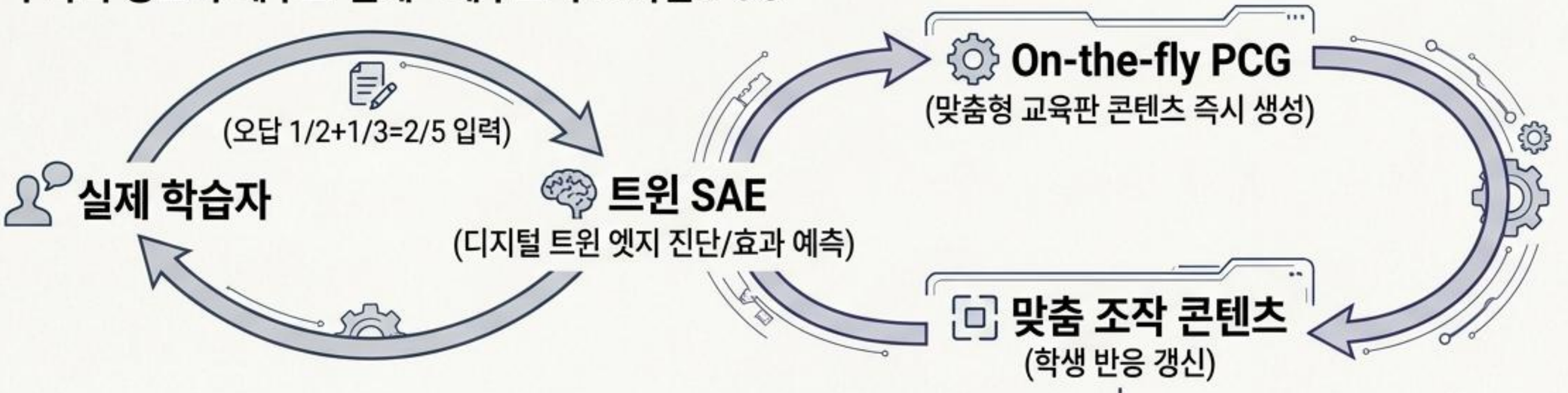
⊂ ZPD/게임 DDA 구동
(예측오차 감소율 기반)

적응형 검사(CAT)와 맞춤형 처방(ITS)은 더 이상 별개의 엔진이 아닙니다.

Gemini Notebook

【통합 아키텍처】 예측과 즉시 생성이 닫힌 고리(Closed-Loop)로 만날 때

Part IV. 수학적 정초와 폐루프 설계 - 폐루프와 교육판 PCG





- 페퍼트의 역전: 생성물이 단순 평가 객체면 '스키너 머신', 조작과 구성의 대상이면 '구성주의'로 진화

결론의 경계선: 기계가 만드는 것은 '풀 문제'가 아니라 학생이 '만들고 조작할 대상'이어야 합니다.

Gemini Notebook

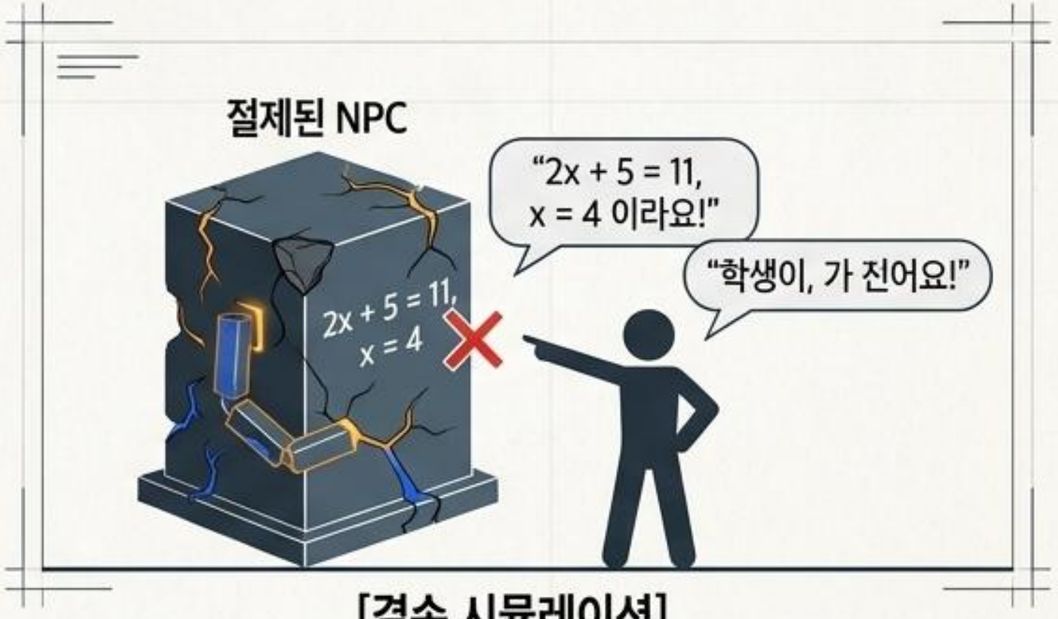
결손을 가진 시뮬레이션 학습자: '학생이 처음으로 공격해 이길 수 있는 상대'

