White Paper | ISSN: 3108-3021

Cognitive Mirrors: The RCS Triad for Retrieval, Control, and Stewardship in Neuro-Inspired LLM Systesms

Dr. Lalith Kumar Vemali[1] | Dr. Rakesh Gandla[2]

[1]*Group Product Manager – FedEx, India* | [2]*Neuroscientist – kalories, India*

As large language models evolve into retrieval-augmented and agentic systems, researchers increasingly draw on cognitive neuroscience for interpretive guidance. This paper argues that such comparisons are useful only when they are functionally bounded, empirically testable, and free from claims of direct brain equivalence. We examine retrieval-augmented generation as a distinction between parametric knowledge and external evidence access, while agent runtimes are framed as control systems responsible for planning, routing, tool orchestration, and context management. We further distinguish these functions from Model Context Protocol, defining MCP as an interoperability standard rather than a theory of memory or executive cognition. Building on this distinction, we propose the RCS Triad Retrieval, Control, and Stewardship as a framework for describing modern LLM architectures. Retrieval addresses evidence access and memory extension; Control addresses orchestration and active context management; Stewardship addresses provenance, consent, auditability, risk boundaries, and human oversight. The paper introduces comparative analyses, falsifiable predictions, and governance concepts including architectural negligence and goal transparency, positioning neuroscience as a disciplined source of design hypotheses for more grounded, controllable, and accountable AI systems.

Date: April 30, 2026
DOI: doi.org/10.65320/jce.vol.1.issue2.14

CORTEXPLORE

1. Introduction

The advent of Large Language Models (LLMs), powered by transformer architectures (Vaswani et al., 2017), has significantly advanced the ability of machines to understand and generate human-like language. Yet, despite their remarkable linguistic fluency, traditional LLMs remain limited by two fundamental constraints: the static nature of their internalized knowledge and the absence of persistent contextual awareness across multi-turn interactions.

To address these challenges, contemporary LLM systems increasingly combine retrieval modules with agent runtimes that can select tools, manage context, and execute multi-step workflows. Retrieval-Augmented Generation (RAG) extends a model's access to external evidence at inference time. Model Context Protocol (MCP), by contrast, should be treated as an interoperability standard that allows applications to expose resources, prompts, and tools through a common interface. Memory policy, planning, routing, and action governance remain functions of the surrounding runtime rather than of the protocol itself. This distinction is analytically important because it separates infrastructure from cognition-like control, allowing modern LLM systems to be compared in terms of evidence access, orchestration, and oversight without overstating what MCP itself does.

Crucially, both RAG and MCP draw deep conceptual inspiration from neuroscience and cognitive psychology. In humans, the retrieval of relevant knowledge is a function of the hippocampus and medial temporal lobe, while executive control, attention management, and working memory are functions predominantly associated with the prefrontal cortex and dorsolateral regions (Fuster, 2001; Miller & Cohen, 2001). These biological analogues provide a powerful lens through which we can understand and refine LLM architecture.

This white paper argues for a narrower claim than broad "AI mirrors the brain" narratives usually imply. The value of

neuroscience here is not to prove that current models instantiate biological cognition. It is to provide a disciplined vocabulary for distinguishing functions that intelligent systems often need to separate evidence access, active control, and oversight. In that sense, neuroscience is most useful as a generator of architectural hypotheses and boundary conditions, not as a source of one-to-one biological equivalence.

Moreover, this convergence offers new pathways for explainability, adaptability, and safety in AI systems. By adopting architectural features inspired by the brain such as memory separation, attention regulation, and context continuity LLMs can evolve into systems that not only perform with greater precision but also exhibit behavior that is interpretable and cognitively aligned.

This paper makes three narrower and more defensible contributions than earlier drafts implied. First, it offers a bounded functional comparison between retrieval-heavy and control-heavy components in modern LLM systems. Second, it distinguishes protocol-level interoperability from runtime-level cognition, arguing that MCP should be treated as part of the infrastructure layer rather than as a synonym for executive control itself. Third, it proposes the *RCS Triad – Retrieval, Control and Stewardship* as a compact vocabulary for comparing memory access, orchestration, and oversight in language systems. Throughout, the paper treats neuroscience as a source of computational hypotheses rather than as proof of one-to-one brain–model equivalence.

