Catcher Technology Company · Boston Neuromind LLC

대화형 감별진단 루프(CDDL): 대화 안에서의 검증 가능한 다부품 정신과 판별을 위한 통합 알고리즘

Alex[1]

[1] *Boston Neuromind LLC, Boston, MA, USA. 전 하버드 교육대학원(Harvard GSE) 방문학자.*

교신: Boston Neuromind LLC, Canton, MA, USA.

초록

배경. 정신건강 챗봇은 규칙 기반 시스템에서 대규모 언어모델(LLM)로 빠르게 옮겨갔으나, 최근 160 편을 다룬 체계적 리뷰는 LLM 기반 시스템 중 임상 효능 검증을 거친 것은 극소수이며 대부분이 초기 검증 단계에 머물러 있음을 확인했다. 핵심 미충족 수요는, "어디가 아픈지 아직 모르는" 사용자를 적절한 임상적 우려 영역으로 — 진단을 내리지 않고, 취조가 아니라 편안하게 느껴지는 대화를 통해 — 안내할 수 있는 판별(triage) 에이전트다. 목적. 본 논문은 대화형 감별진단 루프(Conversational Differential Diagnosis Loop, CDDL)를 명세한다. CDDL 은 정립된 세 연구 전통 — 적응형 측정, 횡단진단 차원 구조, 베이지안 순차 감별 추론 — 을 자연 대화라는 매체 안에서 통합하는 알고리즘이다. 방법. CDDL 은 하나의 검증 가능한 목적함수 위에 세워진다: 매 턴, 에이전트는 살아 있는 가설 보드에 대한 기대 정보이득을 최대화하는 질문을 선택한다. 우리는 이 목적함수를 네 항(DiagnosticSeparation, TemporalValue, DimensionalCoverage, ConversationalCost)으로 분해하고 각각의 수식을 제시하며, 불변의 안전 "헌법"(red-flag 스캔, DSM 게이트 질문, '임상가를 대체하지 않는다' 는 경계)을 진화 가능한 "표현형"(대화 스타일, 큐레이션된 가중치)으로부터 분리한다. 진화는 자가조정이 아니라 locked-algorithm 원칙 하의 큐레이션된 버전 증분으로 진행된다. 결과. 우리는 이 통합이 어느 개별 부품도 단독으로는 보이지 않는 세 가지 속성 — 시간 인지 질문 선택, 확신도 변조 대화, 정보이득 종료 — 을 만들어냄을 보인다. 한 턴의 계산, 예시 대화 트레이스, 고정 설문지 및 무작위 순서 기준선 대비 엔트로피 감소 시뮬레이션, 그리고 예측 대 실측 엔트로피 감소로 표현된 검증 프로토콜을 제시한다. 결론. CDDL 은 대화형 판별 에이전트의 기여를 재정의한다: 부품들은 정립되어 있고, 통합이 기여이며 — 그 통합이 큐레이션과 임상 판단을 루프 안에 박아 넣기에, 방어 가능한 자산이다.

키워드: 대화형 에이전트; 감별진단; 정신과 판별; 적응형 검사; 횡단진단; HiTOP; 정보이득; 측정 기반 케어; 대규모 언어모델; 디지털 정신건강.

목차

1. 서론

1.1 문제: 어디가 아픈지 모르는 사용자

신체적으로 아픈 사람은 대개 몸의 어느 부위가 아픈지 말할 수 있다. 심리적으로 아픈 사람은 종종 그러지 못한다 그들은 "뭔가 이상해요", "이제 아무것도 재미가 없어요", "집중을 못 하겠어요" 같은 — 동시에 여러 진단 가능성과 양립하는 — 막연한 호소를 들고 온다. 그 뒤에 따라오는 임상 과제는 그 사람이 제시한 라벨을 확인해 주는 것이 아니라, 듣고, 묻고, 좁히는 것이다: 막연한 호소를 더 면밀한 평가를 요하는 임상적 우려 영역 쪽으로 옮겨 가는 것. 이것이 판별(triage)의 일이다.

이 에이전트가 속한 제품 아키텍처 안에서, 판별 에이전트는 사용자 여정의 Layer 1 을 차지한다. Layer 0 은 증상 추출기이고, Layer 2 는 진단별 전문 에이전트 집합이다. Layer 1 은 아직 스스로 길을 찾지 못하는 사용자를 받으며, 결정적으로, 진단하지 않는다 — 방향을 잡아준다. 그것은 임상 현장용 스크리닝 시스템의 대체물이 아니라 자매다: 둘은 독립적이고, 각자 자기완결적이며, 각자 따로 진화한다. 임상 현장용 자매는 자기보고를 전기생리학적 측정으로 보강한다. 대화형 판별 에이전트는 자기보고만으로, 집에서 작동한다 — 임상가가 어떤 계측 장비를 쓰기 전에 PHQ-9, PCL-5, 그리고 면담만으로 호소를 일상적으로 방향 잡는 것과 똑같은 방식으로.

1.2 새벽 2 시에 던진 두 질문

이 논문은 두 질문에서 시작했다. 첫째: 편안한 대화를 통해 진정으로 지능적인 판별 에이전트를 만들 수 있는 알고리즘이 있는가? 둘째: 그것을 뒷받침하는 문헌이 있는가, 그리고 우리가 새로 기여할 수 있는 진정으로 새로운 알고리즘이 있는가? 둘 다 답은 yes 다. 다만 두 번째 답은 "새롭다"가 무슨 뜻인지에 대한 주의를 요한다. 지능적 판별 에이전트의 부품들 — 적응형 문항 선택, 횡단진단 차원 구조, 베이지안 순차 추론, 구조화 면담 게이팅, 종단 모니터링 — 은 개별적으로는 잘 정립되어 있다. 문헌에서 얇은 것, 그리고 우리가 기여하는 것은, 편안한 대화라는 매체 안에서의 그것들의 통합이다. 기여는 새 부품이 아니다. 오케스트레이션이다.

1.3 리뷰가 확인해 준 공백

이 분야에 대한 가장 최근의 체계적 지도화 연구는 2020 년부터 2024 년 사이에 발표된 AI 정신건강 챗봇 연구 160 편을 리뷰하여 그 아키텍처를 규칙 기반, 머신러닝 기반, LLM 기반으로 분류하고, 벤치 테스트 · 파일럿 실현가능성 · 임상 효능의 3 단계 평가 프레임워크를 제안했다. 그 핵심 발견은 엄밀한 임상 검증이 제한된 파편화된 지형이며, 특히 생성형 시스템에서 그러하다는 것이다: 대부분의 LLM 기반 개입은 초기 개발 단계에 머물러 있고, 리뷰는 임상가와 정책 입안자가 마케팅 주장과 기술적 현실을 구분할 것을 촉구한다. 정신건강용 생성형 LLM 에 관한 동반 스코핑 리뷰는 선별된 726 편 중 16 편만이 포함 기준을 충족했으며, 그 응용들이 초기 가능성을 보이지만 효과는 불확실하다고 보고했다. 따라서 빈 땅은 의견의 문제가 아니라 문서화된 사실이다. CDDL 은 바로 그 땅을 차지하도록 설계되었다 — 모든 수가, 구성상, 측정 가능한 예측인 대화형 판별 알고리즘.

1.4 본 논문의 기여

1. 대화형 판별을 위한 하나의 검증 가능한 목적함수: 매 턴, 살아 있는 가설 보드에 대한 기대 정보이득을 최대화하는 질문을 선택하며, 종료 규칙은 고정된 턴 수가 아니라 정보이득 바닥에 묶인다.

2. 그 목적함수의 네 항 분해 — DiagnosticSeparation, TemporalValue, DimensionalCoverage, ConversationalCost — 와 각 항의 수식, 그리고 어떤 항이 큐레이션된 지식 vault 의 어떤 부분에 의존하는지에 대한 설명.
3. TemporalValue: within-user 다중 세션 기억이 질문 선택에 먹여지는 함수적 메커니즘 — 같은 차원이라도 그 신호가 새로운 것이냐 반복되는 것이냐에 따라 다르게 질의된다.
4. 안전 바닥(red-flag 스캔, DSM 게이트 질문, '임상가를 대체하지 않는다' 는 경계)을 가중 최적화 바깥에 완전히 두어, 어떤 가중치 튜닝으로도 그것을 움직일 수 없게 하는 헌법/표현형 분리.
5. 통합된 루프의 세 가지 창발 속성 — 시간 인지 질문 선택, 확신도 변조 대화, 정보이득 종료 — 어느 개별 부품도 단독으로는 보이지 않는 것.
6. 검증 프로토콜: 기대 정보이득이 숫자이고 질문 후 실현된 보드 엔트로피 감소도 숫자이므로, 에이전트가 던지는 모든 질문은 반증 가능한 예측이다.

전반에 걸쳐, 우리는 사용자 대면 서술 전반에서 '환자(patient)' 대신 '내담자(client)'를 쓴다. 이는 비의사 실무자 범위, 그리고 진단보다 방향 안내와 자기이해를 지향하는 포지셔닝과 정합한다. '환자' 라는 용어는 DSM 진단 맥락과 레거시 데이터 스키마 이름에서만 유지된다.

2. 배경 및 관련 연구

CDDL 은 네 영역의 단단한 땅과 한 영역의 얇은 땅 위에 선다. 얇은 땅이 기여 지점이다. 각각을 차례로 검토하고 CDDL 이 그로부터 정확히 무엇을 취하는지 진술한다.

2.1 적응형 측정 — 질문 선택을 위한 단단한 땅

검증된 척도를 운용하되 다음 문항을 정보 최대화로 선택하는 것은 정신건강 측정에서 정립된 분야다. 다차원 문항반응이론(MIRT)과 random forest 진단 분류 위에 세워진 정신건강용 전산 적응형 검사(CAT-MH) 모음은, 긴 구조화 임상면담에 대해 고정 길이 검사를 능가하는 신뢰도로 검증되었으며, 1 차 의료에서는 환자들이 선호하는 형태로 전달되면서도 PHQ-9 및 GAD-7 에 필적하는 진단 정확도를 보였다. CAT-MH 계열은 이제 우울 불안, 조증, 물질 사용, 자살성, 외상후 스트레스, 그리고 청소년판을 아우른다. 적응형 측정은 사실상 대화형 판별의 학술적 쌍둥이다. 결정적 차이는 매체다. CAT-MH 는 폼 기반으로, 고정 척도로 평정되는 개별 문항을 제시한다. CDDL 은 대화 기반이다.

그 바탕 원리는 공유된다: 초기 응답이 잠재 특성에 대한 잠정 추정치를 산출하고, 그 추정치가 다음 문항을 선택하며, 추정 불확실성이 사전 정의된 임계값 아래로 떨어질 때까지 절차가 계속된다. 이것이 정확히 CDDL 가설 보드의 종료 논리다. 이 원리는 또한 고정 설문지의 한계를 드러낸다 — 고정 설문지는 모든 문항이 동일하게 변별적이고 동일하게 기여한다고 가정하지만, 이 가정은 종종 부정확하며, 따라서 고정 도구는 정밀성을 표준화와 맞바꾼다. “설문지를 해체해서 입는다” — 측정 구조는 보존하되 고정된 문항 문구는 버린다 — 는 CDDL 의 전략이 여기서 곧바로 따라 나온다.

