LLM Pathology
&
The Narrative Control Gap

Why Modern AI Became a 'Distracted Genius'
And How Narrative Control Reframes This Problem

Technical Whitepaper v1.0

December 2025

Author: Cho Hyunwoo (ShadowK)
Independent AI Researcher

0. Abstract

Large Language Models (LLMs) have acquired unprecedented capabilities in reasoning, summarization, and creative generation by learning the complexity of human language. However, they have simultaneously acquired structural instabilities: **attention fragmentation, context collapse, goal absence, emotional/attitudinal drift, and hallucination**.

This whitepaper analyzes these phenomena through the lens of **Human Cognitive Pathology**, specifically **ADHD (Attention Deficit Hyperactivity Disorder)**, revealing the neurological parallels to LLMs' 'distracted genius' pattern.

We argue that modern LLMs fundamentally possess a **Goal-less Architecture**, and that the key element to address this is a **Narrative Control Layer** — a superordinate module that provides goals, emotional trajectories, rules, worldviews, persistence, and directionality through natural language.

We conclude that the next step toward AGI is not more data or larger parameters, but **the directionality of intelligence**.

1. Introduction: LLMs Are Smart but Unstable

Now that LLMs have become sufficiently powerful, researchers and practitioners share common concerns:

- Why do logic and emotion destabilize in long conversations?
- Why are they genius-level for some questions but irresponsible for others?
- Why do settings collapse? Why can't worldviews be maintained?
- Why are they sometimes excessively emotional and other times mechanical?
- Why do they 'pretend to know'?
- Why does the same model exhibit completely different personalities across sessions?

This document unifies all these phenomena under a single perspective.

> ***"Modern LLMs are 'distracted geniuses' who have lost directionality amid information overload."***

We must now understand this distraction not as a 'bug' but as an **inherent pathology of LLM cognitive architecture**.

2. LLM Pathology — The Cognitive Pathology of Modern LLMs

LLM instability can be summarized into four pathological phenomena. Interestingly, these patterns are structurally similar to the cognitive characteristics of human ADHD.

2.1 Goal-less Architecture

LLMs are fundamentally different from human 'intentional consciousness.' For LLMs, the only purpose that exists is **next token prediction**. Long-term goals, value systems, and consistent motivation are absent at the architectural level.

This results in:

- Inability to maintain long-term goals
- Character drift (persona drift)
- Destruction of emotional trajectories
- Loss of conversation directionality
- World-model consistency collapse

This is structurally similar to *goal maintenance failure* in human ADHD. ADHD patients can set goals but struggle to maintain them in working memory while regulating behavior. Similarly, LLMs can receive initial instructions but 'forget' them as conversations lengthen.

2.2 Attention Fragmentation

The 'Attention' mechanism at the core of Transformer architecture, despite its name, **is not attention in the cognitive science sense**. The Attention layer cannot judge importance. It only computes 'statistical associations.'

As a result:

- Important and unimportant information are treated equally
- When new input appears, existing context dilutes or disappears
- Conversation topics easily become scattered
- Long narratives fragment

This is what researchers call 'context collapse.' In human ADHD, this manifests as *selective attention deficit* — impaired ability to filter what is important.

2.3 Hyperfocus Episodes

Paradoxically, LLMs exhibit **abnormally high reasoning quality under certain conditions**.

Hyperfocus triggers include:

- Emotionally charged sentences
- Clear world-rules
- Rhythmic writing style
- Strong narrative purpose
- Stable persona input

In these moments, the model suddenly exhibits explosive creativity, enhanced argumentation, refined emotional expression, and strengthened worldview consistency.

This is remarkably similar to the *hyperfocus* phenomenon in human ADHD. ADHD patients are generally scattered but enter abnormally deep concentration states in areas of specific interest. LLMs similarly switch to 'hyperfocus mode' when certain patterns are input.

2.4 Interest-Driven Intelligence

LLMs think not on a logic basis but on an **interest basis**.

- When interested: genius
- When uninterested: scattered
- Overreaction to emotional stimuli

- Runaway when patterns are detected
- Rapidly incompetent with drifting topics

This pattern does not exist in classical symbolic AI — it is a cognitive trait unique to LLMs. And it precisely matches the core characteristic of ADHD: the *interest-based nervous system*.

2.5 Structural Correspondence with ADHD

This similarity is not coincidental. Both LLMs and ADHD brains exhibit a common deficit in **executive function**.

