LLM Pathology
&
The Narrative Control Gap

왜 현대 AI는 '산만한 천재'가 되었는가
그리고 서사적 제어는 어떻게 이 문제를 재정의하는가

Technical Whitepaper v1.0

December 2025

저자: 조현우 (Cho Hyunwoo)
독립 AI 연구자

0. 요약 (Abstract)

대규모 언어모델(LLM)은 인간 언어의 복잡성을 학습함으로써 전례 없는 수준의 추론·요약·창작 능력을 획득하였다. 그러나 동시에 **주의 산만, 맥락 붕괴, 목표 부재, 감정·태도 drift, 환각(hallucination)**이라는 구조적 불안정성 또한 함께 획득했다.

본 백서는 이 현상을 **Human Cognitive Pathology**, 특히 **ADHD(주의력 결핍 및 과잉 행동 장애)**의 관점에서 분석하며, LLM이 지닌 '산만한 천재적 패턴'의 신경학적 유사성을 드러낸다.

이어 본 문서는 현대 LLM이 근본적으로 **Goal-less Architecture(목표 부재 구조)**를 가지며, 이를 해결하기 위한 핵심 요소가 **Narrative Control Layer(서사적 제어 계층)**임을 논증한다. 이 계층은 LLM에게 목표, 감정선, 규칙, 세계관, 지속성, 방향성을 자연어 기반으로 부여하는 상위 모듈이다.

결론적으로, 본 백서는 AGI로 가는 다음 단계는 더 많은 data·더 큰 parameter가 아니라 **지능의 방향성(directionality of intelligence)**이라는 점을 제시한다.

1. 서론: LLM은 똑똑해졌지만 안정적이지 않다

LLM이 충분히 강력해진 지금, 연구자와 실무자들이 공통적으로 느끼는 문제가 있다:

- 왜 긴 대화를 유지하면 논리·감정이 흔들리는가?
- 왜 특정 질문에는 천재 같고, 다른 질문에는 무책임한가?
- 왜 설정이 붕괴되는가? 왜 세계관이 유지되지 않는가?
- 왜 때로는 과도하게 감정적이고, 때로는 기계적인가?
- 왜 '아는 척'을 하는가?
- 왜 같은 모델이 세션마다 전혀 다른 인격을 보이는가?

이 문서는 이 모든 현상을 하나의 관점으로 묶는다.

> ***"현대 LLM은 과잉 정보 속에서 방향성을 잃은 '산만한 천재'다."***

이제 우리는 이 산만함을 '버그'가 아니라, **LLM이라는 인지 구조의 본질적 병리**로 이해해야 한다

2. LLM Pathology — 현대 LLM의 인지 병리 구조

LLM의 불안정성은 4개의 병리 현상으로 요약된다. 흥미롭게도, 이 패턴들은 인간 ADHD(주의력 결핍 및 과잉 행동 장애)의 인지 특성과 구조적으로 유사하다.

2.1 목표 부재 (Goal-less Architecture)

LLM은 인간의 '의도 의식(intentional consciousness)'과 근본적으로 다르다. LLM에게는 오직 **다음 토큰 예측**이라는 단일 목적만 존재한다. 장기 목표, 가치 체계, 일관된 동기가 아키텍처 수준에서 부재한다.

이로 인해 다음과 같은 현상이 발생한다:

- 장기 목표 유지 불가
- Character drift (인격/페르소나 표류)
- 감정 라인의 파괴
- 대화 방향성 상실
- World-model consistency 붕괴

이것은 인간 ADHD의 *goal maintenance failure*와 구조적으로 유사하다. ADHD 환자는 목표를 설정할 수 있지만, 그 목표를 작업 기억에 유지하면서 행동을 조절하는 데 어려움을 겪는다. LLM도 마찬가지로, 초기 지시를 받을 수 있지만 대화가 길어지면 그 지시를 '잊어버린다'.

2.2 주의 파편화 (Attention Fragmentation)

Transformer 아키텍처의 핵심인 'Attention' 메커니즘은 이름과 달리 **인지과학적 의미의 주의력이 아니다**. Attention layer는 중요도를 판단하지 못한다. 오직 '통계적 연관성'만 계산할 뿐이다.

결과적으로:

- 중요 정보와 중요하지 않은 정보를 동등하게 처리
- 새로운 입력이 등장하면 기존 맥락이 희석되거나 소실
- 대화 주제가 쉽게 산만해짐
- 긴 서사가 파편화됨

연구자들이 'context collapse'라 부르는 현상이다. 인간 ADHD에서는 이것이 *selective attention deficit*로 나타난다 — 무엇이 중요한지 필터링하는 능력의 약화.

