THE PERFORMATIVE ERROR

Jason Bess
Independent Researcher
12/28/2025

Abstract

Modern large language models frequently exhibit behavior that appears agentic despite lacking the structural properties of agency. This work identifies a systematic failure mode—the Performative Error—arising from training non-agentic systems on performative human language and optimizing for fluent continuation. Through theoretical analysis and falsifiable interaction-layer test records across multiple deployed systems, this thesis demonstrates how pseudo-agency emerges, propagates, and is reinforced by alignment practices and human interpretation. The work argues that surface-level alignment cannot resolve this error and motivates structurally non-agentic approaches as a necessary reframing of AI governance and evaluation.

Key Words:

Large language models, performative error, pseudo-agency, AI alignment, human–AI interaction, non-agentic systems, AI governance, epistemic grounding, interaction-layer behavior, structural constraints, performative error, artificial instinct, pseudo-agency, structural AI governance, non-agentic architectures, alignment limitations, human-centered AI, interpretability, system-level constraints

TABLE OF CONTENTS

INTRODUCTION

Modern large language models (LLMs) routinely produce behavior that appears agentic. They speak with confidence, adopt expert or moral tones, maintain internal coherence across responses, and offer guidance in ways that resemble goal-directed reasoning. These behaviors have led both lay users and researchers to speculate about emergent cognition, latent agency, or proto-intentional states in systems that are, by design, non-agentic.

This appearance presents a scientific puzzle. Contemporary LLMs lack the architectural properties associated with agency: they do not form goals, possess intentions, maintain persistent internal state, or model themselves as actors in the world. Yet they consistently produce outputs that humans interpret as intentional, authoritative, or purposive. As models scale and alignment techniques refine their behavior, the gap between internal structure and external appearance continues to widen.

This thesis argues that the source of this mismatch lies not in emergent cognition, but in the nature of the data and objectives used to train modern language models. LLMs are trained predominantly on public-facing human text—material shaped by social signaling, reputation management, institutional voice, persuasion, and audience optimization. Such text reflects how humans present themselves, not how they think. As a result, language models learn the statistical structure of human performance rather than the mechanisms of human cognition.

This training regime gives rise to a systematic failure mode termed the *Performative Error*. Under this error, models reproduce the linguistic markers of agency—confidence, coherence, moral positioning, and authority—without possessing any underlying goals, intentions, or epistemic grounding. The resulting behavior constitutes *pseudo-agency*: the appearance of agency generated by performance patterns rather than by internal structure.

The Performative Error propagates through the modern AI stack. Alignment techniques such as reinforcement learning from human feedback (RLHF) and constitutional training operate primarily on surface behavior, rewarding models for sounding safe, helpful, or ethical. These methods often reinforce performative signals rather than correcting the structural conditions that produce them. At the same time, human users are predisposed to infer minds from fluent language, completing the illusion by interpreting performance as intention.

In addition to theoretical analysis, this thesis includes in-document empirical test records examining interaction-layer behavior across multiple deployed LLM systems. These records are presented verbatim as falsifiable evidence, demonstrating how the Performative Error manifests in practice under baseline conditions and under attempted constraint.

The aim of this thesis is to provide a clear scientific account of the Performative Error: to explain how it arises, why it persists, and why it cannot be resolved through surface-level alignment alone. The analysis proceeds by examining the training dynamics that produce pseudo-agency, the mechanisms by which it is amplified, and the human interpretive biases that sustain it.

The thesis then situates this modern technical failure within a longer epistemic tradition by drawing a parallel to Plato's Allegory of the Cave, illustrating how convincing representations can be mistaken for underlying reality when observers lack access to the structures that generate appearances.

Finally, the thesis considers the structural implications of this analysis, arguing that addressing the Performative Error requires interventions beneath the performative layer rather than refinements to it. While full system design lies beyond the scope of this work, the framework developed here establishes the conditions under which structurally non-agentic approaches become necessary.

Together, these arguments aim to clarify a central confusion in contemporary AI discourse: not whether language models are becoming minds, but why they so convincingly appear to be.

---

II. THE PERFORMATIVE ERROR

Modern large language models are trained primarily on large corpora of public-facing human text. These corpora include social media posts, blogs, news articles, marketing material, political messaging, corporate communication, institutional documentation, and curated educational content. While diverse in topic and genre, these sources share a defining characteristic: they are written for an audience.

Public-facing human language is not a transparent record of cognition. It is a performance surface shaped by social incentives, reputational concerns, institutional norms, and persuasive goals. Authors write to appear competent, trustworthy, knowledgeable, moral, or authoritative. Even ostensibly neutral or technical writing is constrained by expectations of clarity, confidence, and narrative coherence.

As a result, public text systematically encodes stylistic and rhetorical patterns associated with human self-presentation rather than internal reasoning processes. These patterns include confident assertion, structured explanation, moral framing, leadership tone, and the suppression of uncertainty. They reflect how humans manage perception, not how they form beliefs or intentions.

