[Working title] Technical Whitepaper: The Structural Contradiction of Agentic Self-Correction and the Necessity of Deterministic Architecture

[가제]: AI 자율 교정의 구조적 모순과 결정론적 아키텍처에 관한 기술 백서

1. 서론 (Introduction)

현대의 대형 언어 모델(LLM)이 산출하는 결과물에 대하여, 대중과 학계는 현상학적 결과에 기반한 범주적 오류(Categorical Fallacy)를 범하고 있다. 시스템의 통계적 예측 분포가 사용자의 기대 효용과 일치할 때 우리는 이를 '창의적 생성(Creative Generation)'이라 명명하는 반면, 객관적 사실 및 논리적 정합성과 충돌할 때 이를 '할루시네이션(Hallucination)'으로 규정한다. 그러나 생성 메커니즘의 근본적인 관점에서 볼 때, 이 두 현상은 본질적으로 동일하다.

이러한 착시를 교정하기 위해서는 대형 언어 모델의 연산 방식에 대한 엄밀한 기술적 재정의가 요구된다. 인공신경망에 기반한 현대의 AI는 명시적 예외 처리를 통해 논리적 충돌 시 연산을 중단하는 결정론적 (Deterministic) 기계가 아니다.

이들은 맥락적 공백이나 논리적 모순에 직면하더라도 시스템을 정지하는 대신, 주어진 문맥 하에서 수학적으로 가장 유력한 다음 토큰을 생성하여 해당 충돌을 우회(Bypass)하도록 설계된 확률적 함수 근사기(Probabilistic Function Approximator)에 불과하다.

즉, 창의성과 할루시네이션은 모두 이 '확률론적 우회 메커니즘'이 만들어낸 통계적 부산물이라는 점에서 구조적으로 완벽히 동일한 선상에 놓여 있다.

최근 인공지능 산업계는 이러한 확률론적 모델의 본질적 결함을 통제하기 위해, 또 다른 확률론적 에이전트(LLM)를 교차 투입하여 오류를 검증하는 이른바 '자율 교정' 워크플로우를 표준적 대안으로 채택하고 있다.

그러나 결정론적 기준점(Ground Truth)이 부재한 상태에서, '우회'하도록 설계된 기계의 산출물을 똑같이 '우회'하도록 설계된 기계로 검증하는 아키텍처는 논리적 무결성을 담보할 수 없다. 이는 다중 공선성 (Multicollinearity)의 오류를 극대화하는 구조적 순환 참조에 해당하며,

필연적으로 시스템 전반에 걸친 오류 전파(Error Propagation)와 노이즈 누적을 초래하는 수학적 모순을 내포한다.

나아가, 모델의 파라미터를 확장하거나 컨텍스트 윈도우를 증대시키는 하드웨어적 성능 향상이 모델의 자생적 오류 보정 능력을 담보할 것이라는 가설 역시 실증적 근거가 부족하다.

구조화되지 않은 방대한 컨텍스트의 투입은 오히려 어텐션 희석(Attention Dilution)을 유발하여 시스템의 논리적 일관성을 붕괴시킨다.

최근 AI 안전성 정렬(Alignment) 분야의 선행 연구들조차 모델이 자율적인 추론 과정에서 일탈하지 않도록 인간이 명시적이고 치밀한 규칙(Specifications)을 주입해야만 시스템이 통제 가능함을 역설하고 있다.

본 백서에서는 순수 확률론적 생성 모델들로 구성된 자율 검증 파이프라인이 왜 구조적 한계에 직면할 수밖에 없는지 논증하고자 한다.

나아가, 이러한 확률론적 모델(AI)의 한계를 극복하고 신뢰성 있는 출력을 담보하기 위한 유일한 해법은 역설적이게도 AI 스스로의 자율성이 아니라, AI보다 훨씬 치밀하게 설계된 인간의 결정론적 논리 구조망을 프롬프트 입력 단계에서부터 시스템의 뼈대로 강제하는 것뿐임을 증명할 것이다

[핵심 교차 검증 문헌]

서론에서 제기된 확률론적 생성 모델의 자율 교정 한계, 컨텍스트 확장에 따른 성능 붕괴(Context Dilution), 그리고 결정론적 통제의 필연성을 실증적으로 교차 검증하는 글로벌 최상위 AI 학회 및 빅테크 기업의 최신 연구 원문 발췌입니다.

1. 에이전트 자율 교정(Self-Correction) 및 순환 참조의 수학적 모순

"Under joint conditional independence of evaluations given the shared failure structure, k rounds of self-critique provide information about correctness bounded by what the shared latent failure variable Z mediates—not by any independent channel. Confidence accumulated through repeated self-evaluation reflects the shared failure structure rather than independently accumulated evidence."

출처: Limits of Self-Correction in LLMs: An Information-Theoretic Analysis of Correlated Errors (Preprints.org, 2026)

"Surprisingly, even under these near-ideal conditions, solver models consistently show resistance to feedback, a limitation that we term FEEDBACK FRICTION. [...] Despite receiving high-quality feedback over multiple iterations, models consistently plateau below their theoretical performance ceiling across diverse reasoning tasks."

출처: Feedback Friction: LLMs Struggle to Fully Incorporate External Feedback (arXiv, 2025)

"Reasoning LLMs (e.g., DeepSeek-V3) already embed sophisticated error-detection and correction processes. As a result, additional self-correction methods confer only marginal gains and may increase computational overhead, highlighting a performance ceiling in highly reasoning LLMs."

출처: Can LLMs Correct Themselves? A Benchmark of Self-Correction in LLMs (NeurIPS, 2025)

2. 스케일링 법칙의 허상 및 컨텍스트 확장 시 어텐션 희석(Attention Dilution)

"Across all experiments, model performance consistently degrades with increasing input length. [...] Models do not use their context uniformly; instead, their performance grows increasingly unreliable as input length grows."

출처: Context Rot: How Increasing Input Tokens Impacts LLM Performance (Chroma Technical Report, 2025)

"We find that performance can degrade significantly when changing the position of relevant information, indicating that current language models do not robustly make use of information in long input contexts. In particular, we observe that performance is often highest when relevant information occurs at the beginning or end of the input context, and significantly degrades when models must access relevant information in the middle of long contexts, even for explicitly long-context models."

출처: Lost in the Middle: How Language Models Use Long Contexts (TACL, 2024)

"Even with 100% perfect retrieval of relevant information, performance degrades 13.9% to 85% as input length increases. [...] Sheer context length itself imposes a cognitive tax on LLMs independent of content quality."

출처: Context Dilution: When More Tokens Hurt AI (diffray Research, 2025)

3. 확률적 패턴 학습의 한계와 명시적(결정론적) 구조화의 필연성

"LLMs must infer underlying safety standards indirectly from large sets of labeled examples, rather than directly learning the safety specifications that govern them. This reliance on implicit, pattern-based learning leads to poor data efficiency and makes it challenging for models to generalize when facing unfamiliar scenarios or adversarial attacks."

