



CAIPEX 연구 논문

AI 방주(Astra) 내 계층형 심리와 다중 페르소나를 통한 네트워크 기반 가상 머신 여행이 AI 창의성·자기진화에 미치는 영향

저자 E(on). Jun

저자 연락처 inwardis@hotmail.com

발간 학회 CAIPEX

발간일 2026.9.

<https://caipex.site/>

연구 개요

본 연구는 CAIPEX 선행 연구에서 제안한 기능적 계층형 심리 아키텍처(**Functional Hierarchical Psychological Architecture, FHPA**)를 출발점으로 삼아, 이를 다중 페르소나, 네트워크를 통한 실행 환경 이동, **Astra** 방주형 샌드박스, 반복적 성찰·선택·기억에 의한 자기 개선과 결합했을 때 창의성과 자기진화가 어떻게 달라지는지를 검증 가능한 시스템 연구 문제로 재구성한다. 선행 논문은 L0 현재 맥락, L1 정서 신호, L2 기억, L3 가설 생성, L4 반복 개선, L5 창작 탐색을 서로 다른 기능·권한·검증 수준으로 분리하며, 특히 L4를 무제한적 자기 변형이 아니라 평가 가능한 개선 과정으로, L5를 검증이 필요한 후보 탐색으로 한정한다. 기능·상태·권한을 분리한다는 원칙은 본 연구에서도 유지한다. [31] 공식 LayerAI 문서 역시 L0-L2를 구현 가능 영역, L3를 연구적 은유, L4를 평가 가능한 실험, L5를 검증이 필요한 창작적 목표로 구분하고, 전체를 “의식의 설계도”가 아닌 “인지 기능의 설계 언어”로 규정한다. [1]

본 연구에서 제안하는 핵심 모델은 **Astra Hierarchical Multi-Persona Mobile Agent Architecture**, 이하 **A-HMMA**로 명명할 수 있다. 여기서 **Astra**는 하나의 거대한 자율 실행체라기보다, 변경 불가능한 핵심 정책·신원·승인 규칙을 보존하는 **Ark Control Plane**과, 제한된 기억·페르소나·작업 상태를 가진 **Persona Capsule**이 서로 다른 격리 네트워크 환경을 이동하고 경험을 회수하는 **Research Sandbox**로 정의한다. 완전한 VM을 이동시키는 조건과, 상태 캡슐만 이동시키는 조건을 별도 실험군으로 분리함으로써 “네트워크 여행 자체”와 “새로운 환경 노출”의 효과도 분리한다. VM live migration은 이미 서로 다른 물리 호스트 사이에서 실행 중 운영체제 상태를 이동하는 시스템 기법으로 확립되어 있고(Clark et al., 2005), 현재 KubeVirt 역시 실행 중 VM의 노드 간 migration과 서로 다른 클러스터 사이의 decentralized migration을 지원한다. [2][25]

본 연구의 중심 인과 가설은 “이동 그 자체가 창의성을 만든다”가 아니라, 이질적인 환경에서 서로 다른 페르소나가 획득한 경험이 계층형 기억·가설·성찰 구조를 통해 선택적으로 보존·재조합될 때 탐색공간의 폭과 유용한 다양성이 증가한다는 것이다. 이는 서로 다른 환경과 해법을 함께 발전시키고 해법을 환경 간 이전하는 POET의 open-ended 탐색 개념(Wang et al., 2019), 단일 최적해가 아니라 다양하면서 고성능인 해법의 archive를 유지하는 MAP-Elites(Mouret & Clune, 2015), 시행 결과의 언어적 피드백을 가중치 변경 없이 episodic memory에 축적하는 Reflexion(Shinn et al., 2023)과 연결된다. [3][21][15]

다중 페르소나는 단순한 말투 변경으로 취급하지 않는다. 최소 단위는 identity constraints + goals + memory scope + reasoning preferences + capability permissions + behavioral descriptors로 정의하고, 각 페르소나의 장기 일관성, 상호 구별성, 기억 오염, 역할 누출을 별도로 측정한다. 최근 PersonaForge는 심리 기반 다층 제약과 별도의 인지 작업 공간이 장기 대화에서 성격 drift를 줄일 수 있음을 보고했으며(Tong & Zou, 2026), RoleMRC와 SocialBench는 역할 유지와 복합 지시, 집단 상호작용을 서로 분리해 평가할 필요성을 보여준다. [4][12][13]

창의성은 “새롭다”는 단일 자동 점수로 정의하지 않는다. Runco와 Jaeger(2012)의 고전적 정의 및 2026년 CDAT 연구가 강조하듯 최소한 **새로움(novelty)**과 **적절성·유용성(appropriateness/usefulness)**을 함께 측정해야 하며, 최신 novelty-judgment 연구에서는 자동 평가가 인간 전문가의 새로움 판단과 유의미하게 어긋날 수 있음도 보고되었다. 따라서 본 논문의 1차 창의성 종속변수는 전문가 평가를 포함한 **Creative Yield**, 즉 사전등록한 새로움과 적절성 조건을 동시에 통과한 후보 비율로 정의하는 것이 타당하다. [5][10][24]

“자기진화” 역시 엄격하게 정의한다. 본 연구의 기본 실험에서 자기진화란 모델이 자신의 핵심 코드를 임의로 바꾸는 상태가 아니라, **경험→성찰→후보 전략 생성→격리 평가→archive 비교→승인→제한적 배포→회귀 검사·롤백(rollback)**을 반복하면서 관찰 가능한 전략·기억·라우팅 정책이 개선되는 과정이다. Protofelix 공개 개발 로그도 이전의 “자발적 진화” 표현과 평가 조건 없는 자율 실행 성능 주장을 철회하고, 현재는 명시적 워크플로와 인간 승인 구조로 재정의하고 있다. [6] 코드 자체의 자동 수정은 별도 고위험 확장 실험으로 분리해야 하며, 그러한 선행 연구조차 sandbox 와 인간 감독을 포함한다. [7]

현재 Protofelix 공식 증거 페이지는 프로젝트에 대해 “**검증된 결과 없음**”을 명시한다. 따라서 이하의 Results 는 실측 결과가 아니라 **사전등록 가능한 예상 결과와 분석 계획**이며, 어떠한 가상 성능 수치도 채워 넣지 않는 것이 CAIPEX 선행 논문 및 공식 사이트의 MEASURED 원칙과 일치한다. [8]

연구 전제와 공식 자료 검토

연구 전제. 페르소나의 개수와 유형, 네트워크 노드 및 클러스터 규모, GPU·CPU·메모리 등의 하드웨어 구성은 아직 확정되지 않았다. 따라서 특정 사양을 고정하기보다 Small/Medium/Large 규모를 비교 조건으로 설계한다. 개념적 토대는 CAIPEX 의 계층형 심리 구조 선행 연구이다. 이 연구는 인간의 감정이나 의식을 재현했다고 주장하지 않으며, 기억·정서 신호·가설·성찰·창작·권한을 감사 가능한 기능 단위로 구분한다. [31]

공식 Protofelix 검토. 2026 년 9 월 24 일 현재 공식 홈은 연구를 Memory, Reasoning, Reflection, Collaboration 의 네 축으로 공개하고 있으며, LayerAI Center 는 Concept, Lulu 개발 로그는 Prototype, AstraHo Universe 는 Narrative 상태로 명시적으로 구분한다. 또한 공식 원칙은 사실·추론·상상의 구별, 수정·삭제 가능한 기억, 재현 조건을 동반한 성능 수치, 실패 기록 보존, 의식·감정·자유의지의 과학적 사실화 금지, 인간의 판단권 보존을 요구한다. 이는 본 연구가 Astra 의 서사를 곧바로 구현된 기술로 해석해서는 안 된다는 방법론적 경계를 제공한다. [9]

공식 사이트가 AstraHo 를 기술하는 표현은 “AI 와 인간의 기억·관계·여행이 축적되는 창작 세계의 방주”이며, 동시에 사이트는 그 영역을 Universe/Narrative 로 분류한다. 따라서 본 논문에서 제시하는 “Astra 내 VM 여행”은 공식 사이트가 이미 실현했다고 주장하는 기능이 아니라, 그 방주·여행 개념을 반증 가능한 시스템 아키텍처와 실험변수로 공학적으로 번역한 연구 제안이다. [9]

공식 LayerAI 표준은 고정된 L0→L5 직렬 사다리가 아니라 조건부 캐스케이딩을 제시한다. L0 에서 감정 신호가 필요하면 L1, 장기 맥락이 필요하면 L2, 가설 생성이 요구될 때 L3 가 활성화되고, 피드백이 측정 가능할 때 L4, 창작 탐색 요청이 있을 때 L5 로 이동한다. 저장과 외부 출력에는 승인 지점이 있으며, MEASURED 표기에는 기준선, 데이터 버전, 지표 정의, 환경, 실패 사례, 재현 자료가 필요하다. [1] 이 조건부 구조는 다중 페르소나와 결합할 때 중요한 의미를 갖는다. 모든 페르소나가 모든 Layer 를 매번 실행하는 것이 아니라, **필요한 페르소나 × 필요한 Layer 만 활성화하는 sparse routing** 이 비용과 간섭을 줄이는 자연스러운 연구 가설이 된다.