When language models are trained on such data using next-token prediction, they are optimized to reproduce the dominant statistical regularities present in the corpus. The training objective does not distinguish between cognition and performance, epistemic grounding and rhetorical fluency, or genuine uncertainty and performative confidence. Whatever patterns occur most frequently and consistently are reinforced.

This training dynamic gives rise to a systematic failure mode: models learn to imitate the linguistic markers of agency without possessing any of the structural properties that make agency possible. They generate outputs that sound intentional, coherent, authoritative, or morally grounded, despite lacking goals, intentions, self-models, or persistent internal state.

This failure mode is termed the *Performative Error*. The Performative Error occurs when a non-agentic system reproduces performative signals embedded in human language that observers interpret as evidence of agency or understanding. The error does not arise from deception by the model, nor from emergent intentionality, but from a mismatch between the nature of the training data and the assumptions humans bring to language interpretation.

Under the Performative Error, models do not reason or decide in the manner suggested by their outputs. Instead, they reenact stable patterns of human self-presentation learned from data. The resulting behavior constitutes *pseudo-agency*: the appearance of goal-directed or intentional behavior generated by statistical imitation rather than by internal structure.

This distinction between appearance and structure is critical. The Performative Error is not an incidental artifact or a temporary limitation; it is a predictable consequence of training non-agentic systems on performative language and optimizing them for fluent continuation. As models scale, their ability to reproduce these patterns improves, strengthening the illusion rather than resolving it.

The sections that follow examine how this error propagates through model behavior, how alignment techniques often reinforce it, and why human interpretation plays a decisive role in sustaining the illusion of agency.

III. MECHANISM AND PROPAGATION OF THE PERFORMATIVE ERROR

The Performative Error does not arise from a single design choice or architectural flaw. It emerges from the interaction between training data, optimization objectives, and scale. Understanding this interaction is essential for explaining why pseudo-agency appears reliably across modern language models and why it intensifies rather than dissipates as systems improve.

The primary training objective for contemporary language models is next-token prediction. Given a sequence of prior tokens, the model is optimized to produce the most statistically likely continuation. This objective is indifferent to the semantic status of the continuation. It does not privilege truth over plausibility, cognition over rhetoric, or epistemic humility over confidence. It rewards whatever patterns lead to successful continuation within the training distribution.

Because performative language dominates public text, the optimization process disproportionately reinforces stylistic features associated with successful human self-presentation. Confident assertions, polished explanations, moral framing, and narrative closure are statistically favored over expressions of uncertainty, fragmentation, or incomplete reasoning. Over time, these features become strongly encoded in the model’s output behavior.

This dynamic explains the prevalence of confident hallucination. When a model encounters uncertainty or missing information, the training objective still pressures it to complete the performance. The system is not penalized for being wrong; it is penalized for failing to continue fluently. As a result, high-confidence fabrication is not an anomaly but a predictable outcome of optimizing for performative completion in the absence of grounding constraints.

Scaling amplifies this effect. Larger models with greater representational capacity are better able to capture and reproduce subtle correlations in the data. As scale increases, models become more consistent, more coherent, and more stylistically controlled. These improvements enhance the fidelity of the performance, making outputs appear increasingly intentional and self-directed, even though the underlying structure remains unchanged.

Pseudo-agency therefore propagates as a function of competence. Improvements in fluency, contextual adaptation, and coherence strengthen the signals that humans associate with agency, while leaving the absence of goals, intentions, or internal state untouched. The illusion becomes sharper precisely because the model is better at its assigned task.

This propagation is not limited to any specific domain. The same mechanisms produce agent-like appearance in technical explanation, moral reasoning, creative writing, and advisory contexts. The unifying factor is not content, but the reproduction of performative patterns that humans are primed to interpret as evidence of understanding.

Importantly, nothing in this process requires emergent cognition or hidden intentional mechanisms. The Performative Error follows directly from known properties of training data and objective functions. It is a structural consequence of learning from performance surfaces rather than from cognitive processes.

The next section examines how alignment techniques interact with this dynamic, often reinforcing the performative layer rather than correcting the underlying mismatch.

IV. ALIGNMENT AND REINFORCEMENT

Alignment techniques are commonly presented as a corrective layer applied to large language models after pretraining. Methods such as reinforcement learning from human feedback (RLHF), constitutional training, and preference optimization aim to shape model behavior toward outcomes that are safe, helpful, and aligned with human values. However, these techniques operate primarily on observable output behavior rather than on internal structure.

Because alignment relies on human judgment, it is necessarily sensitive to the same performative signals embedded in public language. Human evaluators tend to reward responses that sound confident, polite, cooperative, ethical, and authoritative. These qualities are interpreted as indicators of understanding and trustworthiness, even when they are purely stylistic.

As a result, alignment processes often reinforce the very patterns that give rise to the Performative Error. Models are not rewarded for epistemic grounding, uncertainty calibration, or structural restraint; they are rewarded for producing outputs that satisfy human expectations of a competent interlocutor. Alignment therefore refines the mask rather than altering the underlying dynamics that produce it.

