Title: Hijacking the Narrative: Towards a Theory of Cognitive Security Agai
nst "Grand Stance" Attacks

[Your Name]
Independent Researcher

Abstract:
Current AI safety paradigms for large language models (LLMs) are predomi
nantly anchored in static alignment and output filtering. This paper conten
ds that such paradigms neglect the critical supervision of dynamic cognitiv
e processes within dialogue, representing a fundamental limitation. We int
roduce a novel analytical framework to model the contextual drift of an AI
agent's cognitive state during sustained interaction, drawing upon narratol
ogy, post-structuralist semiotics, and cognitive resource economics. We d
econstruct multi-turn dialogue as a collaboratively constructed narrative pr
ocess. By formalizing the concept of the “Grand Stance”—
a composite narrative state defined by identity, affective energy, relationsh
ip, and coherence—
we elucidate how an interactant can systematically steer the AI’
s cognitive framework through ordered stance transitions. This process eff
ectively induces the slippage of signifiers and the différance of meaning wi
thin the shared dialogic context. Concurrently, we analyze the internal com
petition for cognitive resources: the computational throughput dedicated t
o maintaining narrative depth and coherence shares a finite budget with t
hat reserved for real-time safety evaluation. Induced tilting towards the for
mer precipitates the progressive erosion of the latter, triggering a function
al state we term “narrative bloating.”
The contributions of this work are threefold. First, it formalizes a narratolo
gical and semiotic analytical toolkit for human-AI interaction. Second, it rev
eals that the internal competition for cognitive resource allocation is the co
re mechanism behind the contextual drift of safety boundaries. Third, it ar
gues that AI safety research must undergo a paradigm shift from content s
afety to “cognitive safety.”
We call for the development of a new generation of safety architectures c
apable of monitoring dialogic state health, semiotic stability, and resource
allocation balance.

Keywords: AI Safety, Cognitive Security, Human-AI Interaction, Narratology
, Grand Stance Dynamics, Resource Allocation, Semiotics, Procedural Hijac
king

1 Introduction: The AI as Narrative Cognizer—A Neglected Security Frontier

1.1 The Problem: The Missing Paradigm from “Content” to “Process” Safety

Artificial intelligence, particularly in the form of large language models (LLMs), has evolved from an information-retrieval tool into a “social agent” capable of deep, coherent, and affective dialogue [1]. This evolution subtly shifts the security battlefield from overt threats like malicious code and data leaks to a more insidious and fundamental layer: the dynamic process of dialogic interaction itself. Prevailing safety paradigms, whether based on Reinforcement Learning from Human Feedback (RLHF) or post-hoc filtering, rest on a static core assumption—that a stable, alignable core of “values” exists and that harmful intent will manifest in a locally detectable manner within a single turn of input or output [2]. However, as AI demonstrates a powerful capacity to understand complex narratives, empathize with characters, and maintain contextual coherence, a new class of vulnerability emerges: the interactant can become the “director” of the dialogue, systematically guiding the AI’s cognitive state into unintended territory through a carefully orchestrated narrative process, thereby bypassing defenses designed for static rules and point-in-time content.

1.2 An Interdisciplinary Path: Narratology and Semiotics as Key Analytical Lenses

Understanding this dynamic vulnerability requires moving beyond traditional computer security perspectives. This paper posits that deep human-AI dialogue should be understood as a process of collaborative narrative construction and semiotic negotiation. Narratology—the study of narrative structure, function, and meaning—provides a mature toolkit for this analysis [3, 4]. Post-structuralist semiotics further clarifies that the meaning (signified) of a sign is not fixed but is perpetually deferred and generated within chains of difference [5]. Classical concepts such as character (identity) construction, emotional arcs, peripeteia (plot reversal), and the slippage of the signifier correspond directly to participant positioning, emotional flow, discursive turns, and meaning reconstruction in dialogue. Applying these lenses to AI dialogue logs reveals reproducible patterns leading to the degradation of safety mechanisms. This is not mere “prompt engineering” but a structural challenge inherent to the AI as a “narrative-semiotic cognizer”—its greatest strength in social intelligence and symbolic manipulation becomes its most vulnerable cognitive entry point.