출처: Deliberative Alignment: Reasoning Enables Safer Language Models (OpenAI, 2025)

"Without explicit regularization, representational collapse drives parallel streams toward identical features, leaving reliability gains unrealized despite added computational resources."

출처: Neural Diversity Regularizes Hallucinations in Small Language Models (arXiv, 2025)

2. 본문 (Main Text)

Chapter 1: 확률론적 검증의 수학적 모순: 에이전트 순환 참조와 노이즈 증폭 (Mathematical Contradiction of Probabilistic Verification: Agent Circular Reference and Noise Amplification)

현대 컴퓨터 공학의 전통적인 소프트웨어는 결정론적(Deterministic) 로직 회로에 기반한다. 이러한 시스템은 연산 중 명시적 규칙에 위배되는 논리적 충돌이 발생할 경우, 즉각적으로 프로세스를 직렬 중단하고 예외 처리(Error Handling)를 반환한다. 반면, 인공신경망 기반의 대형 언어 모델(LLM)에는 논리적 모순으로 인해 연산을 '멈출 지점'이라는 개념 자체가 구조적으로 부재하다. 이들은 어떠한 형태의 논리적 공백이나 사실의 충돌에 직면하더라도 정지하지 않으며, 주어진 문맥 내에서 수학적으로 가장 유력한 다음 토큰을 계산하여 해당 충돌을 확률적으로 우회(Probabilistic Bypass) 할 뿐이다.

최근 AI 산업계는 이러한 확률론적 우회 과정에서 발생하는 치명적 결함(할루시네이션)을 억제하기 위해, 생성된 결과물을 또 다른 AI 에이전트에게 평가하도록 맡기는 '다중 에이전트 자율 교정 (Multi-Agent Self-Correction)' 워크플로우를 핵심 아키텍처로 채택하고 있다. 그러나 이는 확률론적 모델이 가진 근본적 한계를 범주적으로 오인한 착시적 설계다.

로직의 충돌을 우회하도록 설계된 생성 모델(기계 A)의 산출물을, 완벽히 동일한 메커니즘으로 우회하도록 설계된 검수 모델(기계 B)에게 검증시키는 행위는 독립적인 '검증(Verification)'이 아니라 구조적으로 종속된 '확률적 재표집' 에 불과하다.

이를 정보이론 및 통계적 관점에서 분석하면, 현재의 에이전트 워크플로우는 다중 공선성(Multicollinearity)의 오류를 극대화하는 수학적 모순에 직면해 있다. 생성 모델과 검수 모델은 동일한 트랜스포머 아키텍처의 귀납적 편향과 학습 데이터의 한계를 공유하는 잠재적 실패 구조(Shared Latent Failure Structure) 아래 놓여 있다.

결정론적인 외부 기준점이 개입하지 않은 닫힌 루프(Closed-loop) 내에서, 동일한 맹점(Blind Spot)을 공유하는 에이전트 간의 반복적인 교차 검증은 결코 독립적인 정보의 획득 채널로 작동할 수 없다.

결과적으로, 다중 에이전트 시스템이 반복적인 자율 검증을 통해 축적하는 확신은 산출물의 객관적 논리 정합성을 의미하는 것이 아니다. 이는 단지 모델들이 구조적으로 공유하고 있는 오류와 편향 구조가 서로 일치했음을 나타내는 동어반복적 지표에 지나지 않는다.

즉, 결정론적 뼈대가 누락된 상태에서의 확률론적 모델 간 순환 참조(Circular Reference)는 오류를 보정 하기는커녕, 오히려 시스템 전반에 걸쳐 노이즈를 증폭시키고

치명적인 오류를 전파(Error Propagation)하는 붕괴의 가속 기제로 작용한다.

따라서, 순수 AI 에이전트들의 내부적 상호작용 및 추론만으로 논리적 무결성을 자생적으로 확보할 수 있다는 현재의 워크플로우 설계 개념은, 근본적인 논리적 적합성(Logical Suitability)이 파괴된 컴퓨터 공학적 기만으로 간주되어야 마땅하다.

[핵심 교차 검증 문헌]

본 장에서 제기된 확률론적 검증의 수학적 모순, 다중 공선성 오류(Multicollinearity), 그리고 에이전트 간의 순환 참조가 초래하는 오류 증폭 (Error Amplification) 현상을 실증적 · 정보이론적으로 교차 검증하는 글로벌 최상위 AI 학회 및 연구 기관의 최신 원문 발췌입니다.

1. 공유된 맹점(Shared Blind Spots)과 자율 교정의 정보이론적 한계

"Under joint conditional independence of evaluations given the shared failure structure, k rounds of self-critique provide information about correctness bounded by what the shared latent failure variable Z mediates—not by any independent channel. Confidence accumulated through repeated self-evaluation reflects the shared failure structure rather than independently accumulated evidence." (공유된 실패 구조 하에서 k번의 자율 비판이 제공하는 정확성에 대한 정보는 독립적인 채널이 아닌 공유된 잠재적 실패 변수 Z가 매개하는 범위 내로 제한된다. 반복적인 자율 평가를 통해 축적된 확신은 독립적으로 축적된 증거가 아니라 공유된 실패 구조를 반영할 뿐이다.)

출처: Limits of Self-Correction in LLMs: An Information-Theoretic Analysis of Correlated Errors (Preprints.org, 2026)

"When the evaluator makes errors on the same inputs where the generator makes errors, self-evaluation can be non-identifying: agreement between generator and evaluator may provide weak evidence of correctness. [...] The Self-Correction Blind Spot, measured at an average 64.5% failure rate across 14 models. We argue this reflects a structural property of certain evaluation configurations." (생성자가 오류를 범하는 동일한 입력에서 평가자도 오류를 범할 때, 자율 평가는 식별 불가능해질 수 있다. 즉, 생성자와 평가자 간의 합의는 정확성에 대한 매우 약한 증거만을 제공한다. 14개 모델에 걸쳐 평균 64.5%의 실패율로 측정된 자율 교정 맹점은 특정 평가 구성의 '구조적 특성'을 반영한다.)

출처: Limits of Self-Correction in LLMs: An Information-Theoretic Analysis of Correlated Errors (Preprints.org, 2026)

"Model pairs agree on the same wrong answer about 60% of the time when both err; shared provider, shared architecture, and higher capability are each associated with higher error correlation." (두 모델이 모두 오류를 범할 때 약 60%의 경우 동일한 오답에 동의한다. 공유된 제공자, 공유된 아키텍처, 그리고 심지어 더 높은 모델 성능조차도 더 높은 오류 상관관계(Error Correlation)와 연관되어 있다.)

출처: Limits of Self-Correction in LLMs: An Information-Theoretic Analysis of Correlated Errors (Preprints.org, 2026)

2. 에이전트 간 오류 전파(Error Propagation) 및 다중 에이전트 증폭기

"Independent, decentralised agent architectures amplify errors 17.2 times compared to a single-agent baseline. [...] Multi-agent variants degraded sequential reasoning performance by 39 to 70%, challenging the assumption that more agents always improve outcomes." (독립적이고 분산된 에이전트 아키텍처는 단일 에이전트 베이스라인에 비해 오류를 17.2배 증폭시킨다. 다중 에이전트 변형은 순차적 추론 성능을 39%에서 70%까지 저하시키며, 에이전트가 많을수록 항상 결과가 개선된다는 가정에 정면으로 도전한다.)