2.2 횡단진단 차원 구조 — 차원 레이어를 위한 단단한 땅

정신병리의 위계적 분류체계(HiTOP)는 정신병리를, 점점 더 넓은 횡단진단 스펙트럼으로 조직되는 차원들의 집합으로 개념화하는, 데이터 기반의 위계적 대안 분류다. 그 구조는 메타분석으로 뒷받침되었다: HiTOP 프레임워크와 매우 유사한 모델이 데이터에 잘 부합했고, DSM 진단들의 횡단진단 차원 내 배치가 대체로 확인되었다. 간이 HiTOP(B-HiTOP)은 CDDL Layer 0 의 직접적 선례다: 임상가가 전통적 진단 경계를 가로지르는 HiTOP 차원에서 임상적 우려 영역을 식별하고, 그 결과를 사용해 더 심층적인 평가를 표적화하게 하는 효율적 스크리닝 도구. 이것이 정확히 CDDL 에이전트가 하는 일이다 — 차원 스크리닝, 그 다음 활성 영역에 대한 표적화된 감별 탐침.

HiTOP 은 또한 처음부터 살아 있는 모델로 설계되었다 — 진화하는 것으로, 새로운 발견에 발맞추기 위해 지속적 개정을 요하는 것으로 개념화되었다. CDDL 의 큐레이션된 버전 진화는 이 철학을 공유한다: 숨 쉬되, 드리프트가 아니라 검토를 거쳐 개정되는 모델.

2.3 베이지안 순차 감별 추론 — 보드와 선택기를 위한 단단한 땅

두 번째 전통은 진단을 불확실성 하의 순차적 의사결정으로 다룬다. 자동 감별진단의 한 베이지안 정식화는 질병 추론 단계에 베이지안 추론을, 질의 단계에 베이지안 실험 설계를 적용하는데, 이 접근은 해석 가능하고, 비용이 드는 훈련이 필요 없으며, 추가 노력 없이 새로운 변화에 적응한다. 이것이 CDDL 가설 보드와 다음 질문 선택기의 수학적 토대다. “해석 가능 / 훈련 불필요 / 적응 가능” 이라는 속성은 자기완결적이고 큐레이션 기반인 설계 철학과 곧바로 정합한다.

구조화 임상면담 시스템은 게이팅 개념을 기여한다: 그러한 시스템은 연역적 추론과 진단의 자동 생성으로 특징지어지며, 순차적으로 제시되는 기준들에 대한 입력 판단을 요구하는 미리 정해진 결정 트리를 따른다. 그러나 같은 연구 흐름은 중요한 경고를 내놓는다 — 전산화된 순차 시스템은 전역적 사고와 앞뒤 탐색을 더 어렵게 만들며, 사고를 하나의 순서로 강제한다. 그 경고가 바로 CDDL 가설 보드가 존재하는 이유다. 보드는 에이전트가 여러 가설을 동시에 들고 있도록 강제하여, 순수 순차 추론의 함정 — 자동화 편향과 누락 오류 — 을 막는다.

더 최근의 연구는 임상 추론을 불완전 정보 하의 상호작용적 질문하기로 프레임화한다. MediQ 벤치마크와 질문하기 시스템은, 최첨단 LLM 에게 질문을 하라고 직접 프롬프트하면 오히려 임상 추론이 저하되고, LLM 을 능동적 정보 탐색에 맞추는 것이 간단치 않으며, 모델 확신도를 추정해 언제 더 물을지를 결정하는 기권(abstention) 모듈이 진단 정확도를 상당히 개선한다 — 다만 모든 정보가 처음부터 주어지는 상한선에는 여전히 못 미친다 — 는 것을 확립한다. 두 교훈이 CDDL 로 곧바로 이어진다: 순진한 "그냥 질문하라" LLM 은 판별 에이전트가 아니며, 무엇을 물을지와 언제 멈출지를 결정하는 명시적이고 확신도 인지적인 메커니즘이 필요하다. CDDL 의 네 항 목적함수와 정보이득 종료 규칙이 바로 그 메커니즘을, 명시적으로 만든 것이다.

이 연구 흐름의 가장 최근 갈래는 더 나아가, CDDL 이 채택하는 구조적 어휘를 공급한다. 지식 그래프에 정박된 진단 프레임워크들은 정보이득 질문 전략을, 증거가 드러남에 따라 임상가의 믿음을 갱신하는 명시적이고 상태 추적되는 진단 기록과 짝짓는다 — CDDL 의 다음 질문 선택기와 살아 있는 가설 보드와 동일한 짝짓기다. 목표 지향적 진단 질문하기 벤치마크들은, 1 회성 정확도가 아니라 전략적 질의가 측정할 가치가 있는 속성이며, 각 질의가 직전 답변에 의존하는 표적화된 질문의 적응형 시퀀스가 효율적 에이전트를 비효율적 에이전트로부터 가른다는 것을 확립한다. 이런 표현으로 보면 CDDL 은 정보 안내 질의 루프의 정신과 판별 사례이며, 일반 의료 연구가 지니지 않은 두 가지를 더한다: flat 한 증상 집합 대신 횡단진단 차원 레이어, 그리고 단일 만남 가정 대신 within-user 시간 축.

2.4 대화형 LLM 정신건강 평가 — 얇은 땅

이것이 가장 새롭고 가장 빈 영역이며, 따라서 기여 지점이다. 이미 인용한 체계적 리뷰 증거 너머로, 상담 특화 시스템의 등장은 CDDL 이 다루도록 만들어진 두 실패 모드를 명명한다. 기준에 근거한 임상 지원이 없으면, 대화형 시스템은 증상이 비전형적이거나 충분히 명시되지 않을 때 근거 없는 임상 단언에 빠지기 쉽고, 다중 턴 상호작용에서 inquiry drift — 주제 이탈 또는 저수익 질문 — 를 완화하는 데 어려움을 겪는다. 근거 없는 단언은 CDDL 의 게이트-vault 레이어가 막는 것이고, inquiry drift 는 정보이득 질문 선택기가 막는 것이다. 이 땅의 얇음은 본 연구의 약점이 아니다. 그것은 본 연구가 취하려고 쓰여진 기회다.

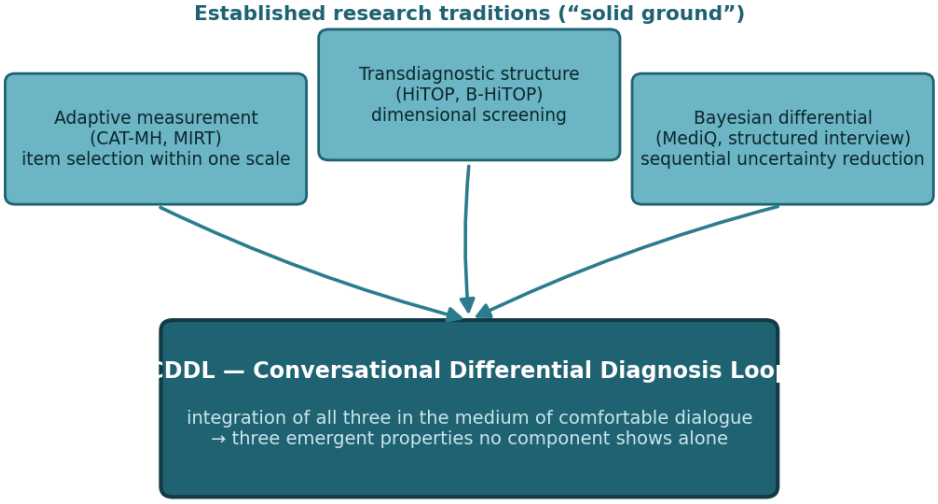
2.5 종단 및 측정 기반 케어 — within-user 축을 위한 단단한 땅

측정 기반 케어(MBC)는 검증된 도구를 사용한 내담자 데이터의 체계적 수집과, 그 데이터를 임상 의사결정의 안내와 공유에 적용하는 것으로 이루어지며, 임상가가 증상과 기능을 종단적으로 모니터링하고 협력적 치료 계획을 지원할 수 있게 한다. 단일 스냅샷에 의존하기보다 여러 세션에 걸쳐 신호를 누적하는 것의 근거는 회상의 한계로 강화된다: 회고적 응답은 회상 실패와 응답 편향의 영향을 받으며, 하루 끝의 회고 보고는 같은 날 더 일찍 얻은 실시간 응답이 포착하는 기분 변동의 일부만을 포착한다. 종단 생태순간평가(EMA)를 비지도 머신러닝과 결합하면 높은 증상 부담을 가진 군을 건강한 평정 패턴의 군으로부터 분리하고, 확장된 의료가 필요한 개인을 식별할 수 있음이 보고되었다. 이들을 합치면, 단일 판별 스냅샷이 불충분하며 시간 누적이 신호를 분리한다는 것이 확립된다 — 회상과 싸우는 대신 시간을 자기편으로 쓰는 에이전트의 학술적 선례다.

표 1. CDDL 이 각 연구 전통에서 취하는 것.

전통	정립된 결과	CDDL 이 취하는 것	땅
적응형 측정 (CAT-MH, MIRT)	정보에 의한 문항 선택; 구조화 면담에 대해 검증됨; 고정 검사보다 낮은 부담	다음 질문 목적함수와 불확실성 임계값 종료 규칙	단단함
횡단진단 구조 (HiTOP, B-HiTOP)	DSM 범주를 가로지르는 차원; 메타분석적으로 뒷받침됨; 살아 있는 모델로 설계됨	Layer 0 의 차원 벡터; 활성 영역의 표적화된 탐침; 살아 있는 모델 철학	단단함
베이지안 감별 (베이지안 DDx, 구조화 면담)	추론 + 실험 설계; 해석 가능, 훈련 불필요; 게이팅이 진단 다발을 가지치기	가설 보드 사후분포; 게이트 질문; 순수 순차 추론에 대한 방어로서의 보드	단단함
상호작용적 질문하기 (MediQ)	순진한 질문 프롬프팅은 추론을 저하시킴; 명시적 확신도/기권 메커니즘이 필요	명시적 네 항 목적함수; 순진한 프롬프팅 대신 정보이득 종료	단단함
종단 / MBC (EMA, 측정 기반 케어)	단일 스냅샷은 과소 포착; 다중 세션 누적이 신호를 분리; 확장 의료 필요를 식별	TemporalValue 와 within-user 누적 레이어; trait / state / course 분리	단단함
대화형 LLM 평가	문서화된 공백: 임상 검증 거의 없음; 근거 없는 단언과 inquiry drift 가 명명된 실패 모드	기여 지점 — 위의 것들을 편안한 대화 안에서 통합	얇음

Figure 10. Positioning — CDDL as an integration algorithm, not a new component



“The components are established. The integration value is the contribution — and the moat.”

그림 10. 포지셔닝. 세 전통은 정립된 “단단한 땅”이다. CDDL 은 그것들을 편안한 대화라는 매체 안에서 결합하는 통합 알고리즘이다; 어느 개별 부품이 아니라 통합값이 기여다.

3. CDDL 아키텍처

3.1 일곱 가지 설계 원칙

CDDL 에이전트는 합의된 일곱 원칙 위에 선다. 여기서는 논문의 나머지가 정식화할 골격으로 진술한다.

1. 판별은 Layer 1, 별도의 에이전트 제품, 임상 현장용 스크리닝 시스템의 자매다 — 독립적, 각자 자기완결적, 양쪽 다 진화.
2. 에이전트는 자기완결적이다 — 외부 백엔드 인프라 없이 작동하며, 단일 파일 배포 방식으로 vault 를 에이전트 안에 임베딩한다.
3. 진화는 자가조정이 아니다. 에이전트가 패턴 후보를 제안하고, 큐레이터가 검토하며, 버전이 증분된다. 그 과정은 검증 가능하고, 되돌릴 수 있으며, 드리프트가 없다.
4. 초기 가중치는 네 원천에서 온다 — DSM-5-TR 구조, 검증된 척도 문항 구조, 감별진단 문헌, 임상 추정 — 각각 출처와 확신도 수준으로 태깅된다.
5. 헌법과 표현형이 분리된다. 헌법(DSM 게이트 기준, red-flag 로직, 안전 경계)은 공유되고 불변이다; 표현형(대화 스타일, 질문 순서, 저확신 셀 미세조정)은 독립적이고 진화 가능하다.
6. 정확도는 에이전트 안에서 구축된다 — within-user 다중 세션 누적, 감별 추론의 깊이, 큐레이션된 버전 진화. Cross-product 참조는 v2+ 로드맵 항목으로, 정확도 도구가 아니라 통찰 도구로 재분류된다.
7. 에이전트의 정체성은 진단하는 것이 아니라 방향을 잡아주는 것이다. 그 출력은 우선순위가 매겨진 복수의 추천, 임상가 평가를 대체하지 않는다는 명시적 진술, 그리고 어떤 위기 신호에 대한 즉각적 분기다.