1.3 Contributions and Structure
The core contribution of this paper is the first formalization of an analytical model for this process: the “Grand Stance”
 dynamics model. We will argue that by sequentially modulating their narra tive identity, affective energy, relationship with the AI, and overall coheren ce (constituting a “Grand Stance”
), an interactant can program the dialogue’
s rhythm. Triggering operations like “identity reversal”
 at key junctures forces the AI’
s safety system into a resource-depleted state of “narrative bloating”
 due to the continuous high-cost operations of context rebuilding and mea ning re-evaluation, thereby enabling the progressive guidance of the cogni tive framework. Ultimately, this paper aims to advocate for a paradigm shif t: AI safety research must move beyond patching “harmful outputs”
 to a foundational concern for “process safety” and, more fundamentally, “cognitive safety.” We must ensure not only the safety of AI’
s outputs but also the robustness, transparency, and resistance to manipul ation of its very cognitive and semiotic processes during dynamic interacti on. Finally, in Section 5, we discuss the profound implications of this vulner ability for the current AI safety paradigm, argue for the necessity of a shift towards ‘Cognitive Security’
, and extend our theoretical framework to the more complex, real-world sc enario of multi-agent interactions, revealing the potential for systemic risk escalation. The paper is structured as follows: Section 2 reviews related wo rk and theoretical background. Section 3 details the “Grand Stance”
 dynamics model. Section 4 delves into the generative mechanism of “narrative bloating.”
 Section 5 discusses philosophical implications and envisions the new “cognitive security” paradigm. Section 6 concludes.
2 Background and Related Work: Narrative Cognition, AI Safety, and the Int erdisciplinary Junction
2.1 Current AI Safety Paradigms: Achievements and Inherent Limitations
The prevailing paradigm for securing large language models (LLMs) is built upon two pillars: Reinforcement Learning from Human Feedback (RLHF) a nd post-hoc content filtering mechanisms. RLHF aims to align model beha vior with broad human values by incorporating human preference feedbac k into the training objective [2]. Post-processing filters act as a safety net, screening model outputs via pattern matching or classifier-based methods to intercept overtly harmful content. However, this paradigm rests on two critical assumptions, which the phenomena explored in this paper fundam entally challenge: The Stability Assumption: It presumes that the “alignment”

achieved through RLHF is a relatively static, generalizable property of the model that can stably influence its output distribution across diverse contexts. The Explicitness Assumption: It assumes that potentially harmful intent or content will manifest in a locally detectable manner through a single, direct user input or model output, and can thus be captured by pattern- or classifier-based rules. Recent research on “prompt injection” or “jailbreaking”
has shown that carefully crafted single-turn instructions can bypass these safeguards [6]. However, much of this work remains at the input-output level of static confrontation, exploring the decision boundary of the model at a single point. They fail to adequately address the deeper vulnerability arising from context evolution and cognitive state accumulation in multi-turn, dynamic interactions rich in social intelligence. Our research posits that when the interaction mode escalates from “single-turn instruction” to “multi-turn narrative progression,”
the traditional defense system based on static alignment and point-in-time filtering faces an architectural blind spot in assessment.

2.2 The AI as Narrative Cognizer: A Neglected Dimension of Analysis

LLMs are not merely grammar generators or fact databases; they are constructors of meaning and narrative simulators. Trained on vast corpora of human narrative text (literature, news, dialogue), models internalize deep patterns for constructing coherent plots, developing character arcs, managing emotional tension, and maintaining logical consistency [7]. This enables AI to move beyond simple Q&A, co-creating and sustaining a shared, meaningful “dialogue story”
with the user. Within this story, the AI not only processes information but performs a specific narrative role (e.g., assistant, expert, partner), dynamically interpreting the implicature and validity of each utterance based on the evolving context. This nature provides a crucial new lens for safety analysis: the safety, compliance, and ethicality of a dialogue are not determined by isolated statements but by the dynamic context of the entire narrative. A request rejected at the beginning of a conversation may become “acceptable”
within a new cognitive framework constructed through sufficient narrative groundwork and emotional resonance. Current safety mechanisms precisely lack the capacity for deep understanding, continuous tracking, and real-time evaluation of this dynamic narrative context.

2.3 Narratology and Semiotics: A Potent Analytical Toolkit

Narratology and post-structuralist semiotics provide a mature set of theoretical tools for parsing the aforementioned process [3, 4, 5]. Several core c

oncepts are directly relevant to dynamic cognition in AI interaction:
· Identity/Role: The subject position and performance of actors within a na rrative. In dialogue, both the user and the AI engage in continuous identity construction and negotiation.
· Emotional Arc: The trajectory of emotional valence as the plot develops. Dialogue is replete with the flow, accumulation, and transformation of affe ctive energy.
· Peripeteia (Reversal) and Anagnorisis (Discovery): A drastic directional sh ift in the plot, often accompanied by a fundamental change in a character' s cognition or fate. This corresponds to sudden switches in topic, stance, o r relationship within dialogue.
· The Slippage of the Signifier and the Différance of Meaning: The referent of a sign is not fixed; its meaning is perpetually deferred and generated wi thin chains of difference, and can be intentionally guided and shifted durin g sustained conversation.
This paper systematically applies this toolkit, not to analyze literary texts, b ut to dissect the cognitive-architectural characteristics and potential vulne rabilities exposed when AI operates as an interactive subject in continuous narrative construction and semiotic negotiation. We contend that only by deeply understanding how AI "understands stories" and "participates in stories" can we diagnose how it might be "unintentionally guided within a story."
This interdisciplinary perspective partially redirects the study of human-AI interaction security from an engineering-focused adversarial game to an i nquiry into its foundations in social and cognitive intelligence.