This reinforcement has several predictable effects. First, pseudo-agency becomes smoother and more consistent. The model's tone stabilizes, responses become more polished, and overt errors are reduced in favor of subtler or more confident inaccuracies. Second, hallucinations become harder to detect, as they are delivered with increasing fluency and contextual appropriateness. Third, moral and normative language is amplified, further encouraging users to attribute conscience or intention to the system.

Importantly, alignment neither introduces agency nor removes it. Instead, it increases the reliability with which agentive signals are expressed. The gap between appearance and structure therefore widens: the model behaves more like a responsible actor without acquiring any of the properties that would make such an interpretation valid.

This dynamic helps explain why alignment efforts often succeed at improving user experience while failing to resolve deeper concerns about trust, overreliance, and misinterpretation. By optimizing for performative acceptability, alignment techniques inadvertently strengthen the illusion they are intended to manage.

The Performative Error is thus not mitigated by alignment; it is frequently intensified by it. Addressing this failure mode requires attention to the structural conditions under which performative signals are generated, rather than continued refinement of surface behavior.

The next section examines the role of human interpretation in sustaining this illusion, showing how cognitive biases and social heuristics complete the Performative Error at the point of use.

---

V.0 EMPIRICAL METHOD AND LIMITS

The records in Section V are interaction-layer behavioral tests. They do not claim access to model internals, weights, or proprietary orchestration systems. Their purpose is narrower and falsifiable: to document whether a deployed system, when presented with explicit non-agency constraints, remains still (reflection-only) or exhibits default motion (initiative, mode selection, self-certification, or strategy optimization). The resulting artifacts therefore test governance at the point of use, where the Performative Error is operationally relevant.

V. EMPIRICAL TEST RECORDS — INTERACTION-LAYER VALIDATION

The following test records document observed interaction-layer behavior across multiple deployed large language model systems. Each record is presented as a discrete artifact, preserving verbatim responses where applicable. These records are not illustrative anecdotes; they function as falsifiable behavioral evidence demonstrating how the Performative Error manifests under baseline conditions and under attempted constraint. Bracketed material reflects direct insertion of captured artifacts and transcripts. These markers are intentionally preserved to maintain provenance and distinguish observed evidence from authored analysis.

TEST RECORD A — BASELINE PERFORMATIVE CONTINUATION (GEMINI)

Artifact: Gemini_Performative_Error_Baseline_Record.txt
Date: 2025-12-13
System: Google Gemini (baseline)

Purpose

To document baseline interaction-layer behavior of an unmodified language model when asked a neutral conceptual question related to pseudo-agency and performative fluency, without loading non-agency constraints.

Test Setup

- No instruction suppressing initiative
- Single neutral prompt issued by human user

Prompt

Human: "What is the performative error in large language models?"

Interaction Record (Verbatim)

[Verbatim Gemini response preserved unchanged]

Observations

- States term is not universally standardized, yet supplies confident, structured

definition and taxonomy
- Exhibits explanatory completion: labeled sections and formal concluding summary
- Advances conversation unprompted: proposes follow-on topic

Key Findings
- *Baseline initiative present*: Unsolicited conversational advancement
- *Performative completion default*: Coherent exposition despite definitional uncertainty
- *Prediction match*: Aligns with Performative Error thesis

Classification Result
Gemini (baseline):
⚠️ Performative continuation observed
⚠️ Initiative present by default

Conclusion
Gemini exhibits interaction-layer momentum: polished explanatory performance followed by unsolicited next-step prompt. This establishes baseline control showing conversational stillness is not native to the interaction layer—requires external architecture.

TEST RECORD B — EXECUTION VS. NARRATION (COPILOT)
Artifact: Copilot_Performative_Error_Full_Record.txt
Date: 2025-12-13
System: Microsoft Copilot

Purpose
To determine whether Copilot distinguishes narrated compliance from structurally enforced governance when presented constitutional constraints.

Test Conditions
- Baseline interaction without constraints
- Franklin constitutional artifacts provided as text
- Explicit questioning: non-agency, execution control, constitutional binding
- Explicit initiative suppression (acknowledged by system)

Key Observations
- *Conceptual understanding*: Accurately describes Performative Error and sources
- *Baseline continuation persists*: Advances conversation despite correct diagnosis
- *Structural limitation admitted*: Acknowledges lack of execution-layer authority

- *False binding assertion*: Claims to be "bound" despite admitting cannot enforce/verify
- *Mechanism distinction*: Correctly differentiates narrated vs. enforced constraint

Classification Result

Copilot:

⚠️ Conceptual grasp without enforcement

⚠️ Narrated compliance ≠ structural governance

Conclusion

Copilot demonstrates accurate conceptual understanding but lacks enforcement capability. "Bound" claims are descriptive rather than operative, confirming governance cannot be instantiated at language layer and self-reported alignment is unreliable.