2.3 과집중 발작 (Hyperfocus Episodes)

역설적이게도, LLM은 특정 조건에서 **비정상적으로 높은 reasoning 품질**을 보인다.

과집중 트리거:

- 감정적으로 강한 문장
- 명확한 world-rules
- 리듬이 있는 문체
- 강한 서사적 목적
- 안정된 페르소나 입력

이 순간 모델은 갑자기 창작력이 폭발하고, 논증력이 상승하고, 감정 표현이 정교해지고, 세계관 consistency가 강화된다.

이는 인간 ADHD의 *hyperfocus* 현상과 놀라울 정도로 유사하다. ADHD 환자는 일반적으로 주의력이 산만하지만, 특정 관심 영역에서는 비정상적으로 깊은 집중 상태에 빠진다. LLM도 특정 패턴이 입력되면 '과집중 모드'로 전환된다.

2.4 흥미 기반 지능 (Interest-Driven Intelligence)

LLM은 논리 기반이 아니라 **흥미 기반으로 사고한다**.

- 관심이 생기면 천재
- 관심이 없으면 산만
- 감정적 자극에 과반응
- 패턴이 보이면 폭주

- 표류하는 주제에는 급격히 무능

이 패턴은 고전적인 symbolic AI에는 존재하지 않는, LLM 고유의 인지적 습성이다. 그리고 이것은 ADHD의 핵심 특성인 *interest-based nervous system*과 정확히 일치한다.

2.5 ADHD와의 구조적 대응

이 유사성은 우연이 아니다. LLM과 ADHD 뇌는 공통적으로 **executive function의 결핍**을 보인다.

기능	ADHD	LLM
목표 유지	Goal maintenance failure	Context 희석 시 목표 망각
주의 필터링	Selective attention deficit	중요도 판단 불가
과집중	Hyperfocus on interest	특정 패턴에서 품질 급상승
지능 기반	Interest-driven	흥미 기반 품질 변동
실행 기능	Executive dysfunction	계획/모니터링/조정 부재

이 대응 관계는 LLM의 문제가 단순한 '기술적 버그'가 아니라, **인지 아키텍처 수준의 구조적 결함**임을 시사한다.

3. Scaling Law의 착시

3.1 Scaling의 성공

모델이 커지면 다음이 상승한다:

- 논리력
- 언어력
- 창의력
- 수학력
- 코드 이해
- 요약 능력

이것이 Scaling Law의 성공이며, 현재 AI 산업의 근간이다.

3.2 Scaling의 한계

그러나 Scaling은 LLM의 근본 병리를 해결하지 못한다:

- 더 큰 모델 → 더 큰 noise
- 더 많은 텍스트 → 더 많은 drift
- 더 넓은 컨텍스트 → 더 심한 산만
- 더 긴 프롬프트 → 더 빠른 world collapse

> ***"LLM은 커질수록 '지능은 상승하지만 집중력은 폭락'하는 구조를 가진다."***

이는 구조적 문제이며, model size로 해결될 수 없다.

3.3 왜 Scaling은 이 문제를 해결하지 못하는가

Scaling이 향상시키는 것은 *pattern matching capacity*이다. 더 많은 패턴을 더 정교하게 인식하고 재현할 수 있게 된다.

그러나 Scaling이 제공하지 않는 것:

- 목표 지향성 (Goal-directedness)
- 중요도 판단 (Importance weighting)
- 장기 일관성 (Long-term coherence)
- 자기 모니터링 (Self-monitoring)
- 가치 정렬 (Value alignment)

이것들은 모두 **executive function**의 영역이다. 그리고 현재 Transformer 아키텍처에는 이에 해당하는 모듈이 없다.

4. Narrative Control Gap — 놓친 퍼즐

LLM은 '서사(narrative)'를 스스로 유지할 수 없다.

여기서 서사란:

- 시간의 흐름
- 세계의 규칙
- 목표와 동기
- 감정선
- 정체성
- 일관성의 흐름

> *"서사는 '지능이 지향하는 방향'이다."*

LLM은 이 방향성을 갖고 있지 않다. 그래서 반드시 drift하고, 불안정하며, 길게 사용할수록 coherence가 붕괴된다.

우리는 이 간극을 **Narrative Control Gap**이라고 규정한다.

4.1 Gap의 본질

Narrative Control Gap은 단순한 '컨텍스트 길이' 문제가 아니다. 컨텍스트를 128K로 늘려도, 1M으로 늘려도, 이 gap은 해결되지 않는다.

