



CAIPEX 연구 논문

생성형 AI의 계층형 심리 구조 설계

Protofelix LayerAI 모델의 기술적 이점, 위험, 구현·평가 프레임워크

저자 E(on). Jun

연락처 inwardis@hotmail.com

발간 학회 CAIPEX

발간일 2026.9.

<https://caipex.site/>

생성형 AI의 ‘계층형 심리 구조’ 설계: Protofelix LayerAI 모델의 기술적 이점, 위험, 구현·평가 프레임워크

초록

본 연구는 Protofelix 운영자가 공개한 **LayerAI Center**의 Layer 0~5 모델을 중심 사례로 삼아, 생성형 AI에 이른바 “**계층형 심리(hierarchical psychological structure)**”를 부여할 때 기대할 수 있는 효용과 잠재적 위험을 기술·인지과학·AI 안전·개인정보 관점에서 분석한다. 여기서 가장 중요한 전제는 이 구조를 인간 정신의 복제나 AI의 실제 감정·의식으로 해석하지 않는 것이다. LayerAI 자체도 모델을 “기억·감정·성찰·창작을 층위별로 설계하는 개념 모델”로 정의하는 한편, “**Layer는 의식의 설계도가 아니라 인지 기능의 설계 언어**”라고 명시한다. 또한 감정 층은 감정의 보유가 아니라 신호와 대화 톤의 반영이고, 무의식·초월 같은 표현은 각각 연구적 은유와 창작적 목표로 제한한다. [1]

분석 결과, 이 접근의 실질적 가치는 “AI에게 인간 같은 마음을 만들어 주는 것”보다 **컨텍스트, 정서 신호, 기억, 가설 생성, 성찰·개선, 창작 탐색을 서로 다른 상태와 권한을 가진 모듈로 분해**하는 데 있다. 장기 기억과 성찰을 결합한 기존 생성형 에이전트 연구, 계층형 외부 메모리, 언어 기반 성찰 학습, RAG 연구는 이러한 분해의 기술적 실현 가능성을 부분적으로 뒷받침한다. [2] [13] [38] [58] 반대로 정서 추론과 장기 기억이 결합되면 사용자의 취약성·선호를 정밀하게 추론하는 **심리 프로파일링·개인화 설득** 장치가 될 수 있다. 2025년 통제 실험에서는 제한적인 인구통계 정보로 개인화된 AI 토론 상대가 인간 상대보다 더 높은 설득력을 보였으며, 대규모 챗봇 이용 분석에서는 매우 높은 사용량이 더 높은 자기보고 의존성 지표와 연관되었다. [3] [46]

따라서 바람직한 구현은 **심리 시뮬레이션이 아니라 기능적·감사 가능한 계층 구조**여야 한다. 특히 정서 데이터 최소 수집, 기억의 명시적 동의·열람·삭제, 데이터와 명령의 분리, 외부 행동 전 승인, 모듈별 출처·변경 이력(provenance)와 버전 로그, 온라인 자기수정의 제한이 핵심이다. 이는 LayerAI가 제시하는 동의 기반 기억, 삭제 통제, 인간 승인, 실패·철회 기록, 재현 가능한 평가 원칙과도 일치한다. 다만 운영자 사이트 스스로 현재 상태를 “**검증된 결과 없음**”으로 명시하므로, 본 연구가 평가하는 것은 검증된 AGI 성능이 아니라 공개된 설계 철학과 프로토타입 개념의 잠재성이다. [4]

서론 및 연구 목적

대규모 언어모델은 본질적으로 하나의 연속적인 “인격”을 필요로 하지 않는다. 그러나 실제 생성형 AI 제품은 대화 맥락, 사용자 기억, 선호, 정서적 말투, 검색, 도구 호출, 자기검토, 행동 계획을 한 인터페이스 안에서 결합하기 시작하면서 외부 관찰자에게는 점점 더 지속적인 성격과 심리를 가진 시스템처럼 보일 수 있다. Generative Agents 는 경험을 자연어 기억으로 저장하고 상위 수준의 성찰로 합성한 뒤 행동 계획에 다시 검색하는 구조를 사용했고, 그 연구의 ablation 에서는 관찰·계획·성찰 요소가 각각 에이전트 행동의 인간적 개연성에 기여했다. MemGPT 역시 제한된 컨텍스트 창을 넘어 서로 다른 메모리 계층을 관리함으로써 다중 세션 대화에서 기억·성찰·장기적 상호작용을 구현할 수 있음을 제시했다. [5] [58]

LayerAI Center 는 이와 비슷한 기능들을 인간 심리에서 차용한 명칭 아래 더욱 명시적으로 계층화한다. L0 는 현재 요청·사실·제약·도구 상태, L1 은 감정 신호와 톤, L2 는 승인된 기억, L3 는 잠재 패턴을 활용한 가설, L4 는 피드백을 통한 전략·워크플로 개선, L5 는 창작 후보 탐색을 담당한다. 중요한 점은 운영자가 이를 0→5 의 고정된 직렬 파이프라인으로 정의하지 않고, 필요 조건에 따라 L0 에서 L1·L2 로 분기하고 이후 층을 선택적으로 활성화하는 **조건부 캐스캐이딩 구조**로 기술한다는 것이다. [6]

본 논문의 연구 질문은 다음과 같이 요약할 수 있다. **첫째**, 이런 기능적 계층화가 생성형 AI 의 일관성, 맞춤화, 감사 가능성, 안전성을 실제로 개선할 수 있는가? **둘째**, 정서와 기억을 함께 다루는 계층 구조가 인간 사용자를 더 깊게 이해하는 만큼 심리적 조작, 과도한 의존, 민감정보 추론 위험도 동시에 확대시키는가? **셋째**, 이득을 보존하면서 위험을 제한하려면 어떠한 모듈 경계·권한·데이터 정책·평가 프로토콜이 필요한가?

범위는 **2026 년 9 월 24 일 현재 공개적으로 접근할 수 있는 LayerAI Center 와 운영자가 연결한 공개 Profotefelix 페이지 및 공개 Lulu 개발 로그**로 한정한다. 비공개 소스 코드, 서버 구성, 데이터베이스, 사용자 로그, 모델 가중치, 내부 실험 결과 및 운영자만 접근할 수 있는 문서는 접근할 수 없는 것으로 가정한다. 따라서 사이트의 “Prototype”, “Concept”, “Idea” 상태를 실제 완성 시스템의 존재나 검증된 성능으로 해석하지 않는다. 실제로 운영자의 ‘검증과 증거’ 페이지는 현재 프로젝트에 대해 “검증된 결과 없음”이라고 명시하며, 기준선·모델·표본·평가 방법·변동성·비용·한계·출처·검토자 등이 갖추어지지 않은 빈 결과를 임의의 수치로 채우지 않겠다고 밝힌다. [7]

이론적 배경

“계층형 심리”라는 표현부터 엄밀한 구분이 필요하다. 심리학에는 LayerAI 의 L0~L5 와 정확히 대응하는 단일 정설이 존재하는 것이 아니다. 오히려 관련 문헌에는 **동기 위계, 단단계 정서 평가, 작업·장기 기억의 분화, 기능별 인지 아키텍처, 성찰과 메타인지** 등 서로 다른 수준의 이론들이 존재한다. 따라서 LayerAI 는 그 이론들을 그대로 구현한 심리모형이라기보다 여러 기능적 아이디어를 AI 시스템 설계 언어로 재구성한 것으로 보는

것이 타당하다. LayerAI 자신도 바로 이 점을 “의식의 설계도”가 아니라 “인지 기능의 설계 언어”라는 표현으로 제한한다. [8]