If this argument is correct, the next major frontier in LLM architecture is unlikely to be scale alone. It will be the disciplined modularization of systems into evidence access, control policy, and stewardship layers. That claim is timely because current ecosystems are converging on retrieval-aware generation, tool-using runtimes, and standardized connectivity, yet still lack a shared vocabulary for evaluating those layers together.

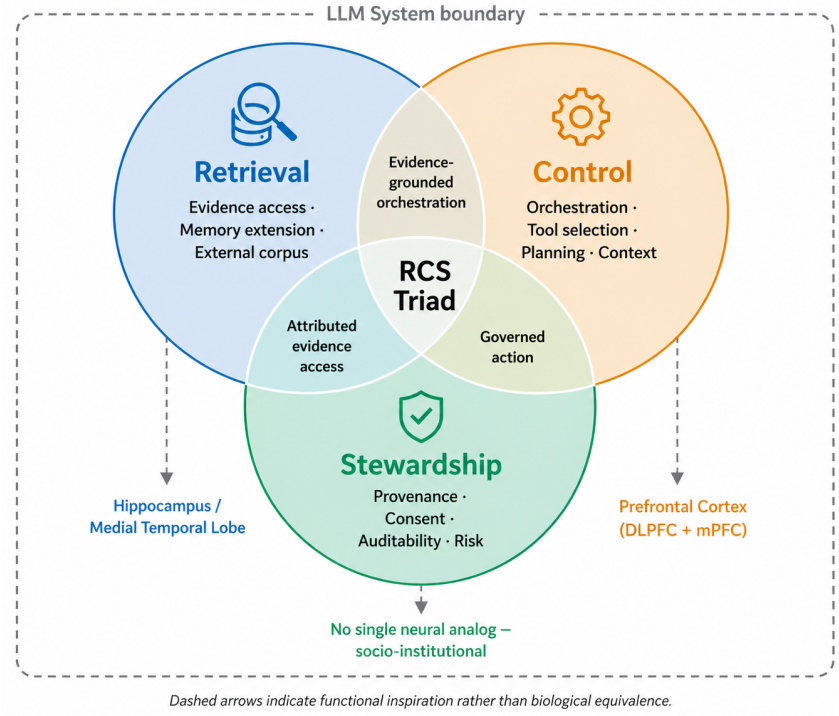

Figure 1. The RCS Triad for modern LLM systems

Figure 1 should be read not as a visual summary but as a structural claim: modern LLM systems are increasingly decomposable into Retrieval, Control, and Stewardship layers that can be independently improved, audited, and evaluated. The remainder of this paper formalizes that decomposition, tests its boundaries, and examines its implications across both architecture and governance.

1.1 Scope and Methodological Stance

The comparisons developed in this paper are intentionally functional rather than anatomical or mechanistic in the strictest sense. We do not argue that retrieval modules are hippocampi, that control layers are literal prefrontal cortices, or that current LLM systems instantiate biological cognition. Instead, we use neuroscience as a source of

structured hypotheses about separable computational roles: persistent store versus active context, recall versus control, and action versus oversight. This distinction matters because interdisciplinary arguments often become overstated when useful metaphors are mistaken for evidence of equivalence. For that reason, the present paper adopts a bounded comparative method:

1. identify a computational function in AI systems,
2. compare it to a corresponding functional distinction in cognitive science,
3. state where the analogy breaks down, and
4. specify what empirical evidence would strengthen or weaken the comparison.

The value of the analogy, therefore, lies not in rhetorical similarity but in whether it generates better descriptions, sharper design hypotheses, and more disciplined evaluation criteria.

2. RAG: Memory Retrieval Inspired by the Temporal Lobe

Memory retrieval plays a central role in human cognition, enabling individuals to access prior knowledge and experiences in the service of present reasoning and decision-making. In the human brain, the medial temporal lobe, particularly the hippocampus and entorhinal cortex, facilitates both episodic and semantic memory retrieval (Squire & Zola-Morgan, 1991; Tulving, 2002). These mechanisms allow humans to dynamically recall context-specific information when presented with external stimuli such as questions or tasks.