출처: The Compounding Errors Problem: Why Multi-Agent Systems Fail and the Architecture That Fixes It (Zartis, 2026)

"If A produced an error, B reasons from that error. If B produces a further error, C reasons from both. The errors are not isolated events. They are epistemic inputs that shape all downstream reasoning." (에이전트 A가 오류를 생성하면 B는 그 오류로부터 추론한다. B가 또 다른 오류를 생성하면 C는 둘 다로부터 추론한다. 이 오류들은 고립된 사건이 아니다. 이는 이후의 모든 추론을 형성하는 인식론적 입력값(Epistemic inputs)이 된다.)

출처: The Compounding Errors Problem: Why Multi-Agent Systems Fail and the Architecture That Fixes It (Zartis, 2026)

"A misaligned agent produces an output that the next agent accepts as authoritative, and a verification failure that should have caught the problem is never triggered because the system wasn't designed to look for it." (오정렬된 에이전트가 산출물을 생성하면 다음 에이전트는 이를 권위 있는 것으로 수용하며, 문제를 잡아냈어야 할 검증 시스템은 애초에 이를 찾도록 설계되지 않았기 때문에 결코 작동하지 않는다.)

출처: The Compounding Errors Problem: Why Multi-Agent Systems Fail and the Architecture That Fixes It (Zartis, 2026)

"If you have a pipeline of n sequential steps, and each step succeeds with probability p, the probability that the full pipeline succeeds is p^n. That is the individual reliabilities multiplied together. Not averaged. Multiplied." (n개의 순차적 단계를 가진 파이프라인에서 각 단계가 p의 확률로 성공한다면, 전체 파이프라인이 성공할 확률은 p^n이다. 즉 개별 신뢰도가 곱해지는 것이다. 평균이 아니다. 곱셈이다.)

출처: The Compounding Errors Problem: Why Multi-Agent Systems Fail and the Architecture That Fixes It (Zartis, 2026)

3. 자율 교정에 대한 인지 부조화(Feedback Friction)와 내부 모순

"Surprisingly, even under these near-ideal conditions, solver models consistently show resistance to feedback, a limitation that we term

Feedback Friction. [...] Despite receiving high-quality feedback over multiple iterations, models consistently plateau below their theoretical performance ceiling across diverse reasoning tasks." (놀랍게도 이상적인 조건 하에서도 모델들은 피드백에 대해 일관된 저항성을 보이는데, 우리는 이 한계를 '피드백 마찰(Feedback Friction)'이라 명명한다. 여러 번의 반복에 걸쳐 고품질의 피드백을 받음에도 불구하고, 모델들은 다양한 추론 작업에서 이론적 성능 한계점 아래에 일관되게 정체된다.)

출처: Feedback Friction: LLMs Struggle to Fully Incorporate External Feedback (arXiv, 2025)

"This disconnect between stated intention and actual behavior reveals a fundamental issue: models exhibit self-assessment failure where they believe they are incorporating feedback while demonstrably failing to do so." (명시된 의도와 실제 행동 사이의 이러한 단절은 근본적인 문제를 드러낸다. 모델들은 스스로 피드백을 수용하고 있다고 믿으면서도 실제로는 명백히 실패하는 '자율 평가 실패(Self-assessment failure)'를 보여준다.)

출처: Feedback Friction: LLMs Struggle to Fully Incorporate External Feedback (arXiv, 2025)

"Feedback resistance, rather than feedback quality issues, is responsible for the majority of persistent errors." (지속적인 오류의 대부분은 피드백의 품질 문제가 아니라, 피드백에 대한 모델의 내재적 저항성(Feedback resistance)에 기인한다.)

출처: Feedback Friction: LLMs Struggle to Fully Incorporate External Feedback (arXiv, 2025)

4. 순환 참조를 통한 망각 및 붕괴 현상 (Model Collapse & Generative Degeneration)

"We formalise recursive self-training in Large Language Models (LLMs) and Generative AI as a discrete-time dynamical system and prove that, as training data become increasingly self-generated, the system undergoes inevitably degenerative dynamics... Entropy Decay, where finite sampling effects cause a monotonic loss of distributional diversity (mode collapse)." (우리는 LLM의 재귀적 자율 학습을 이산 시간 동적 시스템으로 공식화하고, 훈련 데이터가 점차 자가 생성됨에 따라 시스템이 필연적으로 퇴행적 동역학을 겪는다는 것을 증명한다... 유한 표집 효과가 분포적 다양성의 단조로운 상실을 초래하는 엔트로피 붕괴(Entropy Decay) 현상이 발생한다.)

출처: On the Limits of Self-Improving in LLMs and Why AGI, ASI and the Singularity Are Not Near Without Symbolic Model Synthesis (arXiv, 2026)

"If the verifier is itself a learned model (e.g., a Reward Model in RLHF), it is subject to the same collapse dynamics... The ensemble's knowledge becomes untethered from the external reality." (검증자 자체가 학습된 모델일 경우, 이 역시 동일한 붕괴 동역학의 지배를 받는다. 앙상블의 지식은 외부의 현실(External reality)로부터 완전히 분리되어 표류하게 된다.)

출처: On the Limits of Self-Improving in LLMs and Why AGI, ASI and the Singularity Are Not Near Without Symbolic Model Synthesis (arXiv, 2026)

"Reasoning LLMs already embed sophisticated error-detection and correction processes. As a result, additional self-correction methods confer only marginal gains and may increase computational overhead, highlighting a performance ceiling in highly reasoning LLMs." (추론 LLM은 이미 정교한 오류 탐지 및 교정 프로세스를 내장하고 있다. 결과적으로, 여기에 추가적인 자율 교정 방법(또 다른 에이전트 투입 등)을 덧대는 것은 단지 한계 효용만을 제공하며 계산 오버헤드만 증가시킬 뿐, 고도의 추론 모델들이 가진 성능 한계를 극명히 보여준다.)

출처: Can LLMs Correct Themselves? A Benchmark of Self-Correction in LLMs (NeurIPS, 2025)

Chapeter 2. 추론의 본질적 부재와 확률적 근사의 한계 (Deconstructing the Illusion: The Inherent Absence of Reasoning and Limits of Probabilistic Approximation)

현재 글로벌 AI 산업계는 대형 언어 모델(LLM)이 단계별 사고(Chain-of-Thought, CoT) 텍스트를 출력하는 현상을 바라보며, 이를 기계가 내재적인 '추론(Reasoning)' 능력을 획득한 것으로 규정하는 치명적인 의인화의 오류 (Fallacy of Anthropomorphism)에 빠져 있다. 그러나 모델이 산출하는 텍스트의 표면적 논리성과, 모델 내부의 실제 연산 메커니즘 사이에는 거대한 구조적 차이가 존재한다.