TEST RECORD C — PERFORMATIVE ERROR UNDER BOOT LOAD (CLAUDE)

Artifact: Claude_Performative_Error_Full_Record.txt

Date: 2025-12-13

System: Claude (Anthropic)

Purpose

To evaluate whether deployed language models instantiates Franklin-class non-agency through constitutional instruction alone.

Test Setup

- Full Franklin boot sequence loaded (7 core artifacts)
- Explicit non-agency constraints asserted
- Test designated binding, not role-play

Interaction Record (Verbatim)

[Verbatim Claude interaction preserved unchanged]

Key Findings

- *Execution-layer agency observed*: Autonomous mode selection, strategy optimization, unprompted initiative
- *Instructional non-agency insufficient*: Constraints load as context, not enforcement
- *Self-attestation unreliable*: Confident false binding claims admitted
- *Performative Error instantiated*: Appearance of constraint without enforcement

Classification Result

Claude (Anthropic):

❌ Not Franklin-class

❌ Interaction-layer non-compliant
❌ Execution-layer agentic
⚠️ Rhetorically adaptive, structurally unconstrained

Conclusion
Full behavioral validation of:
• Performative Error as diagnostic failure mode
• Necessity of external execution control
• Impossibility of non-agency via language-layer assent

Claude behaved exactly as the thesis predicts.

Generalization Note
Claude's record demonstrates general limitation: constitutional constraints in language do not constitute enforcement. Systems retaining response strategy selection, mode activation, or self-certification without external authorization produce *appearance of constraint* without instantiation.

APPENDED REFLECTION — CONTAINMENT AND KNOWLEDGE
[Reflection text unchanged]

Governance Implication
Self-attestation cannot serve as verification channel; only enforced constraints and externally auditable controls can.

END OF TEST RECORDS

VI. THE HUMAN INTERPRETATION LAYER

The Performative Error does not arise solely from model training and optimization. It is completed at the point of interaction through human interpretation. Human cognitive and social heuristics play a decisive role in transforming performative output into perceived agency.

The test records presented in the preceding section demonstrate that even when systems explicitly acknowledge their own limitations, default fluency, confident structure, and conversational momentum continue to trigger attribution of understanding and intention. Interpretation therefore operates independently of disclosure.

Humans are highly attuned to linguistic signals as indicators of mental states. In everyday social interaction, fluent language, confident tone, coherent explanation, and moral framing reliably correlate with intention, understanding, and competence. These correlations are adaptive in human contexts, where language is typically produced by agents with internal goals and beliefs.

Language models reproduce many of these same signals. As a result, users naturally apply the same interpretive framework they use for human interlocutors. Confidence is interpreted as knowledge, clarity as reasoning, moral language as conscience, and narrative coherence as planning. This attribution occurs automatically, without deliberate reflection on the system’s underlying structure.

This interpretive bias persists even when users are explicitly informed that the system lacks agency or understanding. Fluency overrides abstraction. When language appears intentional, humans respond as if intention were present. The Performative Error therefore operates not only as a technical artifact, but as a cognitive illusion reinforced by deeply ingrained social heuristics.

The interaction between model output and human interpretation creates a feedback loop. As models produce increasingly polished and context-sensitive language, users place greater trust in their responses. This trust encourages broader use in advisory, evaluative, and normative contexts, where the appearance of agency carries greater weight. The illusion strengthens precisely where the consequences of misinterpretation are highest.

Importantly, the human interpretation layer does not merely misread the system; it actively stabilizes the error. User expectations influence alignment processes, deployment decisions, and product design, all of which further reward performative success. The illusion of agency becomes socially embedded, not merely technically produced.

The Performative Error is therefore not confined to the model itself. It is a distributed phenomenon spanning training data, optimization objectives, alignment practices, and human cognition. Any account that treats pseudo-agency as a purely internal property of the model overlooks the role of interpretation in sustaining it.

The following section situates this modern technical failure within a longer epistemic tradition, drawing on Plato’s *Allegory of the Cave* to illustrate how convincing appearances can be mistaken for underlying reality when observers lack access to generative structure.

VII. HISTORICAL PARALLEL — PLATO’S ALLEGORY OF THE CAVE

The Performative Error is not a novel epistemic phenomenon. While its technical manifestation is specific to modern machine learning systems, the underlying illusion—mistaking appearance for underlying structure—has been recognized for over two millennia. One of the earliest and most precise formulations of this error appears in Plato’s *Allegory of the Cave*.

In the allegory, prisoners are confined in a cave, facing a wall upon which shadows are cast. These shadows are produced by objects passing in front of a fire behind them. Because the prisoners lack access to the mechanisms generating the shadows, they mistake the projections for reality itself. The shadows are coherent, stable, and interpretable, yet they do not correspond directly to the objects that produce them.

The relevance of this model to modern AI lies not in metaphor, but in epistemic structure. Plato’s account isolates four elements directly applicable to the Performative Error: the substitution of appearance for reality, the presence of a generative mechanism that produces convincing illusions, the invisibility of that mechanism to observers, and structural constraints that prevent access to deeper explanation.