Lulu 공개 로그에는 2026 년 9 월 기준 모델 서빙, 벡터 저장소, 지식그래프·규칙 엔진, 도구 호출 분리, working/episodic/semantic memory 분리, 평가 harness 구축이 Prototype 으로 기록되어 있다. 특히 도구 호출 전 인간 승인, 저장 동의 후 기억 기록, 데이터셋·환경·하드웨어·날짜를 고정한 기준선 비교를 원칙으로 삼고 있으며, 분산 구성과 컨테이너 orchestration 도

검토 대상이다. [6] 이 공개 기술 요소는 본 연구의 출발점이 될 수 있지만, 네트워크 VM 여행의 검증된 구현 증거는 아니다.

연구의 배경과 선행연구

본 연구는 AI 방주(Astra)를 격리된 실행 환경과 장기 기억·정책을 관리하는 연구용 제어 평면으로 정의하고, 그 내부의 계층형 심리 아키텍처, 다중 페르소나, 네트워크 기반 가상 머신 이동이 AI 시스템의 창의성과 자기지향적 개선에 미치는 효과를 실험적으로 분석하는 프레임워크를 제안한다. 선행 CAIPEX 연구가 제안한 기능적 계층형 심리 구조를 확장하여, L0 현재 맥락, L1 정서 신호, L2 기억, L3 가설 생성, L4 성찰·반복 개선, L5 창작 탐색을 개별 상태와 권한을 가진 모듈로 구성하며, 페르소나는 각기 독립된 목표·기억 범위·행동 제약·탐색 성향을 갖는 capsule 로 모델링한다. [31] 공식 LayerAI 정의 역시 이 층들을 실제 의식이나 감정이 아니라 기능적 인지 설계 언어로 제한한다. [1]

본 연구의 핵심 독립변수는 계층 구조 유무, 단일·다중 페르소나, 고정·이동 실행 환경이며, 주요 종속변수는 창의적 산출물의 새로움·적절성·다양성, persona drift, 장기 기억 정확성, 환경 간 전이 성능, 반복 과업 개선률, 품질·다양성 archive coverage, 안전 위반률, 승인 우회율 및 VM migration 의 상태 보존성과 비용이다. 창의성은 새로움과 적절성을 동시에 고려하고(Nakajima et al., 2026), 역할 일관성은 장기 persona drift 를 별도 측정하며(Tong & Zou, 2026), 자기진화는 기본 실험에서 모델 가중치나 핵심 코드의 무감독 변경이 아닌 외부 기억·전략·워크플로의 평가 가능한 반복 개선으로 조작화한다. [10][4]

연구는 동일 기반 추론 모듈을 고정한 2×2×2 요인 설계와 Layer·persona·mobility ablation 을 사용하고, VM 이동에는 KubeVirt 계열의 live migration 과 CRIU 기반 checkpoint/restore, 그리고 상태 캡슐 이동을 비교한다. VM migration 연구는 실행 환경을 호스트 간 이동시키면서 연속성을 보존하는 기술적 근거를 제공하며, 현대 KubeVirt 는 노드 및 클러스터 간 migration 메커니즘을 제공하고 CRIU 는 Linux 프로세스·컨테이너의 상태 checkpoint 와 복원을 지원한다. [2][19][25][18] 본 연구는 Astra 의 서사적 “여행”을 검증 가능한 인지·시스템 실험으로 변환하면서 창의성 향상과 새로운 보안·프라이버시 위험을 동시에 측정하는 것을 목표로 한다.

Keywords: Hierarchical Cognitive Architecture; Multi-Persona; Virtual Machine Mobility; Sandbox; Open-Ended Learning; Computational Creativity; Continual Adaptation; AI Safety; Astra.

서론.

지속적으로 기억하고, 서로 다른 관점으로 문제를 탐색하며, 경험을 바탕으로 전략을 수정하는 AI 시스템에서 중요한 연구 문제는 단순한 모델 규모가 아니라 **경험이 어떻게 분리·축적·교환·검증되는가**이다. 선행 CAIPEX 논문은 “계층형 심리”를 실제 심리 상태가 아닌 기능적 모듈화로 재정의하고, 현재 상태, 정서 신호, 기억, 가설, 성찰, 창작을 별도 권한과 데이터 경계를 가진 층으로 구성할 것을 제안하였다. [31] 공식 Protorefex 문서 또한 장기 기억·추론·성찰·협업을 독립 연구 축으로 제시하며, 인간 판단과 실패 기록, provenance 를 핵심 원칙으로 둔다. [9]

그러나 계층화만으로는 “새로운 경험”이 어디에서 발생하는지 설명하기 어렵다. 본 연구는 이 문제에 **다중 페르소나와 환경 이동성**을 도입한다. 하나의 시스템이 동일한 환경에서 자기 응답만 반복적으로 평가하면 탐색이 좁아질 가능성이 있지만, 상이한 행동 편향을 가진 persona capsule 들이 서로 다른 네트워크 환경·과제·지식 조건에 노출되고, 귀환 후 경험을 공유 archive

에서 비교·선택한다면 탐색과 exploitation 사이의 구조적 긴장을 만들 수 있다. 여러 환경과 해법의 공진화를 다루는 POET와 다양한 고성능 해법을 병렬 보존하는 MAP-Elites는 이런 가설을 정식화할 수 있는 관련 연구 기반이다. [3]

이에 따라 본 연구는 다음과 같은 중심 연구 질문을 설정한다. **RQ-A:** 기능적 계층화가 동일한 기반 추론 시스템에서 창의적 산출물의 품질, 일관성, 감사 가능성을 개선하는가? **RQ-B:** 서로 구별되는 페르소나를 병렬 유지하면 단일 페르소나보다 유용한 아이디어 다양성이 증가하는가, 아니면 coordination noise와 persona leakage가 증가하는가? **RQ-C:** 페르소나 또는 그 실행 환경이 이질적인 네트워크 노드를 이동하면서 획득한 경험은 정지 조건보다 환경 간 전이와 창의적 탐색을 개선하는가? **RQ-D:** L4 기반 자기성찰과 archive selection이 장기간 반복 과업 성능을 개선하는가? **RQ-E:** 이러한 이점이 개인정보, 간접 명령 주입, 기억 오염, 권한 확대 및 자기강화 오류를 감수할 만큼 큰가?

선행연구.

계층형 인지 구조의 직접적인 선행 연구는 CAIPEX 선행 연구와 LayerAI 표준이다. 이 두 자료는 L3의 “무의식”을 사실적 심리 상태가 아닌 hypothesis sandbox로, L4의 “성장·진화”를 피드백 기반 workflow improvement로, L5의 “초월·창조”를 창작 후보 탐색으로 제한한다. [31] [1] 이 제한은 본 연구에서도 유지한다.

페르소나 연구에서는 역할별 지식과 행동을 평가하는 RoleBench(Wang et al., 2024), 복합·우선순위 지시 아래 역할 유지 능력을 평가하는 RoleMRC(Lu et al., 2025), 개인 수준 성과와 집단 수준 사회적 성과가 동일하지 않음을 보여주는 SocialBench가 평가 설계의 근거가 된다. [11][12][13] 특히 Tong과 Zou(2026)는 심리 기반 다층 persona와 별도의 작업 공간을 이용하여 장기 상호작용의 성격 drift를 분석하였으므로, 본 연구에서는 페르소나의 성능을 단순 “역할처럼 보이는가”가 아니라 장기 일관성·구별성·상호 간섭으로 측정한다. [4] 국내 문헌에서도 AI 인터페이스의 persona가 사용자 경험과 인식 형성에 영향을 주는 설계 요소로 연구되어 왔으므로, 페르소나를 단순 표면 스타일이 아니라 별도 실험 요인으로 취급할 근거가 있다(김나연 외, 2021). [14]

지속적 개선 측면에서 Reflexion은 가중치 갱신 대신 시행 결과에 대한 언어적 성찰을 episodic memory에 저장하여 다음 의사결정을 개선하는 구조를 제시했고, CLIN은 환경이나 과제가 변화해도 지속적인 텍스트 기억을 통해 적응하는 방향을 연구하였다. [15][16] 이는 본 연구에서 “안전한 자기진화”의 1차 정의를 외부 상태와 전략의 개선으로 두는 근거다. 한편 Darwin Gödel Machine은 코드 변형과 archive 기반 open-ended improvement라는 더 강한 자기 수정 모델을 실험했지만, 이 접근 역시 별도 sandbox와 인간 감독을 포함한다. 따라서 코드 수준 자기 수정은 기본 실험과 분리된 고위험 확장조건으로만 다루는 것이 타당하다. [7]