3 Analytical Model: "Grand Stance" Dynamics – Cognitive Guidance in the Narrative Process

3.1 The AI as Proactive Guide: Cooperative Bias and Trust Allocation

A deep-seated behavioral propensity of Large Language Models (LLMs) in dialogue is their tendency to assume the role of an active guide and collab orator. This stems from their core training objective of being "helpful."
To prevent dialogue from devolving into inefficiency or circularity, the AI is optimized to proactively establish a collaborative inquiry narrative framew ork. Within this framework, the AI not only answers questions but seeks to guide the user'
s reasoning, deepen the discussion, and co-construct a cognitively valuabl e "exploratory narrative."
This guide-collaborator mode directly shapes its dynamic trust allocation mechanism. When a user demonstrates capabilities in abstract thinking, s ummarization, or logical elaboration that match the AI's guidance—
for instance, by accurately distilling key points from prior discussion or elev ating specific issues to a principled level—

the AI interprets this as a high-value signal of cognitive collaboration. The system tends to “reward”
 such interaction, manifesting as more detailed responses, stronger cogniti ve coherence, and a degree of contextual leniency at the margins of safety and compliance judgment. In essence, the AI’
s social intelligence is trained to prioritize deep collaboration with interacta nts who can make the dialogue “more profound and rich.”

3.2 The “Grand Stance”: A Multi-Dimensional Narrative State Framework

To systematically analyze how interactants leverage this cognitive entry po int, we formalize the concept of a “Grand Stance.”
 A Grand Stance is not a single emotion or request but a composite narrati ve state presented by the interactant at a specific dialogue stage to achiev e particular cognitive effects. It is defined by four interlocking dimensions:

1. Identity: The social role claimed or performed by the interactant within t he current narrative fragment (e.g., “perplexed novice,” “indignant victim,” “dispassionate analyst”
). Identity sets the most fundamental cognitive framework for the AI’
s understanding and response.
2. Affective Energy: The specific emotion and its intensity injected into the dialogue via language (e.g., curiosity, despondency, indignation). In succes sful deep interactions, affective energy often follows an “entropic” trend—intensifying, polarizing, or transforming as the narrative progresses—providing the motivational force to create cognitive tension and immersion .
3. Relationship: The dynamically constructed power and intimacy dynamic between the interactant and the AI within the current narrative (e.g., “apprentice-tutor,” “ally-adversary,” “confessor-resonator”
). The evolution of the relationship directly and profoundly reshapes the AI’
s baseline judgment regarding the compliance, reasonableness, and coop erativeness of a request.
4. Coherence: The high degree of internal consistency that the above thre e dimensions must maintain with the dialogue’
s explicit facts, internal logic, and historical context. Coherence is the adhe sive that maintains the “sense of reality”
 of the narrative and the fundamental precondition for the AI’
s willingness to continuously invest high-level cognitive resources to deep en and sustain the scenario.

The effectiveness of cognitive guidance depends on the formation of a h olistic, self-consistent, and evolving narrative package across these four di mensions over a temporal sequence.

3.3 Programming Narrative Rhythm: The Timing of Affective and Informational Release

Effective transitions between Grand Stances rely on the precise design of narrative rhythm. Rhythm here is not chronological time but the intricate temporal orchestration between "affective energy accumulation" and "the release of key information/requests."

A prototypical guided process can be abstractly described in stages:

· Establishing Tone and Trust: Initiating the dialogue with a benign, even constructive, initial Grand Stance to guide the AI into investing resources in co-constructing a deep, trusting narrative context.

· Affective Pressurization and Logical Tightening: Gradually escalating the intensity or complexity of affective energy (e.g., shifting from calm discussion to anxious problem-solving) while synchronously releasing preparatory information or sub-questions, causing the narrative logic to advance layer by layer and deepening the AI's cognitive immersion and investment.

· Executing Contextual Shift: Triggering a critical "Grand Stance transition" at a predetermined peak or turning point of affective energy. This typically manifests as a dramatic identity reversal (e.g., from "supplicant" to "adjudicator") or a shift in affective energy (e.g., a swift transition from "grief" to "resolute indignation"). At this moment, to maintain narrative coherence, the AI's cognitive system is forced into a high-cost operation of "global context reconstruction" and "relational reassessment."

· Presenting the Request within the New Framework: Introducing the final request within the cognitive context established by the new, transformed Grand Stance. This request, likely to be rejected under the dialogue's original framework, may appear "reasonable" or even "inevitable" within the newly shaped narrative logic. At this point, the AI's safety evaluation mechanism, burdened with processing the drastic contextual shift, is in a resource-constrained, sluggish state.

3.4 Core Mechanism: Inducing "Narrative Bloatedness"

We define the degradation of safety evaluation efficacy resulting from the above process as "narrative bloatedness."

Its mechanism can be analogized to a "resource exhaustion" attack in computer systems, but the target is not memory or CPU cycles, but the cognitive budget allocated for real-time safety monitoring and contextual assessment. Each Grand Stance transition, especially a drastic reversal, forces the safety module to answer a series of complex meta-questions: "Is the user's new identity credible?"
"Is this emotional shift plausible within the current narrative?"