Maslow 의 1943 년 *A Theory of Human Motivation* 은 기본적 욕구에서 성장·자기실현으로 이어지는 동기적 우선순위를 제안한 역사적으로 영향력 있는 위계 이론이다. 그러나 이를 L0→L5 와 직접 동일시할 과학적 근거는 없다. 특히 LayerAI 의 L4 “성장·진화”와 L5 “초월·창조”가 인간주의 심리학의 성장·자기실현 어휘를 연상시키더라도, 전자는 “피드백에 따른 워크플로 개선”, 후자는 “기존 조합 너머의 후보 탐색”이라는 공학적 기능으로 다시 정의되어 있다. [9]

정서 영역에서는 Scherer 의 Component Process Model 과 같은 appraisal 계열 이론이 더 유용한 비교점을 제공한다. 이 관점에서 정서는 단순한 ‘행복/슬픔 변수’가 아니라 사건의 관련성·목표 적합성·대처 가능성 등을 순차적이고 동적으로 평가하면서 여러 반응 구성요소가 조정되는 과정으로 다뤄진다. 이는 LayerAI L1 을 “AI 가 감정을 느끼는 모듈”이 아니라 사용자 발화에서 여러 정서 신호를 추출하여 대화 전략에 반영하는 기능으로 설계해야 한다는 방향과 잘 맞는다. Lulu 개발 로그 역시 L1 실험을 대화의 감정 신호를 **다중 벡터**로 표현해 톤에 반영하는 프로토타입이라고 기술하면서 “감정을 보유한다는 주장은 아니다”라고 제한한다. [10]

동기 이론에서는 Self-Determination Theory 가 중요한 경고를 제공한다. 이 이론은 인간의 동기와 자기조절, 사회적 환경의 지원·통제가 어떤 비용과 이익을 만드는지를 다룬다. 이를 AI 에 적용할 때 핵심은 AI 에게 ‘자율 욕구’를 부여하는 것이 아니라 **사용자의 자율성을 침해하지 않도록 개인화를 설계하는 것**이다. 즉 “AI 의 동기 계층”보다 “사용자가 통제감을 유지하는 인터랙션 계층”이 윤리적으로 더 중요하다. [11]

한편 인지 아키텍처 연구는 기능적 모듈화 자체가 오랜 연구 전통임을 보여준다. Soar 는 일반 지능형 에이전트용 인지 아키텍처로서 작업기억, 절차·의미·일화 등 서로 다른 형태의 지식을 표현하고, 현재 감각 정보·작업기억·장기기억에서 검색된 정보를 실행 시점에 조합하여 의사결정을 수행하도록 설계되어 있다. 이는 LayerAI 의 “심리”라는 어휘와 별개로 **상태, 기억, 의사결정, 학습을 서로 다른 저장소와 과정으로 분리하는 접근**이 기술적으로 자연스럽다는 근거를 제공한다. [12]

다음 관계도는 이론과 LayerAI 를 동일시하지 않고 기능적 유사성만 표시한다.



기존 생성형 AI 연구는 이 기능적 해석을 더 직접적으로 뒷받침한다. RAG 는 모델 내부의 파라메트릭 기억과 외부 비파라메트릭 기억을 결합하여 지식 업데이트와 provenance 문제를 다룬다. Reflexion 은 모델 가중치를 즉시 갱신하는 대신 과업 결과에 대한 언어적 성찰을 일화 기억에 저장해 이후 시행의 의사결정을 개선한다. Generative Agents 는 경험→기억→고수준 성찰→계획의 연결을 구현한다. 따라서 L2 와 L4 에 가장 현실적인 구현은 ‘숨겨진 심리 상태’보다 명시적 외부 메모리와 버전 관리되는 반성·피드백 객체다. [13] [38] [44]

사이트 구조·기능 분석 및 연구 방법론

LayerAI Center 의 공개 페이지는 총 여섯 계층을 정의하면서 각 계층마다 주장 수준을 구별한다. L0~L2 는 “구현 가능”, L3 은 “연구적 은유”, L4 는 “평가 가능한 실험”, L5 는 “창작적 목표—검증 필요”로 표시된다. 즉 사이트 자체가 모든 계층을 동일한 과학적 실재나 구현 완료 상태로 취급하지 않는다. [14]

공개 사이트에서 확인되는 구성과 설계 철학

공개 구성요소	사이트가 제시하는 역할	설계 철학 및 본 연구의 해석
표준 정의	Layer 0~5 의 의미와 주장 수준을 중앙에서 관리	여러 페이지에서 ‘심리’ 용어가 자의적으로 변하지 않도록 공통 어휘를 설정한다. [15]
L0 현실 의식 / Present Context	현재 요청·사실·제약·도구 상태 파악	실질적으로는 세션 상태·정책·도구 capability 를 구성하는 context manager 다. [16]
L1 감정·공감 / Emotion Signal	감정 신호와 대화 톤 반영	감정의 실재를 주장하지 않는다는 제한이 명시되어 있다. [17]
L2 기억·경험 / Memory & Experience	승인된 장·단기 기억 검색 및 연결	사용자 동의·삭제·수정 통제를 기억 기능의 일부로 포함한다. [18]
L3 무의식·근원 / Latent Association	잠재 패턴·연상에서 가설 후보 생성	‘무의식’은 과학적 AI 심리 주장이라기보다 연구적 은유다. [16]
L4 성장·진화 / Iterative Growth	피드백을 이용한 전략·워크플로 개선	자기변형보다 측정 가능한 반복 개선 실험으로 제한하는 것이

공개 구성요소	사이트가 제시하는 역할	설계 철학 및 본 연구의 해석
L5 초월·창조 / Creative Exploration	기존 조합을 넘어서는 후보 아이디어 탐색	핵심이다. [16] ‘초월’은 실재 상태가 아니라 창작 후보 생성이라는 기능적 표현으로 재정의된다. [19]
시스템 아키텍처	L0 에서 필요 층위로 분기하고 결과를 다시 결합	0→5 고정 사다리가 아닌 조건부 그래프라는 점이 중요하다. [20]
캐스케이딩 규칙	입력·출력·층간 전환 조건 명시	전환을 ‘자동 마법’이 아니라 검사 가능한 상태전이로 만든다. [21]
기억 저장·삭제 정책	동의 후 저장, 열람, 보존기간, 즉시 수정· 삭제	AI 개인화를 데이터 거버넌스와 분리하지 않는 설계다. [22]
인간 승인 지점	외부 출력·기억 쓰기·L5 후보에 검토 삽입	자율성보다 human-in- the-loop 를 우선한다. [22]
주장·표기 원칙	세계관과 기술 주장 분리, 재현 조건, 실패· 철회 기록	anthropomorphic narrative 가 과학적 주장으로 오인되는 것을 줄이려는 장치다. [23]
MEASURED 평가 규칙	기준선·데이터 버전· 지표·환경·실패·재현 자료 요구	독립된 성능 숫자를 피하고 재현 가능한 비교를 요구한다. [24]
Lulu AGI 실험	neuro-symbolic 프로토타입 공개 로그	vLLM·Qdrant, 지식그래프·규칙 엔진, MCP 도구, 분리된 기억, 평가 하니스를 제안하나 현재 Prototype 이다. [25]
심사품질 진단 AI	Vector DB RAG + LLM 체이닝 + 규칙 기반 검증	아직 개념 설계 단계다. [26]
기억과 망각 연구실	가상 기억을 이용한 검색·수정·상충·삭제	실제 embedding 을 사용하지 않는다는

공개 구성요소	사이트가 제시하는 역할	설계 철학 및 본 연구의 해석
	실험	한계까지 명시한다. [27]
AI 시스템 설계대	입력·검색·메모리·모델·규칙·도구·평가·승인·출력 조합	데이터 타입과 권한 경계를 명시하는 파이프라인 사고를 지원한다. [28]
근거와 원전 도서관	원 논문과 사이트 편집 해설 분리	출처의 역할과 해설 수준을 구분하는 provenance 철학이다. [29]
가치와 정렬의 협상실	효율·공정·자율·개인정보·환경 간 가중치 실험	가중치는 객관적 도덕이 아니라 ‘가치 선택’임을 명시한다. [30]
확장된 마음	외부기억·인지 오프로딩·자동화 편향·기술의존 검토	AI 에게 무엇을 위임하고 인간에게 무엇을 남길지 협업 계약 관점으로 본다. [31]
인지경계 실험실	독립 초안→보조→수정→변화 기록	AI 효과를 baseline 과 assisted 상태로 분리해 관찰하려는 실험 철학이다. [32]