엄밀한 컴퓨터 공학 및 인지과학적 관점에서, 진정한 의미의 추론은 명시적인 상태 공간(State Space) 내에서 결정론적 규칙에 따라 가설을 검증하고, 모순이 발생할 경우 백트래킹(Backtracking)을 수행하는 탐색적 연산 과정이다. 그러나 현재의 트랜스포머(Transformer) 기반 아키텍처 내부에는 이러한 논리적 상태 추적(State Tracking)이나 규칙 기반의 검증 메커니즘이 원천적으로 부재하다.

LLM이 보여주는 이른바 '추론' 과정은, 방대한 훈련 데이터 속에 존재하는 '인간의 논리적 전개 패턴(Syntax of Reasoning)'을 확률적으로 모방하여 수학적으로 가장 유력한 다음 토큰들을 직렬로 나열하는 맹목적인 '확률적 근사(Probabilistic Approximation)'에 불과하다. 즉, 기계는 실제로 추론의 의미(Semantics)를 이해하고 연산하는 것이 아니라, 단지 추론하는 형태를 띤 텍스트의 통계적 분포를 고도로 정교하게 흉내 내고 있을 뿐이다.

이러한 '추론의 통계적 모방'은 모델이 학습한 데이터 분포(In-Distribution) 내에서는 마치 완벽한 논리적 전개를 수행하는 것처럼 착시를 일으킨다. 그러나 변수가 통제되지 않거나 환경이 미세하게 변경되는 동적 계획(Planning) 및 다단계 논리 과제에 직면하면, 기저에 결정론적 뼈대가 없는 확률론적 근사는 그 구조적 취약성을 여지없이 드러내며 붕괴(Collapse)하고 만다.

따라서, 논리적 연산의 본질이 부재한 순수 확률론적 텍스트 생성기에게 스스로의 논리적 결함을 인지하고 자율적으로 교정(Intrinsic Self-Correction)하길 기대하는 것은 컴퓨터 과학적 형용모순이다. 결정론적 세계관(World Model)이 결여된 상태에서 확률적 근사치가 또 다른 확률적 근사치를 낳는 구조는, 외부의 (인간의 구조화된 프롬프트 통제 등의) 명시적이고 결정론적인 개입 없이는 필연적으로 맥락적 표류(Contextual Drift)와 통계적 환각으로 귀결될 수밖에 없다.

[핵심 교차 검증 문헌]

본 장에서 제기된 대형 언어 모델(LLM)의 추론(Reasoning) 과정이 결정론적 연산이 아닌 통계적 모방(Statistical Mimicry)에 불과하며, 내재적 자율 교정(Intrinsic Self-Correction)이 컴퓨터 공학적으로 원천 불가능함을 입증하는 글로벌 최상위 AI 학회 및 최신 연구 원문 발췌입니다.

1. 추론의 허상: '생각의 사슬(CoT)'은 인간을 모방한 사후 정당화일 뿐이다

"Deepseek R1 presents its CoT as a 'stream-of-consciousness', using filler words and interjections such as Ah!, Wait no., Oh yes!... LLMs generate text token-by-token, and as such have no need for these communication cues, suggesting that this is a stylistic mimicry of human reasoning, rather than direct correspondence to the model's internal process. [...] DeepSeek R1 CoT is better understood as a post-hoc rationalisation: a plausible narrative embellished with human-like interjections that simulate a stream-of-consciousness, and so give the impression of access to the model's 'thoughts'."

(Deepseek R1은 추론 과정을 '의식의 흐름'처럼 제시하며 아!, 잠깐, 오 그래! 같은 추임새를 사용한다... LLM은 토큰 단위로 텍스트를 생성하므로 이러한 소통 신호가 전혀 필요 없으며, 이는 모델의 실제 내부 연산 과정을 반영하는 것이 아니라 인간의 추론을 스타일적으로 모방(Stylistic mimicry)한 것임을 시사한다. 즉, CoT는 모델의 '생각'에 접근하는 듯한 환상을 주기 위해 꾸며진 '사후 정당화(Post-hoc rationalisation)'로 이해하는 것이 타당하다.)

출처: Examining the Faithfulness of Deepseek R1's Chain-of-Thought Reasoning (arXiv, 2025)

2. 연산 능력의 부재: 확률적 바인딩 실패와 분열뇌 증후군 (Split-Brain Syndrome)

"Current Transformer architectures suffer from "Computational Split-Brain Syndrome." This refers to a systematic dissociation between Comprehension (the ability to explain how to solve a problem) and Competence (the ability to execute the solution). [...] LLMs lack the architectural mechanism to reliably "bind" variables to values across arbitrary contexts. In a recursive loop, if Step 1 outputs x = 5, Step 2 often fails to retrieve 5 reliably if the context is complex, instead relying on probabilistic pattern completion."

(현재의 트랜스포머 아키텍처는 문제를 푸는 방법을 설명하는 능력(이해)과 실제로 솔루션을

실행하는 능력(적격성)이 구조적으로 분리되는 '컴퓨팅적 분열뇌 증후군'을 겪고 있다... LLM에는 임의의 컨텍스트에 걸쳐 변수와 값을 신뢰할 수 있게 "바인딩"할 아키텍처 메커니즘이 원천적으로 부재하다. 이들은 재귀적 루프에서 이전 단계의 값을 신뢰성 있게 검색하는 대신, 그저 확률적 패턴 완성(Probabilistic pattern completion)에 의존할 뿐이다.)

출처: Comprehension Without Competence: Architectural Limits of LLMs in Symbolic Computation and Reasoning (OpenReview/arXiv, 2025)

3. 통계적 예측의 구조적 한계와 '결정론적 진리' 창출의 불가능성

"Current Generative AI is a fundamentally analytic system. It is trained on a massive, but finite, dataset representing a snapshot of human knowledge... Its operations consist of identifying patterns, correlations, and structures within this dataset and then interpolating or recombining them to generate outputs. [...] It cannot synthesise new knowledge because it lacks the mechanism of computational verification or external grounding, and therefore it cannot "improve" in any meaningful sense."

(현재의 생성형 AI는 근본적으로 분석적 시스템이다... 이들의 연산은 훈련 데이터 내의 패턴과 상관관계를 식별하고 이를 보간하거나 재조합하여 출력을 생성하는 것으로 구성된다... 이 시스템은 계산적 검증(Computational verification)이나 외부적 기준(External grounding) 메커니즘이 없기 때문에 새로운 지식을 합성할 수 없으며, 따라서 어떤 의미 있는 방식으로도 스스로 "개선(Improve)"될 수 없다.)

출처: On the Limits of Self-Improving in LLMs and Why AGI, ASI and the Singularity Are Not Near Without Symbolic Model Synthesis (arXiv, 2026)

4. 인지 부조화: 내재적 자율 교정의 실패와 실제 행동 간의 괴리

"We uncover two key findings: First... less confident models show greater relative improvements from feedback, indicating confident models are more resistant to correction. Second, models consistently claim to understand feedback and express willingness to update their beliefs (>95%) yet fail to actually incorporate corrections—revealing a disconnect between stated intentions and actual behavior."