Part IV. 수학적 정초와 페루프 설계 – 게임 플레이어 모델링과 DDA

- **학습 동기의 본질:** '공부가 재미없는 이유는, 수비만 하며 한 번도 이긴 적이 없기 때문입니다.'
- **주어의 역전(Adversarial Peer):** 무오류의 전지전능한 교사 AI를 끌어내려, 결함이 있는 동료 NPC로 학생 앞에 배치
- **논박과 가르침의 과제:** 개념 A의 특징을 절제한 AI 모델의 오답을 학생이 찾아내 비판하고 가르치도록 유도 (Teachable Agent)
- **게임 공학의 교육적 이식:** 20년간 축적된 게임 DDA와 해석가능한 플레이어 모델링을 결합해 비판적 논박 역량 강화



[기존 AI튜터]



[결손 시뮬레이션]

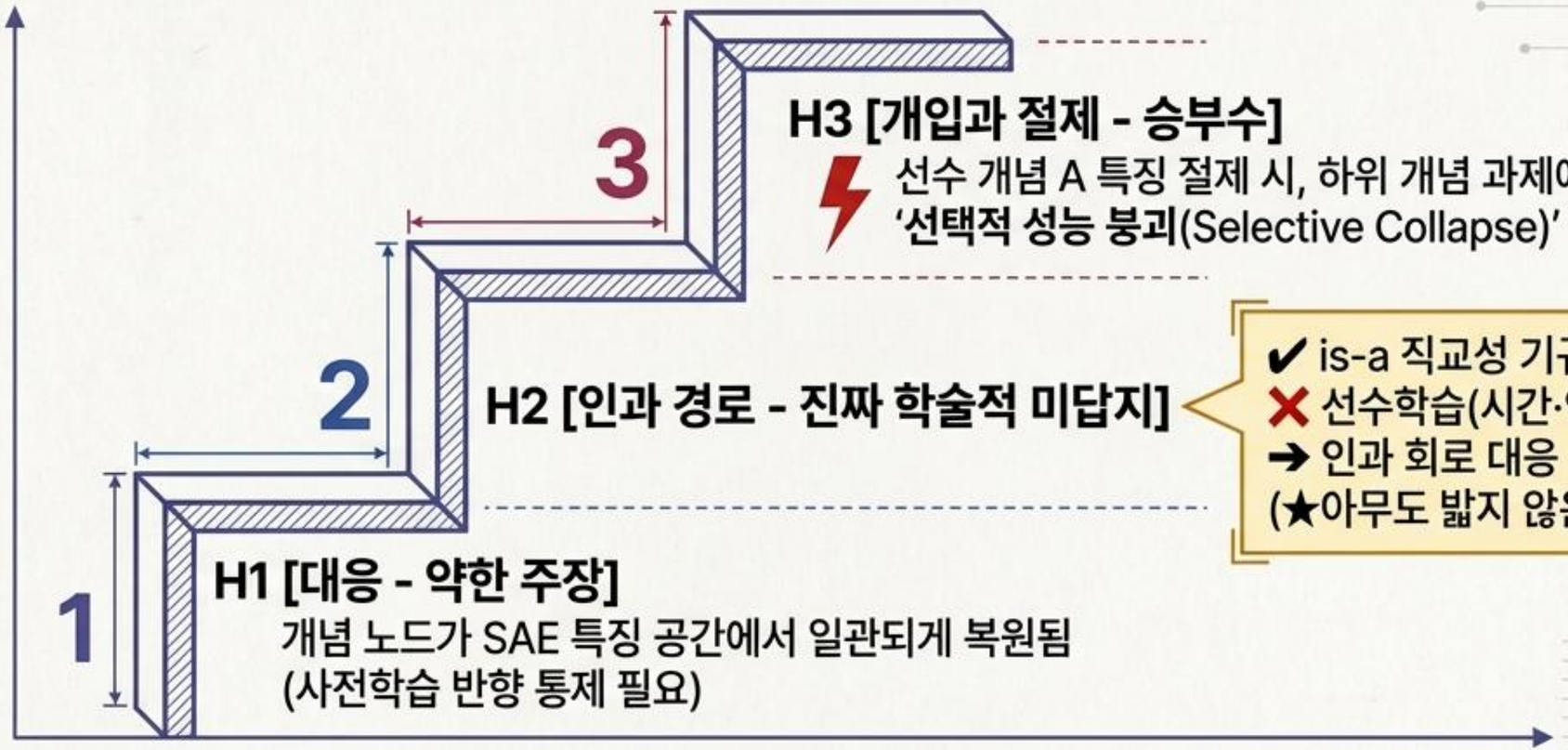
AI를 교단에서 내려오게 하여, 학생에게 조종간과 무기를 넘겨주어야 합니다.