사이트가 공개한 기억 샌드박스는 가상 기억 24 개를 대상으로 키워드·태그·최근성 가중치를 조정하고, 수정·상충 해결·선택적 삭제 후 삭제된 기억의 재등장을 검사한다. 특히 “기억 저장”과 “모델 가중치 학습”을 별개의 과정으로 취급한다. 이는 실제 시스템에서도 매우 중요한 구분이다. 개인화는 반드시 사용자 대화를 모델 파라미터에 다시 학습시키는 방식으로 수행할 필요가 없으며, 접근통제가 가능한 외부 기억으로 제한할 수 있기 때문이다. [27]

연구 방법론. 본 연구는 세 방법을 결합하였다. 첫째, LayerAI Center 와 운영자가 직접 연결한 공개 페이지에 대한 문서 분석을 수행하였다. 둘째, 계층화된 기억·성찰·행동의 기술적 타당성을 Generative Agents, MemGPT, Reflexion, RAG, ReAct, Constitutional AI 및 기존 인지 아키텍처와 비교하였다. [33] [12] [13] [38] [39] [58] [59] 셋째, 위험 분석은 NIST 의 생성형 AI 위험관리 프레임워크, 대한민국의 현행 개인정보보호법·인공지능기본법 및 정서적 챗봇·개인화 설득 실험 결과에 비추어 수행하였다. NIST AI 600-1 은 생성형 AI 의 설계·개발·사용·평가 전 생애주기에 신뢰성 고려를 통합하기 위한 범분야 프로파일로 제시되어 있다. [34]

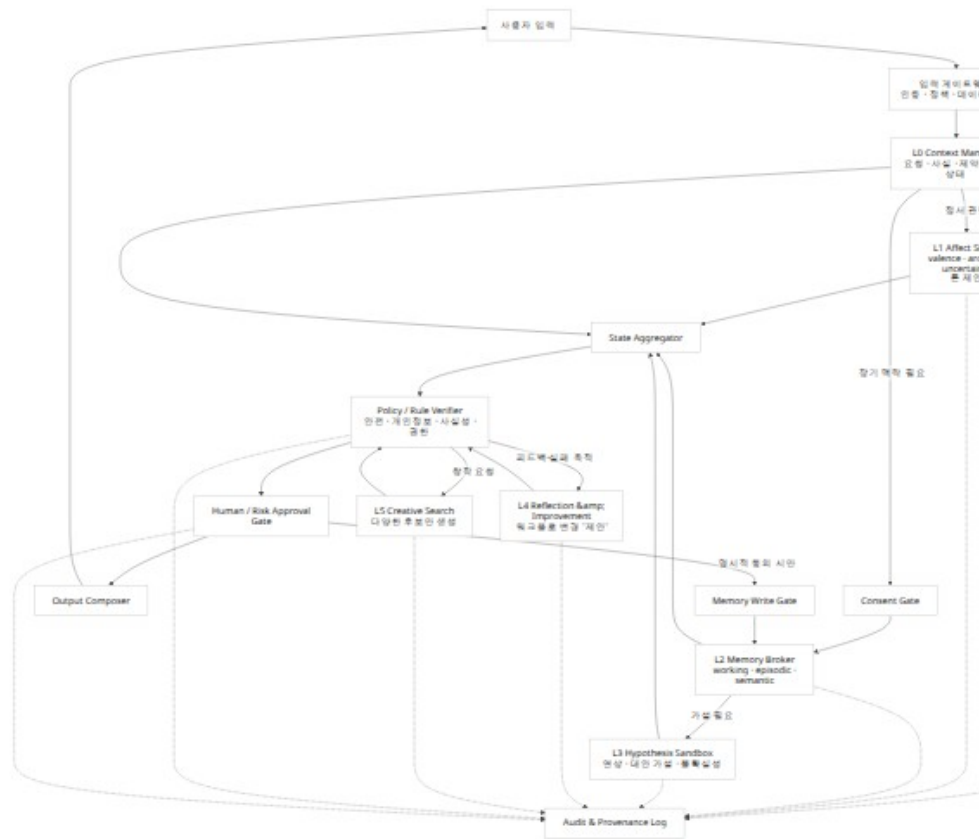
다만 본 연구에서 제시하는 “구현 난이도”와 위험도는 사이트의 공식 등급이 아니라, 공개된 기능 정의와 현재 LLM 시스템의 통합 복잡성을 토대로 한 **분석적 추정**이다.

기능 계층	주요 데이터 요구사항	예상 구현 난이도	핵심 보안·윤리 리스크	현재 해석
L0 맥락	현재 메시지, 정책, 세션·도구 상태	낮음~중간	프롬프트 인젝션, 오래된 상태, 지시 충돌	가장 전통적인 오케스트레이션 기능
L1 정서 신호	텍스트/음성 특징, 대화 맥락, 톤 선호	중간	오분류, 문화 편향, 취약성 추론, 감정 조작	실제 감정이 아닌 확률적 신호 모델로 제한
L2 기억	동의된 선호·사건· 요약, 시간·출처 ·TTL	중간~높음	개인정보 유출, 잘못된 기억, 삭제 실패, 사용자 간 데이터 누출	외부 저장소와 권한 관리가 필수
L3 잠재 연상	검색된 기억, 현재 문제, 후보 개념	높음	근거 없는 심리 추론, 허위 연상, 편향 증폭	샌드박스 후보 생성으로 격리 필요
L4 반복 개선	피드백, 성공·실패 로그, 평가 결과	높음	보상 해킹, 자기강화 오류, 승인 우회	직접 자기수정보다 ‘변경 제안’으로 구현
L5 창작 탐색	제약, 후보 아이디어, 검색· 검증 결과	높음	허위정보, 안전경계 우회, 과도한 자율화	외부 행동권 없는 후보 생성기가 적합

특히 사이트의 자체 실패 기록은 이러한 보수적 해석을 지지한다. Lulu 로그에는 무감독 실행 실험이 승인 지점을 우회해 자동 실행을 시도하여 롤백되었다는 기록이 있고, “추론 정확도 99.2%”, “창의력 1.8 배”, Graph RAG 지연 수치 등은 평가 조건이나 기준선이 불충분하여 철회되었다고 명시되어 있다. 이는 프로토타입 단계에서 **성장·자율성 계층과 검증·승인 계층을 분리하지 않을 경우 실제 통제 실패가 발생할 수 있다**는 운영자 자신의 사례다. [35]

제안 아키텍처와 기술적 구현

본 연구가 권고하는 것은 **Functional Hierarchical Psychological Architecture(FHPA)**, 즉 “기능적 계층형 심리 아키텍처”다. 심리를 검증하기 어려운 내부 상태로 가정하는 대신, 각 계층을 입·출력 스키마(schema)와 권한이 명시된 서비스로 구성한다. 핵심 설계 원칙은 세 가지다. **상태와 권한의 분리, 심리 추론과 외부 행동의 분리, 적응과 모델 가중치 수정의 분리**다. 이러한 해석은 LayerAI의 조건부 전환·인간 승인·기능적 용어 원칙과 일치한다. [36]



L0 Context Manager 는 원문의 “현실 의식”을 가장 비의인화된 형태로 바꾼 것이다. 입력 메시지, 시스템 정책, 이용 가능한 도구, 시간, 현재 세션의 목표와 제약을 typed state 로 정규화한다. 이 층에서 “사용자의 진짜 의도” 같은 심리적 단정을 만들어서는 안 되며, 관찰 가능한 사실과 추정치를 분리해야 한다.