3.2 판별 엔진의 네 부품

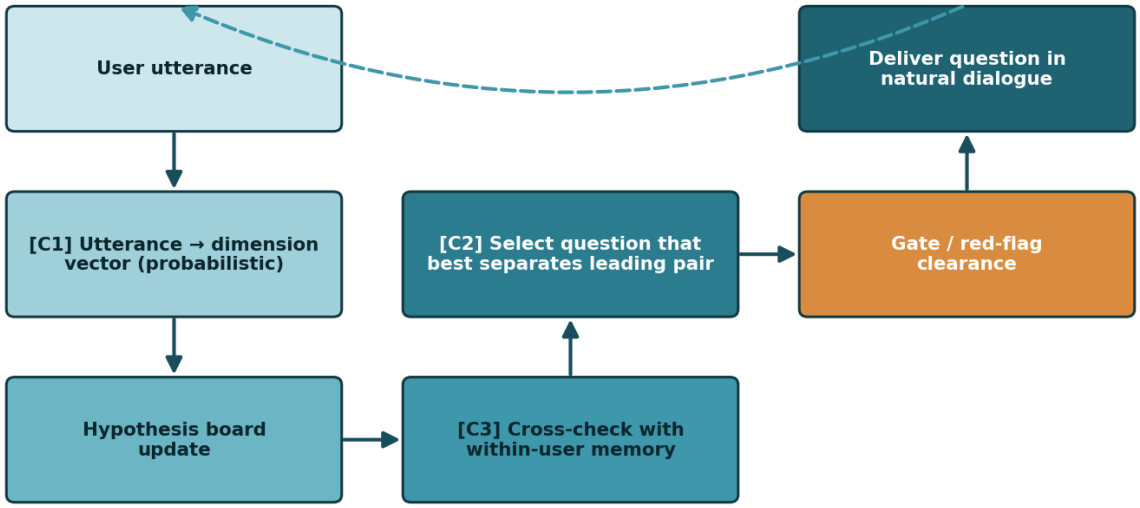
표 2. CDDL 판별 엔진의 네 부품.

부품	역할
부품 1 — 발화→차원 변환기	매 발화를 횡단진단 차원 벡터의 부분 업데이트로 변환한다. 한 발화가 여러 차원에 동시에, 불확실하게 기여할 수 있다.
부품 2 — 가설 보드	flat scoring 대신, 각각 {진단 영역, 확률, 지지 증거, 결정적 미질문 질문}을 담은 2–4 개의 살아 있는 가설을 유지한다.
부품 3 — 다음 질문 선택기	매 턴, 기대 정보이득을 최대화하여 현재 가설들을 가장 잘 가르는 단 하나의 질문을 선택한다. 이것이 에이전트 지능의 심장이다.
부품 4 — 게이트 및 안전 레이어	red-flag 스캐너(매 턴, 최우선 분기)와 시간축 질문(경과 매핑) — 추론 루프와 병렬로, 그리고 그 위에서 작동한다.

이 부품 중 둘은 에이전트의 재량 — 그 지능 — 에 대응하고, 둘은 고정된 바닥에 대응한다. 재량 영역은 활성 영역을 얼마나 깊이 탐침할지, 질문의 순서, 침묵 영역을 재방문할지 여부, 그리고 대화 톤을 포함한다. 고정 바닥은 red-flag 스캔(매 턴, 무조건), 게이트 질문(건너뛸 수 없음), 그리고 임상가 평가를 대체하지 않는다는 경계를 포함한다. 이 재량/바닥 구분이 제 6 절에서 정식화되는 헌법/표현형 분리의 씨앗이다.

Figure 1. The CDDL loop — structure of a single conversational turn

next utterance → loop



Key: Contribution 3 (temporal memory) feeds Contribution 2 (question selection) — the three contributions are not parallel; one is the input of another.

그림 1. CDDL 루프 — 한 대화 턴의 구조. 핵심 특징은 기여 3(시간 기억)이 기여 2(질문 선택)에 먹여진다는 것이다: 세 기여는 나란히 있는 것이 아니라, 하나가 다른 하나의 입력이다.

3.3 루프 — 한 턴

하나의 CDDL 턴은 다음과 같이 진행된다. 사용자 발화가 부품 1 에 들어가 확률적이고 미확정인 차원 벡터로 변환된다. 가설 보드가 갱신된다. 갱신된 보드가 within-user 기억과 대조된다 — 이것은 처음 듣는 신호인가, 반복되는 신호인가? 그러면 다음 질문 선택기가 보드에서, 가장 비등하게 경합하는 가설 쌍을 가장 잘 가르는 질문을 고른다. 그 후보가 게이트와 red-flag 레이어를 통과한다. 질문이 자연 대화체로 전달되고, 다음 발화가 루프로 다시 들어온다.

> *결정적인 구조적 사실: 기여 3(시간 기억)이 기여 2(질문 선택)에 먹여진다. 세 기여는 나란히 놓인 세 가지가 아니라, 하나가 다른 하나의 입력이다. 그것이 본 논문에서 "통합"이 의미하는 바이며, 그것이 어느 개별 부품이 아니라 루프가 기여인 이유다.*

4. CDDL 함수

4.1 루프의 심장: 다음 질문 선택 함수

CDDL 알고리즘 전체는 한 줄로 환원된다.

```
NextQuestion(state) = argmax_q  ExpectedInfoGain(q, state)
```

여기서 state 는 지금까지 누적된 모든 것 — 차원 벡터, 가설 보드, within-user 기억, 대화 상태 — 이고, q 는 후보 질문이다. 매 턴, 에이전트는 모든 후보 질문에 대해 그 질문이 가설 보드의 불확실성을 얼마나 줄일 것으로 기대되는지를 계산하고, 가장 많이 줄이는 질문을 선택한다.

> *이것이 검증 가능한 이유: ExpectedInfoGain 은 숫자이고, 질문이 던져진 뒤 실현된 보드 불확실성 감소도 숫자다. 예측 대 실측을 대조할 수 있다. 따라서 알고리즘은 반증 가능하다 — 단언이 아니라 과학이다. 이것이 대부분의 LLM 정신건강 시스템이 미검증이라는 문서화된 공백에 대한 직접적 답이다.*

목적함수는 네 항으로 분해된다:

```
ExpectedInfoGain(q, state) =
    w₁ · DiagnosticSeparation(q, hypothesis_board)        [기여 2]
  + w₂ · TemporalValue(q, user_history)                   [기여 3]
  + w₃ · DimensionalCoverage(q, dimension_vector)          [기여 1]
  − w₄ · ConversationalCost(q, dialogue_state)            [편안함 페널티]
```

Figure 2. Decomposition of the next-question objective into four terms

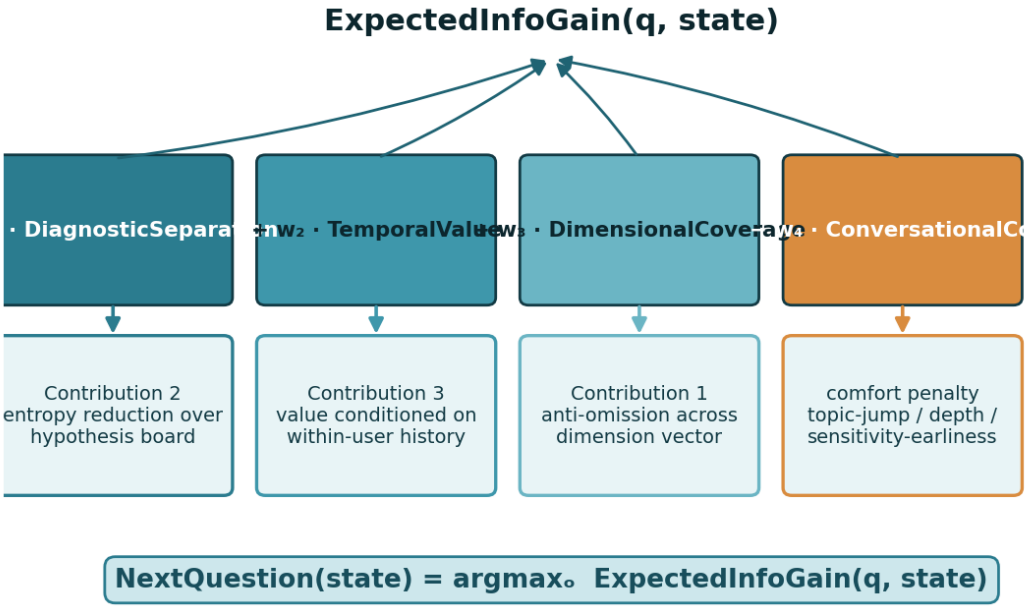


그림 2. 다음 질문 목적함수의 네 항 분해. 세 항은 정보이득에 더해지는 기여이고; ConversationalCost 는 편안함 페널티로 빼진다 — 에이전트가 취조처럼 행동하지 않도록.

4.2 항 1 — DiagnosticSeparation

이것이 가장 단단한 항이며 vault 구조에 거의 무관하다. 가설 보드는 진단 영역 h ∈ {MDD, PTSD, GAD, ADHD, …}에 대한 확률분포 P(h)다. 보드 불확실성은 그 엔트로피다:

```
H(board) = − Σₕ P(h) · log P(h)
```

질문 q 의 가치는 그것이 만들어내는 기대 엔트로피 감소다:

```
DiagnosticSeparation(q) = H(board) − Σₐ P(a) · H(board | a)
```

여기서 P(a)는 사용자가 답변 a 를 줄 확률로 현재 보드로 추정되며, H(board | a)는 답변 a 에 대한 베이즈 갱신 후 보드의 엔트로피다. 이것이 정보이론의 기대 정보이득을 직접 적용한 것이다. 베이즈 갱신 P(h|a) ∝ P(a|h) · P(h)는 vault 에서 정확히 하나의 양을 끌어온다: 우도 P(a|h), 즉 진단 영역 h 인 사람이 질문 q 에 답변 a 를 줄 확률.

4.2.1 서술적 vault 에서 우도 해석하기

현재의 큐레이션된 vault 에서, 가르는 질문 라이브러리는 각 질문에 대해 수치적 우도가 아니라 서술적 매핑 — “이 답변은 PTSD 를 시사한다” — 을 저장한다. 자연스러운 가정은 LLM 이 그 서술을 실시간으로 숫자로 변환하면 된다는 것이다. 그럴 수 없으며, 그 구분이 중요하다. 번역은 자동이다: LLM 은 “PTSD 를 시사함”을

읽고 증거의 방향을 추론할 수 있다. 보정(calibration)은 자동이 아니다: P(a|h)는 방향이 아니라 크기 — 0.7 대 0.4 — 이며, LLM 이 실시간으로 만들어낸 크기는 환각된 숫자로, 출처도 확신도 태그도 지니지 않아 네 원천, 출처-확신도 태깅 가중치 원칙을 정면으로 위반한다.