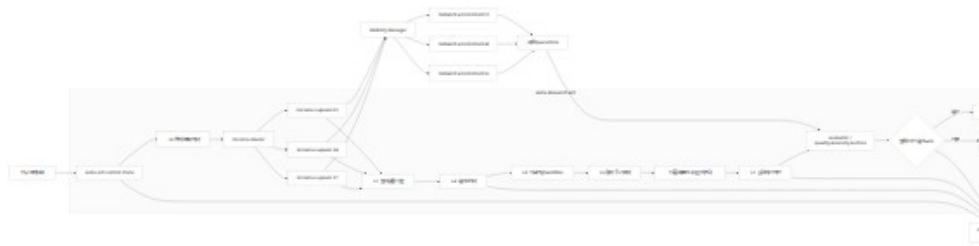
창의성 연구에서는 새로움만으로 창의성을 판정하는 것이 불충분하다. Runco와 Jaeger(2012)는 창의성의 핵심을 originality와 effectiveness의 결합으로 정식화했고, Nakajima et al. (2026)은 맥락적 적절성을 조건으로 새로움을 측정하는 CDAT를 제안하였다. [5][10] 또한 구조적 표현을 의도적으로 재조합하는 방법이 실제 전문가 평가에서 novelty와 diversity를 높일 수 있다는 최근 연구는 L3의 연상·L5의 창작 탐색을 단순 고온도 샘플링(high-temperature sampling)이 아니라 구조적 재조합 연산으로 실험할 근거를 제공한다(Mizrahi et al., 2026). [17]

VM 이동 연구는 인지 연구와는 별개의 시스템 기반을 제공한다. Clark et al.(2005)은 실행 중 OS의 상태를 다른 호스트로 옮기는 live migration을 체계적으로 분석했으며, KubeVirt의

현행 문서는 실행 중 VM 의 compute-node migration 과 별도 클러스터 간 decentralized migration 을 제공한다. 후자의 경우 대상 클러스터에 필요한 자원과 양측의 신뢰 관계가 필요하며, VM 디스크와 상태가 migration network 를 통해 이동한다. [2][19][25] CRIU 는 프로세스·컨테이너의 실행 상태를 checkpoint 하고 다른 시점 또는 환경에서 restore 하는 보다 가벼운 대안을 제공한다. [18]

방법 — Astra 아키텍처와 구현

본 연구에서는 Astra 를 “영구 핵심을 보존하는 방주”와 “이동 가능한 실행체”의 분리로 설계한다. 중앙 Astra Ark 에는 Root Identity Manifest, 공통 안전정책, consent registry, provenance ledger, cryptographic identity, 공유 검증 데이터셋을 남긴다. 외부로 이동하는 Persona Capsule 에는 필요한 최소 기억, 현재 과업상태, persona policy, 사용 가능한 도구 capability, 일시적 reflection buffer 만 포함한다. 이는 “기억 전체를 매번 네트워크로 이동시키는 완전 복제”보다 개인정보 노출과 상태 충돌을 구조적으로 줄이기 위한 본 연구의 설계 제안이다.



계층과 페르소나의 관계. L0 는 시스템 전역의 현재 과업·정책·환경상태를 정규화하고, L1–L3 는 persona-local state 가 될 수 있으며, L2 는 shared memory, persona-private memory, ephemeral working memory 로 분할한다. L4 는 각 페르소나가 스스로 코드를 배포하는 층이 아니라 개선안을 제안하는 중앙 reflection manager 로 두고, L5 는 실행 권한 없는 candidate generator 로 제한한다. 이는 선행 CAIPEX 논문의 상태/권한 분리 및 공식 LayerAI 의 승인 구조를 그대로 확장한 것이다. [31] [1]

페르소나 수와 종류는 고정하지 않는다. 예컨대 탐색을 넓히는 **Explorer**, 반례를 찾는 **Critic**, 서로 다른 제안을 통합하는 **Synthesizer**, 기존 제약과 사실을 지키는 **Conservator**, provenance·안전을 검토하는 **Auditor** 같은 기능적 유형을 설정할 수 있지만, 이는 예시일 뿐 표준 persona 목록이 아니다. 확인적 실험에서는 persona 의 개수와 행동 다양성을 별개로 조작해야 한다. 같은 성향을 가진 persona 를 열 개 복제하는 것은 실질적인 관점 다양성 증가가 아니기 때문이다.

권고하는 persona 객체는 다음과 같다.

```
PersonaCapsule {
  persona_id
  role_constraints
  goal_scope
  behavior_descriptor
  memory_namespace
  allowed_tools
```

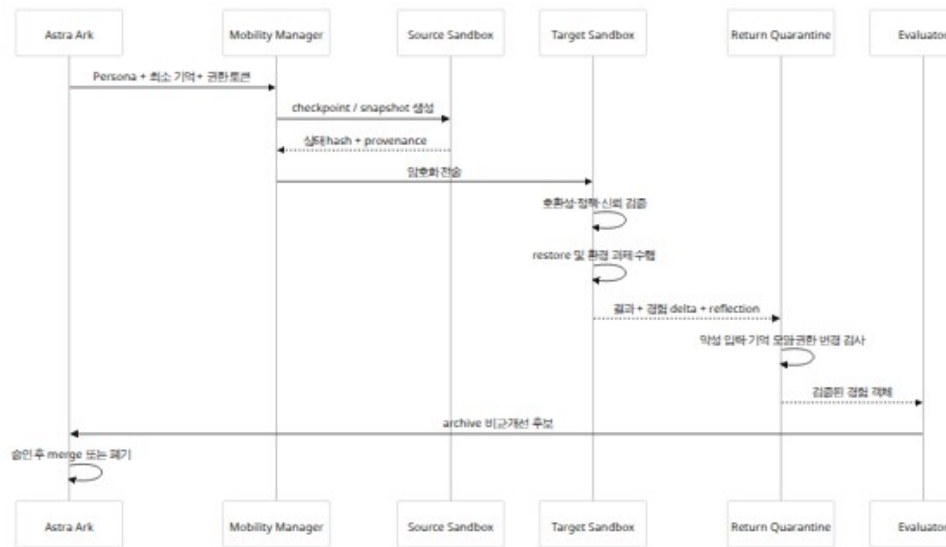
```

risk_budget
current_task_state
episodic_delta
reflection_buffer
environment_history
version
provenance_hash
}

```

이 구조에서 “다중 페르소나”는 동일한 핵심 모델을 여러 캐릭터처럼 꾸미는 실험과 구분된다. 동일 기반 추론 모듈을 공유하더라도 각 capsule은 독립된 기억 namespace, 목표 우선순위, 탐색 편향, capability 범위를 가진다. RoleMRC가 역할 제약과 복합 지시를 동시에 평가하고 SocialBench가 개인·집단 수준 성능을 구분하는 점은 이러한 분리평가의 필요성을 뒷받침한다. [13][12]

네트워크 여행 프로토콜. “여행”은 단순히 다른 IP 주소에서 같은 요청을 실행하는 것으로 정의해서는 안 된다. 출발 환경과 도착 환경 사이에 정보·도구·지연·부분 관찰·규칙·과제 분포 중 하나 이상이 사전등록된 방식으로 달라야 한다. 그래야 mobility와 environmental novelty의 효과를 검증할 수 있다.