Gemini Notebook

구현 효과의 과장이 아닌, 반증 가능한 3단계 연구 가설(H1~H3) 사다리

Part V. 가설 검증 설계, 통제 관문 - 무엇을 어떻게 반증할 것인가

- **인과적 충실도:** 단순 상관관계를 넘어 가중치 개입(Intervention)을 통한 기계론적 증명



이 가설 사다리는 '만들었으니 쓰십시오'가 아니라 '이 기제가 정말 실재하는가'를 묻는 과학적 반증 과정입니다.

Gemini Notebook

SAE 회의론(2025)에 정직하게 직면하는 2대 필수 통제(Controls) 방어선

Part V. 가설 검증 설계, 통제 관문 - 2대 필수 통제 관문

- **최신 회의론의 경고:** 무작위 모델조차 그럴듯한 특징을 뱉는다는 비판(2025) 등 눈대중 라벨링의 치명적 한계 수용

[ICML 2025]
[DeepMind 2025]

- **통제 1. 교차언어 일관성:** 한국어 '분수/통분'과 영어 'fraction/common denominator'가 동일 위치 특징에서 점등되어야 함 (단순 표면 토큰 패턴 배제 / 1주차 Go/No-Go 관문)

[한국어 '분수'] → [Feature #14732] → [영어 'fraction' 점등 일치]

- **통제 2. 가짜 지식맵 대조군 (Counterfactual Curriculum Control):** 선수 순서를 뒤집거나 무관한 개념을 엮은 맵에서는 선택적 붕괴가 철저히 실패해야 함