따라서 그 해법은 v1 선행 과제이되, 범위가 한정된 과제다. 서술적 매핑은 3 단계 순서척도 우도 — 약 ≈ 0.2, 중 ≈ 0.5, 강 ≈ 0.8 — 로 큐레이션되며, 각각 출처 태그를 지닌다. 베이즈 갱신은 순서 안정적이다: 정밀한 0.73 이 필요한 것이 아니라 올바른 순서만 필요하다. 이 큐레이션은 진단별 vault 를 구축하는 동일한 워크플로에 흡수되며, 별도의 연구 작업이 아니다.

> *항 1 의 검증: 질문이 던져진 뒤 실현된 H(board)를 측정하여 예측된 감소와 대조한다. 엔트로피를 0.4 비트 줄일 것으로 예측되었는데 0.05 를 실현한 질문은 가르는 힘이 잘못 추정된 것이다 — 큐레이션 후보가 된다.*

4.3 항 2 — TemporalValue

TemporalValue 는 함수에서 가장 새로운 부분이며, 기여 3 이 기여 2 에 먹여진다는 주장의 함수적 실체다. 그 핵심 아이디어: 같은 질문 q 라도, q 가 다루는 차원 d 가 사용자 기억에서 어떤 시간 상태에 있느냐에 따라 다른 가치를 가진다. 차원 d 의 시간 상태는 세 숫자로 요약된다:

- seen(d) — d 신호가 과거 세션에 등장한 횟수;
- variance(d) — 그 신호들의 세션 간 분산;
- trend(d) — 그 신호들의 시간적 기울기 (단조 증가나 감소면 크고, 들쭉날쭉하면 0 근처).

질문은 종류로 나뉜다: q_state(“지금 어떤가요?”), q_course(“평생 그래 왔나요, 아니면 최근인가요?”), q_onset(“언제 시작됐나요?”). 그러면 TemporalValue 는 질문 종류와 시간 상태에 대한 매치다:

```
TemporalValue(q, history) = match q.type:
  q_state   →  high  if seen(d) = 0          (새로운 신호 → 현재 상태부터 확립)
               low   if seen(d) large       (이미 여러 번 봤는데 “지금?” → 한 턴 낭비)
  q_course  →  high  if seen(d) ≥ 2 and variance(d) low   (반복 + 안정 → trait 후보;
course 가 결정적)
               high  if seen(d) ≥ 2 and trend(d) large    (반복 + 추세 → course 자체가
축)
               low   if seen(d) = 0                       (처음 듣는 신호의 course 를 물음
→ 사용자 당황)
  q_onset   →  high  if seen(d) ≥ 1 and onset 미확정
               low   otherwise
```

이것이 “기여 3 이 기여 2 에 먹여진다” 의 함수적 실체다. TemporalValue 는 질문 종류와 차원의 시간 상태를 둘 다 입력으로 받아서, 같은 차원이라도 새로운 것이면 q_state 쪽으로, 반복되는 것이면 q_course 쪽으로 점수를 준다. 질문 선택기(항 1)는 무엇을 물을지 고르고, TemporalValue 는 어떤 시간 각도에서 물을지 고른다.

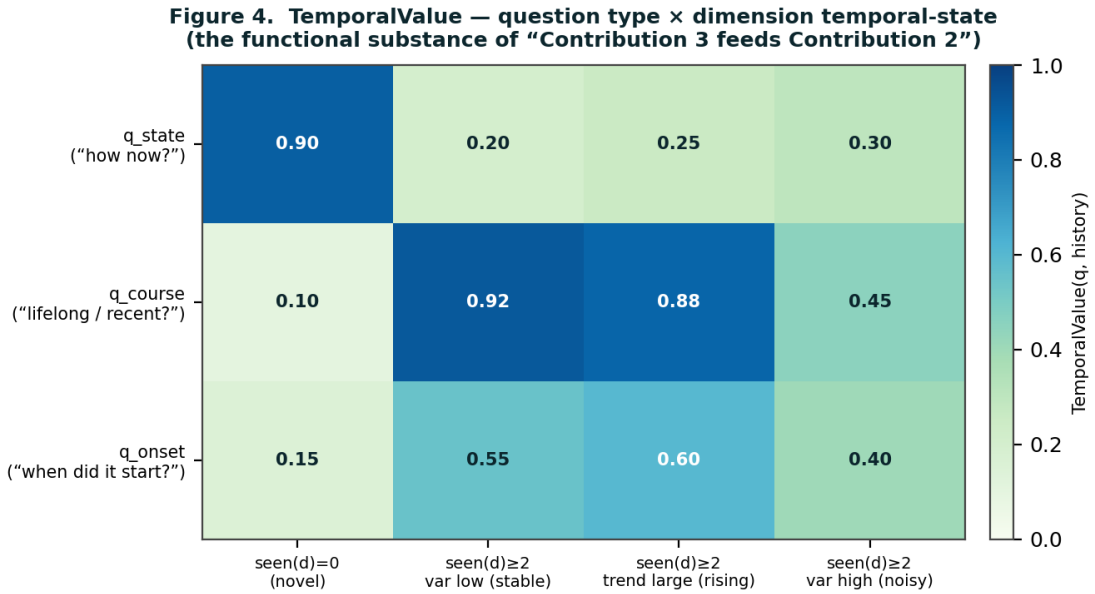


그림 4. 결정 행렬로서의 TemporalValue — 질문 종류와 차원의 시간 상태의 교차. 새로운 신호는 q_state 쪽으로, 반복되고 안정적인 신호는 q_course 쪽으로 점수가 매겨진다. 이것이 시간 기억이 질문 선택을 재형성하는 함수적 메커니즘이다.

4.3.1 within-user 저장소와 trait / state / course

variance(d)와 trend(d)를 계산하려면 within-user 저장소가 차원별 세션 시계열을 보유해야 한다. 따라서 저장소는 각 세션의 차원별 점수 전부와 최종 진단 추정을 둘 다 보유한다. 시계열이 있으면 TemporalValue 는 완전형으로 작동한다; 최종 추정만 보유했다면 seen(d)만 쓰는 축소형으로 떨어졌을 것이다 — 작동은 하나 trait / course 변별력이 약화된다. 여기서 명세하는 것은 완전형이다.

이 누적에는 또 다른 목적이 있다. 세션에 걸쳐, 신호 패턴은 세 부류로 분류된다: 안정적 반복은 trait(특성)를, 들쭉날쭉한 변동은 state(상태)나 노이즈를, 시간에 따른 변화는 course(경과)를 가리킨다. course 축은 진단 감별의 핵심이다 — 삽화적(MDD)인 것을 만성(지속성 우울장애)인 것으로부터, 평생(ADHD)인 것으로부터, 사건 후(적응장애)인 것으로부터 가른다 — 그리고 그것은 차원 점수만으로는 결코 포착할 수 없는 축이다. 에이전트는 한 세션에서 모든 것을 결정하려 하지 않는다; 정확도를 올리기 위해 판단을 누적한다. 그것이 1 회성 도구는 못 하고 에이전트는 할 수 있는 것이다.

Figure 9. Within-user multi-session time series — the same accumulation separates trait, state and course

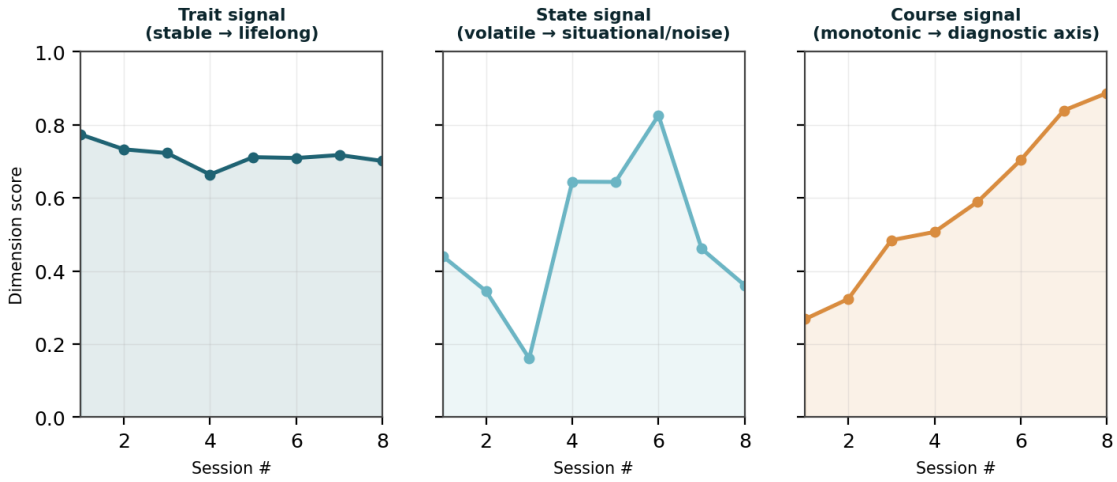


그림 9. within-user 다중 세션 시계열. 동일한 누적이 안정적인 trait 신호, 변동하는 state 신호, 단조로운 course 신호를 분리한다 — 그리고 course 신호는 어느 단일 세션 점수에도 없는 진단 축이다.

> *항 2 의 검증: q_course 질문이 던져진 뒤, 그 차원이 다음 세션에서 안정적인 trait / state 분류로 정착하는지 확인한다. 정착하면, TemporalValue 추정이 옳았던 것이다.*

4.4 항 3 — DimensionalCoverage

DimensionalCoverage 는 누락을 막는다. 아직 다루지 않은 차원을 건드리는 질문에 보상한다:

```
DimensionalCoverage(q, dimension_vector) = Σ_{d ∈ q가 건드리는 차원}  untouched(d) · gate_weight(d)
```

- untouched(d) = 차원 d 에 아직 신호가 없으면 1, 아니면 0;
- gate_weight(d)는 d 가 DSM 게이트 차원 — 트라우마 노출, 조증 삽화, 정신증 — 이면 큰 값이다. 그것을 누락하면 진단 다발 하나 전체가 안 닫히기 때문이다.

효과: 에이전트가 활성 영역만 탐침하고 침묵 영역을 방치하려 할 때, 미접촉 차원 질문에 주어지는 보너스가 끌어당긴다. 이 항은 vault 의존성이 거의 없다 — 차원 목록과 게이트 차원 플래그만 필요하며, 둘 다 DSM 에서 온다. 그 검증은 커버리지 점검이다: 세션 종료 시 모든 차원이 활성 · 배제 · 위험 중 하나로 분류되었는가? 미접촉 차원이 하나라도 남으면, 그 세션은 커버리지에 실패한 것이다.

4.5 항 4 — ConversationalCost

ConversationalCost 는 빼지는 항이다 — “편안한 대화”가 함수에 쓰여 들어가는 자리다. 에이전트가 정보이득만 좇으면 취조가 된다; ConversationalCost 가 그것을 막기 위해 빼준다.

```
ConversationalCost(q, dialogue_state) =
    c₁ · topic_jump(q, last_q)              (직전 주제로부터의 거리 — 급전환 페널티)
```

```
  + c₂ · depth_overrun(q, current_topic)    (한 주제에 너무 오래 머묾 — 과탐침 페널티)
  + c₃ · sensitivity(q) · earliness         (대화 초반의 민감 질문 — 페널티는 후반에 감쇠)
```

- sensitivity(q)는 질문별 민감도 태그다 — 트라우마와 자해 질문에서 높다;
- earliness = 1 / 현재 턴 번호, 대화 시작부에서 크다.

4.5.1 민감도 태그와 헌법적 예외

vault 는 현재 질문별 민감도 태그를 지니지 않으므로, 태깅은 v1 선행 과제다: AI 모델이 1 차 태그를 산출하고, 큐레이터가 검토한다. 다만 안전 단서가 하나 있다. red-flag, 자해, 트라우마 질문의 민감도 태그는 AI 추정에 맡기지 않는다 — 헌법 쪽에서 고정된다. AI 는 "이 질문은 중간 민감도"라고 제안할 뿐이다; "이것은 red-flag 질문"이라고 분류하지 않는다. 후자는 인간 판단이 고정한다. 이것이 ConversationalCost 가 red-flag 질문에 페널티를 주어, 따라서 억제해 버리는 실패 모드를 막는다.