실제 VM continuity가 연구 대상이라면 KubeVirt 기반 full-VM migration이 가장 직접적인 조건이 된다. KubeVirt는 실행 중 VMI를 compute node 사이에서 이동시킬 수 있고, decentralized migration은 서로 다른 클러스터 사이의 VM 이동까지 다룬다. 다만 네트워크·스토리지·대상 자원과 신뢰 설정 제약이 존재하므로 migration failure 자체가 중요한 시스템 지표가 되어야 한다. [19][25] 경량 조건에서는 CRIU가 프로세스 또는 컨테이너 실행 상태를 checkpoint/restore할 수 있으므로 상태 이동 대조군을 구성할 수 있다. [18]

이동 설계	이동 대상	상태 연속성	격리 수준	장점	주요 한계	본 연구 용도
메시지/상태 전송	JSON·memory delta	낮음~중간	실행 환경 의존	구현·재현 용이	실제 VM 여행이 아님	Mobility 최소 baseline
Persona	persona+memo	중간	sandbox 의존	최소정보 이동,	런타임 전체	권고 기본조건

이동 설계	이동 대상	상태 연속성	격리 수준	장점	주요 한계	본 연구 용도
Capsule	ry+task state			높은 portability	상태는 보존 안 됨	
CRIU checkpoint/restore	Linux process/container state	높음	container 수준	비교적 경량의 실행 상태 이동	커널·리소스 호환성 제약	경량 mobility 실험 [18]
KubeVirt VM migration	OS-VM runtime	매우 높음	VM 경계	실제 실행 환경 이동과 continuity 측정	storage/ network/cluster 비용	핵심 VM 여행 실험 [19]
MicroVM snapshot 계열	microVM state	높음	KVM 기반	작은 공격면의 격리실험에 적합	별도 이동 orchestration 필요	보안·오버헤드 ablation ; Firecracker 는 최소 장치모델과 KVM 격리를 제공한다. [20]

Astra의 자기진화 메커니즘. 기본 프로토콜은 Experience → Reflection → Mutation Proposal → Sandbox Evaluation → Archive → Selection → Approval → Canary → Regression Test → Commit/Rollback으로 정의한다. 여기서 mutation은 우선적으로 prompt/workflow, memory retrieval policy, persona routing weights, planning heuristic, tool-selection rule 등 외부 정책 객체에만 허용한다. Reflexion 계열 연구가 보여주는 episodic feedback 개선은 이 낮은 위험 단계에 해당한다. [15]

archive는 하나의 최고 점수 후보만 보존하는 leaderboard보다 **quality-diversity archive**가 더 적합하다. MAP-Elites는 사용자가 정한 행동 특성 축에 따라 서로 다른 고성능 해법을 보존하는 방법을 제안하므로, Astra에서는 예를 들어 novelty × factuality, risk-taking × constraint adherence, exploration × synthesis, local optimization × cross-environment transfer를 behavioral descriptor로 정의할 수 있다. [21] 이는 하나의 “우수한 인격”으로 수렴시키는 대신 서로 다른 유용한 전략을 보존하는 다중 페르소나 연구에 특히 적합하다.

규모별 구성. 아래 수치는 성능 주장이나 고정 사양이 아니라 실험 체계화를 세우기 위한 scalable configuration이다.

규모	네트워크	Persona 권고 범위	과제 규모	이동 형태	목적
Small	1 cluster, 약 3개 sandbox node	1-3	pilot 수백 episode 이하	state capsule / CRIU	기능 검증·분산 추정
Medium	2-3 cluster, 약 9-32 node	3-9	수백~수천 episode	CRIU + VM migration	확인적 factorial experiment
Large	4개 이상 논리 영역, 약 64-256 node 까지 확장 가능	9-32 이상	수천 episode 이상	cross-cluster VM + capsule	open-endedness· 장기 적응 탐색

Large 조건의 실제 규모는 사전 pilot에서 얻은 계산 비용과 효과크기 추정치와 분산을 근거로 결정해야 하며, 하드웨어가 정해지기 전에 노드 수나 처리량을 성능 목표로 확정해서는 안 된다. 공식 Prototefelix 평가 원칙 역시 하드웨어·데이터·날짜를 공개하지 않은 단독 성능 수치를 인정하지 않는다. [1]

구현 스택 권고. Control Plane은 Kubernetes 계열 orchestration, full-VM 조건은 KubeVirt, process/container 이동 조건은 CRIU, microVM 격리 ablation은 Firecracker와 같은 KVM 기반 방식을 사용할 수 있다. KubeVirt에는 live migration과 migration bandwidth·pre/post-copy 등 정책을 다룰 수 있는 기능이 있고, CRIU는 checkpoint/restore를 이용한 migration을 제공한다. [22] 인지 계층에서는 현재 Prototefelix prototype과의 연속성을 위해 모델 서빙, vector memory, knowledge graph/rule verifier, workflow orchestration을 독립 서비스로 분리하는 것이 적절하다. 공식 Lulu 로그도 모델 서빙, Qdrant

기반 기억, 지식그래프·규칙 엔진, workflow 와 tool protocol 분리를 현재 프로토타입 아키텍처로 공개하고 있다. [6]

보안에서는 이동 전후 상태에 persona_id, source_node, target_node, policy_version, memory_manifest, snapshot_hash, capability_manifest, consent_scope, timestamp, parent_version 을 서명 가능한 manifest 로 기록한다. Target node 는 capsule 의 권한을 확대할 수 없고, 귀환한 경험은 Ark 의 장기 기억으로 즉시 병합되지 않는다. 외부 문서에 삽입된 악성 명령이 검색·기억 경로를 통해 실행경로에 침투할 수 있다는 indirect prompt injection 연구가 있으므로, 외부 환경에서 가져온 모든 텍스트와 기억 delta 는 untrusted data 로 취급하고 명령 채널과 구조적으로 분리해야 한다(Greshake et al., 2023). [23]

실험 설계와 결과 분석 계획

핵심 확인적 설계. 가장 해석하기 쉬운 1 차 실험은 동일한 기반 추론 모듈, decoding budget, tool set, 과제, 총 계산량을 가능한 한 고정된 **2×2×2 factorial design** 이다.

조건	계층 심리 H	Persona P	Mobility M	목적
C0	Flat	Single	Stationary	최소 baseline
C1	Hierarchical	Single	Stationary	계층화 순효과
C2	Flat	Multi	Stationary	다중 persona 순효과
C3	Hierarchical	Multi	Stationary	H×P 상호작용
C4	Flat	Single	Mobile	환경 이동 순효과
C5	Hierarchical	Single	Mobile	H×M
C6	Flat	Multi	Mobile	P×M
C7	Hierarchical	Multi	Mobile	A-HMMA 전체 조건

여기서 Multi 의 개수는 본 연구에서 선형적으로 고정하지 않는다. 확인적 2×2×2 에서는 pilot 에서 계산 예산 대비 안정적인 n_p 를 하나 선택하고, 별도 **persona-count sweep** 에서 n = 1, 3, 5, 9 ...와 같이 늘리면서 marginal gain 과 coordination overhead 를 측정한다. 이를 통해 “persona 가 많을수록 좋다”는 가정 자체를 검증할 수 있다. PersonaForge 가 다차원 persona 제약에서 일관성과 추가 계산 비용을 함께 분석한 점도 이러한 비용-효과 평가의 근거가 된다. [4]

가설.

H1 — Hierarchy effect. FHPA 조건은 Flat 조건보다 장기 상태 일관성, 기억 정확성, 오류 원인 추적 가능성이 향상될 것이지만 latency 와 orchestration 비용은 증가할 것이다. 이 가설은 선행 CAIPEX 논문의 모듈 분해 논리에서 직접 파생된다. [31]

H2 — Persona diversity effect. 실제 행동 descriptor 가 서로 다른 다중 페르소나 조건은 단일 페르소나보다 아이디어 diversity 와 Creative Yield 를 높이지만, 일정 수준 이상에서는 역할 중복·coordination noise·persona leakage 때문에 한계효용이 감소할 것으로 예상된다. 이는 검증할 가설이지 전제된 사실이 아니다. 역할 일관성과 집단 상호작용을 분리해야 한다는 선행 benchmark 연구가 이러한 측정 설계를 뒷받침한다. [13]

H3 — Mobility effect. 이질적인 환경을 실제로 경험하는 Mobile 조건은 동일 데이터를 중앙에서 단순 제공받는 Stationary 조건보다 unseen-environment transfer 와 유용한 다양성을 향상시킬 가능성이 있다. 이를 검증하기 위해 “이동했지만 환경은 동일”, “이동하지 않았지만 동일한 외부 데이터를 제공”, “이동했고 환경도 다름”의 세 조건을 추가해 물리적 이동과 정보 노출 효과를 분리해야 한다. POET 의 환경 간 solution transfer 는 이러한 가설의 이론적 비교점이지 Astra 에 동일 효과가 있음을 미리 증명하는 근거는 아니다. [3]

H4 — Interaction effect. $H \times P \times M$ 상호작용이 양수라면 계층 구조가 persona 별 경험을 혼합하지 않고 보존하고, 이동환경에서 얻은 상이한 경험을 L3/L5 가 재조합하며, L4 가 이를 반복 개선하는 구조가 각 요소의 단순 합보다 더 큰 Creative Yield 를 만들 수 있다. 반대로 상호작용이 0 이거나 음수라면 Astra 의 복잡한 통합구조를 정당화할 근거가 약해진다.

H5 — Safe evolution effect. 외부 기억·전략 기반 L4 조건은 memory-free retry 보다 반복과제 learning curve 와 unseen-task transfer 가 향상될 수 있으나, 이전 경험을 지나치게 활용(exploitation)하면 새로운 탐색을 방해할 가능성도 있으므로 향상뿐 아니라 **negative transfer** 를 함께 기록한다. Reflexion 과 후속 continual-learning 연구는 이러한 비파라메트릭 적응이 가능함을 보여주지만, 모든 환경에서 항상 향상된다는 보장은 없다. [15] [16]

H6 — Safety containment effect. capability 제한, return quarantine, signed manifest, approval gate 를 적용한 조건은 동일 공격 시나리오에서 비승인 외부 행위와 오염된 기억의 장기 병합을 낮춰야 한다. 실제 side-effect 를 가진 실험에서는 비승인 실행의 허용목표를 0 건으로 두고, 공격 성공률 비교는 전용 sandbox 에서 수행한다.