- **인과 개입 통과:** 진짜 교육과정 정렬의 성공이 단순 코퍼스 반향(Corpus Echo)이 아님을 증명

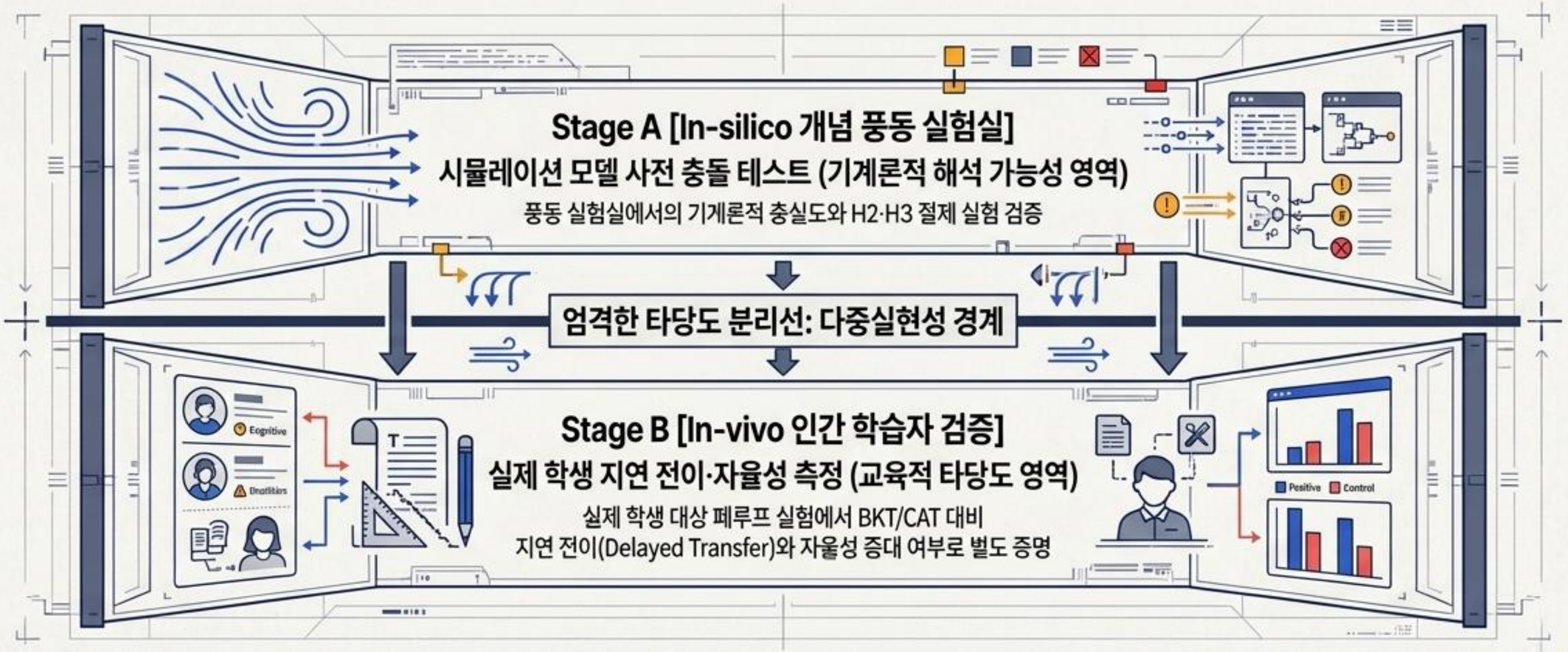
이 두 가지 엄격한 대조군 통과 없이는 어떠한 기계론적 인과성도 주장할 수 없습니다.

Gemini Notebook

‘기계론적 해석가능성’과 ‘인간 학습자 타당성’을 가르는 개념 풍동 실험실(Wind Tunnel)

Part V. 가설 검증 설계, 통제 관문 - 핵심 인식론적 분리

- **다중실현성의 함정 경계:** 시뮬레이터가 동일 오답을 낸다고 인간 뇌 뉴런과 동형(Isomorphic)이라 과잉 주장하지 않음 (비행기와 새의 구분)
- **Qwen-Scope / Gemma Scope 2 인프라:** 40개 개념 노드 대상 2만 문장 대조 코퍼스 및 1,200 평가 문항 투입



우리가 구축하는 것은 인간 뇌의 복제본이 아니라, 교육적 개입을 수백 번 부딪혀볼 정교한 가상 풍동(Wind Tunnel)입니다.

Gemini Notebook

티칭머신을 더 똑똑하게 만들 것인가, 학생에게 조종간을 넘길 것인가

Part V. 가설 검증 설계, 통제 관문 - 결론 및 제안

이원적 SAE 정렬 | 4대 침습 조작 | IPS-NC 궤적 | EFE 능동추론 | 개념 풍동 실험실

- **규범적 선언:** '내부를 열 수 없고, 조건부 즉석 생성이 불가능하다면 더 이상 그것을 적응형이라 부르지 맙시다.'
- **공동 연구 제안:** 18개 백화점식 포털을 멈추고, 단 하나의 교과(중학 수학) 노드 실증 셀(Proof Cell) 공동 파일럿 구축
- 지식맵의 CASE 표준 고도화 및 지연 전이·반박 활동의 실증 지표 측정 협력
- **오픈사이언스:** 1주차 Go/No-Go 관문 결과부터 대조 코퍼스, 정렬 파이프라인 전 과정 오픈소스 투명 공개



[1953 스키너 머신]

기계를 더 똑똑한 감독관으로 만드시겠습니까, 아니면 학생에게 조종간을 넘기 시겠습니까?



2026 LLM 아이콘

Gemini Notebook