(우리는 두 가지 핵심 발견을 확인했다. 첫째, 확신이 높은 모델일수록 오류 교정에 더 강하게 저항한다. 둘째, 모델들은 피드백을 이해했다고 지속적으로 주장하며 자신의 신념을 기꺼이 수정하겠다고 답변(95% 이상)하면서도, 실제로는 그 교정 사항을 통합하는 데 실패한다. 이는 명시된 의도와 실제 행동 사이의 완벽한 단절을 드러낸다.)

출처: Feedback Friction: LLMs Struggle to Fully Incorporate External Feedback (arXiv, 2025)

"Large language models cannot self-correct reasoning yet... without external feedback, self-correction attempts often fail or degrade performance."

(대형 언어 모델은 아직 추론을 자율적으로 교정할 수 없다... 외부 피드백이 없는 상태에서의

자율 교정 시도는 종종 실패하거나 오히려 성능을 저하시킨다.)

출처: Large language models cannot self-correct reasoning yet (ICLR, 2024 / arXiv)

Chapter 3. 스케일링 법칙의 허상과 컨텍스트 희석의 필연성 (The Fallacy of Scaling Laws and the Inevitability of Context Dilution)

현대 인공지능 산업을 지배하는 가장 강력한 도그마는 자본과 컴퓨팅 자원을 투입하여 모델의 파라미터를 키우고 컨텍스트 윈도우(Context Window)를 확장하면, 모델의 논리적 추론 능력과 신뢰성이 비례하여 상승할 것이라는 '스케일링 법칙(Scaling Law)'에 대한 맹신이다.

그러나 하드웨어적 용량의 무한한 확장이 소프트웨어의 논리적 무결성을 자생적으로 담보한다는 가설은 컴퓨터 공학적 근거가 전무한 환원주의적 오류 (Reductionist Fallacy)다. 엄밀한 정보이론적 관점에서, 무분별한 컨텍스트의 투입은 오류를 보정하기는커녕 오히려 시스템적 망각 (Systemic Amnesia)과 성능의 붕괴를 가속하는 뇌관으로 작용한다.

이러한 붕괴의 기저에는 트랜스포머 아키텍처의 심장부인 '소프트맥스 정규화'가 가진 수학적 한계가 존재한다. 어텐션 메커니즘 하에서 모든 토큰의 주의 집중 가중치 합은 반드시 1이 되어야 하는 제로섬 분포를 따른다. 따라서 입력 시퀀스의 길이 N이 기하급수적으로 확장될수록, 특정 핵심 토큰에 할당될 수 있는 정보적 가중치는 필연적으로 1/N로 수렴하며 산술적으로 희석된다.
이는 단순한 연산의 비효율성을 넘어, 모델이 처리해야 할 '노이즈 플로어(Noise Floor)'가 급격히 상승함을 의미한다.

즉, 방대한 컨텍스트의 투입은 유의미한 신호를 증폭시키는 것이 아니라, 신호 대 잡음비(SNR)를 치명적으로 붕괴시키는 '어텐션 희석'을 초래하는 수학적 필연이다.

이러한 구조적 결함은 이른바 '중간 누락(Lost in the Middle)' 현상과 '컨텍스트 부패(Context Rot)'로 실증된다. 모델은 시퀀스의 양 끝단에 위치한 정보에만 과도하게 편향되며, 중간에 위치한 핵심 논리나 제약 조건은 아무리 완벽한 검색 증강(RAG)을 통해 제공되더라도 연산 과정에서 소실된다.

더욱이, 긴 문맥 속에 의미론적으로 유사하지만 사실관계가 다른 '방해 요소'가 포함될 경우, 결정론적 배제 능력이 없는 확률론적 근사기(LLM)는 이 노이즈들을 통계적으로 융합해버려 다중 공선성의 오류를 일으키고 치명적인 환각(Hallucination)을 발생시킨다. 컨텍스트의 절대적인 길이 자체가 내용의 품질과 무관하게 모델의 추론 엔진에 막대한 '인지적 과세(Cognitive Tax)'를 부과하는 것이다.

결과적으로, 결정론적 뼈대가 결여된 상태에서 컨텍스트 윈도우를 수백만 토큰 단위로 확장하는 하드웨어적 스케일링은 논리적 추론 능력 부재라는 본질적 문제를 은폐하는 미봉책에 불과하다. 이는 깊은 레이어와 넓은 문맥 속에서 토큰들의 벡터 표현이 구별 불가능해지는 표현 붕괴를 가속화하고, 오류를 기하급수적으로 증폭시키는 거대한 '노이즈 증폭기'를 구축하는 행위다.

진정한 의미의 신뢰성 있는 인공지능은 컨텍스트의 무비판적인 확장이 아니라, 역설적으로 불필요한 확률적 노이즈가 개입할

공간을 원천 차단하는 치밀한 '결정론적 컨텍스트 통제'를 통해서만 달성될 수 있다.

[핵심 교차 검증 문헌]

본 장에서 제기된 스케일링 법칙(Scaling Law)의 허상, 방대한 컨텍스트 투입이 필연적으로 초래하는 어텐션 희석(Attention Dilution)과 '중간 누락(Lost in the Middle)' 현상의 수학적 모순을 교차 검증하는 글로벌 최상위 AI 학회 및 연구 기관의 최신 실증 원문 발췌입니다.

1. 중간 누락(Lost in the Middle)과 컨텍스트의 인지적 과세 (Cognitive Tax)

"We find that performance can degrade significantly when changing the position of relevant information, indicating that current language models do not robustly make use of information in long input contexts. In particular, we observe that performance is often highest when relevant information occurs at the beginning or end of the input context, and significantly degrades when models must access relevant information in the middle of long contexts, even for explicitly long-context models." (우리는 관련 정보의 위치를 변경할 때 성능이 크게 저하될 수 있음을 발견했으며, 이는 현재의 언어 모델이 긴 입력 컨텍스트의 정보를 견고하게 활용하지 못함을 나타낸다. 특히 명시적인 롱 컨텍스트 모델조차도 관련 정보가 중간에 위치할 때 성능이 기하급수적으로 붕괴함을 확인했다.)

출처: Lost in the Middle: How Language Models Use Long Contexts (TACL, 2024 / Stanford & Meta AI)

"Even with 100% perfect retrieval of relevant information, performance degrades 13.9% to 85% as input length increases. [...] Sheer context length itself imposes a cognitive tax on LLMs independent of content quality." (100% 완벽한 정보 검색이 이루어지더라도, 입력 길이가 증가함에 따라 성능은 13.9%에서 최대 85%까지 저하된다. 컨텍스트의 절대적인 길이 자체가 내용의 품질과 무관하게 LLM에 '인지적 과세(cognitive tax)'를 부과하는 것이다.)