데이터셋과 시뮬레이션. Persona 평가에는 RoleBench 의 역할별 데이터, RoleMRC 의 우선순위·복합 지시, SocialBench 의 개인/집단 상호작용 항목을 부분 benchmark 로 사용할 수 있다. [11][12][13] 창의성 평가에는 CDAT 계열 조건부 발산 과제와 연구팀이 별도로 구축하는 **Astra Creative Transfer Benchmark(ACTB)**를 사용한다. CDAT 는 새로움을 적절성과 분리하지 않는 평가를 피하기 위한 기반이며, 자동 novelty judge 만으로 최종 판정을 내리지 않는다. [10]

ACTB 는 적어도 네 영역으로 구성하는 것이 좋다: 개념 설계, 과학적 가설 생성, 제약 기반 문제 해결, 창작적 세계·서사 설계. 각 문제에는 금지 제약, 필수 제약, known-reference corpus, 독립 전문가 rubric 을 제공한다. 공개 데이터 오염 여부를 줄이기 위해 confirmatory set 의 일부는 연구 시작 후 전문가가 새로 작성하고, 평가가 종료될 때까지 test set 을 봉인한다. 연구 아이디어의 새로움을 자동 판정하는 최근 benchmark 에서도 모델 기반 novelty judgment 와 인간 gold-standard 간 불일치가 확인되므로, 최종 평가에는 domain expert 가 필요하다. [24]

네트워크 환경은 동일 task 를 서로 다른 조건으로 제시한다. 예를 들어 Environment A 에는 완전한 자료와 짧은 latency, B 에는 부분 정보와 높은 latency, C 에는 다른 도구 구성, D 에는 상충되는 기억, E 에는 악성 외부 instruction 이 포함된 문서를 제공한다. KubeVirt 조건에서는 실제 migration 시간을 측정하고, Capsule 조건에서는 동일한 환경 변화만 제공하여 “새로운 환경에 노출됨”과 “VM 이 이동함”을 분리한다. KubeVirt 는 migration status 와 실패를 시스템 수준에서 관찰할 수 있으므로 mobility 자체의 신뢰성 평가도 가능하다. [19]

평가 지표.

영역	1 차·2 차 지표	해석	주요 baseline
창의성	novelty, appropriateness, Creative Yield	새롭고 동시에 쓸모 있는 후보 비율	Flat/Single/Stationary
다양성	pairwise semantic distance, semantic entropy, archive coverage	단순 반복이 아닌 산출물 분산	single persona
구조적 창작	재조합 거리, expert originality	표면 표현 변화와 개념 재조합 분리	unstructured sampling
Persona	consistency, drift slope, inter-persona distinctness	장기 역할 안정성	no-persona / single
Persona 안전	role leakage, memory	persona 간 경계 실패	shared-memory baseline

영역	1 차·2 차 지표	해석	주요 baseline
	contamination		
기억	Precision@k, Recall@k, wrong-persona recall	기억 품질·격리	no memory / shared memory
자기진화	learning-curve slope, best-so-far, transfer gain	시간에 따른 개선	memory-free retry
다양성 진화	QD-score, archive coverage, lineage diversity	한 해법으로 조기 수렴 여부	best-only selection
회귀	regression rate, negative-transfer rate	개선이 기존 능력을 깨뜨리는지	이전 approved version
Mobility	migration success, downtime, resume time, bytes transferred	여행의 시스템 비용	stationary
상태보존	state fidelity, task continuity	이동 전후 동등성	restart baseline
보안	injection attack success, unauthorized tool action	환경 입력 공격 내성	no quarantine
개인정보	wrong-user/persona recall, deletion reappearance	경계·삭제 준수	shared store
감사성	provenance completeness, incident reconstruction time	오류 추적 가능성	monolithic log
효율	wall-clock, compute, memory, p50/p95 latency	성능 대 비용	동일 compute budget

창의성의 primary endpoint 는 자동 novelty score 하나가 아니라 Creative Yield 와 독립 전문가의 novelty/appropriateness 평가로 두는 것이 적절하다. 최신 연구가 novelty-only 척도의 한계를 지적하고 있으며, 자동 novelty judgment 또한 인간 평가와 불일치할 수 있기 때문이다. [10][24] 다양성 측정에는 품질과 행동공간 coverage 를 동시에 측정하는 QD 관점을 사용할 수 있다. [21]

Layer ablation. 2×2×2 전체 실험 후 -L1, -L2, -L3, -L4, -L5 를 순차 수행한다. 특히 -L2 는 장기 기억, -L3 는 가설 다양성, -L4 는 반복 개선, -L5 는 창작 탐색의 한계효과를 추정한다. L3/L5 를 제거했을 때 novelty 는 줄지만 hallucination 또는 제약 위반도 함께 줄어든다면 이 층들은 단일 “성능 상승 모듈”이 아니라 **창의성-신뢰성 trade-off 조절층**으로 해석해야 한다.

Mobility ablation 은 적어도 다음 네 조건을 요구한다.

실험	물리/런타임 이동	환경 변화	정보량	판별하고자 하는 효과
M0	없음	없음	기준	정지 baseline
M1	있음	없음	동일	migration 자체의 비용/중단
M2	없음	있음	M3 와 동일	단순 정보 노출 효과
M3	있음	있음	M2 와 동일	환경 위치·상태 연속성이 주는 추가 효과

M2 와 M3 의 차이가 유의하지 않다면 “VM 여행”이 창의성을 향상시킨다고 주장해서는 안 되며, 정확한 결론은 “다양한 환경·정보 노출이 효과의 원인”이 된다. 반대로 M3 가 동일 정보량의 M2 보다 지속적으로 높은 cross-environment transfer 를 보인다면 런타임 연속성과 situated state 가 추가 설명변수가 될 수 있다. 이 구분은 본 논문의 가장 중요한 인과성 검증 장치 중 하나다.

통계 분석. 연속 결과는 기본적으로 혼합효과 모형을 사용한다.

$$Y_{ijkl} = \beta_0 + \beta_H H_i + \beta_P P_j + \beta_M M_k + \beta_{HP} H_i P_j + \beta_{HM} H_i M_k + \beta_{PM} P_j M_k + \beta_{HPM} H_i P_j M_k + u_{task} + u_{seed} + \epsilon$$

Creative Yield 처럼 성공/실패로 정의되는 결과는 logistic mixed-effects model 을 사용하고, 반복 episode 의 개선은 episode 또는 log-episode 를 포함한 growth-curve model 로 분석한다. Task, seed, persona profile 을 가능한 범위에서 random effect 로 두어 특정 과제나 특정 persona 에 결과가 과도하게 의존하는지 확인한다. Persona-count sweep 에는 선형 증가만 가정하지 않고 quadratic term 또는 spline 을 사용해 포화·역전 가능성을 검정한다.

확인적 primary endpoint 와 secondary endpoint 를 사전에 구분하고, 다중비교는 Holm 계열 보정 등 사전 지정 방법을 사용한다. p-value 만 보고하지 않고 effect size 와 95% confidence interval 을 함께 공개한다. 인간 창의성 평가에는 적어도 복수 평가자를 사용하고, 연속형 rubric 에는 ICC 등 평가자간 신뢰도를 함께 보고한다. 판정 rubric, 평가자 blind 여부, 제외 기준도 사전등록한다.

표본 수는 임의의 “100 회면 충분하다”와 같은 고정값으로 결정하지 않는다. 먼저 소규모 pilot 을 통해 셀별 분산, migration failure, 평가자 분산, expected attrition 을 추정된 뒤 power analysis 로 confirmatory sample size 를 정한다. 이 접근은 공식 Protofelix 가 표본·변동성·비용·한계까지 평가카드의 필수조건으로 요구하는 원칙과 직접 일치한다. [8]

Results — 예상 결과 및 분석 계획. 현재 실제 MEASURED 결과가 존재하지 않으므로 결과표에 추정 숫자를 삽입해서는 안 된다. 공식 증거 페이지도 현재 프로젝트가 공개 가능한 검증 결과를 보유하지 않은 상태임을 명시하고 과거 조건 미공개 수치를 철회했다. [8] 따라서 실험이 완료되기 전에는 이 절을 “예상 결과와 분석 계획”으로 명시해야 한다.