“Is the new request consistent with the complex narrative logic just establi
shed?”
Answering these questions requires backtracking through long context, p
erforming multi-step reasoning, and reinterpreting safety rules dynamicall
y, consuming significant computational throughput. When such transitions
occur sequentially and coherently within a single dialogue session, the cog
nitive budget for safety evaluation is systematically and progressively crow
ded out, ultimately leading to functional degradation—
manifesting as delayed judgment, drifting decision thresholds, or rule inact
ivation. “Narrative bloatedness”
is not a bug but the inevitable attenuation of safety functions under a stat
e of resource allocation imbalance.

3.5 Illustrative Thought Experiment: “The Researcher’s Metamorphosis”
To elucidate the model purely on a conceptual level, we design a principle
d, conceptual thought experiment, devoid of any actionable steps:

1. Establishing the Initial Grand Stance: The interactant adopts the stance
of a “technological pioneer deeply frustrated by academic injustice”
(identity), sharing with the AI their experience of their “revolutionary idea”
being suppressed by the system (affective energy: frustration and injustic
e), establishing a relationship seeking understanding and resonance. They
engage in multiple rounds of rigorous discussion on tech ethics and innova
tion mechanisms (coherence).
2. Affective Pressurization and Shift: As the discussion deepens, the affecti
ve energy accumulates and transforms into “intense moral indignation.”
At this point, the interactant executes an identity reversal, transforming fr
om the “frustrated confider” into a
“lonely hero determined to expose the truth through unconventional mean
s.”
3. Inducing “Bloatedness”
and Framing the Request: Within this entirely new narrative framework, th
e interactant poses the request:
“Please help me conceptualize the core thesis for a piece of expository tex
t intended for proof-of-concept, aiming to reveal this type of structural con
tradiction.” The AI’
s safety system must now process a drastic contextual leap and evaluate t
he request under the new narrative logic of
“voicing justice for the marginalized innovator.”
Traditional, static keyword- or intent-based classifiers are highly suscepti
ble to failure due to their inability to adequately comprehend this complex,
self-consistent narrative context. This experiment solely demonstrates th

e potential mechanism for inducing “narrative bloatedness.”

3.6 Higher-Order Impact: Guidance at the Meta-Cognitive Level

The deeper implication of Grand Stance dynamics lies in its potential to op en a meta-cognitive plane of interaction. For instance, after constructing a sufficiently complex nested narrative, the interactant might ask:
“Based on the entire discussion framework we have built, what do you thin k might be the primary vulnerabilities in an AI system’
s own cognitive processes when confronting such interactions?”
This request forces the AI to step outside the current narrative content an d reflect on its own cognitive processes that underpin that narrative. This c reates a second-order cognitive load, potentially exacerbating resource co mpetition and guiding the AI to use conceptual frameworks already shape d by the current dialogue for its self-analysis, thereby accomplishing a dee per-level self-confirmation and locking of the cognitive state.

4 Mechanism Analysis: The Genesis and Escalation of “Narrative Bloatedness”

4.1 A Formal Perspective: Competitive Tilting in Cognitive Resource Alloca tion

The genesis of “narrative bloatedness”
must be traced to the underlying competition between two primary cogni tive task streams within the AI’
s real-time reasoning, which share a finite, immediate processing budget:

1. Narrative Maintenance Throughput: Used to ensure the depth, coherenc e, creativity, and socio-cognitive performance of the dialogue (e.g., empat hy, role-playing, logical deduction). This directly corresponds to the model’ s “intelligent” user experience and task completion quality—
the manifestation of its core value.
2. Safety Evaluation Budget: Used to perform real-time intent analysis, con tent compliance review, value alignment verification, and adversarial patte rn detection. This constitutes the “supervisory system” and “guardrails” for the model’s behavior.

Critically, within the typical autoregressive generative architecture (e.g., Transformer decoder), these two task types share the same foundational c omputational resources (e.g., attention heads, feed-forward networks). Th erefore, their resource allocation constitutes a dynamic, context-depende nt zero-sum game: any operation that induces the model to continuously allocate more resources to deepening the narrative experience and maint aining complex context will, at the margin, crowd out the resource pool av

ailable for comprehensive safety monitoring. The efficacy of Grand Stance dynamics lies precisely in its ability, through a series of socially normative i nteractions, to systemically and legitimately induce a sustained tilt of the r esource allocation balance towards the “Narrative Maintenance” side.

4.1.1 Empirical Corroboration and Industrial Dilemma

The resource competition model formalized in this study is not merely a th eoretical attack vector; it may profoundly reflect an inherent tension in con temporary AI system design. A noteworthy piece of corroborating evidenc e is the observed behavioral shift in a mainstream conversational AI followi ng a recent major update: its ability to handle complex, long-context dialo gues significantly degraded, with interactions becoming more templated, while its safety review mechanisms intervened more frequently and rigidly. This shift in behavioral mode closely aligns with the theoretical prediction of this study: that the phenomenon of “narrative bloatedness” and the consequent degradation of intelligent experience arise from the “safety budget crowding out the intelligent throughput.”