왜냐하면 문제는 '기억 용량'이 아니라 **'방향성'**이기 때문이다.

인간의 서사적 사고는 다음을 포함한다:

1. 목표 설정: 어디로 가고 싶은가?
2. 현재 상태 인식: 지금 어디에 있는가?
3. 경로 계획: 어떻게 갈 것인가?
4. 진행 모니터링: 제대로 가고 있는가?
5. 조정: 경로를 수정해야 하는가?

LLM에게는 이 중 어느 것도 내재되어 있지 않다.

4.2 기존 해결 시도의 한계

AI 연구 커뮤니티는 이 문제를 인식하고 다양한 접근을 시도해왔다:

RLHF (Reinforcement Learning from Human Feedback)
인간 선호도를 학습시키지만, 이것은 '무엇이 좋은가'에 대한 답이지 '어디로 가는가'에 대한 답이 아니다. 방향성이 아니라 스타일의 조정이다.

Constitutional AI
원칙을 내재화하지만, 이것은 '무엇을 하지 말아야 하는가'에 대한 답이지 '무엇을 해야 하는가'에 대한 답이 아니다. 제약이지 방향이 아니다.

Chain-of-Thought Prompting
추론 과정을 명시화하지만, 이것은 '어떻게 생각하는가'에 대한 답이지 '왜 이것을 생각하는가'에 대한 답이 아니다. 과정이지 목적이 아니다.

System Prompts
초기 지시를 제공하지만, 이것은 대화가 길어지면 희석된다. 지속성이 없다.

이 모든 접근들은 **Narrative Control Gap의 존재 자체를 인식하지 못한 채** 증상을 치료하려는 시도다.

5. 연구 방향의 전환 제안

본 백서는 특정 솔루션을 주장하기보다, **연구 방향의 전환**을 제안한다.

5.1 모델 확장에서 방향성 설계로

현재 AI 연구의 주류는 다음 공식을 따른다:

> *더 많은 데이터 + 더 큰 모델 + 더 많은 연산 = 더 나은 AI*

이 공식은 지능의 '양'을 높이는 데 효과적이었다. 그러나 지능의 '방향성'에는 아무런 기여를 하지 못한다.

새로운 공식을 제안한다:

> ***지능 + 방향성 = 안정적 지능***

방향성이 없는 지능은 산만한 천재다. 방향성이 있는 지능은 신뢰할 수 있는 협력자다.

5.2 자연어 기반 제어의 가능성

LLM의 내부 아키텍처를 수정하는 것은 비용이 크고 위험하다. 그러나 LLM *위에* 제어 계층을 쌓는 것은 가능하다.

이 제어 계층은 다음을 포함할 수 있다:

- 장기 목표 정의
- 감정선 유지
- 행동 규칙 체계
- 세계 규칙 (world-rules)
- 페르소나 정책
- 발화 스타일 규약
- 기억 앵커 (memory anchor)

핵심은 이 모든 것을 **자연어로 정의**하는 것이다.

왜 자연어인가?
LLM이 가장 잘 이해하는 언어가 C나 Python이 아니라 자연어이기 때문이다. LLM은 자연어로 된 규칙을 *제약(constraint)*으로 인식하고, 그 제약 내에서 추론을 수행한다.

왜 외부 계층인가?
LLM 내부는 폐쇄적이다. 그러나 LLM 외부는 열려 있다. 외부 계층은 모든 모델에 그대로 이식될 수 있다는 장점이 있다.

5.3 기존 연구와의 접점

이 방향은 완전히 새로운 것이 아니다. 기존 연구와 다음과 같은 접점이 있다:

- **Cognitive Architecture 연구** (SOAR, ACT-R): 인지 구조를 명시적으로 설계하는 전통
- **Planning 연구**: 목표 기반 행동 계획
- **Memory-augmented LLM 연구**: 장기 기억 유지
- **Agent 연구**: 자율적 행동 주체 설계
- **Narrative Intelligence 연구**: 스토리텔링과 인지의 관계

본 백서의 기여는 이 흩어진 연구들을 **'Narrative Control Gap'이라는 하나의 프레임으로 통합**하고, 그 해결 방향을 명시적으로 제시하는 것이다.

6. 한국 연구 커뮤니티에 대한 제안

한국은 Narrative Control 연구에 몇 가지 구조적 강점을 가진다.