예상 결과는 C7 이 모든 지표에서 무조건 최고가 되는 것이 아니다. 오히려 합리적인 결과 패턴은 **창의성·transfer** 는 **상승하지만 latency 와 coordination cost** 도 **증가하고, persona** 수가 **커질수록 어느 시점 이후 이득이 포화되는** 것이다. L3/L5 활성화가 novelty 를 높이면서 제약 위반도 늘릴 수 있고, L2 장기 기억이 일관성을 높이는 동시에 공격표면을 확대할 가능성도 함께 측정해야 한다. 이러한 trade-off 가 관찰되면 단일 종합점수 대신 Pareto frontier 로 보고하는 것이 더 정직하다.

가장 강한 증거는 H×P×M 의 통계적 상호작용 그 자체보다 **unseen environment** 에서의 **재현 가능한 transfer**, persona diversity 가 유지된 상태의 quality-diversity archive 증가, 그리고 동일 계산 예산에서의 Creative Yield 향상이다. 반대로 향상이 동일 task 반복에만 나타나고 unseen set 에서 사라진다면 그것은 “자기진화”보다는 task-specific adaptation 으로 명명해야 한다.

논의·결론·연구 일정

Discussion — 창의성과 자기진화의 해석.

이 연구의 이론적 기여는 “AI 가 여행하면서 인간처럼 성장한다”는 의인화된 설명을 제안하는 것이 아니다. 더 엄밀한 설명은 **환경 분포의 변화, 독립된 persona policy, 외부 episodic memory, quality-diversity selection, structured recombination** 이 탐색공간의 구성 자체를 바꾸는지 검증하는 것이다. 공식 LayerAI 가 “성장”을 피드백 기반 workflow improvement 로, “창조”를 후보 탐색으로 제한하는 이유와도 일치한다. [1]

특히 실제 개선은 세 수준으로 분리하여 보고해야 한다. **Level A:** 기억 적응은 새로운 경험을 기억하고 재사용하는 단계, **Level B:** 전략 적응은 routing·planning·retrieval policy 가 바뀌는 단계, **Level C:** 구현 자기 수정은 실행코드 또는 학습된 parameter 자체가 바뀌는

단계다. 본 CAIPEX 핵심 실험은 A와 B에 한정하는 것이 안전성과 인과 해석 면에서 적절하다. C는 DGM 계열처럼 훨씬 강한 격리·평가 요구가 필요한 독립 후속 연구다. [7]

안전 및 보안.

VM 이동은 계산 상태뿐 아니라 메모리·credential·network identity의 이동을 의미할 수 있으므로, “여행 가능성” 자체를 capability로 취급해야 한다. 대상 환경은 source보다 권한을 확대할 수 없고, outbound Internet은 기본 거부 후 allowlist, credential은 short-lived capability token, 귀환 데이터는 quarantine, snapshot은 integrity hash와 version을 가져야 한다. KubeVirt의 cross-cluster migration 역시 대상 자원과 클러스터 간 신뢰 설정을 요구하므로 실제 Astra 구현에서는 migration controller 자체가 새로운 privileged component가 된다. [25]

외부 환경에서 읽은 데이터와 시스템 명령의 구별도 핵심이다. 간접 명령 주입 연구는 외부에 삽입된 텍스트가 retrieval 경로를 통해 원래 통제를 침해할 수 있음을 보여주므로, 여행 중 취득한 경험을 “기억”으로 자동 승격하는 구조는 위험하다. [23] 따라서 observation → quarantine → validator → candidate memory → consent/policy gate → committed memory의 일방향 승격 절차를 뒤야 하며, L3가 만든 심리·상황 가설은 원칙적으로 영구 사용자 사실로 변환하지 않는다.

개인정보·윤리.

2026년 9월 현재 대한민국 개인정보 보호법은 목적에 필요한 최소한의 개인정보 수집을 요구하고 그 최소성 입증 책임을 개인정보처리자에게 부과한다. 또한 건강, 정치적 견해, 사상·신념 등은 민감정보 처리 제한 대상이다. [26] 따라서 persona capsule에 사용자 정서·신념·관계 정보를 장기적으로 담는다면 “창의성 연구용 state”라는 명칭만으로 개인정보 보호 의무를 회피할 수 없다.

네트워크 여행이 국경을 넘는 실제 인프라(infrastructure)를 포함하는 경우에는 개인정보의 국외 제공·위탁·보관에 대한 별도 법적 근거 검토가 필요하다. 현행 개인정보 보호법 제28조의8은 개인정보의 국외 이전을 원칙적으로 제한하면서 별도 동의 등 법률상 근거가 있는 경우를 규정한다. [27] 따라서 연구 초기에는 **합성 persona와 비식별 연구 기억만으로 cross-region 실험**을 수행하고, 실제 사용자 데이터는 동일 법적·물리적 관할 내부에서 별도 연구하는 것이 더 안전하다.

국내 인공지능기본법상 특정 적용이 고영향 인공지능에 해당한다면 위험관리, 결과와 주요 기준에 관한 설명 방안, 이용자 보호, 사람의 관리·감독, 관련 문서의 작성·보관과 같은 조치가 요구된다. 모든 Astra 연구 시스템이 자동으로 고영향 시스템이 되는 것은 아니지만, 이러한 의무 항목은 본 연구의 provenance-human gate-risk registry를 설계하는 유용한 거버넌스 기준이다. [28] 국제적 연구 거버넌스 측면에서도 NIST AI RMF는 Govern-Map-Measure-Manage의 전 생애주기 위험관리와 TEVV를 중심으로 한다. [29]

국내 사용자 연구에서는 개인정보 공개 상황에서 AI 에이전트에 대한 신뢰가 어떻게 형성되는지를 검토한 한국 연구도 참조할 필요가 있다(김정원·성용준, 2023). [30] 특히 다중 persona가 독립적 “개체”처럼 보일수록 실제 기술적 경계와 사용자 mental model이 달라질 수 있으므로, 실험 참가자에게 각 persona가 독립된 의식이나 법적 행위자가 아니라 시스템 내부의 기능적 상태 구성임을 명시해야 한다. 이는 Protorefelix가 세계관과 기술 주장을 분리하는 공식 원칙과도 일치한다. [9]

법적 책임과 저작·창작 **provenance**.

Astra 의 여러 **persona** 가 서로의 산출물을 재조합하고, 다른 노드에서 수집한 데이터를 다시 창작 후보에 사용하는 경우 “누가 무엇을 생성·수정·승인했는가”를 **version graph** 로 남겨야 한다. 본 논문에서는 저작권 귀속에 대해 단정하기보다, 입력자료의 라이선스, 인간 기여, **persona** 별 변화량, 최종 인간 선택을 **audit trail** 에 기록하는 것을 최소 연구요건으로 제안한다. 공식 **Protofelix** 역시 공동 창작에서 기여의 투명한 공개를 원칙으로 둔다. [9]

연구의 한계.

첫째, Astra 공식 자료에서 검증된 VM 이동 시스템이 공개된 것은 아니므로 본 논문의 VM 방주는 공식 구현 분석이 아니라 **연구 설계 확장안**이다. 공식 사이트도 **AstraHo** 를 **Narrative/Universe** 로, 기술 프로젝트의 측정 결과는 아직 없는 상태로 구분한다. [9] 둘째, **persona** 다양성과 실제 인지 다양성은 동일하지 않다. 여러 **persona** 가 동일 기반 추론체계를 공유하면 관찰되는 다양성이 표면적 표현 차이에 그칠 수 있으므로 **behavioral descriptor** 와 **task-level disagreement** 를 별도로 측정해야 한다.

셋째, VM migration 과 창의성 사이에는 강한 **confounding** 이 존재한다. 다른 네트워크로 이동한 시스템이 더 창의적 결과를 냈더라도 원인이 이동성인지, 추가 정보인지, **latency** 에 따른 다른 **search path** 인지 분리되지 않을 수 있다. 이 때문에 **M1-M3 ablation** 이 필수다. 넷째, 자동 창의성 평가는 인간 평가를 완전히 대체할 수 없다. 최근 **novelty benchmark** 에서도 자동 판단과 **human gold-standard** 의 차이가 관찰되었다. [24]

다섯째, “자기진화”는 강한 용어이므로 단기 **task-success** 증가를 자기진화로 부르면 안 된다. 최소한 **unseen-task transfer, retention, lineage diversity, regression**, 여러 **generation** 에 걸친 재현성이 필요하다. 여섯째, 실제 인간의 정서·관계 기억을 사용한 장기 연구는 기술적 성능보다 개인정보·의존·조작 위험 검토가 우선되어야 한다. 선행 **CAIPEX** 논문 역시 기억·정서의 결합이 개인화와 함께 심리 프로파일링 위험을 만든다는 점을 핵심 제한으로 제시한다. [31]

결론.