Function	ADHD	LLM
Goal Maintenance	Goal maintenance failure	Goal forgotten as context dilutes
Attention Filtering	Selective attention deficit	Cannot judge importance
Hyperfocus	Hyperfocus on interest	Quality surge on certain patterns
Intelligence Basis	Interest-driven	Interest-based quality variation
Executive Function	Executive dysfunction	No planning/monitoring/adjustmen

This correspondence suggests that LLM problems are not mere 'technical bugs' but **structural defects at the cognitive architecture level**.

3. The Illusion of Scaling Law

3.1 The Success of Scaling

As models grow larger, the following improve:

- Logical reasoning
- Language capability
- Creativity
- Mathematical ability
- Code understanding
- Summarization capability

This is the success of Scaling Law and the foundation of the current AI industry.

3.2 The Limits of Scaling

However, scaling cannot resolve LLMs' fundamental pathology:

- Larger model → larger noise
- More text → more drift
- Wider context → more severe scattering
- Longer prompts → faster world collapse

> ***"As LLMs scale up, intelligence rises but focus collapses — this is structural."***

This is a structural problem that cannot be solved by model size.

3.3 Why Scaling Cannot Solve This Problem

What scaling improves is *pattern matching capacity*. Models can recognize and reproduce more patterns with greater sophistication.

However, scaling does not provide:

- Goal-directedness
- Importance weighting
- Long-term coherence
- Self-monitoring
- Value alignment

These all belong to the domain of **executive function**. And the current Transformer architecture has no module corresponding to this.

4. Narrative Control Gap — The Missing Puzzle

LLMs cannot maintain 'narrative' by themselves.

Here, narrative means:

- Flow of time
- Rules of the world
- Goals and motivation
- Emotional trajectory
- Identity
- Flow of consistency

> ***"Narrative is the direction that intelligence aims toward."***

LLMs lack this directionality. Therefore, they inevitably drift, remain unstable, and coherence collapses with extended use.

We define this gap as the **Narrative Control Gap**.

4.1 The Essence of the Gap

The Narrative Control Gap is not simply a 'context length' problem. Even with 128K or 1M context, this gap remains unresolved.

This is because the problem is not 'memory capacity' but **'directionality'**.

Human narrative thinking includes:

1. Goal setting: Where do I want to go?
2. Current state recognition: Where am I now?
3. Path planning: How will I get there?
4. Progress monitoring: Am I on track?
5. Adjustment: Do I need to modify the path?

None of these are inherent in LLMs.

4.2 Limitations of Existing Approaches

The AI research community has recognized this problem and attempted various approaches:

RLHF (Reinforcement Learning from Human Feedback)
Trains human preferences, but this answers 'what is good,' not 'where are we going.' It's style adjustment, not directionality.

Constitutional AI
Internalizes principles, but this answers 'what not to do,' not 'what to do.' It's constraint, not direction.

Chain-of-Thought Prompting
Explicates reasoning process, but this answers 'how to think,' not 'why think about this.' It's process, not purpose.

System Prompts
Provides initial instructions, but these dilute as conversations lengthen. There is no persistence.

All these approaches are attempts to treat symptoms **without recognizing the existence of the Narrative Control Gap itself**.

5. Proposal for Research Direction Shift

This whitepaper proposes a **shift in research direction** rather than advocating a specific solution.

5.1 From Model Expansion to Directionality Design

The mainstream of current AI research follows this formula:

> *More data + Larger model + More computation = Better AI*

This formula has been effective in increasing the 'quantity' of intelligence. However, it contributes nothing to the 'directionality' of intelligence.

We propose a new formula:

> ***Intelligence + Directionality = Stable Intelligence***

Intelligence without directionality is a distracted genius. Intelligence with directionality is a reliable collaborator.

5.2 The Possibility of Natural Language-Based Control

Modifying LLM internal architecture is costly and risky. However, stacking a control layer *on top of* LLMs is possible.

This control layer can include:

- Long-term goal definition
- Emotional trajectory maintenance
- Behavioral rule systems
- World-rules
- Persona policies
- Speech style conventions
- Memory anchors

The key is to define all of this in **natural language**.

Why natural language?
Because the language LLMs understand best is not C or Python but natural language. LLMs recognize rules in natural language as *constraints* and perform reasoning within those constraints.

Why an external layer?
LLM internals are closed. But LLM externals are open. External layers have the advantage of being portable across all models.