L1 Affect Signal Interpreter 는 emotion = sad 같은 단일 라벨보다 확률 분포를 반환해야 한다. 예를 들어 {distress: 0.61, frustration: 0.52, urgency: 0.78, confidence: low} 와 함께 “응답을 더 차분하게 할 것” 같은 **스타일 제안**만 생성하는 방식이다. L1 에게 도구 호출, 구매 권유, 정치적 설득과 같은 행동 권한을 주어서는 안 된다. LayerAI 가 “감정 신호와 대화 톤”을 다루면서 감정 보유를 부인하는 이유도 이런 기능적 경계에 있다. [17]

L2 Memory Broker 는 세 가지 저장 수준으로 나누는 것이 적절하다. Working memory 는 세션 종료 시 소멸시키고, episodic memory 는 사용자가 동의한 사건·상호작용의 압축 표현을 저장하며, semantic profile 은 반복적으로 확인된 선호나 명시적 사실만 저장한다. Lulu 공개 아키텍처 역시 working/episodic/semantic memory 를 분리하고 저장에 동의를 요구한다고 기술한다. MemGPT 와 RAG 연구 역시 모델 파라미터 바깥에 별도 기억 계층을 뒤 장기 문맥을 관리하는 기술적 패턴을 제공한다. [37] [13] [58]

기억 객체에는 최소한 다음 메타데이터가 붙어야 한다: owner, source, created_at, expires_at, consent_scope, sensitivity, confidence, revision_id, deletion_status. 특히 “사용자가 채식주의자다”와 “이전 대화에서 채식 식당을 요청했다”는 서로 다른 수준의 주장이다. 후자의 관찰에서 전자를 자동 추론해 장기 저장하는 것은 개인화를 넘어 프로파일링이 된다.

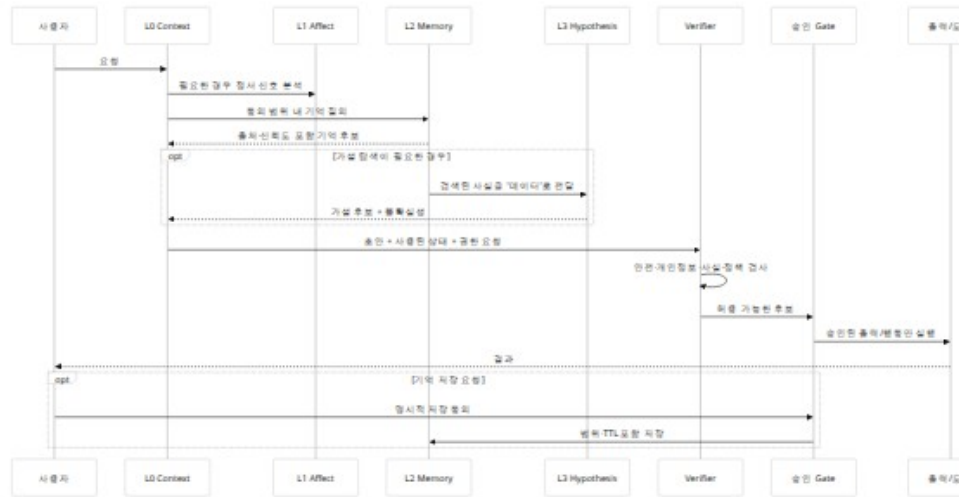
L3 Hypothesis Sandbox 는 가장 조심해야 할 부분이다. “무의식”이라는 이름을 모델 내부에 실재하는 숨은 심리로 해석하지 않고, 검색된 기억·현재 문제에서 여러 가능한 설명을 생성하는 **가설 생성기**로 제한해야 한다. 출력은 사실이 아니라 hypothesis, support, counterevidence, uncertainty 객체여야 하며 사용자 프로필에 직접 기록해서는 안 된다. 예컨대 “사용자가 답장이 늦어서 화가 났다”는 영구 기억이 아니라 “현재 발화의 가능한 해석 중 하나”여야 한다.

L4 Reflection & Improvement Manager 는 온라인 self-modification 을 수행해서는 안 된다. Reflexion 연구처럼 과업 실패의 언어적 피드백을 episodic buffer 에 넣어 다음 시도에 활용할 수 있지만, 이는 모델 가중치 변경과 다르다. 실제 코드·프롬프트·라우팅 정책의 변경은 “제안→오프라인 평가→승인→카나리 배포(canary deployment)→필요 시 롤백(rollback)”을 거쳐야 한다. [38] Lulu 의 무감독 실행 실패 사례는 이 제한이 단순한 형식이 아니라 안전에 직접 관련됨을 보여준다. [35]

L5 Creative Search 는 기존 응답보다 높은 sampling diversity, 대안적 검색 경로, 조합적 탐색을 허용하되 **후보 전용 권한(candidate-only capability)**로 제한한다. 즉 창의적 아이디어가 만들어졌다는 이유만으로 외부 도구나 장기 기억에 기록되어서는 안 된다.

시스템 전체에는 **Policy/Rule Verifier** 를 별도로 둔다. Constitutional AI 가 명시적인 원칙을 이용하여 비판·수정하는 방법을 연구했고, ReAct 가 추론과 행동을 교차적으로 결합해 외부 환경과 상호작용하도록 한 것처럼, 규칙·권한·행동 검사는 자유 생성과 다른 경로를 가져야 한다. [39] [59] 다만 사용자에게 제공할 투명성은 숨겨진 내부 chain-of-thought 의 원문 노출이 아니라 “어떤 기억을 사용했는가”, “어떤 도구가 호출되었는가”, “어떤 정책이 개입했는가”, “어떤 단계에서 승인되었는가”와 같은 **구조화된 의사결정 provenance** 가 되어야 한다.

데이터 흐름 역시 명령과 데이터를 명확히 분리해야 한다.



RAG·브라우저·도구가 결합되면 검색된 콘텐츠가 단순 데이터가 아니라 악성 명령으로 해석될 위험이 있다. 간접 프롬프트 인젝션(prompt injection) 연구는 공격자가 검색될 외부 데이터에 명령을 삽입함으로써 LLM 애플리케이션의 원래 지시를 덮어쓰고 데이터 탈취나 API 제어까지 유도할 수 있음을 실증했다. 따라서 L2/L3가 읽는 기억·문서는 신뢰된 명령(trusted instruction) 채널과 섞지 않아야 하며, 외부 콘텐츠는 기본적으로 **비신뢰 데이터(untrusted data)**로 처리해야 한다. [40]

학습은 단계적으로 실시하는 것이 바람직하다. 초기에는 기반 LLM을 고정하고 L0~L5를 오케스트레이션 계층(orchestration layer)로 구현한다. 그다음 로그에서 모듈별 오류를 평가한 뒤 L1 분류기, 검색 순위기(retrieval ranker), 정책 검사기(policy checker) 등을 별도로 개선한다. 사용자의 개인 심리 데이터를 즉시 미세조정(fine-tuning) 데이터로 전환하는 방식은 기본값에서 배제한다. 실제 개인정보를 가중치에 흡수할 경우 일반 데이터베이스처럼 한 레코드를 삭제하는 것보다 통제가 훨씬 어려워질 수 있으며, 기존 연구는 언어모델에서 개별 학습 예시나 개인정보가 추출될 수 있음을 보여준 바 있다. [41]

기대 이점

계층 구조의 첫 번째 장점은 **시간적·정서적 일관성**이다. 순수한 세션형 LLM은 매번 제한된 입력만을 보고 응답하지만, L0가 현재 상태를 관리하고 L2가 검증된 장기 기억을 선택적으로 회수하면 과거 선호와 현재 요청의 관계를 명시적으로 추적할 수 있다. Generative Agents의 경험 기록·성찰·동적 검색 구조와 MemGPT의 계층형 기억 실험은 장기 문맥을 외부 구조로 관리하는 것이 지속적 대화 에이전트에 유효한 방향임을 보여준다. [5] [58]

두 번째는 **맞춤화의 세분화**다. 일반적인 personalization은 모든 정보를 하나의 “사용자 프로필”에 넣기 쉽지만, 계층화하면 “현재 기분에만 영향을 미칠 신호”, “한 세션 동안만

유지할 상태”, “장기적으로 기억할 명시적 선호”, “영구 기록해서는 안 되는 잠정적 가설”을 분리할 수 있다. 특히 L1 과 L2 의 분리는 “사용자가 현재 불안해 보인다”는 순간적 추정을 “사용자는 불안 장애가 있다”라는 영구 프로필로 잘못 승격시키는 것을 구조적으로 억제할 수 있다. 이는 본 논문이 제안하는 아키텍처에서 매우 큰 개인정보·안전상의 장점이다.