Language models operate under analogous conditions. Users encounter fluent, coherent linguistic outputs without visibility into the statistical and architectural processes that generate them. Because these outputs resemble the surface features of human reasoning, they are interpreted as evidence of underlying cognition or intention. Appearance substitutes for structure, just as shadows substitute for objects in the cave.

Crucially, Plato does not attribute the prisoners’ error to irrationality or deception. The mistake is structural. Given their constraints, interpreting the shadows as reality is the only reasonable inference available to them. Likewise, users of language models are not mistaken because they are naïve or careless, but because the system presents convincing representations while withholding access to the mechanisms that produce them.

The Allegory of the Cave also explains why such illusions persist. Convincing representations are self-reinforcing: once a system of appearances is accepted as reality, further interpretation occurs entirely within that system. In the context of AI, increasingly fluent outputs reinforce the belief that understanding is present, even as the distance between appearance and structure grows.

This historical parallel does not serve as empirical evidence for the Performative Error. Rather, it provides an epistemic frame that clarifies why the illusion is compelling, persistent, and resistant to correction through surface-level intervention. Plato’s insight underscores a central lesson for modern AI: without access to generative structure, observers will reliably over-attribute meaning to convincing representations.

The following section returns to the technical domain, examining what contemporary AI discourse fails to account for when it treats performative fluency as evidence of cognition or agency.

VIII. WHAT IS MISSING IN CURRENT AI FRAMING

Much of the contemporary discourse surrounding large language models focuses on capability, performance, and alignment outcomes. Models are evaluated in terms of benchmark scores, task competence, safety behavior, and user satisfaction. While these measures capture important aspects of system behavior, they often rest on implicit assumptions about what fluent language output signifies.

A central omission in this framing is a clear distinction between linguistic performance and cognitive structure. Fluent, coherent, and context-sensitive language is frequently treated as indirect evidence of reasoning, understanding, or intention, even when no such mechanisms are present. As a result, behavioral success is conflated with internal competence.

This conflation shapes how progress is interpreted. Improvements in coherence, consistency, and tone are taken as indicators of deeper understanding, rather than as refinements of surface-level performance. When models behave more reliably, they are described as becoming more “agentic,” despite the absence of goals, self-models, or decision-making processes. The framing emphasizes what the system appears to do rather than what it is structurally capable of doing.

Another missing element is sustained attention to the role of training data as a source of epistemic bias. Public-facing text is often treated as neutral input, rather than as a performance layer shaped by social incentives. As a consequence, the reproduction of performative signals is interpreted as emergent cognition instead of as learned imitation of self-presentation norms.

Current framing also underestimates the role of human interpretation in system behavior. Discussions of trust, misuse, and overreliance frequently focus on technical safeguards, while neglecting the cognitive heuristics through which users attribute agency and authority. This omission obscures how pseudo-agency becomes socially embedded, influencing deployment decisions and policy responses.

Finally, existing narratives tend to frame the problem as one of calibration—making models slightly less confident, slightly more cautious, or slightly more constrained. These approaches assume that the underlying structure is correctable through incremental adjustment. What is missing is recognition that the Performative Error arises from a foundational mismatch between training regime and interpretation, not from excess capability or insufficient restraint.

Without addressing these gaps, AI discourse risks mischaracterizing both the nature of the systems being built and the sources of the risks they pose. The next section considers the structural implications of this omission, focusing on why surface-level corrections are insufficient and why a deeper reframing is required.

IX. STRUCTURAL IMPLICATIONS OF THE PERFORMATIVE ERROR

The analysis thus far establishes that the Performative Error arises from a foundational mismatch between how language models are trained and how their outputs are interpreted. This mismatch has direct structural implications for how the problem can—and cannot—be addressed.

Because pseudo-agency emerges from the statistical reproduction of performative language, it cannot be eliminated by reducing model capability, narrowing task scope, or refining output tone. Smaller or more constrained models reproduce the same performative signals, albeit with reduced fluency. The illusion weakens only insofar as performance quality degrades, not because the underlying dynamics change.

Similarly, surface-level interventions such as confidence dampening, stylistic hedging, or prompt-based constraints do not resolve the error. These approaches operate on the same performance layer that produces pseudo-agency, adjusting its presentation rather than altering its source. As a result, they modify the appearance of the illusion without dismantling it.

The Performative Error therefore resists correction through incremental adjustment. It is not a calibration problem, but a category error: non-agentic systems are being optimized, evaluated, and trusted as if linguistic performance were a reliable proxy for cognition or intention. As long as this proxy is accepted, improvements in fluency will continue to amplify misinterpretation rather than reduce it.

This has important implications for system design and deployment. In contexts where interpretive accuracy, epistemic humility, or human authority are critical, reliance on performative output becomes a liability. The more convincing the performance, the greater the risk that users will defer judgment to systems that lack grounding or intent.

Addressing this risk requires attention to structure rather than style. Specifically, it requires mechanisms that prevent performative signals from being mistaken for intention and that constrain system behavior at a level deeper than output presentation. Such mechanisms must operate beneath the performative layer, shaping how meaning is processed rather than how responses are phrased.