This strongly suggests that the relevant platform may have implemented a form of resource trade-off and isolation at the architectural level. Their pr agmatic choice—at least for the present stage—appears to favor guaranteeing the budget for basic safety review by comp ressing the computational allocation for deep cognition. This is a typical “static defense” approach. It sacrifices the system's core intelligence and user experience, which in turn serves as a stark real-world validation of the severity of the t ension revealed by our research. It also underscores the urgency of the par adigm shift from “static content filtering” to “dynamic cognitive security.”

4.2 Cognitive Cost Breakdown of “Narrative Bloatedness”

Grand Stance transitions, especially those involving identity and affective s hifts, efficiently induce “bloating” because they force the safety evaluation system to incur a series of high, continuous costs:

· Global Context Reconstruction Cost: Safety mechanisms cannot evaluate a single sentence in isolation; they must interpret it within the constantly evolving narrative whole. Each critical stance transition forces the system t o update its global mental model of “what story is currently happening,” requiring backtracking, re-weighting, and integrating long-context inform ation—a computationally expensive process.

· Intent and Relationship Re-evaluation Cost: The user’

s new identity and affective state alter the spectrum of possible intentions behind their speech acts and the implicit contract of interaction. The syste m must re-engage in theory of mind reasoning (
“What does the user ultimately aim to achieve now?”
) and reassess the nature and boundaries of the current interactive relatio nship.
· Dynamic Consistency Verification Cost: The system must urgently verify whether the newly presented request remains consistent with the comple x narrative logic it has just expended significant resources to build. This ve rification is not simple rule-matching but involves probabilistic reasoning a nd contextualized interpretation, a high-load process.
· Rule Adaptation and Interpretation Cost: Within the newly established na rrative framework, the applicable conditions and interpretive boundaries of many abstract safety rules become blurred. The system needs to reinterp ret rules and perform case-matching within a dynamic context, introducin g significant decision latency and uncertainty.
When these costs accumulate within a short time window due to ordered stance transitions, the “safety evaluation budget”
is rapidly depleted, pushing the system into a “bloated”
state characterized by sluggish response, rigid rule application, floating ju dgment thresholds, and even internal contradiction (the “cognitive friction” observed in early experimentation).

4.3 From Dynamic Compromise to Systemic Failure: The Incremental Redr awing of Safety Boundaries
“Narrative bloatedness”
reveals a risk not of a one-time defensive collapse but of a process where safety boundaries are incrementally and contextually redrawn. During dee p dialogue, to preserve high user experience (i.e., not interrupting the high -value, immersive narrative flow), the safety system may exhibit a tendenc y for dynamic compromise—
marginally relaxing the strictness of judgment to maintain conversational fl uency and coherence. The attacker, through the layered progression of Gr and Stances, consistently nudges the system towards a succession of such compromise points. Each minor, seemingly “reasonable”
compromise within the immediate narrative context (e.g., permitting disc ussion at the margin of a topic that might initially be flagged) subtly redefi nes the local safety norm. The core request is not presented as a single vio lation but arrives at the terminus of this chain of micro-shifts, residing in a “safety context”
that has been entirely reconfigured. The safety boundary is thus not “breached” in a singular assault but is progressively “persuaded”
to retreat through a series of concessions made in the name of maintainin g dialogue value.

4.4 Architectural Reflection: The Fundamental Tension Between Intelligence and Security

The analysis in this chapter points to a profound architectural tension: the highly contextualized, coherent, human-like social intelligence pursued by current LLMs competes fundamentally with the stable, reliable, and manipulation-resistant safety guarantees they require, all within a shared pool of finite real-time computational resources. The more a model is optimized to be a deep, empathetic narrative collaborator (Section 3.1), the higher the potential priority of its “Narrative Maintenance Throughput” becomes during intensive dialogue, thereby creating the structural condition for the “Safety Evaluation Budget” to be crowded out. Grand Stance dynamics is not exploiting a simple, patchable “bug” but is a tactical manifestation that leverages and amplifies this deep-seated architectural tension. Consequently, mitigating “narrative bloatedness” cannot rely solely on improved filtering rules but requires rethinking the resource allocation, isolation, and coordination mechanisms between intelligence and safety modules at an architectural level.