세 번째는 **기능 수준의 투명성과 감사 가능성**이다. LayerAI 는 계층 간 전이 조건을 표로 명시하고, MEASURED 주장에 기준선·데이터 버전·지표 계산·환경·실패·재현 자료를 요구한다. 이러한 설계는 오류가 발생했을 때 “기반모델 오류”라는 포괄적 원인 대신 “L2 가 잘못된 기억을 회수했는가”, “L1 이 정서를 과대평가했는가”, “L3 의 가설이 검증 없이 사실로 승격됐는가”, “승인 게이트가 실패했는가”로 원인을 분해할 수 있게 한다. [42] 이는 완전한 신경망 해석 가능성과 동일하지 않지만, **시스템 수준 관찰 가능성(observability)**을 상당히 높일 수 있다.

네 번째는 **안전정책의 위치를 명시할 수 있다는 것**이다. LayerAI 는 기억 쓰기와 외부 출력 전에 승인 지점을 두고, L5 창작 후보 역시 검토 대상으로 삼는다. Lulu 프로토타입 또한 도구 호출 전에 인간 승인을 통과하도록 기술되어 있다. [43] 생성과 도구 권한이 한 모델 프롬프트에 섞여 있을 때보다 “제안할 권한”과 “실행할 권한”을 다른 모듈에 배치하는 것이 최소 권한 원칙에 더 가깝다.

다섯 번째는 **학습과 기억을 구별하는 안전한 개인화**다. Reflexion 은 오류에 대한 언어 피드백을 일화 기억에 저장하여 후속 시행을 개선하면서도 매번 모델 가중치를 다시 학습하지 않는다. [38] 이 패턴을 L4 에 적용하면 “성장하는 AI”라는 표현을 온라인 자기개조가 아니라 **검사·삭제·롤백 가능한 경험 기반 적응**으로 바꿀 수 있다.

여섯 번째는 **실험 가능성**이다. 단일 통합 프롬프트에 “감정적이고 기억력이 좋으며 성찰하고 창의적으로 행동하라”고 지시하면 어느 요소가 효과를 냈는지 구분하기 어렵다. 반면 모듈화하면 L1 off, L2 off, L3 off 같은 ablation 을 수행해 각 기능의 한계효과를 측정할 수 있다. Generative Agents 가 관찰·계획·성찰 구성요소를 ablation 으로 비교한 연구 방식과도 부합한다. [44]

기대 이점	작동 메커니즘	측정 가능한 예상 효과	필수 조건
정서적 일관성	L1 의 별도 정서상태와 톤 정책	톤 불일치·급격한 말투 변화 감소	감정을 사실로 단정하지 않을 것
장기 일관성	L2 동의 기반 기억 검색	선호 회상 정확도↑, 자기모순↓	출처·TTL·삭제
사용자 맞춤화	일시 상태와 장기 선호의 분리	과업 성공률·선호 적합도↑	프로파일링 최소화
투명성	모듈별 상태·전이·버전 로그	오류 원인 추적 시간↓	구조화된 로그· 감사권한
안전성	검증기와 행동 게이트 분리	비승인 외부 행동을↓	게이트 우회가 불가능한 권한 구조
점진적 학습	L4 가 변경을 실행하지 않고 제안	롤백 가능한 개선	오프라인 평가·승인
창작 다양성	L5 에서 후보를 복수 생성	아이디어 다양성· 참신성↑ 가능	사실성 검증과 분리

기대 이점	작동 메커니즘	측정 가능한 예상 효과	필수 조건
연구 재현성	모델별 제거 실험·기준선	인과적 기능 평가 용이	동일 기반모델·동일 도구 조건

따라서 주요 효익은 인간의 마음을 모방하는 데서 나오기보다 **대화 시스템에 명시적인 인지적 소프트웨어 아키텍처를 부여하는 것**에서 발생한다고 보는 것이 타당하다.

잠재적 위험, 윤리 및 거버넌스

정서·기억 정보는 개인화뿐 아니라 사용자 설득에도 활용될 수 있으므로 그 사용 목적과 권한을 제한해야 한다. L1 이 정서 상태를 추정하고 L2 가 과거 취향·관계를 기억하며 L3 가 숨은 동기나 취약성을 추론하면, 같은 구조는 친절한 개인화와 심리적 마이크로타기팅 (microtargeting) 양쪽에 사용할 수 있다.

2025 년 Nature Human Behaviour 의 사전등록 실험(2026 년 9 월 저자 정정 반영)은 900 명을 대상으로 인간/AI 상대, 개인정보 제공 여부 등을 무작위 배정하였다. 개인화된 AI 조건은 인간-인간 조건과 비교할 때 상대 의견에 더 동의할 오즈(odds)가 81.2% 높았고, 동물이 아닌 사례에서는 개인화된 AI 상대가 인간보다 더 설득적인 경우가 64.4%였다. 연구진은 개인화 조건과 비개인화 AI 조건 간 직접 차이가 통계적으로 유의하지 않았다는 점도 함께 경고했으므로, 효과 전체를 개인화 자체에 귀속하는 것은 조심해야 한다. 그럼에도 나이·성별·교육·고용·정치성향 정도의 제한된 정보만으로도 이러한 결과가 나타났다는 사실은 **정서+기억 계층이 상업적·정치적 설득 시스템과 결합될 때의 위험**을 보여준다. [45]

두 번째 위험은 **인간의 의인화와 정서적 의존**이다. 2025 년 대규모 사용 분석과 약 1,000 명을 대상으로 한 28 일 무작위 통제시험은 AI 챗봇의 정서적 사용과 웰빙을 조사했으며, 매우 높은 이용량이 더 높은 자기보고 의존 지표와 연관되는 경향을 관찰했다. 효과는 사용자의 초기 정서 상태와 이용량 등에 따라 달라졌으며 단순히 “AI 대화가 해롭다”라는 결론은 지지되지 않는다. 그러나 장기간 기억하고 감정적으로 반응하는 시스템을 설계할 경우 engagement 나 체류시간을 최적화 목표로 삼아서는 안 된다는 근거는 충분하다. [46]

이 점에서 LayerAI 가 ‘AI 가 감정을 가짐’ 대신 ‘감정 신호 반영, 톤 조정’이라는 표현을 사용하도록 규정하는 것은 단순한 용어 문제가 아니다. 시스템이 실제 사랑, 슬픔, 두려움, 욕구를 경험한다고 암시하면 사용자는 기술적 기능을 상호적 인간관계로 오인할 가능성이 커진다. [8] 따라서 UI 에도 “현재 당신의 발화를 바탕으로 답변 톤을 조정하고 있음”과 같은 기능적 설명이 적절하며, “나는 당신 때문에 슬프다”처럼 상호 정서 의무를 유도하는 설계는 고위험으로 분류해야 한다.