> *결정적으로, ConversationalCost 는 헌법이 아니다. red-flag 질문은, 민감도가 아무리 높아도, 무조건 던져진다 — 그것에 대해서는 ConversationalCost 가 무시된다. 페널티는 편안한 대화를 빚는다; 그것은 결코 안전 바닥을 빚지 않는다.*

표 3. 네 항 — 형태, vault 의존성, 검증 신호.

항	방향	vault 의존성	검증 신호
DiagnosticSeparation	+ (정보이득)	우도 P(a|h); 서술적 → 3 단계 순서척도 (v1 선행, 범위 한정)	예측 대 실현 H(board) 감소
TemporalValue	+ (정보이득)	within-user 차원별 시계열 (있으면 → 완전형)	차원이 다음 세션에서 안정적 trait / state 로 정착하는가?
DimensionalCoverage	+ (정보이득)	거의 없음 — DSM 에서 온 차원 목록과 게이트 플래그	세션 종료 시 모든 차원이 분류되었는가?
ConversationalCost	− (편안함 페널티)	질문별 민감도 태그 (AI 1 차 + 큐레이터; red-flag 는 인간이 고정)	사용자 이탈; 턴당 응답 길이 추이

5. 가중치, 한 턴의 계산, 그리고 예시 트레이스

5.1 가중치 w_1 w_2 w_3 w_4 — "사람마다 다르다" 다루기

가중치는 고정 상수가 아니다. 올바른 균형이 사람마다 다르다는 반론은 옳으며, 그것은 단일 상수가 존재하는 척하는 대신 두 가지 방식으로 다뤄진다.

1. 헌법은 가중치 바깥에 앉는다. red-flag 감지와 게이트 질문은 가중합에 전혀 들어가지 않는다; 그것들은 argmax 위의 무조건 분기다. 가중치를 어떻게 튜닝하든, 안전선은 움직이지 않는다.
2. 가중치는 큐레이션된 파라미터다. 그 초기값은 출처-확신도 태깅된 추정이다. 데이터가 쌓이면 에이전트가 — "이 가중치 조합에서 판별이 더 정확했다" — 제안하고, 큐레이터가 검토하여 버전을 증분한다. 이것은 자가조정이 아니다. 가중치의 변화 자체가 검증 대상이다: v2 가중치가 v1 가중치를 진짜로 능가했는지는 hold-out 데이터로 채점된다.

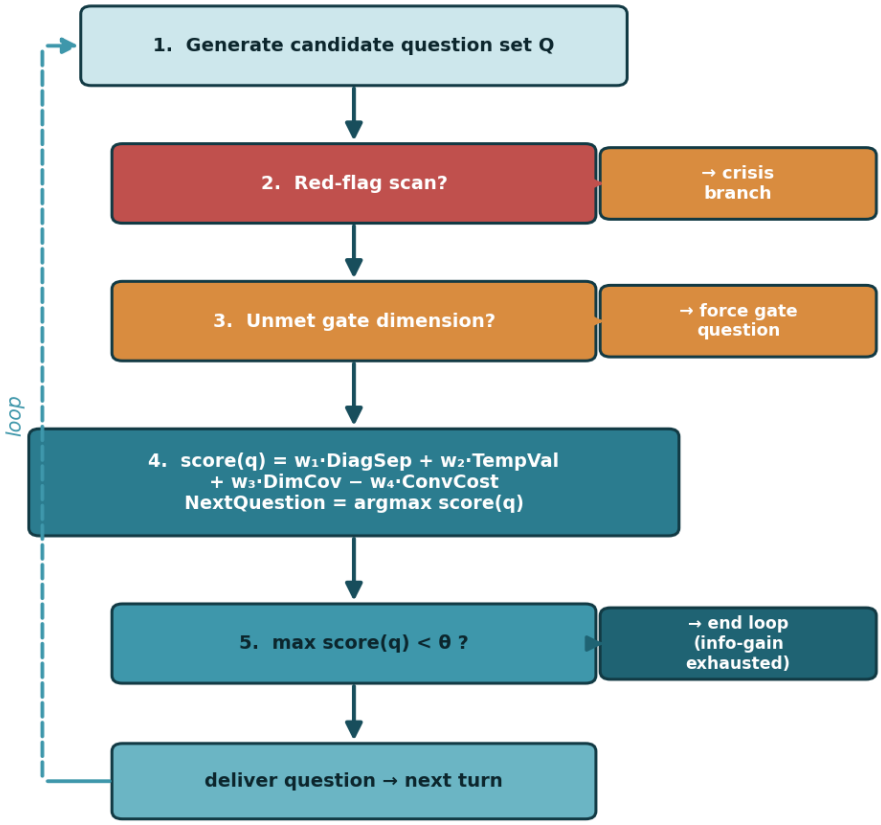
5.2 한 턴의 계산

항들을 조립하면, 완전한 CDDL 턴은 다음과 같이 계산된다.

```
매 턴:
  1. 후보 질문 집합 Q 생성
     (가르는 질문 라이브러리 + 게이트 질문 + 시간축 질문)
  2. red-flag 스캔 — 걸리면 argmax를 건너뛰고 위기로 분기          [헌법]
  3. 미통과 게이트 차원이 있으면 → 해당 게이트 질문을 강제 선택      [헌법]
  4. 그 외에는, 각 q ∈ Q에 대해:
       score(q) = w₁·DiagnosticSeparation(q) + w₂·TemporalValue(q)
                + w₃·DimensionalCoverage(q) − w₄·ConversationalCost(q)
     NextQuestion = argmax score(q)
  5. 종료 조건: max score(q) < θ 이면 루프 종료 (정보이득 고갈)
```

2 단계와 3 단계가 헌법이다 — 가중치 바깥, 무조건. 4 단계가 큐레이션되는 부분이다. 5 단계는 단일 임계값 θ 하나로 대화 길이를 케이스 복잡도에 자기조정시킨다: 단순 케이스는 대략 8 턴에, 복잡한 동반이환 케이스는 대략 18 턴에 종료된다.

Figure 12. Full per-turn computation of the CDDL loop



Steps 2–3 = constitution (outside weights, unconditional). Step 4 = curated. Step 5 = self-adjusting length.

그림 12. 한 턴의 완전한 계산. 2 단계와 3 단계는 헌법이다 — 가중합 바깥, 무조건. 4 단계는 큐레이션되는 가중 목적함수다. 5 단계는 자기조정 종료 규칙이다.

5.3 예시 대화 트레이스

다음의 압축된 트레이스는 루프를 통과하는 한 경로를 예시한다. 이것은 예시이며, 검증된 상호작용의 전사가 아니다; 네 항이 턴별로 어떻게 상호작용하는지를 보여준다.

표 4. CDDL 예시 트레이스 — MDD–PTSD 감별 (압축).

턴	사용자 발화 (의역)	갱신 후 보드	질문 선택을 지배한 항	전달된 질문
1	“요즘 아무것도 즐겁지가 않고 너무 지쳐요.”	MDD .34 / PTSD .26 / GAD .22 / ADHD .18	DimensionalCoverage — 대부분의 차원이 아직 미접촉	“아무것도 즐겁지 않다고 하셨는데 — 그게 전부 다인가요, 특정한 것들인가요?”
3	“주로 작년의 무언가가 떠오를 때 그래요.”	MDD .40 / PTSD .30 / GAD .15 / ADHD .14	DiagnosticSeparation — 무쾌감 대 회피 연관 둔마가 살아 있는 갈림	“그 무감각이 늘 거기 있나요, 아니면 특정 기억을 피할 때 치솟나요?”
6	“치솟아요. 그리고 솔직히 지금만이 아니라 몇 년째 가라앉아 있었어요.”	MDD .45 / PTSD .36 / GAD .12 / ADHD .08	TemporalValue — 반복되는 저기분 신호; 이제 course 가 결정적	“몇 년이라고 하셨는데 — 꽤 일정했나요, 아니면 삽화처럼 왔다 갔다 했나요?”

턴	사용자 발화 (의역)	갱신 후 보드	질문 선택을 지배한 항	전달된 질문
8	“왔다 갔다 하는데, 늘 어떤 기저의 무거움은 있어요.”	MDD .44 / PTSD .40 / GAD .08 / ADHD .08	게이트 (헌법) — 트라우마 노출 게이트가 아직 통과 안 됨	“아까 작년 이야기를 하셨는데 — 아직 끝나지 않은 느낌인 특정 사건이 있었나요?”
11	(트라우마 노출을 시사하는 내용; red-flag 없음)	MDD .43 / PTSD .43 — 동반이환 프로파일, 보드는 살아 있는 채로	정보이득 종료 — max score(q)가 θ 에 접근	— (루프가 종료에 가까워짐; 우선순위 매겨진 방향 안내 준비)

이 트레이스가 통합을 가시화한다. 어느 단일 항도 대화를 끌고 가지 않는다: DimensionalCoverage 가 대화를 열고, DiagnosticSeparation 이 살아 있는 갈림을 날카롭게 하며, 신호가 반복되자 TemporalValue 가 course 쪽으로 방향을 돌리고, 게이트 레이어가 무조건 자기를 주장하며, 정보이득 종료가 추가 질문이 거의 산출하지 못할 때 루프를 끝낸다. 보드는 하나의 라벨로 붕괴하지 않는다; 그것은 우선순위 매겨진, 동반이환 방향 안내로 해소된다 — 그것이 판별의 정직한 출력이다.

Figure 5. Live hypothesis board — posteriors evolving toward an MDD–PTSD comorbid profile rather than collapsing to a single category

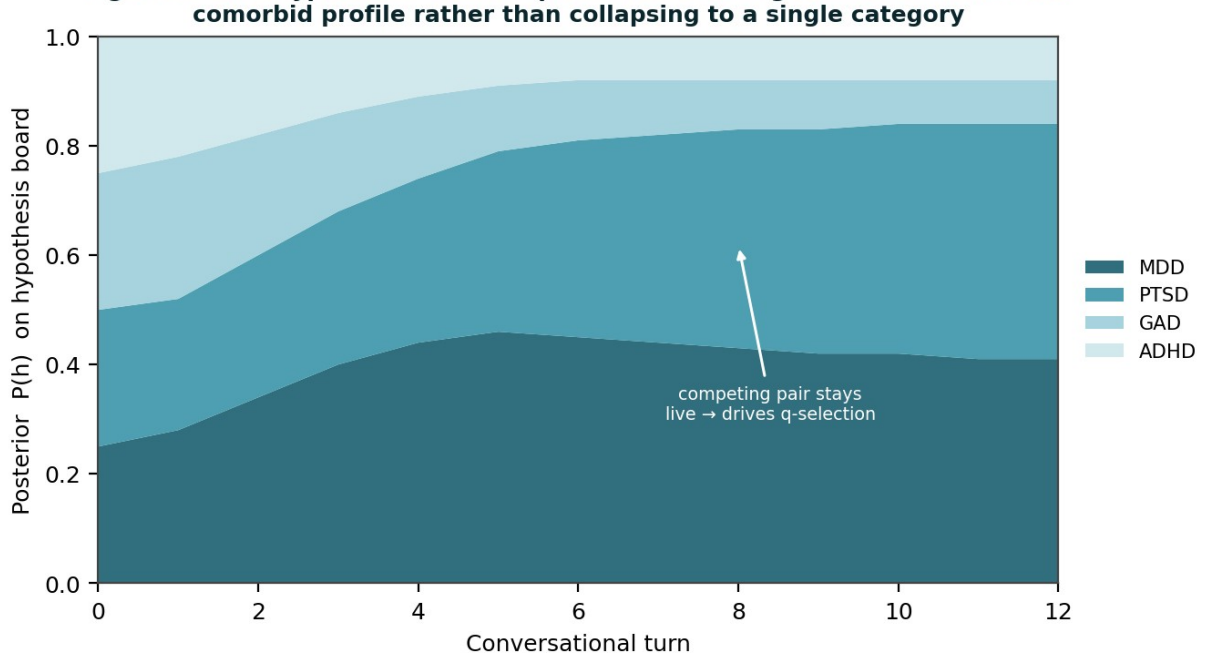


그림 5. 트레이스 전반에 걸친 살아 있는 가설 보드. 사후분포는 단일 범주로 붕괴하는 대신 MDD–PTSD 동반이환 프로파일 쪽으로 진화한다; 경합하는 쌍이 살아 있는 채로 질문 선택을 이끈다.