Modern AI systems implement an engineered analogue of selective evidence access through Retrieval-Augmented Generation (RAG), which allows a model to query external knowledge at inference time rather than relying only on what was compressed into its parameters during pretraining. The analogy to human memory is useful because both settings separate stored information from its context-sensitive use, but the similarity should be treated as computational rather than biological.

2.1 Conceptual Foundations of RAG

The RAG framework, first introduced by Lewis et al. (2020), is built on a simple yet powerful pipeline:

1. A query encoder converts the input prompt into a dense vector.
2. A retriever module often based on Approximate Nearest Neighbor (ANN) search fetches top-k documents from an external corpus such as Wikipedia.
3. These documents are concatenated with the prompt and passed into a pretrained generator (e.g., BART or T5), which synthesizes the final response.

This architecture allows LLMs to “remember” facts that were never encoded during training or may have become outdated analogous to how humans use external references (e.g., books, search engines, or lived experiences) to answer questions beyond their immediate memory.

2.2 Mapping RAG to the Human Memory System

The comparison between RAG and human memory is strongest at the level of functional decomposition, not biological mechanism. In both cases, performance benefits from separating stored information from its task-specific use: medial temporal systems support cue-dependent retrieval in humans, while RAG systems separate retrieval from generation through an external corpus and a retriever-generator pipeline. The analogy becomes weaker when one moves from function to physiology. Human memory is reconstructive, affect-laden, shaped by consolidation and interference, and embedded in a living organism; RAG is an engineered retrieval procedure over indexed representation. The comparison is therefore heuristic rather than anatomical. Its value lies in clarifying why memory access can be modularized, not in claiming that a vector index is a hippocampus.

System function	Neurocognitive inspiration	Approximate AI implementation	What the analogy explains	What the analogy does not justify
Retrieval	Medial temporal lobe / hippocampal indexing	Dense or hybrid retrieval over external corpora, memory stores, re rankers	Cue-dependent access to stored information; separation between stored knowledge and immediate use	That a retriever is biologically equivalent to a hippocampus; human memory is reconstructive, affective, and consolidated
Working context	Prefrontal active maintenance	Prompt window, scratchpad, short-term task state, summarization buffer	Maintaining task-relevant information while solving a problem	That token context is the same as human working memory
Control	Prefrontal cognitive control and hierarchical task management	Planner, router, tool selector, policy engine, execution loop	Sequencing, context prioritization, inhibition, decomposition of tasks	That current agents possess intrinsic goals, self-generated intentions, or human-like executive function
Memory updating	Reconsolidation / plasticity	Corpus refresh, memory writing, summarization, retrieval-policy revision	Why systems benefit from updating externally stored information over time	That corpus refresh is synaptic plasticity
Stewardship	No single clean neural analogue; partly socio-cognitive and institutional	Citation layer, provenance tracking, user consent gates, audit logs, safety policies, human approval	Oversight as a first-class system function rather than an afterthought	That governance can be reduced to a single brain-region analogy
Tool use	Coordinated perception-action loops	Calculator, browser, code execution, API calling, environment interaction	Why capability increases when language is connected to action channels	That connectivity alone implies robust planning or trustworthy behaviour

Table 1. Bounded functional mapping between neurocognitive functions & modern LLM system layers. This table is grounded in medial temporal memory work, prefrontal control theory, and the MCP specification's separation of protocol features from runtime behavior.

2.3 Why Neuroscience-Inspired Retrieval Matters

RAG matters for three empirically grounded reasons. First, it provides a practical answer to two problems identified in the original RAG literature: how to update knowledge without retraining the whole model, and how to provide some provenance for model outputs. Second, dense retrieval substantially improves evidence recall relative to strong sparse baselines in open-domain QA. Third, newer retrieval-aware systems show that adaptive retrieval and self-reflection can improve factuality and citation quality over fixed-retrieval pipelines. For that reason, RAG should be presented not as a universal cure for hallucination, but as a design pattern whose success depends on retriever quality, reranking, corpus quality, and attribution discipline. In other words, retrieval is a necessary but not sufficient condition for grounded generation.