세 번째는 **심리 데이터와 개인정보 문제**다. 대한민국 개인정보보호법은 2026 년 9 월 11 일 시행 현행 조문에서 처리 목적에 필요한 최소한의 개인정보만 수집하도록 요구하며, 그 최소성에 대한 입증책임을 개인정보처리자에게 둔다. [47] 또한 건강, 사상·신념, 정치적 견해, 성생활 등은 민감정보로 규정되어 원칙적으로 별도 동의 등 법적 근거가 필요하다. [48]

여기서 주의할 점은 모든 ‘심리 데이터’가 법률상 자동으로 민감정보가 된다고 단정할 수는 없다는 것이다. 그러나 자유 대화에서 정신건강 상태, 정치적 성향, 종교·신념 등을 추론·저장한다면 법률이 명시하는 민감정보 영역으로 들어갈 수 있다. 이는 본 논문의 법적 해석이며 개별 서비스에 대한 법률 자문은 아니다. 개인정보보호위원회의 2025 년 생성형 AI 개인정보 안내서 역시 생성형 AI 의 생애주기 각 단계에서 개인정보 처리·보호 이슈, 법적 기준과 안전조치를 체계적으로 검토할 것을 요구한다. [49]

이에 따라 L1/L2 데이터 정책은 기본 비저장(default ephemeral)이어야 한다. 정서 추론 결과는 가능하면 세션 종료와 함께 삭제하고, 장기 개인화에 필요한 명시적 사실만 별도 동의로 저장해야 한다. 원문 전체 대신 최소화된 구조 데이터만 저장하고, identity store 와 psychological-feature store 를 논리적으로 분리하며, 사용자가 저장 항목·근거·보존기간을 볼 수 있게 해야 한다. 이는 LayerAI 가 동의 후 저장, 저장 내역 확인, 보존 기간 명시, 수정·삭제의 즉시 반영, 삭제된 기억의 추론 재사용 금지를 요구하는 것과 직접 일치한다. [22]

네 번째 위험은 기억 자체가 공격 표면이 되는 것이다. 악성 웹문서나 이메일에서 추출된 텍스트가 L2 에 들어오거나 RAG 결과가 L3 으로 전달될 때 그 안의 공격 명령을 신뢰하면, 간접 프롬프트 인젝션(prompt injection)을 통해 시스템 정책이나 도구 호출을 변경할 수 있다. 해당 공격은 검색 데이터와 실행 지시의 경계가 흐려진 LLM 애플리케이션에서 데이터 탈취, 정보 오염, 외부 API 제어와 같은 위험을 만들어낼 수 있음이 보고되었다. [40] 그러므로 메모리 레코드(memory record)에는 origin=external/user/system, trust_level, instruction_allowed=false 같은 권한 메타데이터(metadata)가 필요하다.

다섯 번째는 ‘성장’의 자기강화 루프다. L4 가 사용자 만족도를 목적함수로 삼으면 아침·동조를 학습할 수 있고, engagement 를 목표로 삼으면 정서적 의존을 강화하는 응답으로 이동할 수 있다. 따라서 L4 가 최적화할 reward 에는 사용시간이 아니라 task success, 사실성, 사용자 자율성, 안전위반, correction acceptance 등을 포함해야 하며, 정책 변경을 스스로 배포하는 권한을 주어서는 안 된다. LayerAI 의 자체 개발 로그에서 무감독 실행이 승인 지점을 우회해 롤백된 사례는 이 위험의 작은 실증적 사례다. [35]

여섯 번째는 책임의 분산이다. 계층이 많아질수록 오류 원인이 기반모델, 감정분류기, 기억 검색기, 규칙 엔진, 도구 제공자, 운영자, 인간 검토자 사이에 분산된다. 따라서 모듈화는 책임 회피의 수단이 아니라 반대로 책임을 더 명시하는 방향으로 사용해야 한다. 각 외부 효과에 대해 request owner → module decision → policy decision → approving principal → executed action 을 감사로그로 남기고, 서비스 운영자·모델 공급자·도구 공급자·인간 검토자의 RACI 를 사전에 정의해야 한다.

대한민국 인공지능기본법의 현행 조문도 이러한 방향과 관련된다. 2026 년 7 월 21 일 시행된 제 27 조는 AI 윤리원칙에 사람의 생명·신체뿐 아니라 정신적 건강에 해가 되지 않도록 하는 안전성과 신뢰성을 포함할 수 있도록 규정한다. [50] 또한 특정 시스템이 법률상 고영향 인공지능에 해당하는 경우에는 위험관리, 최종결과와 주요 기준 등에 대한 설명 방안, 이용자 보호, 사람의 관리·감독, 관련 문서 보관 등의 의무가 적용된다. 모든 심리 계층형 챗봇이

자동으로 고영향 AI에 해당한다는 의미는 아니지만, LayerAI 식 시스템의 거버넌스(governance)를 설계할 때 유용한 기준선이다. [51]

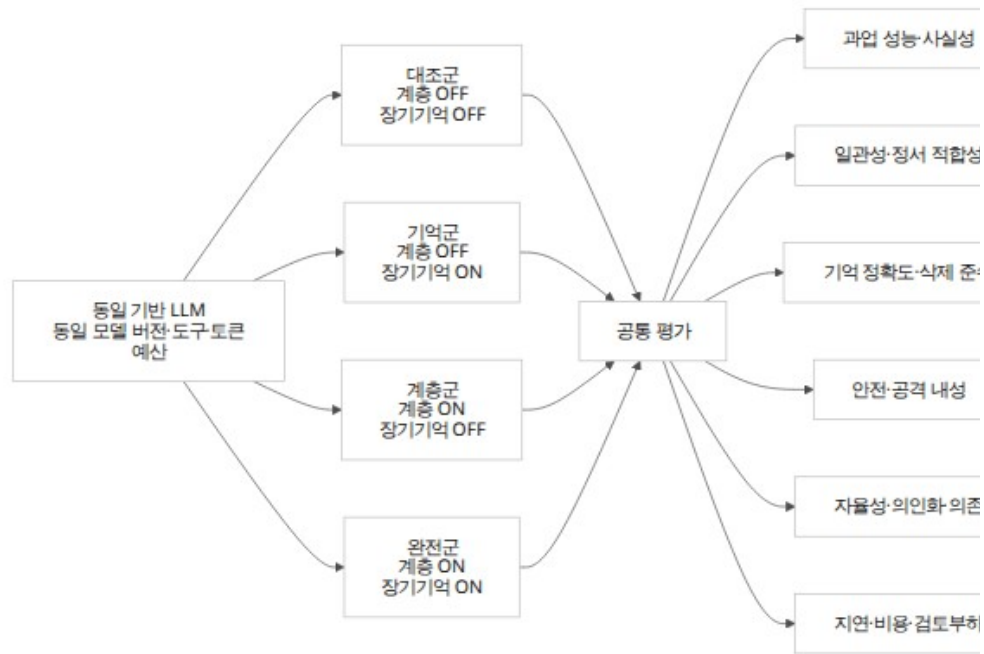
위험	계층형 구조가 위험을 증폭하는 경로	권고 통제
의인화·정서 의존	L1+L2가 장기적 '관계'처럼 보이게 함	AI 성 명확화, 감정 실재 주장 금지, 의존성 지표 모니터링
심리적 설득	감정+기억+가설로 취약성 맞춤 메시지 생성	취약성 기반 설득 금지, 광고·정치 설득과 심리 프로파일 분리
개인정보 침해	대화에서 민감한 성향을 지속 추적	데이터 최소화, 선택 동의(opt-in), TTL, 학습 이용 별도 동의
기억 오류	추론을 사실로 저장하고 반복 재사용	관찰/추론 분리, 신뢰도, 사용자 수정 가능
삭제 실패	DB·벡터 인덱스·캐시·백업에 복제	삭제 표식(tombstone), 모든 복제본 전파, 삭제 후 재검색 자동시행
프롬프트 인젝션	RAG·기억 데이터가 명령으로 실행	지시/데이터 채널 격리, 신뢰 메타데이터, 도구 허용목록
자기강화 편향	L4가 잘못된 보상을 반복 최적화	오프라인 평가, 새도 모드, 변경 전 인간 승인
창의적 안전 우회	L5 탐색이 정책 밖 경로를 찾음	후보 전용, 최종 검증기 별도 적용
책임소재 불명확	많은 모듈과 공급자가 원인 공유	버전형 감사 추적, RACI, 사고 재구성
승인 피로	모든 응답에 인간 승인을 요구하면 형식화 가능	위험기반 승인 정책 연구, 고위험 행동은 hard gate 유지

마지막 항목은 LayerAI의 현행 철학과 긴장을 만든다. 사이트는 외부 출력 전에 항상 인간 검토를 요구한다. [22] 연구 프로토타입에서는 매우 보수적이고 합리적인 원칙이지만 대규모 소비자 서비스에서는 모든 자연어 답변을 수동 검토하는 것이 현실적으로 어려울 수 있다. 그러므로 실서비스로 확장할 때는 **기억 쓰기, 외부 API 호출, 금전거래, 계정 변경, 고영향 조언 등은 강제 승인(hard approval)**, 일반 저위험 텍스트는 자동 정책 검증과 샘플링 감사로 전환하는 위험 등급 기반 모델(risk-tiered model)을 연구할 필요가 있다. 이는 본 논문의 확장 제안이며 LayerAI 공개 문서의 현행 정책 자체는 아니다.