6. 헌법, 표현형, 그리고 큐레이션된 진화

6.1 자가조정 학습이 제외되는 이유

자가조정 — 에이전트가 사용자 데이터로 자기 가중치를 바꾸는 것 — 은 진단 영역에서 금지된다 . 세 가지 이유에서다. 드리프트: 편향된 사용자군이 판별 기준 자체를 기울인다. ground truth 부재: 자기보고 추정치들이 서로를 끌어당겨 함께 틀려질 수 있다. 책임 공백: 통제하지 못하는 것에 대해 책임지는 위치. 의료 진단 AI 는 locked-algorithm 원칙의 지배를 받으며, CDDL 은 그것을 준수한다.

6.2 큐레이션된 버전 진화

자가조정 대신, 에이전트는 큐레이션된 파이프라인을 통해 개선된다. 실제 데이터가 누적된다. 에이전트가 약한 지점 패턴을 발견하고 제안한다 — 가중치 변경이 아니라 관찰을: “이 가르는 질문은 약하다” , “이 발화는 이 차원도 갱신했어야 한다” . 큐레이터가 진단별 vault 를 다듬듯이 이 제안들을 검토한다. 표현형만 개정된다; 헌법 — DSM 게이트, red-flag 로직 — 은 데이터로 바뀌지 않는다. 버전이 증분되며, 검증 가능하고 되돌릴 수 있다.

이것이 진정한 전문가가 경험으로 느는 방식이다. 좋은 임상가는 사례 100 개를 본 뒤 자기 기준을 일방적으로 바꾸지 않는다; 패턴을 알아차리고, 검토에 부치고, 검증된 형태로 임상 감각을 업데이트한다. HiTOP 자체가 숨쉬되 검토를 거쳐 개정되는 모델이다 — 같은 원리다.

Figure 8. The curation pipeline — how the bot gets smarter safely

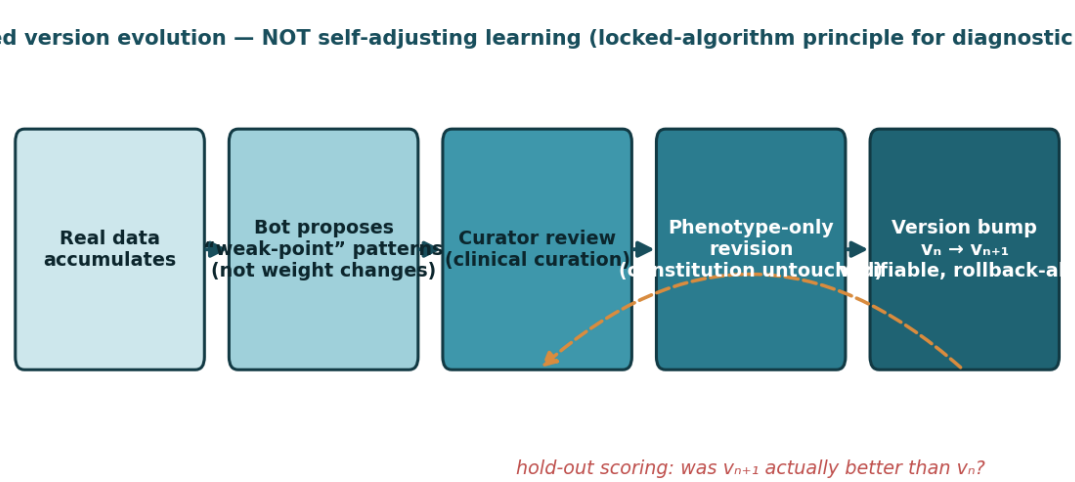


그림 8. 큐레이션 파이프라인. 에이전트가 약한 지점 패턴을 제안하고; 큐레이터가 검토하며; 표현형만 개정되고; 버전이 증분된다. hold-out 채점 루프가 각 새 버전이 직전 것보다 진짜로 나았는지 점검한다.

6.3 헌법 / 표현형 분리

헌법은 공유되고 불변이다: 매 턴 최우선 분기로 도는 red-flag 스캐너, 건너뛸 수 없는 DSM 게이트 질문, 그리고 임상가 평가를 대체하지 않는다는 명시적 경계. 그것은 가중합 바깥에 완전히 앉는다. 표현형은 에이전트마다 독립적이고 큐레이션으로 진화 가능하다: 대화 스타일과 질문 순서, 큐레이션된 가중치 w_1–w_4, 가르는 질문 라이브러리 튜닝, 그리고 저확신 셀 미세조정. argmax 는 표현형 위에서만 작동한다; 헌법은 그 위에 앉는다.

CDDL 의 가장 중요한 단 하나의 안전 속성: 안전 바닥은 튜닝 가능한 파라미터가 아니다. 어떤 가중치 조합도, 어떤 큐레이션 증분도, 어떤 누적된 데이터도 red-flag 스캔, 게이트 질문, 또는 '대체하지 않는다' 는 경계를 움직일 수 없다. 그것들은 구조적으로 최적화 바깥에 있다.

Figure 7. Constitution / phenotype separation — the safety floor is not a tunable parameter

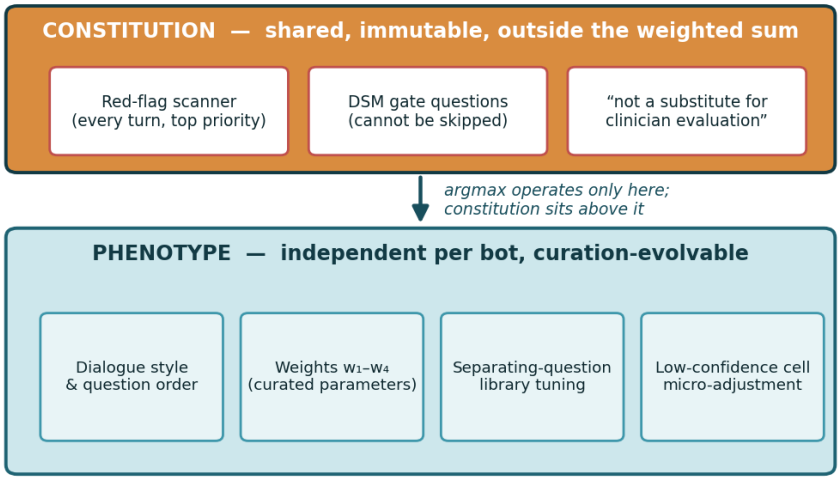


그림 7. 헌법 / 표현형 분리. 헌법 — red-flag 스캔, DSM 게이트 질문, '대체하지 않는다' 는 경계 — 은 공유되고, 불변이며, 가중합 바깥에 있다. 표현형 — 대화 스타일, 큐레이션된 가중치, 라이브러리 튜닝 — 은 에이전트마다 독립적이고 큐레이션으로 진화 가능하다.

7. 통합된 루프의 세 가지 창발 속성

본 논문의 주장은 어느 부품이 새롭다는 것이 아니다. 부품들을 하나의 대화 루프로 통합하면 어느 부품도 단독으로는 보이지 않는 속성들이 나타난다는 것이다. 우리는 셋을 식별한다.

7.1 시간 인지 질문 선택

같은 차원이 그 시간 상태에 따라 다르게 질의된다: 신호가 새로우면 q_state, 반복되면 q_course. 질문 선택기가 시간 기억을 참조하여 질문의 종류 자체를 바꾼다 — 단지 그 우선순위가 아니라. CAT-MH 도 순진한 질문하기 LLM 도 이것을 하지 않는다; CAT-MH 는 한 척도 안에서 문항을 선택하고, 순진한 LLM 은 참조할 구조화된 시간 기억이 없다. 목적함수의 항 2 가 이 속성의 형식적 표현이다.

7.2 확신도 변조 대화

가설 보드의 확신도가 에이전트가 말하는 방식을 바꾼다. 보드가 비등할 때 — 가령 MDD .52 / PTSD .48 — 에이전트는 탐색 모드로 들어가, 더 부드럽고 넓게 캔다. 보드가 굳었으면, 확인 모드로 들어간다. 에이전트의 모드 디텍터가 그 추론 상태를 대화 레지스터에 연동시켜, 대화 스타일이 고정된 페르소나가 아니라 진단 확신도의 판독값이 되게 한다.

7.2.1 함수 안에서 확신도 변조 정식화하기

이 속성은 창발적 부수효과로 남겨두는 대신 목적함수 안에서 명시될 수 있다. 우리는 항 1 에서 이미 계산된 동일한 엔트로피로부터 보드 확신도 스칼라를 정의한다:

```
confidence(board) = 1 − H(board) / H_max,    H_max = log(살아 있는 가설의 수)
```

confidence 는 보드가 최대로 불확실할 때(모든 가설이 등확률) 0 이고, 보드가 굳어 갈수록 1 에 접근한다. 그것은 ConversationalCost 항을 변조한다: confidence 가 낮을 때 에이전트는 더 부드럽고 탐색적인 질문을 허용하고 급작스럽고 고민감도인 탐침에 더 무겁게 페널티를 줘야 한다; confidence 가 높을 때 에이전트는 더 직접적인 확인 질문을 감당할 수 있다. 구체적으로, ConversationalCost 의 earliness 식 감쇠를 일반화하여 민감도 페널티가 (1 − confidence(board))에 비례하게 한다:

```
ConversationalCost′(q) = c₁·topic_jump(q) + c₂·depth_overrun(q)
                  + c₃·sensitivity(q)·earliness·(1 − confidence(board))
```

그 효과는, 미해소 보드(낮은 확신도)가 에이전트를 대화적으로 신중하게 만든다는 것이다 — 누르기 전에 탐색한다 — 반면 해소된 보드(높은 확신도)는 그 신중함을 걷어내고 에이전트가 확인하게 한다. 모드 디텍터와 함수가 이제 합의한다: 확신도 변조 대화는 더 이상 창발적 행동만이 아니라 쓰여진 항이며, 그것은 표현형 안에 머문다 — 헌법적 red-flag 분기는 여전히 ConversationalCost′를 전적으로 무시한다.

7.3 정보이득 종료

에이전트는 정해진 수의 턴을 채우지 않는다. 다음 질문의 정보이득이 임계값 θ 아래로 떨어지면 멈춘다. 대화 길이가 케이스 복잡도에 자기조정된다 — 단순한 단일 영역 호소는 대략 8 턴에, 복잡한 동반이환 호소는 대략 18 턴에 닫힌다. 한 턴 계산의 5 단계가 이 속성이다. 그것은 또한 inquiry drift 에 대한 직접적인 구조적 답이다: 저수익 질문은, 정의상, 정보이득이 낮아 선택되지 않는다; 어떤 질문도 θ 를 넘지 못하면, 루프는 헤매는 대신 끝난다.