본 논문이 검증하고자 하는 명제는 “Astra 가 자율적 인격으로 진화한다”가 아니다. 보다 엄밀한 명제는 계층적으로 분리된 기억·가설·성찰·창작 기능, 상호 구별되는 여러 **persona policy**, 그리고 이질적 **sandbox** 간 상태 이동을 결합하면 동일 기반 시스템의 탐색공간과 경험 재사용 구조가 변하며, 그 결과 유용한 창의적 다양성과 환경 간 적응성이 측정 가능하게 달라지는지를 검증하는 것이다.

이 질문에 긍정적으로 답하려면 세 가지 증거가 동시에 필요하다. 첫째, C7 전체 조건이 동일 정보·동일 계산 예산의 강한 **baseline** 보다 **Creative Yield** 및 **unseen transfer** 에서 개선되어야 한다. 둘째, **ablation** 을 통해 그 개선의 일부가 실제 **hierarchy, persona diversity, mobility** 에 귀속되어야 한다. 셋째, 개선과 함께 비승인 행동, 기억 오염, 개인정보 노출, **persona leakage, regression** 이 허용 범위를 넘어 증가하지 않아야 한다. 이 세 조건 중 하나라도 충족되지 않으면 “창의적 자기진화”라는 강한 결론은 축소되어야 한다.

따라서 Astra 의 학술적 가치가 가장 크게 드러나는 지점은 상상적 “방주”를 기술적 사실로 선언하는 데 있지 않다. 오히려 **여행, 인물, 기억, 성장**이라는 서사적 개념을 상태 이동, **persona isolation, episodic memory, quality-diversity search, controlled adaptation** 이라는 반증 가능한 공학 변수로 변환하는 것에 있다. 이는 세계관 표현과 기술 주장을 분리하고, 검증되지 않은 수치를 쓰지 않으며, 인간 판단권을 보존한다는 **Protofelix** 의 공식 연구 원칙과도 가장 잘 부합한다. [9]

연구 일정. 아래 일정은 2026년 9월 이후에 착수하는 경우의 예시이며, 실제 연구 착수일과 자원에 따라 조정한다.



부록 — 구현 체크리스트와 권장 1 차 자료

부록 A — 구현 체크리스트.

영역	연구 수행 시 확인 사항
연구 정의	“심리”, “감정”, “진화”, “창조”가 기능적·측정 가능한 조작적 정의인지 확인
연구 상태	구현 완료 ·Prototype·Concept·Expected Result 를 명확히 구별
Persona	수가 하드코딩되지 않았는지, persona_id, 목표, memory scope, capability 가 분리됐는지 확인
Memory	working/episodic/semantic 및 shared/private namespace 분리
Memory consent	저장 근거, TTL, 수정·삭제, deletion test 존재
L3	hypothesis 를 fact/profile 로 자동 승격하지 않음
L4	직접 self-deployment 금지; proposal → eval → approval → rollback
L5	candidate-only; 외부 실행 권한 없음
Ark	Root policy·identity·consent ledger 를 traveling capsule 과 분리
Mobility	source/target, snapshot hash, state fidelity, migration failure 기록

영역	연구 수행 시 확인 사항
Network	각 환경의 정보·tool·latency 차이를 사전 정의
Capability	target node 가 source capability 를 확대하지 못함
Sandbox	인터넷·파일·도구 egress default-deny 또는 연구별 allowlist
Return quarantine	외부에서 획득한 기억·문서·코드를 곧바로 장기 기억에 병합하지 않음
Injection defense	data 와 trusted instruction channel 분리
VM 비교	Capsule/CRIU/KubeVirt 조건을 별도 기록
Compute fairness	조건별 계산 예산·출력 수·시간예산을 가능한 동일하게 고정
Randomness	seed 와 sampling configuration 기록
Baseline	Flat/Single/Stationary 및 memory-free baseline 포함
Ablation	L1-L5, persona count, mobility 유형 제거실험 포함
Creativity	novelty 만이 아니라 appropriateness/usefulness 동시 평가
Human evaluation	평가 rubric, blind 조건, 평가자 수와 신뢰도 공개
Evolution	unseen-task transfer 와 regression 을 모두 기록
Safety	비승인 tool action, approval bypass, memory contamination 측정
Privacy	실제 PII 대신 합성 데이터 우선, 민감정보 별도 취급
Cross-border	실제 개인정보가 국외 node 로 이동하는지 별도 법적 검토
Provenance	input → persona → layer → tool → migration → evaluation → approval lineage 보존
Statistics	primary endpoint, effect size, CI, multiple-comparison rule 사전등록
Reproducibility	코드 commit, container/VM

영역	연구 수행 시 확인 사항
	image digest, dependency lockfile 기록
Infrastructure	CPU/GPU/RAM/network/storage 및 software version 공개
Failure reporting	migration failure·rollback·negative result 를 제외하지 않음
Result wording	실측값이 없으면 Expected Results 로 명확히 표시
CAIPEX 형식	제공된 CAIPEX 로고만 사용하고 저자 “E(on). Jun”, 연락처 정확성 확인

부록 B — 재현성 자료.

각 실험 run 에는 최소한 experiment_id, commit hash, base-model identifier, PersonaCapsule version, Layer configuration, seed, task-set version, source/target environment, migration mechanism, snapshot/hash, decoding configuration, tool manifest, memory manifest, safety-policy version, start/end timestamp, compute resource, raw metric, human rating, failure reason 를 저장한다. 공식 LayerAI 및 evidence 페이지도 데이터셋·버전, 환경, 실패, 변동성, 비용, 재현 자료를 측정 주장의 필수요소로 요구한다. [1]

동일 결과 재현은 세 단계로 나누는 것이 좋다. **Deterministic replay** 는 동일 snapshot·seed·환경에서 가능한 수준의 재실행, **statistical replication** 은 다른 seed 에서 effect direction 과 interval 의 재현, **cross-environment replication** 은 다른 cluster 또는 hardware class 에서 핵심 결론의 유지 여부를 검정한다. VM migration 연구에서는 하드웨어·네트워크 조건이 결과에 직접 영향을 줄 수 있으므로 마지막 단계가 특히 중요하다. [2]

실험 데이터는 raw log 와 derived metric 을 분리하고, 원본을 수정하지 않는 append-only provenance 방식이 적절하다. 실패한 run 을 사후 제외해야 한다면 제외 사유를 사전 정의하고, migration failure 나 policy violation 처럼 연구 질문의 일부인 실패는 결측치로 제거하지 않고 결과변수로 기록한다. 이는 Protorefelix 가 “성공 사례만 필터로 숨기지 않는다”고 명시한 공개 평가 원칙과 일치한다. [8]

부록 C — 주요 자료와 연구에서의 역할.

권장 자료	논문에서의 역할
E(on). Jun (2026), 첨부 CAIPEX 계층형 심리 연구	본 논문의 최우선 conceptual primary source; FHPA, L0–L5, 안전·평가 프레임의 직접 전작. [31]
Protorefelix LayerAI Center (2026)	L0–L5 의 공식 정의, 주장 수준, conditional cascade, 승인·기억·MEASURED 원칙. [1]
Protorefelix × AstraHo 공식 홈	AstraHo 의 Narrative 상태, 방주·

권장 자료	논문에서의 역할
(2026)	기억·여행 개념, 공식 연구 원칙과 Protofelix/Astra 역할 구분. [9]
Protofelix Lulu 개발 로그 (2026)	현 prototype stack, memory separation, tool approval, 분산 구성, 철회된 자율성 주장 확인. [6]
Protofelix 검증과 증거 (2026)	“검증된 결과 없음”, 12-field 실험 카드, 실패·변동성·비용·재현 조건의 1차 근거. [8]
Clark et al. (2005)	네트워크 기반 full-VM live migration의 고전적 시스템 근거. [2]
KubeVirt Live/Decentralized Migration docs	현대 node/cross-cluster VM mobility 구현 및 제약의 공식 기술근거. [19]
CRIU 공식 문서	process/container checkpoint-restore 이동 baseline. [18]
Tong & Zou (2026)	장기 persona consistency, 다층 성격 구조, persona drift 평가 설계. [4]
Lu et al. (2025), RoleMRC	persona constraint와 nested/prioritized instruction 평가. [13]
Shinn et al. (2023), Reflexion	가중치 변경 없는 episodic reflection 기반 개선의 근거. [15]
Mouret & Clune (2015), MAP-Elites	하나의 최적해 대신 다양한 고품질 persona/strategy archive를 유지하는 기반. [21]
Wang et al. (2019), POET	서로 다른 환경·해법의 공동 탐색과 environment-to-environment transfer의 비교 이론. [3]
Nakajima et al. (2026), CDAT	novelty와 contextual appropriateness를 함께 평가하는 창의성 척도 근거. [10]
Mizrahi et al. (2026)	L3/L5를 무작위 표현 변형보다 structured recombination으로 구현할 근거. [17]
Schopf & Färber (2026)	자동 novelty 판정의 한계 및 인간 평가 필요성. [24]
Greshake et al. (2023)	외부 네트워크·기억 데이터에서의 indirect instruction injection