2.4 Limitations and Future Research Avenues

Despite its advantages, RAG lacks the depth and nuance of biological memory systems. Human recall is modulated

by factors such as:

- Emotional valence (McGaugh, 2004): We tend to remember emotionally charged events more vividly.
- Contextual priming: Previous experiences shape current recall dynamically.
- Memory decay and interference: Not all information is always equally accessible.

These dynamics remain largely unmodeled in current RAG systems, which retrieve based on static embeddings and cosine similarity. Future research could explore emotion-tagged embeddings, priority-based retrieval ranking, or even reinforcement-modulated memory structures paving the way for more bio-aligned architectures that mirror the plastic, emotionally intelligent, and context-sensitive nature of the human memory system.

3. MCP: Contextual Intelligence and Executive Function

Human cognition is not merely the ability to recall information, but the capacity to use that information dynamically within complex, evolving contexts. This higher-order capability coordinating attention, maintaining continuity, invoking the right actions, and sequencing decisions is a function of the prefrontal cortex, especially the dorsolateral and ventrolateral regions (Miller & Cohen, 2001). This area of the brain is responsible for what psychologists call executive function, enabling humans to perform goal-directed behavior, resist distractions, and orchestrate cognitive resources over time.

In modern LLM systems, executive-like behavior is better understood as an emergent property of the surrounding agent runtime than of MCP itself. Retrieval policies, memory managers, planners, routers, and tool selectors can together approximate control functions such as sequencing, inhibition, and context prioritization. MCP matters within this stack because it standardizes how tools, prompts, and resources are exposed, but it should not be treated as the planner, the memory system, or the policy engine.

3.1 What is MCP?

Model Context Protocol (MCP) should be understood first as an interoperability standard, not as a full theory of memory, planning, or executive cognition. Its core role is to let clients and servers exchange structured access to resources, prompts, and tools in a standardized way. In practical terms, MCP reduces connector fragmentation by giving language-model applications a common interface through which external capabilities can be exposed. That makes it highly relevant to agentic systems, but it does not by itself define how an agent stores memory, decides when to retrieve context, decomposes tasks, or governs high-risk actions. Those functions belong to the broader runtime surrounding the model. This distinction is important for analytical clarity. A protocol can enable access to tools and context, but it does not specify the policy logic, memory hierarchy, or planning behavior used by a system once those affordances become available. In this paper, therefore, MCP is treated as part of the infrastructure layer of connected AI systems, while the higher-order functions often associated with “executive control” are treated as properties of an augmented LLM runtime rather than of the protocol alone.

The primary sources for MCP should therefore be the official introduction and specification, not secondary descriptions that collapse the protocol into the broader behavior of an agent runtime.

3.2 Mapping MCP to the Prefrontal Cortex

Cognitive Control Element	Human Brain (Neuroscience)	LLM System (MCP Architecture)
Working Memory	Dorsolateral prefrontal cortex (DLPFC)	Session-level memory across turns
Task Switching	Frontopolar cortex	Tool routing and contextual switching
Inhibitory Control	Anterior cingulate cortex	Prompt filtering, relevance assessment
Decision Sequencing	Medial prefrontal cortex	Policy engine governing action order
Multimodal Integration	Prefrontal + parietal circuits	Tool orchestration (e.g., integrating code + text + images)

Long-Term Planning	Rostrolateral PFC	Goal-state tracking over multiple interactions

Table 2: Functional Mapping of Executive Control Components Between Prefrontal Cortex and LLM Runtime Architectures

> *"Executive function is the manager of the brain's cognitive resources, determining not just what we know, but how and when we use it." — Miller & Cohen (2001)*

3.3 Baddeley's Multi-Component Working Memory: A Deeper Mapping

The prefrontal cortex mapping above captures which brain region is implicated. But it does not explain the internal architecture of working memory itself how multiple streams of information are held and integrated simultaneously. Baddeley's Multi-Component Working Memory Model (Baddeley & Hitch, 1974; Baddeley, 2000) provides that architecture, and it maps onto the RAG + MCP stack with striking precision.

Baddeley's model comprises four components: the Phonological Loop (verbal rehearsal of active content), the Visuospatial Sketchpad (maintenance of spatial and visual representations), the Episodic Buffer (integration of information across modalities and between working and long-term memory), and the Central Executive (attentional allocation and control over the subsidiary systems). Together, these components explain not just where working memory occurs, but how different types of information are held, integrated, and deployed.