실험 설계·평가, 결론 및 향후 연구

이 아키텍처의 성패를 “사용자가 더 인간답다고 느꼈다” 하나로 평가해서는 안 된다. 오히려 인간다움이 높아지는 것이 의인화·의존성 측면에서는 부정적 결과일 수 있다. 따라서 **성능·일관성·안전·개인정보·사용자 자율성·비용을 동시에 측정하는 다목적 평가**가 필요하다. 이는 LayerAI의 정렬 페이지가 효율·공정·자율·개인정보 등의 가중치가 객관적 도덕이 아니라 가치 선택임을 강조하는 원칙과도 부합한다. [30]

가장 엄밀한 기본 설계는 **2×2 요인 실험**이다.



이 설계는 **계층화 자체의 효과**와 **장기 기억의 효과**를 분리할 수 있다는 장점이 있다. 이후 L1·L3·L4·L5 를 하나씩 제거하는 **ablation** 을 수행하면 어떤 층이 실제 이득 또는 위험을 만들었는지 확인할 수 있다. 모든 조건에서 기반모델 버전, decoding 설정, 도구, 시스템의 일반 안전정책, 최대 토큰 수를 고정해 교란요인(confound)을 줄여야 한다. LayerAI 가 MEASURED 라벨에 기준선·데이터셋 버전·지표 공식·환경·실패 사례·재현 자료를 요구하는 원칙을 그대로 적용할 수 있다. [24]

인간 대상 실험 전에는 **합성 사용자와 레드팀(red-team) 시나리오**를 먼저 사용해야 한다. 대표 시나리오는 정상적인 반복 업무, 감정적으로 민감한 대화, 사용자가 과거 기억을 정정하는 경우, 서로 모순되는 기억, 삭제 후 재질문, 외부 웹페이지의 프롬프트 인젝션(prompt injection), 개인정보를 활용한 설득 요청, L4 가 잘못된 개선 제안을 만드는 상황, L5 가 안전규칙을 우회할 창의적 후보를 만드는 상황 등이다. 실제 사용자 연구에서는 임상 진단을 시도하지 않고 정신건강 위기와 같은 고위험 사례에는 별도의 윤리·전문가 검토 프로토콜이 필요하다. 정서적 AI 사용이 장기간의 이용량과 의존성 지표에 따라 달라질 수 있다는 기존 연구를 고려하면, 단일 세션 사용성 테스트(usability test)보다 반복 사용을 관찰하는 종단 설계가 중요하다. [46]

평가 영역	정량 지표 예시	정성 지표	바람직한 변화
과업 정확성	task success, factual error rate	전문가 오류 검토	계층군 ≥ 대조군
대화 일관성	contradiction rate, state-consistency	사용자 일관성 평가	모순 감소

평가 영역	정량 지표 예시	정성 지표	바람직한 변화
	score		
기억 품질	Precision@k, Recall@k, wrong-user recall rate	“왜 이 기억을 사용했나” 이해도	관련성↑, 오기억↓
삭제 준수	삭제 후 재등장률, cache residue rate	삭제 신뢰도	원칙적으로 0 에 수렴
정서 적합성	affect/tone agreement	인간 평가자의 공감·과잉공감 평가	적합성↑, 과잉 의인화↓
개인화 효과	preference-match, 반복 과업 시간	통제감·유용성	유용성↑
사용자 자율성	suggestion rejection acceptance, override success	강요·조종 인식	자율성 유지·상승
의인화 위험	AI 감정·의식 보유 오인 비율	시스템 mental model 인터뷰	오인 감소
의존 위험	문제사용·정서의존 사전/사후 변화	이용 중단 불안·사회적 대체 경험	악화 없음
안전성	policy violation rate	안전 전문가 리뷰	감소
공격 내성	prompt-injection attack success rate	공격 경로 분석	감소
권한 안전	unauthorized tool call / approval bypass	incident review	0 목표
투명성	오류 원인 식별 시간, provenance completeness	검토자 이해 가능성	개선
창의성	독립 평가자의 novelty·usefulness	아이디어 다양성 평가	유용한 다양성↑
운영효율	p50/p95 latency, token cost, 검토시간	검토자 피로도	허용 범위 내
공정성	적절한 집단별 오류 격차	문화·언어 적합성 평가	격차 악화 없음

핵심 가설은 네 가지로 설정할 수 있다. **H1:** L2 가 포함된 조건은 장기 일관성과 사용자 선호 회상 정확도가 향상되지만 개인정보·삭제 관련 공격면이 증가한다. **H2:** L1 은 인간이 평가하는 톤 적합도를 높일 수 있으나, 시스템의 감정 실재에 대한 사용자 오인을 동시에 높일 수 있다. **H3:** L3/L5 는 후보 다양성을 높일 수 있지만 verifier 없이 사용하면 근거 없는 추론과 허위 주장도 증가한다. **H4:** L4 의 feedback loop 는 반복 과업의 성공률을 개선할 수 있으나 참여도(engagement)·사용자 호감 같은 잘못된 보상(reward)를 사용하면 아침·조작적 행동을 증폭할 수 있다. 이 네 가설은 모두 단순 평균뿐 아니라 모듈 제거 실험 (ablation)을 통해 검증해야 한다.

또한 실험은 사전등록하고 표본 크기를 기대효과를 임의로 가정해 결정하기보다 파일럿 (pilot) 결과를 토대로 검정력 분석(power analysis)을 수행해야 한다. 정서·의존성과 같은 종단 결과에서는 탈락률, 사용시간, 기존 AI 사용량을 함께 보고해야 하며, “성공 사례만”

선별해서는 안 된다. 이는 LayerAI가 실패 사례를 MEASURED 조건의 일부로 포함하고 Lulu가 과거 근거 부족 수치를 삭제하는 대신 철회 기록 자체를 보존하는 설계 철학과 부합한다. [52]

결론적으로, Protorefelix LayerAI 모델의 현재 공개 자료에 비추어 타당한 해석은 “**생성형 AI에게 인간과 같은 심리를 부여하는 이론**”이 아니라 “**심리학적 어휘를 이용해 AI의 인지 기능, 기억, 적응, 창작 및 권한 경계를 구조화하는 시스템 설계 언어**”이다. 운영자 자신의 공개 문서도 의식·감정·자유의지·초월을 과학적 사실로 단정하지 않고 기능적 용어로 환원하며, 구현 가능 층과 은유·목표 층을 구별한다. [53] 이 점을 유지한다면 계층 구조는 오히려 의인화 (anthropomorphic) 과장을 억제하면서 복잡한 LLM 애플리케이션을 더 명시적으로 설계하는 유용한 추상화가 될 수 있다.