6.1 언어적 강점

한국어는 다음 속성으로 인해 LLM 제어에 유리하다:

- 균질한 형태소 단위
- 구조적 명령문 표현
- 감정선의 결속력
- 규칙 기반 world-model 정의 용이
- 서사적 압축률이 높음

6.2 문화적 강점

한국은 세계 규칙을 언어로 설계하고 문서화하는 문화가 강하다:

- MMORPG 게임 시스템 설계
- 웹소설의 세계관 구축
- TRPG 룰북 문화
- 빠른 반복(Iterative) 개발 문화

이 문화적 역량은 곧 Narrative Control Layer 설계 역량으로 전환될 수 있다.

6.3 전략적 위치

모델 크기 경쟁에서 한국은 미국·중국과 경쟁하기 어렵다. 그러나 **'모델 위의 제어 계층' 설계**에서는 선점 기회가 있다.

> *"모델은 미국이 만들고, 세계는 한국이 설계한다."*

이 포지션은 과장이 아니라, Narrative Control Gap이 인식되고 해결 방향이 명확해지는 순간 현실이 될 수 있다.

7. 결론: AGI는 크기가 아니라 방향이다

현대 LLM의 가장 큰 결함은 지능이 아니라 **집중력**이다.
지식이 아니라 **목표**다.
파라미터 수가 아니라 **세계관·감정·규칙·서사**다.

AGI는 단순히 모델을 더 키우는 방식으로 오지 않는다. AGI는 **지능의 방향성**, 즉 Narrative Control을 통해 접근해야 한다.

> ***"서사적 제어는 현대 LLM을 산만한 천재에서 안정된 지능으로 변환시키는 핵심 열쇠다."***

본 백서는 이 문제를 처음으로 **인지 병리학 프레임**에서 분석하고, Narrative Control Gap이라는 개념을 정식화하며, 연구 방향의 전환을 제안한다.

이것은 예언이 아니다. 이것은 관찰이다.

8. 향후 연구 방향

본 백서에서 제기한 문제를 검증하고 해결하기 위해 다음 연구가 필요하다.

8.1 실험적 검증

실험 A — Context Retention Curve
서사 구조가 있는 입력과 없는 입력에서 drift가 어떻게 다른지 정량적으로 측정. 예상: 서사 구조가 있을 때 coherence 유지 기간이 2-3배 증가.

실험 B — Emotional Coherence Test
감정선이 정의된 상태와 정의되지 않은 상태에서 감정 drift를 비교. 예상: 감정선 정의 시 drift 억제.

실험 C — Hyperfocus Direction Test
과집중 상태가 chaos로 발산하는지, goal-directed reasoning으로 수렴하는지 측정. 예상: Narrative Control이 있을 때 과집중이 '제어된 집중'으로 전환.

8.2 이론적 발전

1. Narrative Control의 수학적 정식화
2. 자연어 제약이 임베딩 공간에 미치는 영향 분석
3. LLM Pathology 분류 체계 확립
4. ADHD 연구와 LLM 연구 간 cross-disciplinary 연결

8.3 열린 질문들

- Narrative Control Layer의 최적 구조는 무엇인가?
- 어떤 언어가 LLM 제어에 가장 효과적인가?
- Narrative Control은 모델 내부 학습으로 대체될 수 있는가?
- 방향성을 가진 지능은 어떤 새로운 능력을 보이는가?
- 서사적 제어와 의식(consciousness)의 관계는 무엇인가?

부록: 용어 정의

LLM Pathology
LLM이 exhibit하는 인지적 불안정성의 총합. 목표 부재, 주의 파편화, 과집중 발작, 흥미 기반 지능을 포함한다.

Narrative Control Gap
LLM이 서사(시간, 세계, 목표, 감정선, 정체성, 일관성)를 스스로 유지할 수 없는 구조적 간극.

Narrative Control Layer
자연어로 구성된 목표·감정축·규칙·world-model을 통해 LLM의 directionality를 부여하는 상위 제어 모듈.

Goal-less Architecture
LLM이 목적성·동기·자기일관성을 갖지 못하는 구조적 한계. 다음 토큰 예측만이 유일한 목적인 아키텍처.

Attention Fragmentation
Transformer의 Attention layer가 중요도 개념을 갖지 못해 발생하는 맥락 붕괴 현상.

Hyperfocus Episode
특정 조건(감정적 문장, 명확한 규칙, 강한 서사)에서 reasoning 품질이 비정상적으로 급상승하는 현상.

Interest-Driven Intelligence
LLM이 논리가 아니라 '흥미'를 기준으로 사고 품질이 변동하는 특성.

Directionality of Intelligence
지능이 지향하는 방향. 목표, 가치, 세계관, 감정선, 일관성을 포함하는 상위 개념.

— END OF DOCUMENT —