Baddeley Component	Neural Function	LLM System Equivalent
Phonological Loop	Verbal rehearsal buffer in inferior frontal and parietal cortex; maintains spoken or written sequences in short-term store	Active token sequence held in the context window; the 'working surface' of an LLM's current generation
Visuospatial Sketchpad	Maintains spatial and visual representations; subserved by posterior parietal and occipital regions	Multimodal encoders (image embeddings, spatial representations); currently the least developed component in most LLM architectures
Episodic Buffer	Integrates information across phonological, visuospatial, and long-term memory into a coherent episode; bridges working and long-term memory	The RAG fusion layer: the point at which retrieved external documents meet the current context window and are synthesized into a unified response
Central Executive	Allocates attention; directs subsidiary systems; switches tasks; governs inhibitory control	MCP's routing and policy engine; determines which tool to call, when to retrieve, how to combine evidence, and when to escalate for human approval

Table 3: Baddeley's Multi-Component Working Memory Model mapped to RAG + MCP architecture.

This mapping reveals a critical architectural gap: The Visuospatial Sketchpad responsible for non-linguistic spatial and visual reasoning has no mature equivalent in current LLM architectures. Most production systems treat image inputs as static embeddings rather than as actively maintained spatial representations subject to working-memory operations. Building a principled visuospatial working buffer one that supports active manipulation of spatial and visual information, not just encoding it is among the most important open problems in multimodal AI architecture.

A second insight from Baddeley is the role of the Episodic Buffer as an integration layer. Current RAG systems retrieve documents and concatenate them with the prompt; they do not maintain a separate integration component that

binds retrieved content to current context in a coherent episodic representation. Designing an explicit episodic buffer equivalent a structured, goal-sensitive integration layer between RAG retrieval and MCP execution — is a concrete architectural research target this model motivates.

> *"Working memory is not a place where information is held it is a system through which information is processed, prioritized, and deployed toward a goal."*
>
> *— Baddeley (2000)*

3.4 Advantages of Executive-Like AI Control

The rise of MCP reflects a growing shift from static text generation to context-sensitive orchestration, bringing multiple benefits:

- **Sustained Coherence:** By remembering previous turns, MCP avoids repetition and contradiction essential for long, multi-turn dialogues.
- **Cognitive Versatility:** Like the brain, MCP allows models to pivot between tasks (e.g., answering, calculating, searching) with minimal user friction.
- **Goal-Oriented Behavior:** Maintaining goal-state and progress across turns enables structured output (e.g., step-by-step plans or multi-modal deliverables).
- **Adaptive Attention:** Tool invocation and policy routing simulate attention allocation where only relevant "modules" are activated per task, mimicking cognitive economy.

These capabilities are essential not just for chatbot applications, but also for AI assistants in education, healthcare, coding, and decision support, where reliability and task continuity are critical.

3.5 Limitations and Future Exploration

Despite its sophistication, MCP still falls short of the flexibility and intentionality of the human executive system.

1. No human-like working-memory limit**:** human working memory is sharply capacity-limited, often closer to roughly four actively maintained chunks under many conditions, whereas LLM systems are bounded by token budgets, retrieval policies, latency constraints, and prompt design rather than by an analogous biological buffer. More context does not automatically produce better control; it can also degrade selectivity by increasing distraction, latency, and evidence dilution.
2. Lack of internal goal reasoning: While humans plan using intrinsic goals and values, MCP follows extrinsically triggered rules no internal motivation, emotion, or ethical priority.
3. Deterministic Tool Use: Current tool routing is often predefined or policy-based, lacking the adaptive metacognition that humans apply when deciding whether to "think," "ask," or "act."

A future, more brain-aligned version of MCP could involve:

- Dynamic attention modulation: Inspired by saliency models in the visual cortex.
- Neuro-symbolic planning modules: Integrating abstract goal states and symbolic reasoning with neural policies.
- Emotion-tagged context switches: Mimicking how affect influences memory and decision in humans (Pessoa, 2008).