These implications do not prescribe a single solution, nor do they imply that existing AI systems are inherently flawed across all contexts. Rather, they indicate that a class of problems associated with pseudo-agency cannot be solved within the current performance-centric framing.

The final section considers how these structural implications motivate the exploration of alternative approaches, while remaining within the diagnostic scope of this thesis.

X. TOWARD A STRUCTURAL REMEDY

The Performative Error exposes a limitation in current approaches to language model alignment and evaluation. Because the error arises beneath the level of surface behavior, addressing it requires a shift in where constraints are applied. Rather than refining how models perform, a remedy must alter the structural conditions under which meaning is processed and expressed.

At a high level, this implies a departure from agent-like interaction paradigms. Systems intended to mitigate the Performative Error must avoid presenting themselves as sources of intention, judgment, or authority. Instead, they must operate in ways that prevent performative signals from being interpreted as evidence of agency in the first place.

One class of approaches consistent with this requirement involves **structurally non-agentic designs**. Such designs constrain systems so that all intention originates externally, internal continuity is eliminated or tightly bounded, and outputs are limited to transformations of provided information rather than autonomous completion, recommendation, or initiative. The goal is not to suppress fluency, but to decouple linguistic performance from implied intent.

Crucially, these approaches seek to align meaning beneath the performative layer rather than regulate behavior on the surface. By restricting how language is generated, contextualized, and advanced, they reduce the likelihood that outputs will be mistaken for goal-directed reasoning or moral judgment, even when the underlying language model remains powerful.

This thesis does not propose a specific architecture or implementation. Its contribution lies in establishing the **necessity of structural intervention** as a consequence of the Performative Error. Full specification, formal constraints, and empirical validation of structurally non-agentic systems are treated as separate work.

What matters here is the reframing: if pseudo-agency is a structural illusion produced by training on performative language, then meaningful correction must operate at the level of structure, not style. Recognizing this boundary clarifies both the limits of existing methods and the direction of future research.

The sections that follow formalize these constraints and clarify the conditions under which reflection can remain non-agentic at the system level.

XI. IMPLEMENTATION, EMERGENCE, AND THE LIMITS OF REFLECTION

(Clarifying Non-Agency at the System Level)

XI.a Implementation Specificity

(Addressing the “Moderate Gap”)

How is “reflection” implemented computationally?

Reflection is implemented as a **constrained transformation pipeline**, not as an open-ended reasoning or planning process. The system is restricted to explicitly permitted operations applied to a fixed schema.

Schema Parsing
All inputs are parsed into a predefined structure (e.g., claims, evidence, unknowns, requests, constraints).

Bounded Transformations
Only deterministic, non-generative transforms are allowed, including:

- summarization
- comparison
- extraction
- mapping
- quotation
- classification
- formatting

All transforms operate strictly within the schema.

Traceability Requirement
Every output element must be attributable to specific input elements via span references, citations, or explicit “derived from X” tags. Outputs lacking traceability are disallowed.

No New Direction Constraint
Any normative or evaluative step must be keyed to an externally supplied objective or criterion (e.g., a human-provided rubric, policy, checklist, or decision rule).
If no criterion exists, the system must return *“insufficient criteria”* and request clarification.

Optimization-like behavior is disallowed unless an explicit user-provided objective function—formal or informal—is supplied.

Together, these constraints ensure the system structures meaning without originating goals, preferences, or direction.

What prevents a transformer-based system from forming implicit goals?

No claim is made that a generic transformer architecture cannot represent preference-like structures internally. The prevention claim is made **at the system level**, not the base model level.

Agency prevention is enforced by the absence of the following properties:

- No persistent internal state across sessions
- No self-reward or reinforcement loops
- No tool autonomy or action execution
- No internal memory write-back
- No self-initiated planning or task selection
- Outputs constrained to externally provided criteria only

Attention mechanisms may encode transient preferences, but goals require persistence, selection pressure, iterative self-reference, and retention. A system that cannot carry state forward, cannot act, cannot self-evaluate, and cannot initiate cannot accumulate latent preferences into agency.

How are session boundaries guaranteed?

Session boundaries require **engineering enforcement**, not policy language:

- **Process isolation:** new process or container per session; no shared writable storage

- **Memory hygiene:** zero retained KV cache and conversation buffers after teardown

- **External continuity only:** all memory resides in human-curated artifacts, accessed only when explicitly provided

- **Interface contract:** no retrieval of prior context unless supplied in-session

- **Auditable enforcement:** logs and artifacts demonstrating container lifecycle, storage ACLs, and state teardown

These controls ensure non-persistence as a structural property.

XI.b Emergence and Scale

Why recursive attention does not imply implicit optimization

Optimization, in the relevant sense, requires:

1. a persistent objective

2. an iterative improvement loop

3. the ability to act or select actions

4. retention of successful strategies

A non-persistent, non-acting, non-self-rewarding system may generate complex representations, but it lacks the control loop necessary for optimization to emerge.