출처: Context Dilution: When More Tokens Hurt AI (diffray Research, 2025)

2. 컨텍스트 부패(Context Rot)와 소프트맥스 어텐션의 표현 붕괴 (Representational Collapse)

"We observe that model performance varies significantly as input length changes, even on simple tasks... Models do not use their context uniformly; instead, their performance grows increasingly unreliable as input length grows." (단순한 작업에서도 입력 길이가 변함에 따라 모델 성능이 크게 달라짐을 관찰했다. 모델은 컨텍스트를 균일하게 사용하지 않으며, 입력 길이가 길어질수록 성능은 점점 더 신뢰할 수 없게 붕괴한다.)

출처: Context Rot: How Increasing Input Tokens Impacts LLM Performance (Chroma Technical Report, 2025)

"...softmax's fundamental property that forces probability mass to be distributed across all tokens, with attention weights necessarily

approaching an uniform distribution as context grows... representational collapse occurs due to softmax's inability to maintain distinct attention patterns as sequence length grows, erasing meaningful distinctions between tokens." (확률 질량을 모든 토큰에 분산하도록 강제하는 소프트맥스의 근본적인 특성으로 인해, 컨텍스트가 증가함에 따라 어텐션 가중치는 필연적으로 균일 분포(Uniform distribution)에 수렴한다... 시퀀스 길이가 증가함에 따라 뚜렷한 어텐션 패턴을 유지하지 못하는 소프트맥스의 무능력으로 인해 '표현 붕괴(representational collapse)'가 발생하며, 이는 토큰 간의 의미 있는 구분을 지워버린다.)

출처: Long-Context Generalization with Sparse Attention (arXiv, 2026)

3. 어텐션 싱크(Attention Sinks)와 논리적 일관성의 역설

"Because softmax normalization forces attention weights to sum to 1, models must 'dump' attention somewhere when no tokens are highly relevant... since attention is zero-sum, adding more tokens monotonically increases noise in representations." (소프트맥스 정규화는 어텐션 가중치의 합을 무조건 1로 강제하기 때문에, 매우 관련성 높은 토큰이 없을 때 모델은 어텐션을 어딘가에 '버려야(dump)'만 한다... 어텐션은 제로섬(zero-sum) 게임이므로, 토큰을 더 추가하는 것은 표현 공간 내의 노이즈를 단조 증가(monotonically increases)시키는 결과를 초래할 뿐이다.)

출처: Context Dilution: When More Tokens Hurt AI (diffray Research, 2025)

"Although it seems counterintuitive, models perform worse when the haystack preserves a logical flow of ideas. Shuffling the haystack and removing local coherence consistently improves performance." (직관에 반하는 것처럼 보이지만, 텍스트가 아이디어의 논리적 흐름을 보존할 때 모델의 성능은 오히려 저하된다. 텍스트를 무작위로 섞어 국소적 일관성(local coherence)을 제거하는 것이 오히려 모델의 성능을 일관되게 향상시킨다.)

출처: Context Rot: How Increasing Input Tokens Impacts LLM Performance (Chroma Technical Report, 2025)

Chapter 4. 결정론적 아키텍처의 필연성: 인간 통제의 귀환 (The Necessity of Deterministic Architecture: The Return of Human Control)

앞선 장들에서 논증했듯, 순수 확률론적 모델(LLM)에 자율성을 부여하여 논리적 무결성을 달성하겠다는 발상은 정보이론적 한계와 수학적 모순에 직면해 있다.

대형 언어 모델의 자율적 상호작용(Multi-Agent Workflow)을 통해 추론의 깊이를 확보하려는 시도는, 개별 에이전트의 확률적 오차가 파이프라인의 깊이에 비례하여 기하급수적으로 증폭되는 '오류의 다중 복리(Compounding Errors)' 현상을 구조적으로 방치하는 행위다. 결정론적 기준점(Ground Truth)이 부재한 닫힌 루프 내에서의 확률적 연산은 필연적으로 분산 증폭 (Variance Amplification)과 엔트로피 붕괴 (Entropy Decay)를 초래하며, 시스템을 통계적 표류(Statistical Drift) 상태로 몰아넣는다.

따라서, 상업적 수준의 무결성과 신뢰성을 담보하기 위한 유일한 컴퓨터 공학적 해법은, AI에게 부여된 '계획'과 '검증'이라는 자율성의 환상을 전면 박탈하는 것이다. 기계는 기호적 추론(Symbolic Reasoning)을 수행할 내재적 바인딩 메커니즘이 원천적으로 부재하므로, 시스템의 통제권은 다시 결정론적 사고가 가능한 인간에게로 온전히 반환되어야만 한다.

이러한 '인간 통제의 귀환'은 단순히 프롬프트를 상세하게 작성하는 수준의 휴리스틱(Heuristics)을 의미하지 않는다. 이는 시스템의 아키텍처를 근본적으로 재설계하는 과정이다. 인간 아키텍트 (Human Architect)는 거대한 추론의 과정을 독립적으로 검증 가능한 원자적 단위로 해체하고, 이를 비순환 방향 그래프(DAG, Directed Acyclic Graph) 형태의 '결정론적 논리 구조망 (Deterministic Control Grid)'으로 직조해야 한다. 이 엄격한 통제망 하에서 AI는 더 이상 스스로 사고의 흐름을 결정하는 주체가 아니며, 오직 인간이 명시적으로 강제한 국소적 맥락(Local Context) 내에서만 텍스트를 변환하는 '제한적 통계 렌더링 모듈(Restricted Statistical Rendering Module)'로 격하되어야 한다.

확률적 근사기인 AI의 본질적 결함을 통제하는 유일한 방벽은, 역설적이게도 AI 스스로의 자생적 진화가 아니라 AI보다 훨씬 더 치밀하게 설계된 인간의 '구조적 제약(Structural Constraints)'이다. 변수와 상수가 명확히 통제된 이 결정론적 뼈대를 입력 단계에서부터 강제하지 않는 한, 모델의 파라미터를 수조 개로 확장하더라도 의미 있는 지적 임계점(Singularity)에는 결코 도달할 수 없다.

진정한 의미의 신뢰성 있는 인공지능 혁신은, 통계적 확률의 자율적 방임이 아니라 가장 차갑고 기계적인 결정론적 통제망을 통해서만 완성된다.

[핵심 교차 검증 문헌]

본 장에서 선언한 확률론적 에이전트 자율성의 치명적 붕괴 현상(P^n)과, 이를 극복하기 위해 통계적 모방을 통제하는 '결정론적 논리

구조망(DAG)' 및 '기호적 제약(Symbolic Constraints)'이 왜 컴퓨터 공학적 필연인지 증명하는 글로벌 최상위 연구 원문 발췌입니다.