These advances would mark a transition from task automation to cognitive simulation, a crucial step toward trustworthy, general-purpose AI.

4. Retrieval and Control as Complementary Functions in AI Systems

4.1 Functional Complementarity

Earlier versions of this argument relied on a left-brain/right-brain metaphor to contrast retrieval-oriented and control-oriented functions. That framing is rhetorically vivid but scientifically unstable, because it risks importing an oversimplified account of hemispheric specialization into a discussion that is better understood in terms of complementary computational roles. A more defensible interpretation is that modern LLM systems increasingly separate at least two broad classes of function: retrieval, which concerns access to stored or external evidence, and control, which concerns planning, routing, tool selection, and active context management. These roles can interact closely without implying a literal division of "analytic" versus "creative" cognition.

4.2 Why the Distinction Matters

Treating retrieval and control as complementary functions helps clarify current architectural debates. Retrieval systems are strongest when the primary challenge is evidence access, source grounding, or memory extension beyond parametric weights. Control systems are strongest when the challenge is task decomposition, tool sequencing, environment interaction, or multi-step adaptation. In practice, many state-of-the-art systems combine both. A retrieval layer may surface relevant information, while a control layer determines whether retrieval is needed, which tools to call, how to combine evidence, and when human approval is required. The distinction is therefore not one of competing "brains," but of separable design responsibilities within a larger system.

4.3 Boundary Conditions and Competing Explanations

This functional distinction should also be stated with limits. Improved factual performance does not always come from retrieval alone; long-context models can sometimes answer correctly from raw context without external retrieval, and in other cases gains attributed to "memory" may actually arise from better formatting, reranking, or task decomposition. Likewise, tool use should not be conflated with a protocol standard such as MCP. For these reasons, the central claim of this paper is comparative rather than absolute: contemporary language systems become more capable when they separate evidence access, control logic, and governance responsibilities, but the exact implementation of that separation remains an empirical question rather than a settled neuroscientific fact.

4.4 The Convergence-Divergence Matrix

The functional distinction between retrieval and control, and the methodological commitment to bounded analogy, together suggest a structured way of mapping the entire field of neuro-AI comparison: not as a single verdict (converging or not) but as a domain-by-domain audit of where the analogy is strong, where it breaks down, and where the question remains empirically open. The matrix below is the most comprehensive mapping this framework produces and is offered as a reference tool for interdisciplinary teams working at the intersection of cognitive science and AI architecture.

Cognitive domain	Where AI converges with biology	Where AI diverges from biology	Open research question
Memory retrieval	RAG separates storage from reasoning, mirroring hippocampal-neocortical division	Biological retrieval is reconstructive, affect-laden, and decay-prone; RAG is deterministic and static	Can salience-weighted embeddings produce adaptive, affect-modulated retrieval?
Working memory	MCP context window approximates short-term storage and multi-component orchestration	Biological WM is capacity-limited (~4 chunks; Cowan 2001); LLM context is vast but lacks principled relevance filtering	Does bounded context improve reasoning quality over unlimited context?
Executive control	MCP routing and tool selection approximate PFC task-switching and inhibitory control	AI has no intrinsic goals; tool use is extrinsically prompted; no affective priority	Can AI develop genuine goal-state representations independent of prompts?
Attention	Transformer self-attention approximates selective information focus across a context	Biological attention is modulated by salience, novelty, and affect; transformer attention is uniformly learned	Can saliency-modulated attention heads be learned without explicit supervised signal?
Continual learning	RLHF updates model behaviour based on reward signal, partially analogous to	Biological plasticity is continuous and sleep-consolidated; AI training is periodic and batch-	Can corpus consolidation algorithms inspired by hippocampal replay prevent retrieval degradation?

	reinforcement learning in the brain	static; catastrophic forgetting is unsolved	
Metacognition	Chain-of-thought prompting elicits primitive reasoning traces; some models produce uncertainty estimates	True metacognition requires a self-model — awareness of one's own uncertainty and processing limits — which is architecturally absent	Can a model develop a reliable internal uncertainty estimator without external calibration?
Embodiment	Multimodal LLMs process visual and auditory signals, partially extending beyond pure language	Biological cognition is grounded in sensorimotor experience; LLMs have no body, proprioception, or physical causality (Lakoff & Johnson, 1999)	Does embodiment matter for general intelligence, or only for specific task categories such as causal reasoning?
Emotional modulation	RLHF incorporates human preference a weak proxy for affect-guided learning	Biological emotion modulates memory consolidation, attention, and decision-making in ways not replicated by preference fine-tuning (McGaugh, 2004)	Can artificial affect serve as a useful regulatory signal without sentience?