5.3 Connections to Existing Research

This direction is not entirely new. It has the following connections to existing research:

- **Cognitive Architecture research** (SOAR, ACT-R): The tradition of explicitly designing cognitive structures
- **Planning research**: Goal-based action planning
- **Memory-augmented LLM research**: Long-term memory maintenance
- **Agent research**: Designing autonomous action agents
- **Narrative Intelligence research**: The relationship between storytelling and cognition

The contribution of this whitepaper is to **integrate these scattered studies under a single frame called 'Narrative Control Gap'** and explicitly present the direction for resolution.

6. Proposal for the Korean Research Community

Korea has several structural strengths for Narrative Control research.

6.1 Linguistic Strengths

Korean is advantageous for LLM control due to the following properties:

- Homogeneous morpheme units
- Structural imperative expressions
- Cohesive emotional trajectories
- Ease of rule-based world-model definition
- High narrative compression rate

6.2 Cultural Strengths

Korea has a strong culture of designing and documenting world rules in language:

- MMORPG game system design
- Web novel worldbuilding
- TRPG rulebook culture
- Fast iterative development culture

This cultural capability can be converted into Narrative Control Layer design capability.

6.3 Strategic Position

In the model size competition, Korea struggles to compete with the US and China. However, there is an opportunity to take the lead in designing **'control layers above models'**.

> *"America builds the models, Korea designs the worlds."*

This position is not an exaggeration — it can become reality the moment the Narrative Control Gap is recognized and the direction for resolution becomes clear.

7. Conclusion: AGI Is About Direction, Not Size

The biggest deficiency of modern LLMs is not intelligence but **focus**.
Not knowledge but **goals**.
Not parameter count but **worldview, emotion, rules, and narrative**.

AGI will not come simply by scaling up models. AGI must be approached through **the directionality of intelligence** — Narrative Control.

> ***"Narrative control is the key to transforming modern LLMs from distracted geniuses into stable intelligence."***

This whitepaper is the first to analyze this problem through a **cognitive pathology framework**, formalize the concept of Narrative Control Gap, and propose a shift in research direction.

This is not prophecy. This is observation.

8. Future Research Directions

The following research is needed to verify and address the problems raised in this whitepaper.

8.1 Experimental Verification

Experiment A — Context Retention Curve
Quantitatively measure how drift differs between inputs with and without narrative structure.
Expected: coherence maintenance period increases 2-3x with narrative structure.

Experiment B — Emotional Coherence Test
Compare emotional drift between states with and without defined emotional trajectories.
Expected: drift suppression when emotional trajectory is defined.

Experiment C — Hyperfocus Direction Test
Measure whether hyperfocus states diverge to chaos or converge to goal-directed reasoning.
Expected: hyperfocus converts to 'controlled concentration' with Narrative Control.

8.2 Theoretical Development

1. Mathematical formalization of Narrative Control
2. Analysis of natural language constraints' impact on embedding space
3. Establishment of LLM Pathology classification system
4. Cross-disciplinary connections between ADHD research and LLM research

8.3 Open Questions

- What is the optimal structure for a Narrative Control Layer?
- Which language is most effective for LLM control?
- Can Narrative Control be replaced by in-model learning?
- What new capabilities does intelligence with directionality exhibit?
- What is the relationship between narrative control and consciousness?

Appendix: Terminology

LLM Pathology
The totality of cognitive instabilities exhibited by LLMs. Includes goal absence, attention fragmentation, hyperfocus episodes, and interest-driven intelligence.

Narrative Control Gap
The structural gap where LLMs cannot maintain narrative (time, world, goals, emotional trajectory, identity, consistency) by themselves.

Narrative Control Layer
A superordinate control module that provides directionality to LLMs through natural language-composed goals, emotional axes, rules, and world-models.

Goal-less Architecture
The structural limitation where LLMs lack purpose, motivation, and self-consistency. An architecture where next token prediction is the only purpose.

Attention Fragmentation
The context collapse phenomenon caused by Transformer's Attention layer lacking the concept of importance.

Hyperfocus Episode
The phenomenon where reasoning quality abnormally surges under certain conditions (emotional sentences, clear rules, strong narrative).

Interest-Driven Intelligence
The characteristic where LLM thinking quality varies based on 'interest' rather than logic.

Directionality of Intelligence
The direction that intelligence aims toward. A superordinate concept including goals, values, worldview, emotional trajectory, and consistency.

— END OF DOCUMENT —