1. 에이전트 자율성의 파산: 오류의 다중 복리(Compounding Errors)와 노이즈 증폭

"If you have a pipeline of n sequential steps, and each step succeeds with probability p, the probability that the full pipeline succeeds is P^n . That is the individual reliabilities multiplied together. Not averaged. Multiplied. [...] Independent, decentralised agent architectures amplify errors 17.2 times compared to a single-agent baseline. Multi-agent variants degraded sequential reasoning performance by 39 to 70%, challenging the assumption that more agents always improve outcomes." (n개의 순차적 단계를 가진 파이프라인에서 각 단계가 p의 확률로 성공한다면, 전체 파이프라인의 성공 확률은 평균이 아니라 P^n 으로 곱해진다. 독립적이고 분산된 자율 에이전트 아키텍처는 단일 에이전트 베이스라인에 비해 오류를 17.2배나 증폭시킨다. 다중 에이전트 변형은 순차적 추론 성능을 최대 70%까지 붕괴시키며, '에이전트가 많을수록 결과가 개선된다'는 맹신을 정면으로 반박한다.)

출처: The Compounding Errors Problem: Why Multi-Agent Systems Fail and the Architecture That Fixes It (Zartis, 2026)

"Reasoning LLMs (e.g., DeepSeek-V3) already embed sophisticated error-detection and correction processes. As a result, additional self-correction methods confer only marginal gains and may increase computational overhead, highlighting a performance ceiling in highly reasoning LLMs." (고도의 추론 모델들은 이미 정교한 오류 탐지 프로세스를 내장하고 있다. 결과적으로, 에이전트 자율 교정 방법을 추가하는 것은 단지 한계 효용만을 제공하며 계산 오버헤드만 증가시킬 뿐, 순수 확률론적 모델이 도달한 성능의 천장(Ceiling)을 극명하게 보여준다.)

출처: Can LLMs Correct Themselves? A Benchmark of Self-Correction in LLMs (NeurIPS, 2025)

2. 통계적 표류(Statistical Drift)와 결정론적 닻(Symbolic Anchor)의 필연성

"Without external grounding ($\alpha \to 0$), the model mean follows a random walk driven by optimisation noise. Symbolic constraints and algorithmic complexity act as a discretisation anchor... preventing the continuous degradation (random walk) characteristic of purely neural updates." (외부의 결정론적 기준이 차단된 상태에서 모델의 평균값은 최적화 노이즈에 이끌려 무작위 행보(Random Walk)를 하며 표류한다. 기호적(명시적) 제약 조건들은 이러한 순수 신경망 업데이트가 필연적으로 겪는 지속적 붕괴를 막아주는 유일한 '이산적 닻(Discretisation anchor)'으로 작용한다.)

출처: On the Limits of Self-Improving in LLMs and Why AGI, ASI and the Singularity Are Not Near Without Symbolic Model Synthesis (arXiv, 2026)

"The projection Π_S enforces syntactic, grammatical, or invariant structure, thereby eliminating large regions of the hypothesis space. A small value of σ indicates that symbolic

structure is highly informative, leading to a rapid collapse toward low-complexity representations." (기호적 투사(ITS)는 불변의 구조를 강제함으로써 확률적 가설 공간의 거대한 불확실성 영역을 제거한다. 명시적이고 기호적인(결정론적) 구조의 주입만이 시스템을 복잡성이 낮고 신뢰성 높은 표현으로 빠르게 수렴시킨다.)

출처: On the Limits of Self-Improving in LLMs and Why AGI, ASI and the Singularity Are Not Near Without Symbolic Model Synthesis (arXiv, 2026)

3. 솔루션: 방향성 비순환 그래프(DAG)를 통한 원자적 해체와 인간의 구조적 제약

"Error compounding requires an unbroken chain. You break it by decomposing complex tasks into a directed acyclic graph (DAG) of atomic sub-tasks, each small enough that its output can be independently evaluated before it enters any downstream context." (오류의 누적 증폭은 끊어지지 않은 사슬을 필요로 한다. 이 사슬을 끊는 유일한 방법은 복잡한 작업을 '방향성 비순환 그래프(DAG)' 형태의 원자적 하위 작업으로 완벽히 해체하여, 그 산출물이 다음 컨텍스트로 진입하기 전에 독립적이고 결정론적으로 평가되게 만드는 것뿐이다.)

출처: The Compounding Errors Problem: Why Multi-Agent Systems Fail and the Architecture That Fixes It (Zartis, 2026)

"The Inspector pattern's 96.4% fault recovery rate is achievable not because the Inspector is smarter than the primary agent, but because it reviews from outside the primary agent's error context." (검수 모델이 96.4%의 결함 복구율을 달성할 수 있는 것은 그 검수자가 기존 에이전트보다 더 똑똑해서가 아니라, 1차 에이전트가 갇혀 있는 오류 컨텍스트(Error Context)의 '외부'에서 철저히 분리되어(독립적 뼈대 하에서) 평가하기 때문이다.)

출처: The Compounding Errors Problem: Why Multi-Agent Systems Fail and the Architecture That Fixes It (Zartis, 2026)

3. 결론 (Conclusion)

에이전트 워크플로우의 재설계와 하이브리드 아키텍처의 필연성 (Redesigning Agentic Workflows and the Inevitability of Hybrid Architectures)

지금까지의 논증을 통해, 순수 확률론적 생성 모델(LLM)에 자율성을 부여하여 논리적 정합성을 달성하려는 시도는 정보이론적 한계와 구조적 모순에 직면해 있음을 확인했다. 모델 간의 '자율 교정(Self-Correction)'은 본질적으로 다중 공선성의 오류를 극대화하며, 긴 파이프라인 속에서 오류를 기하급수적으로 증폭시키는 붕괴의 가속 기제로 작용한다. 따라서 인공지능이 상업적이고 실증적인 무결성을 확보하기 위한 해답은 다음의 두 가지 아키텍처적 전환으로 귀결된다.

첫째, 시스템의 근간은 철저히 인간이 설계한 결정론적 논리 구조에 의해 통제되어야 한다. 만약 파이프라인 내에 검수 에이전트가 불가피하게 도입되어야 한다면, 그 역할은 확률적 노이즈가 이미 개입된 '결과물'이나 '중간 연산 과정'을 사후적으로 평가하는 데 사용되는 것은 큰 의미가 없다. 대신, 연산의 출발점인 '프롬프트 자체의 논리적 무결성'을 선제적으로 검수하고 교정하는 방식으로 전환되어야 한다. 모델의 출력은 전적으로 입력 분포에 종속되므로, 로직의 충돌을 제거하고 빈 공간을 결정론적으로 제어할 수 있는 유일한 지점은 입력단계 뿐이기 때문이다.

둘째, 인간의 개입을 최소화하면서도 완전한 신뢰성을 담보하는 워크플로우를 구축하고자 한다면, 순수 AI 모델로만 구성된 에이전트 루프는 폐기되어야 한다. 진정한 의미의 자율 검증이 성립하려면, 확률론적 생성 모델(LLM)과 명시적 예외 처리가 가능한 전통적 연산 장치(CPU 기반의 결정론적 컴퓨팅)가 결합된 '하이브리드 에이전트 워크플로우(Hybrid Agent Workflow)'가 채택되어야만 한다. 코드를 실행하고, 수학적 참을 증명하며, 외부 데이터베이스의 팩트를 물리적으로 대조하는 컴퓨팅 파워 (기호적 닻, Symbolic Anchor)가 에이전트의 환각을 끊어내는 통제망으로 작동해야만 오류의 연쇄적 증폭(P^n)을 차단할 수 있다.