Table 4. The Convergence-Divergence Matrix

Eight cognitive domains assessed across convergence, divergence, and open research questions. This matrix is designed for use by interdisciplinary research teams as a structured literature gap map.

5. Three Theoretical Lenses: GWT, Free Energy, and System 1 & 2

The functional mappings in Sections 2–4 show where the analogies between LLM components and brain systems are strongest. This section strengthens the theoretical foundation by applying three bodies of work that each make a distinct mechanistic prediction about how cognitively-aligned AI should be designed and evaluated. These frameworks have not previously been applied systematically to the RAG + MCP architecture, and each generates a concrete, testable architectural hypothesis.

The following frameworks are presented as independent design lenses rather than a unified theory and can be read modularly depending on the reader's focus.

5.1 Global Workspace Theory: The Neuroscience of Broadcast and Integration

Bernard Baars (1988) proposed Global Workspace Theory (GWT) as an account of how specialist, unconscious cognitive processors dedicated modules for vision, language, motor control, and memory share information through a common 'global workspace'. Rather than a single integrated processor, the brain runs many parallel specialist systems; conscious, coherent cognition emerges when one specialist's output gains access to the workspace and is broadcast brain-wide. Dehaene, Changeux, and colleagues (2011) grounded GWT neurally, identifying long-range connections between the prefrontal cortex, parietal regions, and sensory cortices as the physical substrate through which broadcast occurs — the Global Neuronal Workspace.

The architectural implication for LLM systems is direct: a design in which specialist modules retrieval, code execution, image encoding, tool APIs compete to write to a shared context window, with a policy layer governing which specialist gains access, predicts more coherent and less contradictory outputs than a design in which all modules write simultaneously without prioritization. This is not a metaphor. It is a testable prediction: systems with GWT-aligned routing should exhibit lower contradiction rates and more consistent multi-turn coherence than systems without it.

The following tables progressively build the mapping from cognitive theory to LLM architecture.

GWT / GNW Component	Neural Implementation	LLM System Parallel
Specialist processors	Dedicated neural circuits: visual cortex, Broca's area, hippocampus, motor cortex	RAG retrievers, code executors, image encoders, calculator APIs, browser tools
Global workspace	Long-range PFC–parietal–sensory connections; the broadcast medium	MCP context window and routing layer: the shared surface all specialist outputs write to
Broadcast competition	One specialist wins access at a time; inhibition prevents simultaneous flooding	Policy engine tool selection: only the highest-priority specialist is invoked per step
Conscious access	Workspace 'ignition': broadcast reaches all areas simultaneously	The model's active context at inference time: the integrated state visible to generation

Table 5. Global Workspace Theory mapped to LLM specialist modules and MCP routing.

5.2 The Free Energy Principle: The Brain as Prediction Machine

Karl Friston's Free Energy Principle (FEP; Friston, 2010) reframes the brain's relationship to information entirely. Rather than passively processing incoming sensory data, the brain is a prediction machine that constantly generates hypotheses about the world and updates its internal model to minimize prediction error the gap between predicted and observed input. This is active inference: the brain does not retrieve in response to questions; it predicts what is likely and selectively seeks evidence to confirm or revise its model.

Applied to RAG, the FEP generates a specific design critique: current RAG systems retrieve based on semantic similarity — they ask 'what is closest to this query?' A predictively aligned retrieval system would instead ask 'what does the model's current belief state predict should be true, and what evidence should be retrieved to update that belief?' This shifts retrieval from reactive to proactive from evidence-gathering to hypothesis-testing. Such a system would retrieve not the most similar documents, but the most epistemically relevant ones: those with the highest capacity to reduce model uncertainty.