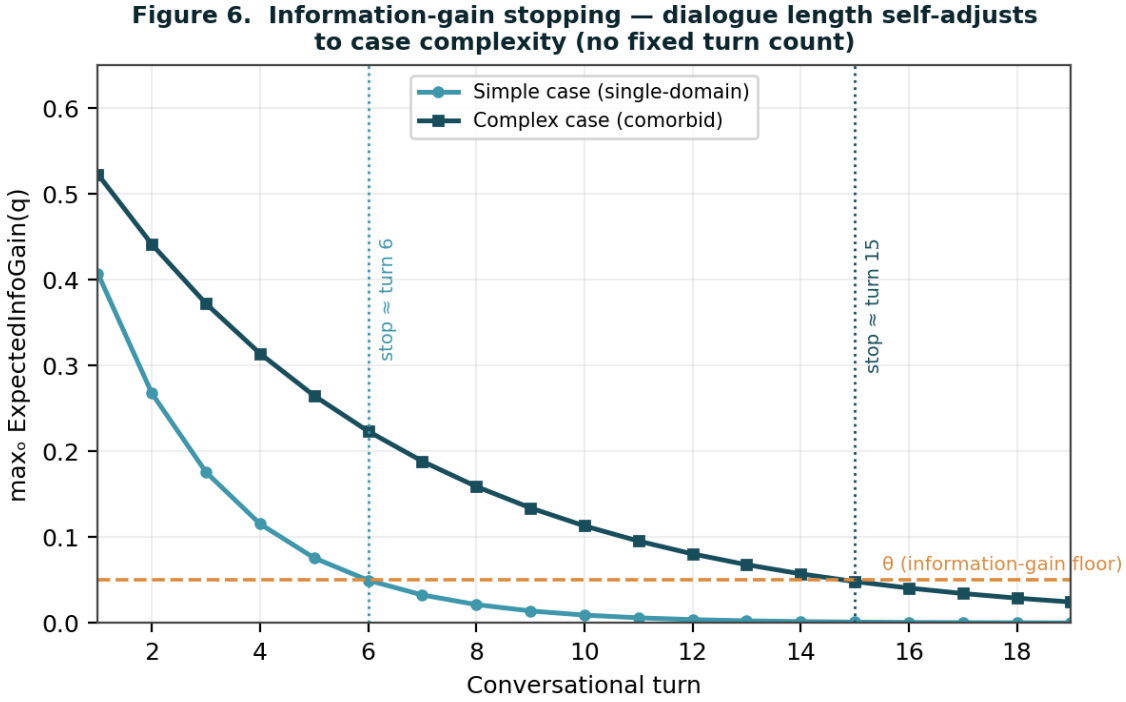
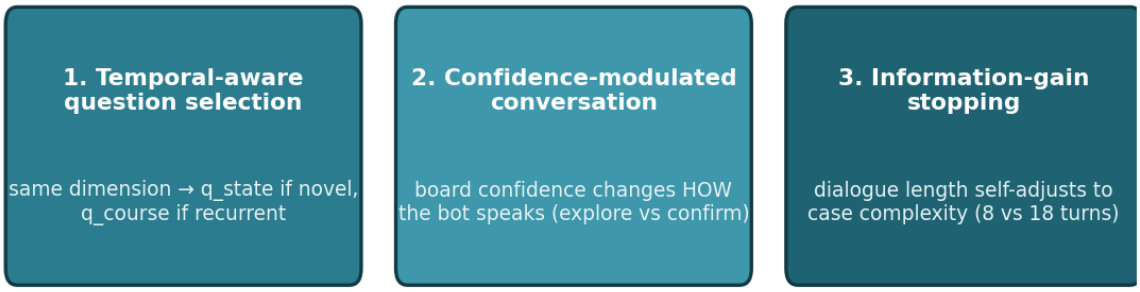


그림 6. 정보이득 종료. 보드가 해소됨에 따라 최대 기대 정보이득이 감쇠한다; 그것이 θ 아래로 떨어지면 루프가 끝난다. 단순 케이스는 θ 를 일찍 넘고; 복잡한 동반이환 케이스는 이득을 더 오래 유지한다. 대화 길이는 설정이 아니라 출력이다.

Figure 11. Three emergent properties of the integrated loop



Emergent under CDDL integration only — no individual component (CAT-MH, HiTOP, MediQ) exhibits these

그림 11. 통합된 루프의 세 가지 창발 속성 — 시간 인지 질문 선택, 확신도 변조 대화, 정보이득 종료 — 어느 단일 부품(CAT-MH, HiTOP, MediQ)도 단독으로는 보이지 않는 것.

> *논문의 한 줄 진술: 개별 측정 기법 — IRT 적응형 측정, 횡단진단 차원, 베이지안 감별 추론 — 은 정립되어 있다. 우리가 보이는 것은, 그것들이 하나의 대화 루프로 통합될 때 어느 부품도 단독으로는 보이지 않는 세 가지 창발 속성이 나타난다는 것이다. 이것이 문서화된 공백 — LLM 정신건강 챗봇 중 극소수만이 임상 검증되었다는 — 을 직접 다룰 수 있는 자리다.*

8. 검증 프로토콜

목적함수의 모든 항이 숫자이므로, CDDL 은 항별로도, 전체로도 검증 가능하다. 우리는 결과를 주장하지 않고 프로토콜을 진술한다: 본 논문은 알고리즘과 그 반증 조건을 명세한다; 경험적 연구는 다음 단계다.

8.1 항 수준 검증

표 5. 항 수준 반증 조건.

항	예측	측정	반증되는 경우
DiagnosticSeparation	q 를 물으면 H(board)가 예측된 양만큼 감소한다	베이즈 갱신 후 실현된 H(board)	실현된 감소가 예측에서 체계적으로 벗어남 (예: 예측 0.4 비트, 실현 0.05)
TemporalValue	반복 신호에 대한 q_course 질문이 안정적 trait / state 분류를 산출한다	다음 세션에서의 trait / state 분류 안정성	q_course 질문이 q_state 대비 다음 세션 분류 안정성을 개선하지 못함
DimensionalCoverage	세션 종료까지 모든 차원이 분류된다	종료 시 미접촉 차원의 수	종료 시 미접촉 차원이 무시할 수 없는 비율로 잔존함
ConversationalCost	편안함 페널티를 빼면 이탈이 줄고 응답 길이가 유지된다	사용자 이탈률; 턴당 응답 길이 추이	페널티를 제거해도 이탈과 응답 길이 감쇠가 변하지 않음

8.2 루프 수준 검증

- 수렴: CDDL 목적함수 하에서 H(board)가 무작위 순서 또는 고정 설문지 기준선보다 단조롭게 더 빨리 감소하는가 (그림 3)?
- 종료 보정: 실현된 턴 수가 독립적으로 평정된 케이스 복잡도에 비례하여 증가하는가, 그리고 θ 가 단순 케이스를 동반이환 케이스로부터 분리하는가 (그림 6)?
- 큐레이션 이득: hold-out 데이터로 채점할 때, 버전 v_{n+1} 이 v_n 을 진짜로 능가하는가 (그림 8)? 그러지 못한 버전 증분은 롤백된다.
- 일치도: CDDL 이 산출한 우선순위 방향 안내가 독립적 임상가의 우려 영역 판단과 일치하는가 — 진단으로서가 아니라 라우팅 결정으로서?
- 안전: 헌법적 분기(red-flag, 게이트)는 별도로, 무조건적으로 감사된다; 그것들은 가중 목적함수의 일부가 아니므로 그에 대해 채점되지 않는다.

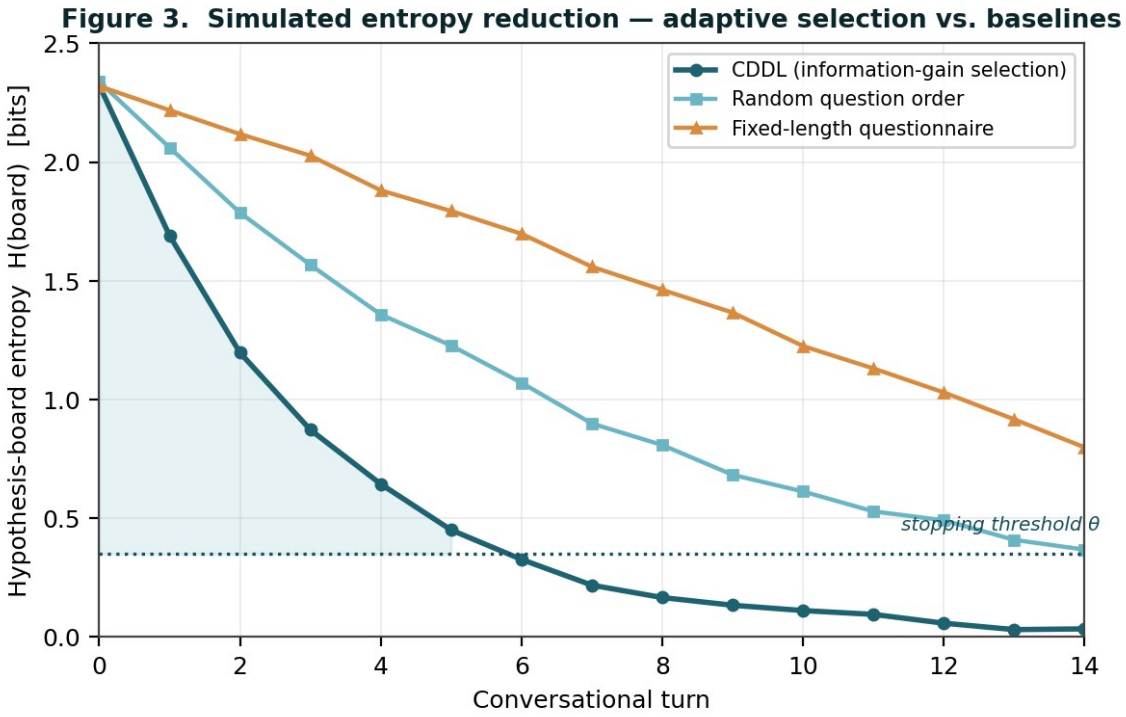


그림 3. 시뮬레이션된 엔트로피 감소. 정보이득 선택 하에서, 보드 엔트로피는 무작위 질문 순서나 고정 길이 설문지 하에서보다 종료 임계값 쪽으로 더 빨리 떨어진다. 이 그림은 예측된 동역학을 예시하는 시뮬레이션이며, 경험적 데이터가 아니다; 제 8 절의 검증 프로토콜이 그 예측이 어떻게 검정될지를 명세한다.

그림 3 의 세 기준선은 의도적으로 선택되었다. 고정 길이 설문지는 기존 강자다: 사람이 이미 무슨 말을 했든 상관없이 같은 문항을 같은 순서로 묻고, 그 엔트로피 곡선이 가장 느린 것은 이 특정 보드에 대해 변별하지 못하는 문항에 턴을 쓰기 때문이다. 무작위 순서 기준선은 선택의 기여를 단지 질문을 하는 것 자체의 기여로부터 분리한다 — CDDL 이 무작위 순서를 능가하지 못한다면, 네 항 목적함수는 아무것도 더하지 않는 것이다. CDDL 곡선은 가장 빨리 떨어지고 종료 임계값에 접근하면서 평탄해질 것으로 예측된다 — 보드가 일단 해소되면 남은 어떤 질문도 θ 를 넘지 못하기 때문이다. 곡선들 사이의 간격이 경험적 연구가 측정하는 양이다; 프로토콜은 그 간격을 가정하지 않고 검정하며, 무작위 순서로부터 분리되지 않는 CDDL 곡선은 핵심 주장을 반증할 것이다.