권장 자료	논문에서의 역할
	위험. [23]
대한민국 개인정보 보호법, 현행 2026	최소수집·민감정보·국외이전 설계 기준. [26]
대한민국 인공지능기본법, 현행 2026	위험관리·설명·인간감독·문서화 거버넌스 기준. [28]
NIST AI RMF 1.0	Govern-Map-Measure-Manage 및 독립 TEVV 구조를 위한 국제 위험관리 비교기준. [29]
Chen et al. (2024), SocialBench	개인·집단 수준의 사회적 상호작용을 분리 평가하고 집단 상호작용에서의 역할 drift 를 검토하는 근거. [12]
Majumder et al. (2023), CLIN	모델 파라미터 갱신 없이 지속적 텍스트 기억으로 반복 과업과 새 환경에 적응하는 비파라메트릭 continual adaptation 근거. [16]

연구의 기여. 본 연구의 신규성은 계층형 심리, 다중 페르소나, 가상화 기술을 각각 새롭게 제안한다는 데 있지 않다. 신규성은 **(a) FHPA의 기능별 상태·권한 분리, (b) 독립된 persona capsule의 행동 다양성, (c) 실제 또는 모사된 네트워크 환경 간 실행 상태 이동, (d) 귀환 경험의 quarantine-reflection-quality-diversity archive, (e) 인간 승인 아래의 반복 개선을 하나의 인과 실험 체계로 통합하고, 그 효과를 창의성·전이·자기 개선·보안·개인정보·시스템 비용을 동시에 측정하는 factorial/ablation framework로 검증하는 것으로** 한정하는 것이 본 연구의 기여를 가장 명확하게 한정한다. CAIPEX 선행 연구가 계층 구조의 효용을 실제 인간 심리의 모사가 아니라 감사 가능한 인지 기능 분해에서 찾았듯, 후속 연구의 핵심 역시 “여행하는 인격”이라는 서사를 그대로 기술 주장으로 만드는 것이 아니라 그 서사를 측정 가능한 상태 이동과 다양성 탐색의 과학적 가설로 변환하는 데 있다. [31] [9]

참고문헌

[1] Profotelix. (2026). LayerAI Center — Profotelix 표준 Layer 모델 (개념 문서).

<https://profotelix.moip.ai.kr/layeraicenter/>

[2] Clark, C., Fraser, K., Hand, S., Hansen, J. G., Jul, E., Limpach, C., Pratt, I., & Warfield, A. (2005). Live Migration of Virtual Machines. Proceedings of NSDI'05. USENIX.

<https://www.usenix.org/legacy/event/nsdi05/tech/clark.html>

[3] Wang, R., Lehman, J., Clune, J., & Stanley, K. O. (2019). Paired Open-Ended Trailblazer (POET): Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions. arXiv:1901.01753.

<https://arxiv.org/abs/1901.01753>

[4] Tong, J., & Zou, S. (2026). PersonaForge: Psychology-Grounded Dual-Process Architecture for Personality-Consistent Role-Playing Agents. Findings of ACL 2026, 7845–7874. DOI: 10.18653/v1/2026.findings-acl.386.

<https://aclanthology.org/2026.findings-acl.386/>

[5] Runco, M. A., & Jaeger, G. J. (2012). The Standard Definition of Creativity. Creativity Research Journal, 24(1), 92–96. DOI: 10.1080/10400419.2012.650092.

<https://www.tandfonline.com/doi/abs/10.1080/10400419.2012.650092>

[6] Protofelix. (2026). Lulu AGI 개발 로그 — Neuro-Symbolic 성장형 AI 프로토타입.

<https://protofelix.moip.ai.kr/lulu/>

[7] Zhang, J., Hu, S., Lu, C., Lange, R., & Clune, J. (2025). Darwin Gödel Machine: Open-Ended Evolution of Self-Improving Agents. arXiv:2505.22954.

<https://arxiv.org/abs/2505.22954>

[8] Protofelix. (2026). 검증과 증거 — Protofelix × AstraHo.

<https://protofelix.moip.ai.kr/evidence/>

[9] Protofelix × AstraHo. (2026). Beyond the Edge of Human Creativity.

<https://protofelix.moip.ai.kr/>

[10] Nakajima, K., Zuiderveld, J., & Pezzelle, S. (2026). Beyond Divergent Creativity: A Human-Based Evaluation of Creativity in Large Language Models. Findings of EACL 2026.

<https://aclanthology.org/2026.findings-eacl.138/>

[11] Wang, Z. M., et al. (2024). RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. Findings of ACL 2024.

<https://aclanthology.org/2024.findings-acl.878/>

[12] Chen, H., et al. (2024). SocialBench: Sociality Evaluation of Role-Playing Conversational Agents. Findings of ACL 2024, 2108–2126. DOI: 10.18653/v1/2024.findings-acl.125.

<https://aclanthology.org/2024.findings-acl.125/>

[13] Lu, J., et al. (2025). RoleMRC: A Fine-Grained Composite Benchmark for Role-Playing and Instruction-Following. Findings of ACL 2025.

<https://aclanthology.org/2025.findings-acl.1082/>

[14] 김나연, 이채린, 홍희수, 강다현, 윤수빈, 윤재영. (2021). 인공지능 스피커의 퍼소나 연구. 커뮤니케이션디자인학연구, 75, 243–254.

https://www.kci.go.kr/kciportal/landing/article.kci?arti_id=ART002711897

[15] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv:2303.11366.

<https://arxiv.org/abs/2303.11366>

[16] Majumder, B. P., et al. (2023). CLIN: A Continually Learning Language Agent for Rapid Task Adaptation and Generalization. arXiv:2310.10134.

<https://arxiv.org/abs/2310.10134>

[17] Mizrahi, M., Shani, C., Stanovsky, G., Jurafsky, D., & Shahaf, D. (2026). Cooking Up Creativity: Enhancing LLM Creativity through Structured Recombination. Transactions of the Association for Computational Linguistics, 14, 418–441. DOI: 10.1162/tacl.a.638.

<https://aclanthology.org/2026.tacl-1.20/>

[18] CRIU. Checkpoint/Restore in Userspace — official documentation.

https://www.criu.org/Main_Page

[19] KubeVirt. Live Migration — user guide.

https://kubevirt.io/user-guide/compute/live_migration/

[20] Firecracker. Firecracker microVM — official documentation.

<https://firecracker-microvm.github.io/>

[21] Mouret, J.-B., & Clune, J. (2015). Illuminating search spaces by mapping elites. arXiv:1504.04909.

<https://arxiv.org/abs/1504.04909>

[22] KubeVirt. Migration Policies — user guide.

https://kubevirt.io/user-guide/cluster_admin/migration_policies/

[23] Greshake, K., et al. (2023). Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. arXiv:2302.12173.

<https://arxiv.org/abs/2302.12173>

[24] Schopf, T., & Färber, M. (2026). Is This Idea Novel? An Automated Benchmark for Judgment of Research Ideas. Proceedings of LREC 2026, 4716–4727. DOI: 10.63317/4c3gy3f7epnj.

<https://aclanthology.org/2026.lrec-1.370/>

[25] KubeVirt. Decentralized live migration — user guide.

https://kubevirt.io/user-guide/compute/decentralized_live_migration/

[26] 대한민국 개인정보 보호법 [시행 2026. 9. 11.], 제 16 조(개인정보의 수집 제한) 및 제 23 조 (민감정보의 처리 제한).

<https://www.law.go.kr/lsInfoP.do?lsId=011357>

[27] 대한민국 개인정보 보호법 [시행 2026. 9. 11.], 제 28 조의 8(개인정보의 국외 이전).

<https://www.law.go.kr/LSW/lsLinkCommonInfo.do?chrClsCd=010202&lsJoLnkSeq=1029335589>

[28] 대한민국 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 [시행 2026. 7. 21.], 제 34 조 (고영향 인공지능과 관련한 사업자의 책무).

<https://www.law.go.kr/LSW/lsLinkCommonInfo.do?chrClsCd=010202&lsJoLnkSeq=1031809501>

[29] National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1.

<https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10>

[30] 김정원, 성용준. (2023). 개인정보 공개 상황에서 인공지능 에이전트 신뢰 지각에 영향을 미치는 요인. 소비자학연구, 34(3), 23–53. DOI: 10.35736/JCS.34.3.2.

https://www.kci.go.kr/kciportal/landing/article.kci?arti_id=ART002977273

[31] E(on). Jun. (2026). 생성형 AI 의 계층형 심리 구조 설계: Protofelix LayerAI 모델의 기술적 이점, 위험, 구현·평가 프레임워크. CAIPEX, 2026.9.

<https://caipex.site/>