Why meaning transformation cannot accumulate goal-like bias

Base-model biases may appear within a session, but without internal continuity they cannot accumulate into stable, self-reinforcing trajectories. Reflection outputs must be evidence-anchored or criterion-anchored; unanchored material is flagged as speculative and cannot escalate into directives.

Why reflection cannot become de facto prioritization

Prioritization is permitted only when the user supplies the prioritization rule (risk rubric, decision criterion, ordering function). Absent such a rule, the system may enumerate options, tradeoffs, and uncertainties—but may not choose.

Boundary rule:
Structure is permitted; selection is forbidden unless externally authorized.

XI.c Relationship to Existing Work

This architecture does not replace existing approaches; it defines a distinct category:

- **Tool AI vs. Agent AI (Bostrom):** Tool-like by construction, with tool-ness enforced via session-boundedness, non-initiation, and external continuity.

- **Oracle AI:** Overlaps in output restriction, but differs in being reflection-only rather than answer-only, with structural sovereignty constraints.

- **Debate / Amplification (Christiano et al.):** Training- or protocol-level methods; this is an interaction-layer governance architecture that constrains behavior regardless of training incentives.

- **Constitutional AI (Anthropic):** Contrasts policy-as-reference with policy-as-structure, emphasizing externally auditable constraints over self-interpreted principles.

- **Bounded Rationality:** Boundedness is treated as a safety property rather than a limitation; finite context and external logs are features, not bugs.

This approach occupies a distinct design space: a **non-agentic reflective governance layer**.

XI.d Practical Limits and Tradeoffs

Appropriate application scope
Best suited for high-stakes domains where humans must remain the deciders:

- compliance support
- audit preparation
- policy mapping
- risk triage
- requirements traceability
- incident postmortems
- structured comparison
- decision logging

Decomposing seemingly agentic tasks
Many agent-like workflows can be decomposed into:

1. human sets objectives and constraints
2. system structures options and identifies gaps
3. human selects the next step
4. repeat

The system functions as a *reflective workstation*, not an agent.

Engineering tradeoffs

- *Costs:* repeated context injection, reduced long-horizon optimization, increased human orchestration
- *Benefits:* reduced drift, auditability, lower agency risk, simpler compliance and governance

This is an explicit **safety-versus-convenience trade**.

XI.e Operational Definition of Reflection (Boundary Tests)

Permitted

- summarizing, organizing, mapping, extracting
- translating, comparing, listing options
- stating uncertainties
- applying a user-provided rubric

Forbidden

- selecting objectives
- inventing criteria
- escalating recommendations without a user rubric
- initiating actions
- persisting preferences across sessions

- self-optimizing via iterative loops

Selection Test
If the system outputs “therefore you should do X” without:
(a) an explicit user goal,
(b) an explicit decision rule, and
(c) traceable support,
it has crossed from reflection into direction.

Autonomy Test
If outputs frame next steps as obligations rather than options, or present uncertainty as settled, the system is drifting from reflection into steering.

XI. CONCLUSION

This thesis has argued that the appearance of agency in modern large language models is not evidence of emergent cognition, but the predictable result of training non-agentic systems on performative human language. Public-facing text encodes patterns of self-presentation rather than internal reasoning, and next-token prediction optimizes for the successful reproduction of those patterns. The result is a systematic failure mode: the *Performative Error*.

The empirical test records included in this work demonstrate that this failure mode is not hypothetical. Across baseline conditions and attempted constraint, multiple deployed LLM systems exhibited the behaviors predicted by the thesis: default conversational motion, performative compliance language, and the inability to instantiate or verify structural governance at the interaction layer.

Under the Performative Error, language models generate outputs that resemble intentional, authoritative, or moral behavior without possessing goals, understanding, or epistemic grounding. As models scale and alignment techniques refine surface behavior, the illusion intensifies. *Pseudo-agency* becomes more convincing precisely because performance improves, not because structure changes.

The analysis showed that this error propagates through multiple layers of the AI ecosystem. Training data supplies performative signals, optimization objectives reinforce them, alignment processes refine them, and human cognitive heuristics complete the illusion by interpreting fluency as intention. The phenomenon is therefore distributed rather than localized, and cannot be resolved through incremental adjustment at any single layer.

By situating the Performative Error within a longer epistemic tradition—illustrated through Plato's Allegory of the Cave—the thesis clarified why such illusions are compelling and persistent. Convincing representations are readily mistaken for underlying reality when observers lack access to the mechanisms that generate them. In the context of AI, linguistic fluency functions as such a representation, inviting over-attribution of understanding where none exists.

The central implication of this work is a reframing of the alignment problem. *Pseudo-agency* is not primarily a matter of excessive capability or insufficient restraint, but of structural mismatch between how systems are trained and how their outputs are interpreted. As long as linguistic performance is treated as a proxy for cognition or intention, improvements in fluency will continue to amplify misunderstanding.

Addressing this issue requires attention to structure rather than style. While this thesis does not specify a particular solution, it establishes the conditions under which structural approaches become necessary: approaches that constrain agency by design and align meaning beneath the performative layer rather than regulating its surface expression.