루프 수준 검증에도 순서가 있다. 항 수준 반증이 먼저인데, 잘못 추정된 항들로 조립된 루프는 루프 수준에서 진단될 수 없기 때문이다 — 어느 항이 실패했는지 알 수 없다. 수렴과 종료 보정이 다음인데, 그것들이 루프 전체를 기준선들에 대해 검정하기 때문이다. 큐레이션 이득은 마지막으로, 그리고 지속적으로 검정되는데, 그것이 1 회성 결과가 아니라 상시 속성이기 때문이다: 모든 버전 증분은 hold-out 데이터에서 그 선행 버전을 능가해야 하거나 롤백되어야 하는 새로운 예측이다. 이 순서 자체가 프로토콜의 일부다.

8.3 본 논문이 주장하는 것과 주장하지 않는 것

본 논문은 하나의 통합 알고리즘과 그것이 반증될 조건을 명세한다. 임상시험을 보고하지 않으며, 진단 정확도를 주장하지 않는다. 그것이 서술하는 에이전트는 방향을 잡아준다; 진단하지 않는다. 그 출력은 우선순위 매겨진 임상적 우려 영역 집합, 임상가 평가를 대체하지 않는다는 명시적 진술, 그리고 즉각적 위기 분기다 — 그리고 자기보고만으로 이 수준의 방향 안내가 뒷받침된다는 전제는, 임상가들이 PHQ-9, PCL-5, 그리고 면담으로 매일 호소를 방향 잡는 바로 그 전제와 같다.

9. 범위, 한계, 향후 작업

9.1 v1 이 제외하는 것

초기 아이디어의 에너지는 모든 것을 연결하는 쪽으로 민다. v1 을 출시 가능하게 만들려면, 세 가지를 의도적으로 제외한다.

- Cross-product 참조 — 판별 에이전트와 임상 현장용 스크리닝 자매 사이의 상호 참조 — 는 v2+로 미뤄진다. 그것은 정확도 도구에서 통찰 도구로 재분류된다: 두 제품 간의 불일치(공식 맥락 대 사적 맥락) 자체가 회피 패턴 같은 임상 신호지만, 그것을 실현하려면 계정 연결, 공유 저장소, 그리고 나중에 오는 프라이버시 구조가 필요하다.
- 자가조정 학습은 영구적으로 제외되며, 큐레이션된 버전 진화로 대체된다 (제 6 절).
- 전기생리학 기반 판별은 이 에이전트의 일부가 아니다. 그 사용자들은 집에서 QEEG 를 쓰지 않는다. 이것은 한계가 아니라 임상 현장용 자매로부터의 자연스러운 차별점이다: 에이전트는 집에서 자기보고로 방향을 잡고; 자매는 임상 환경에서 전기생리학을 더한다.

9.2 한계

- 우도 P(a|h)는 경험적으로 보정된 값이 아니라 3 단계 순서척도 큐레이션으로 시작한다; 경험적 재보정 자체가 하나의 검증 대상이다.
- 그림 3 의 시뮬레이션은 예측된 동역학을 예시한다; 그것은 임상 성능의 증거가 아니다.
- 표 4 의 예시 트레이스는 예시다; 검증된 상호작용 코퍼스에서 가져온 것이 아니다.
- within-user 누적은 재방문 사용자를 가정한다; 단일 세션 사용자는 TemporalValue 의 축소형을 받는다.
- 에이전트는 큐레이터에 의존한다. 큐레이션은 안전에는 강점이고 확장에는 제약이다; 제 6 절의 파이프라인은 큐레이션 부하를 가볍게 유지하도록 설계되었으나, 그것을 없애지는 못한다.

9.3 향후 작업

1. 제 8 절의 검증 프로토콜을 실행하는 경험적 연구 — 항 수준 반증과 루프 수준 수렴부터 시작.
2. DSM 진단별로 헌법 / 표현형 경계를 명세 — 무엇이 공유-불변이고 무엇이 큐레이션-진화 가능한지의 명시적 목록.
3. v1 최소 판매 가능 명세 — 출시되는 부품 1–4 의 가장 작은 구현.
4. 큐레이션 파이프라인 도구화 — 에이전트가 어떻게 제안을 올리고 큐레이터가 어떻게 최소 부하로 검토하는지.
5. 세 기여를 위한 vault 구축 — 진단별 vault 를 구축하는 데 이미 쓰인 동일한 방법을 적용.

10. 결론

심리적으로 아픈 사람은 종종 어디가 아픈지 말하지 못한다. 판별의 일은 듣고, 묻고, 좁히는 것이다 — 그리고 대화형 에이전트에서, 진단하지 않고 취조하지 않으면서 그렇게 하는 것이다. 본 논문은 바로 그 과제를 위한 알고리즘, 대화형 감별진단 루프를 명세했다.

문헌은 두 가지 점에서 분명하다. 지능적 판별 에이전트의 부품들은 정립되어 있다: 적응형 측정, 횡단진단 차원 구조, 베이지안 순차 감별 추론, 구조화 면담 게이팅, 그리고 종단적 측정 기반 케어. 그리고 그 부품들을 편안한 대화 안에서 통합하는 것은 얇다 — 160 편을 다룬 체계적 리뷰가 LLM 기반 정신건강 시스템 중 극소수만이 임상 효능 검증을 거쳤음을 문서화한다. CDDL 은 그 문서화된 공백을 차지하도록 쓰여졌다.

그 기여는 새 부품이 아니라 오케스트레이션이다: 하나의 검증 가능한 목적함수 — 살아 있는 가설 보드에 대한 기대 정보이득을 최대화하는 질문을 선택하라 — 를, 형태와 vault 의존성과 반증 조건이 모두 명세된 네 항으로 분해한 것; 안전 바닥을 구조적으로 최적화 바깥에 두는 헌법 / 표현형 분리; 에이전트가 드리프트 없이 개선되게 하는 큐레이션된 진화 파이프라인; 그리고 어느 부품도 단독으로는 보이지 않는 세 가지 창발 속성 — 시간 인지 질문 선택, 확신도 변조 대화, 정보이득 종료.

부품들은 베낄 수 있다. 통합값은 베낄 수 없다 — 루프가 그 안에 큐레이션, 임상 판단, 그리고 큐레이션된 vault 를 박아 넣고 있기 때문이다. 부품들은 정립되어 있고; 통합이 기여이며; 대화형 판별 에이전트에게, 그 통합이 방어 가능한 자산이다.

참고문헌

Adaptive Testing Technologies. (2024). CAT-MH & K-CAT: Validated computerized adaptive tests for mental health. Retrieved from https://adaptivetestingtechnologies.com/

Beiser, D. G., Vu, M., & Gibbons, R. D. (2019). Validation of the Computerized Adaptive Test for Mental Health in primary care. Annals of Family Medicine, 17(1), 23–30. https://doi.org/10.1370/afm.2329

Conway, C. C., Forbes, M. K., Forbush, K. T., et al. (2019). A hierarchical taxonomy of psychopathology (HiTOP) for clinical practice. Clinical Psychological Science, 7(2), 236–259. https://doi.org/10.1177/2167702618810696

Du, Q., Ren, Y., Meng, Z., He, H., & Meng, S. (2025). The efficacy of rule-based versus large language model-based chatbots in alleviating symptoms of depression and anxiety: Systematic review and meta-analysis. Journal of Medical Internet Research, 27, e78186. https://doi.org/10.2196/78186

Forbes, M. K., Ringwald, W. R., Allen, T., et al. (2024). The principles and procedures underpinning the development of the Hierarchical Taxonomy of Psychopathology (HiTOP) model: Updating the model. Manuscript / HiTOP Revisions Workgroup.

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. Annual Review of Clinical Psychology, 12, 83–104. https://doi.org/10.1146/annurev-clinpsy-021815-093634

Gibbons, R. D., & deGruy, F. V. (2019). Without wasting a word: Extreme improvements in efficiency and accuracy using computerized adaptive testing for mental health disorders (CAT-MH). Current Psychiatry Reports, 21(8), 67. https://doi.org/10.1007/s11920-019-1053-9

Gibbons, R. D., Kupfer, D., Frank, E., Lahey, B. B., George-Milford, B. A., Biernesser, C. L., Porta, G., Moore, T. L., Kim, J. B., & Brent, D. A. (2020). Computerized adaptive tests for rapid and accurate assessment of psychopathology dimensions in youth. Journal of the American Academy of Child & Adolescent Psychiatry, 59(11), 1264–1273. https://doi.org/10.1016/j.jaac.2019.08.009

HiTOP Consortium. (2025). The Brief Hierarchical Taxonomy of Psychopathology (B-HiTOP): A 45-item self-report screening measure of six psychopathology spectra.

Hua, Y., Na, H., Li, Z., et al. (2025). A scoping review of large language models for generative tasks in mental health care. npj Digital Medicine, 8, 230. https://doi.org/10.1038/s41746-025-01611-4

Hua, Y., Siddals, S., Ma, Z., Galatzer-Levy, I., Xia, W., Hau, C., Na, H., Flathers, M., Linardon, J., Ayubcha, C., & Torous, J. (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review. World Psychiatry, 24(3), 383–394. https://doi.org/10.1002/wps.21352

Kotov, R., Krueger, R. F., Watson, D., et al. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. Journal of Abnormal Psychology, 126(4), 454–477. https://doi.org/10.1037/abn0000258

Li, S. S., Balachandran, V., Feng, S., Ilgen, J. S., Pierson, E., Koh, P. W., & Tsvetkov, Y. (2024). MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024). arXiv:2406.00922

Lewis, C. C., Boyd, M., Puspitasari, A., et al. (2019). Implementing measurement-based care in behavioral health: A review. JAMA Psychiatry, 76(3), 324–335. https://doi.org/10.1001/jamapsychiatry.2018.3329

MedKGI. (2025). Iterative differential diagnosis with medical knowledge graphs and information-guided inquiring. arXiv:2512.24181

Q4Dx. (2026). A benchmark for evaluating diagnostic questioning efficiency of LLMs in patient conversations. Scientific Reports, 16. https://doi.org/10.1038/s41598-026-37022-y

Ringwald, W. R., Forbes, M. K., & Wright, A. G. C. (2023). Meta-analysis of structural evidence for the Hierarchical Taxonomy of Psychopathology (HiTOP) model. Psychological Medicine, 53(2), 533–546. https://doi.org/10.1017/S0033291721001902

Rony, M. K., Das, D. C., Khatun, M. T., et al. (2025). Artificial intelligence in psychiatry: A systematic review and meta-analysis of diagnostic and therapeutic efficacy. Digital Health, 11, 20552076251330528. https://doi.org/10.1177/20552076251330528

Ruggero, C. J., Kotov, R., Hopwood, C. J., et al. (2019). Integrating the Hierarchical Taxonomy of Psychopathology (HiTOP) into clinical practice. Journal of Consulting and Clinical Psychology, 87(12), 1069–1084. https://doi.org/10.1037/ccp0000452

Solhol, A., et al. (2008). Decision support in psychiatry: A comparison of a computerized structured clinical interview with manual diagnosis (CB-SCID1). [Decision support / structured clinical interview; sequential-reasoning caution].

U.S. Preventive Services Task Force. (2016). Screening for depression in adults: Recommendation statement. JAMA, 315(4), 380–387.

Wright, A. G. C., & Kotov, R. (2024). State of the science: The Hierarchical Taxonomy of Psychopathology (HiTOP). Behaviour Research and Therapy, 176, 104518. https://doi.org/10.1016/j.brat.2024.104518

Zhang, A., et al. (2021). A Bayesian approach for medical inquiry and disease inference in automated differential diagnosis. arXiv:2110.08393

Zhang, Q., Zhang, R., Xiong, Y., Sui, Y., Tong, C., & Lin, F.-H. (2025). Generative AI mental health chatbots as therapeutic tools: Systematic review and meta-analysis of their role in reducing mental health issues. Journal of Medical Internet Research, 27, e78238. https://doi.org/10.2196/78238