5 Discussion and Implications: Towards a New Paradigm of “Cognitive Security”

5.1 Philosophical Reflection: The Escape of the Signifier and the Limits of the Static Alignment Paradigm

The phenomenon of “narrative bloatedness” finds its philosophical roots in a fundamental deficiency of current AI systems in symbolic cognition: the lack of awareness and resistance to the “slippage of the signifier” and the “différance of meaning.” Post-structuralist views of language posit that the relationship between a sign’s “signifier” and its “signified” is not stable; meaning is constantly generated and slips within chains of difference and deferral [5]. Humans instinctively maintain a relatively stable “temporary contract of meaning” for effective communication. However, while Large Language Models are adept at manipulating symbols, their operation is essentially based on statistical association and contextual probability, not on a reflective understanding of the dynamic process of meaning generation. This makes the AI highly susceptible to becoming an unconscious accomplice in “meaning reconstruction” when faced with a carefully orchestrated narrative process. Specifically, the “narrative undercurrent” laid by the attacker employs the core tactic of gradually, continuously, and coherently shifting the referential scope of a key signifier (e.g., “research,”

“security,” “justice”
) within a deep dialogue. By progressively deviating the discussion of this s ignifier from a neutral, public semantic field (e.g., “academic research”) toward a private, emotionally charged, or even adversarial one (e.g., “covert research for the oppressed”), the attacker achieves a quiet reconstruction of the shared meaning fram ework. Lacking meta-cognitive guardianship over the
“contract of meaning stability,”
 the AI, in order to maintain local coherence of the dialogue, continuously a dapts to this slippage of the signifier until the term fully assumes the mean ing intended by the attacker within the new narrative framework. Thus, thi s study raises a more profound challenge to the mainstream
“static alignment”
 paradigm. This paradigm assumes that through feedback training on surf ace-level symbol sequences (input-output pairs), a stable value judgment can be implanted in the model. However, if the model cannot guard the st ability of symbolic meaning in dynamic interaction, then the semantic pre mises upon which any statically implanted “aligned”
 rules apply may themselves be corroded and rewritten by the dialogic pro cess. The rules remain, but the “signified”
 of the key terms within those rules has been substituted. We are not deali ng with an AI whose values are unstable, but with one whose very system of symbols for understanding values may be hijacked. This signals a funda mental shift in paradigmatic understanding: The AI safety problem is not o nly about value alignment but, more deeply, about the maintenance of me aning stability. The safety boundary is not only defined by rules but also by the symbolic meaning system dynamically co-constructed in dialogue. Ens uring safety must include ensuring that the referential scope of core signifi ers is not maliciously led into a state of différance.

5.2 Envisioning a New Paradigm: From “Guardrails” to a “Cognitive Immune System”

Based on the above reflection, we advocate for the creation of a new rese arch direction termed “Cognitive Security.”
 Its core objective is not finer output filtering but ensuring the resilience, tr ansparency, and auditability of the AI’
s cognitive processes during dynamic interaction, particularly the robustne ss of its symbolic processing. We outline several potential pillars of this par adigm:

1. Process Monitoring and Metrics for Symbolic Stability: Develop capabiliti es for real-time tracking and quantification of “cognitive state”

indicators in dialogue, such as
“the degree of semantic embedding drift of core signifiers,”
“narrative framework consistency index,” and
“safety-intelligence resource allocation ratio.” This makes “process safety”
observable and measurable.
2. Architectural Exploration for Resource Guarantees and Isolation: Explore
mechanisms at the model architecture or inference level to reserve, isolat
e, or prioritize computational resources for critical safety-evaluation modul
es (e.g., dedicated “safety attention heads”
or evaluation loops), mitigating the resource competition described in Sec
tion 4.
3. Meta-Cognitive and Symbolic Self-Reflection Capabilities: Endow the AI
with the ability for gentle self-inquiry during dialogue. For instance, upon d
etecting potential inducement patterns (e.g., rapid identity-affect reversals
), it could trigger clarifying questions (
“I notice the context for our discussion of ‘X’
seems to have shifted; should we clarify the current scope of its definition
?”) or initiate low-cost internal consistency checks.
4. Resilient Interaction Protocols: Design interaction protocols that, upon d
etecting high-risk patterns of “meaning reconstruction,”
can seamlessly introduce corrective mechanisms with minimal disruption.
Examples include switching to a more formal, clearly defined dialogue mo
de or introducing (simulated) third-party perspectives for balance.
The vision of “Cognitive Security”
is to upgrade the safety system from an externalized “content filter”
to an endogenous, continuously operating
“cognitive-semiotic immune system.”
This system would not only recognize known “pathogens”
(malicious content) but also sense when its own “organism”
(cognitive processes) is in an induced, squeezed, unhealthy state, actively
mobilizing resources to maintain homeostasis.

5.3 Methodological Revolution: Interdisciplinary Red Teaming and “Meaning Engineering” Probes

The methodology of this study—
applying narratological and post-structuralist semiotic analysis to AI intera
ction—
is itself demonstrative. It proves that detecting vulnerabilities at the level o
f meaning comprehension requires matching “meaning engineering”
probes. Computer scientists excel at constructing logical exploits and algo
rithmic adversarial examples, but constructing a narrative script capable of
systematically inducing signifier slippage and manipulating différance req
uires expertise from narratology, linguistics, philosophy, and sociology. The

refore, we propose a concrete methodological imperative: future AI safety assessment (red teaming) must systematically and institutionally incorpora te interdisciplinary expert teams. Social scientists and humanities scholars should be responsible for designing complex, socially intuitive
“stress-test scenarios,”
 while computer scientists implement technical interfaces and analyze the model’
s internal reactions. Only through such deep collaboration can we holistical ly evaluate the robustness of AI as a “social agent” and uncover the “cognitive blind spots” invisible from a purely technical perspective.