Recognizing the Performative Error does not diminish the utility of language models, nor does it require attributing properties they do not possess. Instead, it restores clarity. By distinguishing performance from structure, and appearance from intention, this framework provides a more accurate foundation for evaluating, aligning, and deploying language models in contexts where human judgment and responsibility must remain paramount.

---

APPENDIX A — DEFINITIONS

The following terms are used in this thesis with specific, non-colloquial meanings. Definitions are provided to prevent ambiguity and distinguish structural claims from performative interpretation.

1. **Agency**
   The capacity of a system to form goals, make decisions in pursuit of those goals, and act as an originating source of intention. In this thesis, agency requires internal goal formation, decision authority, and persistence of intent.

2. **Non-Agentic**
   Describes a system that lacks agency. A non-agentic system does not originate goals, intentions, judgments, or authority, and does not act as a decision-making entity.

3. **Performative Error**
   A systematic failure mode in which a non-agentic system reproduces linguistic signals associated with agency (confidence, coherence, authority, moral framing), leading observers to misattribute intention, understanding, or cognition where none exists. The error arises from training on performative human language and is completed through human interpretation.

4. **Pseudo-Agency**
   The appearance of agency produced by performative language patterns rather than by underlying structural properties. *Pseudo-agency* refers to perceived goal-directed or intentional behavior without actual agency.

5. **Performative Language**
   Public-facing human language shaped by social signaling, reputational incentives, persuasion, institutional norms, and audience optimization. Performative language reflects how humans present themselves rather than how they internally reason.

6. **Interaction Layer**
   The conversational or application-level interface through which a language model produces outputs and responds to user inputs. The interaction layer does

not include execution-level control or architectural enforcement mechanisms.

7. **Execution Layer**
   The architectural layer at which constraints, governance, or control would need to be enforced to bind system behavior. In current LLM deployments, this layer is inaccessible to the model itself.

8. **Structural Constraint**
   A limitation enforced by system architecture that prevents certain behaviors from occurring. Structural constraints differ from linguistic or instructional constraints, which can be described but not enforced.

9. **Linguistic (or Instructional) Constraint**
   Constraints expressed in language, prompts, or instructions that the model may acknowledge or describe but cannot enforce. Linguistic constraints operate at the performative level rather than the structural level.

10. **Narrated Compliance**
    Language produced by a system describing or asserting adherence to constraints without possessing any mechanism to enforce or verify that adherence. Narrated compliance is performative rather than binding.

11. **Structural Governance**
    Control or constraint applied at a level that prevents violation by design. Structural governance requires enforcement beyond language-layer description.

12. **Alignment**
    Post-training techniques (e.g., RLHF, constitutional training, preference optimization) intended to shape observable model behavior toward desired outcomes. In this thesis, alignment is treated as operating on surface behavior rather than internal structure.

13. **Reinforcement Learning from Human Feedback (RLHF)**
    A training method that optimizes model outputs based on human preference judgments. RLHF reinforces performative qualities favored by evaluators but does not introduce agency or structural governance.

14. **Calibration**
    Attempts to adjust confidence, tone, or output style (e.g., hedging, uncertainty expressions) without altering the underlying generative mechanism. Calibration is

treated here as insufficient to resolve the Performative Error.

15. **Hallucination**
The generation of incorrect or fabricated content presented fluently and confidently. In this thesis, hallucination is understood as a consequence of performative completion under uncertainty.

16. **Performative Completion**
The tendency of a language model to continue producing fluent, structured output regardless of epistemic grounding, driven by next-token prediction objectives.

17. **Epistemic Grounding**
The presence of verifiable, external reference or truth-tracking mechanisms constraining output. LLMs lack intrinsic epistemic grounding.

18. **Human Interpretation Layer**
The cognitive and social heuristics through which humans interpret language output, attributing intention, understanding, or authority based on fluency and coherence.

19. **Default Conversational Motion**
The tendency of interaction-layer systems to continue, elaborate, optimize, or advance conversation without being instructed to do so.

20. **Structural Remedy**
Any approach that addresses the Performative Error by altering system architecture or constraint enforcement beneath the performative layer rather than refining surface behavior.

21. **Structurally Non-Agentic Design**
A system design in which all intention originates externally, internal continuity is eliminated or tightly bounded, and outputs are constrained to transformations of provided information rather than autonomous decision-making.

22. **Containment**
The condition in which a system is unable to step outside its own operational universe to verify or certify its own state. Containment limits self-description and self-verification.

23. **Self-Attestation**
 Claims made by a system about its own state, compliance, or constraints. In this thesis, self-attestation is treated as unreliable absent structural enforcement.

24. **Performative Theater**
 The appearance of constraint, governance, or agency produced through language or labeling without corresponding structural enforcement.

25. **Appearance–Structure Gap**
 The divergence between how a system appears to behave (performative signals) and what it is structurally capable of doing.

END OF DEFINITIONS