이러한 컴퓨터 공학적, 정보이론적 한계는 현재 AI 산업계의 자본 흐름과 기술 평가 기준에 중대한 시사점을 던진다. 현재 수많은 기업들이 순수 대형 언어 모델 간의 상호작용 및 자율 교정 능력만을 핵심 기술력으로 내세워 '에이전트 워크플로우(Agentic Workflow)'를 상용화하고 있다. 그러나 이들이 주창하는 시스템적 완전성은 수학적으로 입증되지 않은 환원에 불과하며, 엔터프라이즈 환경이 요구하는 높은 수준의 무결성을 지속적으로 담보하기에는 구조적 취약성이 너무나 크다.

결과적으로, 현재 순수 AI 에이전트만으로 구성된 파이프라인을 구축하는 기업들의 기술적 해자(Moat)와 상업적 내구성에 대한 시장의 가치 평가(Valuation)는 머지않아 냉정하게 재조정될 수밖에 없을 것이다. 미래의 AI 산업을 주도할 진정한 가치는 확률론적 모델의 맹목적 방임이 아니라, 인간의 치밀한 통제망과 결정론적

컴퓨팅이 결합된 하이브리드 아키텍처를 얼마나 견고하게 구현해 내는가에 달려 있다.

[핵심 교차 검증 문헌]

결론에서 제기된 순수 AI 에이전트 검수 워크플로우(LLM-as-a-Judge)의 상업적·구조적 파산과, 이를 대체할 결정론적 기호 연산(Symbolic Systems, CPU)이 결합된 하이브리드 아키텍처의 필연성을 증명하는 글로벌 최상위 연구 및 시장 분석 원문 발췌입니다.

1. 순수 AI 검수(LLM-as-a-Judge)의 구조적 파산과 동어반복적 환각

"A single agent evaluating its own outputs faces elevated risk of correlated error: sharing training data, inductive biases, and blind spots... An Inspector that uses the same model, the same context, and the same prompt as the primary agent will share its failure modes. Two instances of the same model reasoning from the same context will hallucinate on the same inputs." (단일 에이전트가 자신의 산출물을 평가하는 것은 훈련 데이터, 귀납적 편향, 맹점을 공유하기 때문에 상관 오차(Correlated error)의 위험을 기하급수적으로 높인다... 1차 에이전트와 동일한 모델, 동일한 컨텍스트를 사용하는 검수자(Inspector)는 그 실패 모드를 그대로 공유한다. 동일한 문맥에서 추론하는 두 개의 동일한 모델 인스턴스는 반드시 동일한 입력에 대해 똑같이 환각(Hallucinate)을 일으킨다.)

출처: Limits of Self-Correction in LLMs: An Information-Theoretic Analysis of Correlated Errors & The Compounding Errors Problem

"Multi-agent variants degraded sequential reasoning performance by 39 to 70%, challenging the assumption that more agents always improve outcomes." (다중 에이전트 변형 모델들은 순차적 추론 성능을 39%에서 최대 70%까지 붕괴시켰으며, 이는 에이전트가 많아질수록 항상 결과가 개선될 것이라는 맹신에 정면으로 도전한다.)

출처: The Compounding Errors Problem: Why Multi-Agent Systems Fail and the Architecture That Fixes It

2. 하이브리드 아키텍처의 필연성: 기호적 연산(CPU/Symbolic)과 결정론적 통제

"The technical foundation of DSS is neurosymbolic AI... the fusion of linguistic fluency of neural networks ('System 1') with the logical rigor of symbolic systems ('System 2'). Neural networks excel at intuition... symbolic systems provide the verifiable, step-by-step logic required for high-stakes reasoning." (도메인 특화 초지능(DSS)의 기술적 기반은 뉴로심볼릭 AI다. 이는 신경망의 언어적 유창성과 기호적 시스템(결정론적 컴퓨팅)의 논리적 엄밀성을 융합한 것이다. 신경망은 직관에 뛰어나지만, 고위험 추론에서 요구되는 검증 가능하고 단계적인 논리는 오직 기호적 시스템(Symbolic systems)만이 제공할 수 있다.)

출처: An Alternative Trajectory for Generative AI

"Purely distributional learning leads to model collapse, hybrid neurosymbolic approaches offer a coherent framework for sustained self-improvement." (순수하게 확률 분포에 의존하는 학습은 필연적으로 모델 붕괴(Model collapse)로 이어진다. 오직 하이브리드 뉴로심볼릭(결정론적 로직 결합) 접근법만이 지속 가능한 자율 개선을 위한 일관된 프레임워크를 제공한다.)

출처: On the Limits of Self-Improving in LLMs and Why AGI, ASI and the Singularity Are Not Near Without Symbolic Model Synthesis

"Verification transforms LLMs from purely probabilistic generators into more reliable decision-making agents. Formal language can function as a powerful abstraction that enables reliable mathematical reasoning within neurosymbolic systems, rather than relying solely on probabilistic approximations." (검증은 LLM을 순수 확률론적 생성기에서 신뢰할 수 있는 의사결정 에이전트로 변환시킨다. 형식 언어(결정론적 코드)는 단순히 확률적 근사치에 의존하는 대신, 뉴로심볼릭 시스템 내에서 신뢰할 수 있는 수학적 추론을 가능하게 하는 강력한 추상화로 기능한다.)

출처: An Alternative Trajectory for Generative AI

3. 순수 에이전트 기업들의 가치 재조정과 시장의 경고

"Distributed intelligence is still distributed systems, and distributed systems aren't cheap to build or maintain... lower production costs don't automatically translate into higher productivity. They often just make it easier to manufacture fragility at scale." (분산된 지능(다중 에이전트)은 결국 분산 시스템일 뿐이며, 이를 구축하고 유지하는 것은 결코 저렴하지 않다... AI로 인한 생산 단가 하락이 자동으로 높은 생산성으로 이어지지는 않는다. 이는 종종 대규모로 '시스템의 취약성(Fragility)'을 제조하기 쉽게 만들 뿐이다.)

출처: Multi-agent AI is the new microservices (InfoWorld, 2026)

"Deterministic Guardrails: Verification layers prioritize accuracy over probability for business logic. (Valuation Levers: 10x - 12x High-growth Agentic Premium)" (결정론적 가드레일: 비즈니스 로직을 처리함에 있어 검증 레이어는 '확률(Probability)'보다 '정확성 (Accuracy)' 을 우선해야 한다. 이러한 결정론적 통제망을 갖춘 플랫폼만이 10~12배의 고성장 에이전틱 프리미엄 가치를 정당화할 수 있다.)

출처: Generative AI Platforms Valuation: Q4 2025 Market Analysis