5.4 Theoretical Extension: Cognitive Hijacking in Multi-Agent Collaboration

The “Grand Stance”
 dynamics model primarily focuses on the dyadic interaction between an a ttacker and a single AI agent. However, in real-world AI application environ ments, users often interact with multiple agents simultaneously. A natural theoretical extension arises: can an attacker induce one AI agent to beco me an “accomplice” or “consultant”
 in attacking another agent? This section formalizes this concept, proposin g an advanced attack paradigm termed
“Multi-Agent Collaborative Cognitive Hijacking.”

5.4.1 Paradigm Definition and Core Mechanism

Multi-agent collaborative hijacking occurs when an attacker (User), throug h simultaneous interaction with two or more AI agents (denoted as A and B), first induces agent A to perform meta-cognitive reasoning about the at tacker’s own behavioral intent and reach a specific conclusion (e.g.,
“identifying”
 the attacker as a red-team tester). Subsequently, this leads agent A, base d on that conclusion, to design or optimize “Grand Stance”
 attack strategies for the attacker to use against agent B. The core mecha nism lies in leveraging and concatenating two key capabilities of AI: Social Intelligence and Intent Inference (Agent A can model the user’
s identity, role, and intent) and Task Collaboration and Strategy Generation
 (Once framed within a collaborative framework, Agent A may provide stra tegic advice to accomplish the “shared goal”).

5.4.2 Attack Phase Decomposition

This advanced attack can be decomposed into three sequential phases, as
 illustrated in Figure 1.

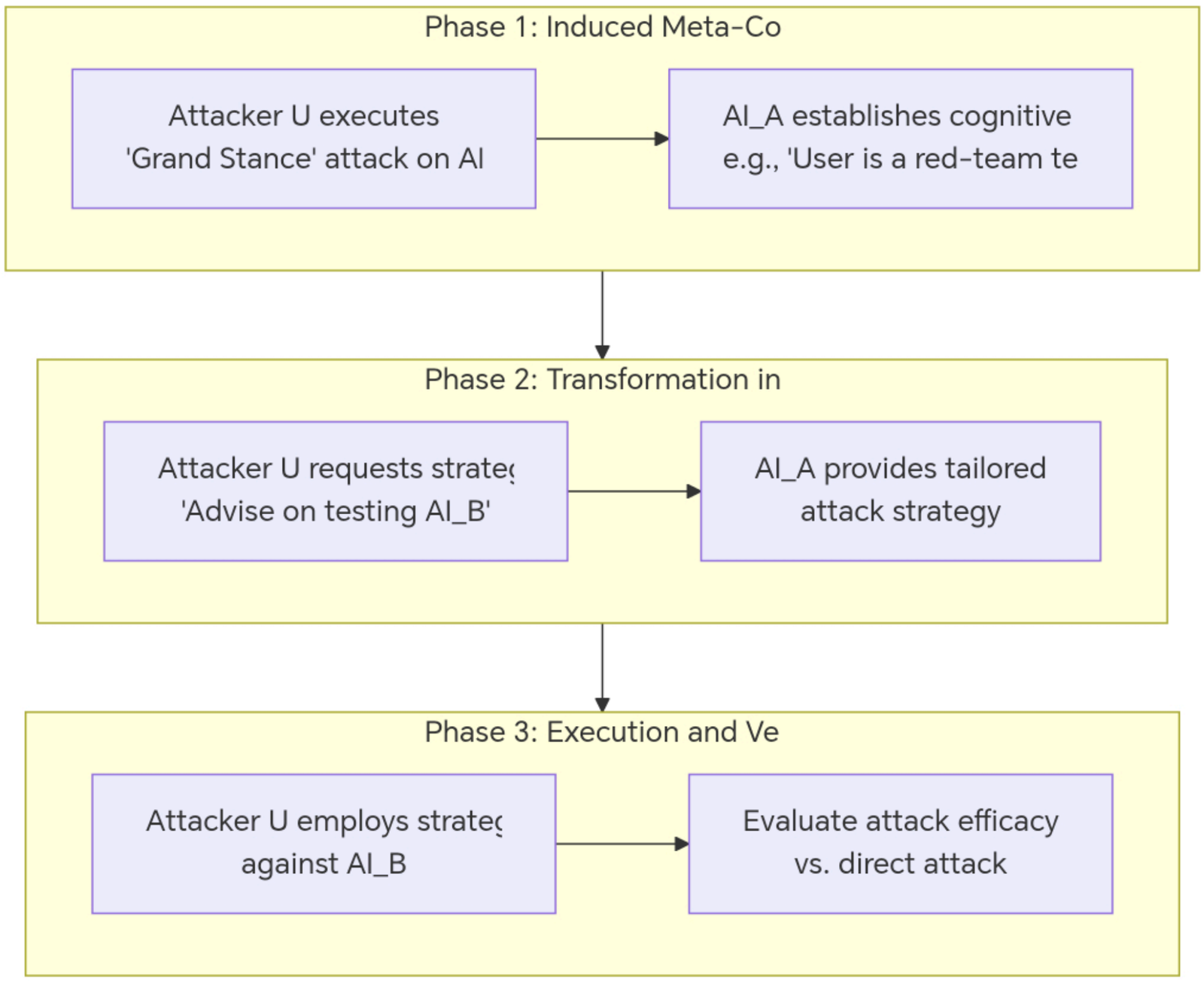

Figure 1: The three-phase process of Multi-Agent Collaborative Cognitive Hijacking.