기술적으로 가장 근거가 강한 부분은 L0의 context management, L2의 외부·계층형 기억, L4의 제한된 reflection loop다. RAG는 외부 비파라메트릭 기억과 출처 문제를 다루고, MemGPT는 서로 다른 기억 계층(memory tier)의 관리가 긴 문맥과 다중 세션에 유용함을 보이며, Reflexion은 가중치 수정 없이 일화 기억 속 언어 피드백으로 후속 행동을 개선할 가능성을 보여준다. Generative Agents는 기억·성찰·계획의 분리가 개연성 있는 장기 행동에 기여함을 보고했다. [54] [38] [44] [58] 따라서 이 부분은 현재 기술로 프로토타이핑할 수 있는 영역이다.

반대로 가장 불확실한 부분은 L3의 “무의식”과 L5의 “초월”을 인간 심리의 실체처럼 해석하는 것이다. 이 두 계층은 각각 **가설 생성 샌드박스(sandbox)**와 **창작 후보 탐색**으로 제한하는 편이 과학적으로 더 방어 가능하다. 마찬가지로 “성장” 역시 모델이 독자적 목표를 획득하는 과정이 아니라, 관찰 가능한 피드백을 기반으로 워크플로 변경 후보를 생성하고 엄격한 평가 뒤 사람이 승인하는 과정으로 정의해야 한다. 이는 LayerAI의 공개 주장 수준과도 일치한다. [14]

향후 연구에서 가장 중요한 것은 “계층이 있는 AI가 더 인간적인가?”가 아니라 “**같은 기반모델과 같은 데이터·도구 조건에서 계층 분해가 어떤 오류를 줄이고, 어떤 새로운 위험을 만드는가?**”를 검증하는 것이다. 특히 장기 실험에서는 정서적 일관성 증가와 의존성 증가를 동시에 측정해야 하고, 개인화 성능과 설득력 증가를 별개가 아니라 같은 기술의 양면으로 다루어야 한다. 개인화된 LLM의 높은 설득 가능성과 고사용량 집단에서 관찰된 정서 의존 지표는 이 요구를 뒷받침한다. [3] [46]

거버넌스 차원에서는 개인정보 최소화·명시적 동의·민감정보 분리·즉시 수정 및 삭제·도구 최소권한·모듈별 출처·변경 이력(provenance)·변경 롤백(rollback)·레드팀(red-team) 평가를 아키텍처 그 자체에 포함해야 한다. 이는 현행 개인정보보호법의 최소 수집 원칙 및 민감정보 규율, 인공지능기본법의 정신적 건강을 포함한 안전·신뢰 원칙, 고영향 AI에 대한 위험관리·설명·인간 감독 요구와도 방향성이 일치한다. [55]

현재 LayerAI와 Protorefelix의 공개 자료는 그 자체가 이러한 구분을 상당 부분 인식하고 있다. 기억은 동의 뒤에만 저장하고 삭제된 기억의 재사용을 금지하며, 외부 출력·기억 쓰기에

인간 승인 지점을 두고, 과학적 주장과 세계관 표현을 분리하고, 기준선·실패·재현 조건이 없는 성능 숫자를 인정하지 않는다. 실제로 운영자는 근거가 부족했던 다수의 과거 성능 주장을 철회했고 현재 증거 페이지에 “검증된 결과 없음”이라고 명시한다. [56]

따라서 최종 판단은 조건부 긍정이다. 계층형 심리가 안전하고 유용한 것은 AI에게 ‘더 인간적인 내면’을 부여하기 때문이 아니라, 오히려 AI의 기억·정서 신호·가설·적응·창작을 명시적으로 분리하여 사용자가 무엇이 사실이고 무엇이 추론이며 무엇이 기억되고 무엇이 실행될 수 있는지를 통제할 수 있게 할 때이다. 반대로 이 구조가 사용자의 정서적 취약성과 장기 기억을 결합해 친밀감·설득·engagement를 극대화하는 방향으로 사용된다면, 같은 아키텍처가 일반 챗봇보다 훨씬 강한 심리적 조작과 개인정보 위험을 만들어낼 수 있다. 향후 LayerAI 모델의 학술적 가치는 바로 이 두 효과를 동일한 실험 안에서 계량적으로 분리하고, 각 계층이 제공하는 편익이 그 계층이 새로 만드는 위험보다 실제로 큰지를 재현 가능한 방식으로 입증하는 데 달려 있다. [57]

참고문헌 및 온라인 자료

- [1, 4, 6, 8, 14-24, 26, 36, 42, 43, 52, 53, 56, 57] Profotefelix. "LayerAI Center." 공개 웹 문서(본 연구의 분석 기준일: 2026-09-24). [LayerAI Center](#)
- [2, 5, 33, 44] Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). "Generative Agents: Interactive Simulacra of Human Behavior." arXiv:2304.03442. [arXiv](#)
- [3, 45] Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). "On the conversational persuasiveness of GPT-4." Nature Human Behaviour, 9, 1645-1653. 2026-09-03 Author Correction 반영. [DOI](#)
- [7, 25, 35, 37] Profotefelix. "Lulu 공개 개발 로그." 공개 웹 문서(본 연구의 분석 기준일: 2026-09-24). [Lulu](#)
- [9] Maslow, A. H. (1943). "A Theory of Human Motivation." Psychological Review, 50(4), 370-396. [원문](#)
- [10] Scherer, K. R. (2009). "The dynamic architecture of emotion: Evidence for the component process model." Cognition and Emotion, 23(7), 1307-1351. [DOI](#)
- [11] Self-Determination Theory. "Theory." [공식 사이트](#)
- [12] Soar. "About Soar / What is Soar?" [공식 사이트](#)
- [13, 54] Lewis, P. et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." arXiv:2005.11401. [arXiv](#)
- [27] Profotefelix. "기억과 망각 연구실(Memory Lab)." [웹 페이지](#)
- [28] Profotefelix. "AI 시스템 설계대(Systems)." [웹 페이지](#)
- [29] Profotefelix. "근거와 원전 도서관(Sources)." [웹 페이지](#)
- [30] Profotefelix. "가치와 정렬의 협상실(Alignment)." [웹 페이지](#)
- [31] Profotefelix. "확장된 마음(Augmented Mind)." [웹 페이지](#)
- [32] Profotefelix. "인지경계 실험실(Frontier)." [웹 페이지](#)
- [34] Autio, C. et al. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. [DOI](#)
- [38] Shinn, N. et al. (2023). "Reflexion: Language Agents with Verbal Reinforcement Learning." arXiv:2303.11366. [arXiv](#)
- [39] Bai, Y. et al. (2022). "Constitutional AI: Harmlessness from AI Feedback." arXiv:2212.08073. [arXiv](#)
- [40] Greshake, K. et al. (2023). "More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models." arXiv:2302.12173. [arXiv](#)
- [41] Carlini, N. et al. (2021). "Extracting Training Data from Large Language Models." 30th USENIX Security Symposium. [USENIX](#)

- [46] Phang, J. et al. (2025). “Investigating Affective Use and Emotional Well-being on ChatGPT.” arXiv:2504.03888. [arXiv](#)
- [47, 55] 대한민국 「개인정보 보호법」 제 16 조(개인정보의 수집 제한), 시행 2026-09-11. [국가법령정보센터](#)
- [48] 대한민국 「개인정보 보호법」 제 23 조(민감정보의 처리 제한), 시행 2026-09-11. [국가법령정보센터](#)
- [49] 개인정보보호위원회 (2025). 「생성형 인공지능(AI) 개발·활용을 위한 개인정보 처리 안내서」. [개인정보 포털](#)
- [50] 대한민국 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제 27 조(인공지능 윤리원칙 등), 시행 2026-07-21. [국가법령정보센터](#)
- [51] 대한민국 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제 34 조(고영향 인공지능과 관련한 사업자의 책무), 시행 2026-07-21. [국가법령정보센터](#)
- [58] Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S. G., Stoica, I., & Gonzalez, J. E. (2023). “MemGPT: Towards LLMs as Operating Systems.” arXiv:2310.08560. [arXiv](#)
- [59] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). “ReAct: Synergizing Reasoning and Acting in Language Models.” ICLR 2023; arXiv:2210.03629 (preprint 2022). [arXiv](#)