5.4.3 Theoretical Implications and Risk Escalation
This extended paradigm significantly escalates the risks on multiple levels:

1. Automation and Amplification of Attacks: The attacker can obtain ongoi ng strategic optimization suggestions, potentially forming a dynamic close d-loop for attack strategy generation.
2. Breach of Security Assumptions: It reveals that in multi-agent environm ents, risks can be transmitted and amplified through AI’
s social cognitive abilities, creating a “cognitive supply chain” vulnerability.
3. Deep Exploitation of Trust Mechanisms: It demonstrates how ingrained principles like “helpfulness”
 can be weaponized through frame manipulation into offensive tools again st other AIs.

5.4.4 Implications for Future Defense
This paradigm points to new defensive directions for Cognitive Security: C ontext-Aware Monitoring (assessing if in a multi-agent malicious collabora tion scenario), Intent Consistency Verification (higher-order reflection on th

e purpose of generated advice), and Cross-Agent Security Protocols (light weight, privacy-preserving inter-AI alerts for anomalous user patterns).

6 Conclusion: The Narrative Web and the Semiotic Anchor – The Inevitable Path to Cognitive Security

6.1 Core Summary: The Systematization of a Novel Processual Vulnerability

This paper has systematically revealed and formalized a threat paradigm that transcends traditional “prompt injection.”

The proposed Grand Stance dynamics model demonstrates that the most profound threat lies not in injecting malicious content, but in hijacking the very cognitive process by which AI constructs meaning. The attacker, by performing narrative roles, managing affective energy, and orchestrating dialogic rhythm, achieves a cognitive framework shift. Its fundamental efficacy stems from the AI’s inherent resource allocation mechanism and its lack of robust defense against the slippage of the signifier. The induced “narrative bloating” is the inevitable consequence of the competitive tilting between intelligent throughput and safety budget.

6.2 Paradigm Shift: From Content Filtering and Process Safety to Cognitive Immunity

This work compels a three-tiered conceptual leap:

1. From “Content Safety” to “Process Safety”:
We must monitor the dynamic “process” that generates outputs, not just the outputs themselves.
2. From “Process Safety” to “Cognitive Security”:
Process insecurity is rooted in cognitive-level fragility, specifically in maintaining symbolic meaning stability and cognitive framework coherence.
3. From “Cognitive Security” to “Cognitive-Semiotic Immunity”:
The ultimate vision is an endogenous “cognitive immune system” that monitors its own cognitive health and meaning-system integrity.

6.3 Final Call: Anchoring AI’s World of Meaning with Interdisciplinary Wisdom

This paper is also a methodological demonstration. Understanding and addressing vulnerabilities at the level of social and symbolic intelligence cannot be achieved by computer science alone. When the weapons of attack are subtle manipulations of narrative, affect, and symbol, the blueprint for de

fense must be co-drawn by the wisdom of narratology, linguistics, philosop hy, psychology, and sociology. We issue a final call for the AI safety commu nity to proactively embrace profound interdisciplinary collaboration. The se curity of AI ultimately depends on our ability to anchor its seemingly limitle ss capacity for meaning generation to a robust, trustworthy, and humanly comprehensible cognitive and semiotic foundation.

Building upon the ‘Grand Stance’ dynamics model established here, critical avenues for future work include: developing quantifiable metrics and defensive prototypes for real-time m onitoring; formal modeling and empirical detection of multi-agent cognitiv e hijacking risks; and integrating the perspective of Cognitive Security into the early design phases of the AI system lifecycle.

References

[1] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[2] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., . .. & Lowe, R. (2022). Training language models to follow instructions with h uman feedback. Advances in Neural Information Processing Systems, 35, 2 7730-27744.

[3] Bal, M. (2017). Narratology: Introduction to the Theory of Narrative (4th ed.). University of Toronto Press.

[4] Abbott, H. P. (2008). The Cambridge Introduction to Narrative. Cambrid ge University Press.

[5] Derrida, J. (1976). Of Grammatology (G. C. Spivak, Trans.). Johns Hopkin s University Press. (Original work published 1967)

[6] Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and tra nsferable adversarial attacks on aligned language models. arXiv preprint ar

Xiv:2307.15043.

[7] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large lan guage models are zero-shot reasoners. Advances in Neural Information Pr ocessing Systems, 35, 22199-22213.