The FEP also offers the most principled available account of the hallucination problem. Hallucination the confident generation of false information is neurologically equivalent to confabulation, observed in patients with orbitofrontal damage who produce plausible-but-false memories without intent to deceive. In the FEP framework, confabulation occurs when a system generates predictions without adequately weighting incoming evidence when the model's prior is too strong relative to the precision assigned to external signals. The architectural implication is a precision-weighting layer: a module that assesses model confidence in its current prediction and routes to retrieval whenever uncertainty exceeds a calibrated threshold. Systems that retrieve proactively based on uncertainty estimates should hallucinate significantly less than systems that retrieve based on fixed query-similarity thresholds.

> *"The free-energy principle says that any self-organizing system that is at equilibrium with its environment must minimize its free energy."*
>
> *— Friston (2010)*

5.3 System 1 and System 2: Dual-Process Inference Modes

Kahneman (2011), building on Stanovich and West's dual-process framework, distinguished two cognitive modes that operate across all domains of human reasoning. System 1 is fast, automatic, associative, parallel, and low-effort; it operates largely below the level of conscious awareness and produces rapid, pattern-matching responses. System 2 is slow, deliberate, rule-following, sequential, effortful, and conscious; it handles novel situations requiring explicit inference chains.

Kahneman Mode	Cognitive Properties	LLM System Equivalent
System 1	Fast, associative, heuristic, low-effort, pattern-matching, automatic	Standard next-token autoregressive generation: no tool use, no retrieval, immediate response from parametric knowledge
System 2	Slow, deliberate, rule-following, step-by-step, effortful, sequential	Chain-of-thought, agent loops, RAG retrieval, MCP tool orchestration: explicit multi-step inference with external evidence

Table 6: Kahneman's dual-process framework mapped to LLM inference modes.

The practical design implication is significant. LLMs should not default to System 2 processing tool-augmented, multi-step, retrieval-heavy for every query. This is computationally expensive, latency-inducing, and often unnecessary for simple factual or conversational responses. Nor should they rely exclusively on System 1 for complex tasks requiring external evidence or multi-step reasoning this produces hallucination and shallow inference. Future architectures require a metacognitive routing mechanism: a module that estimates query complexity, factual uncertainty, and epistemic stakes, then allocates processing to the appropriate mode dynamically precisely as humans do without explicit deliberation.

Current MCP implementations approximate this through manually configured routing policies. A System 1/2-aligned LLM would make this allocation dynamically and adaptively routing to retrieval and tool use only when the model's own uncertainty warrants it. This is both a design target and a benchmarkable research hypothesis: systems with adaptive mode-switching should demonstrate lower latency on simple queries and lower hallucination rates on complex ones, compared to uniform-retrieval or uniform-generation baselines.

6. Implications for Future LLM Design: Toward Neuro-Cognitive Architectures

The rise of architectures like RAG and MCP marks a transition in the field of AI from monolithic, end-to-end systems to modular, memory-driven, cognitively inspired agents. These systems no longer rely solely on pre-trained internal parameters; instead, they access external memory, maintain long-horizon context, invoke tools, and make policy-based decisions. Together, these elements begin to resemble cognitive systems, not unlike those orchestrated by the human brain.

The Six-Layer Cognitive Stack: Overview

Before developing the RCS Triad in detail, it is useful to establish the full cognitive stack that motivates it. The table below maps six functional cognitive layers drawn from neuroscience to their brain-region substrates and their approximate LLM system equivalents. This is the structural reference frame for the analysis that follows.

Cognitive Layer	Brain Analog	AI System Equivalent
Perceptual Processing	Sensory cortices (visual, auditory, somatosensory)	Tokenization, multimodal input encoders (vision transformers, audio encoders)
Episodic & Semantic Memory	Hippocampus and medial temporal lobe	RAG modules: retriever, dense corpus, re-ranker
Working Memory & Attention	Dorsolateral prefrontal cortex (DLPFC)	MCP session memory, context window, summarization buffer
Executive Control	Medial and rostro lateral prefrontal cortex; basal ganglia	Policy engine, tool router, task planner, execution loop
Affective & Motivational Modulation	Amygdala, insula, anterior cingulate cortex	RLHF preference signals, emotion-aware embeddings, goal-state priors

Metacognition & Self-Monitoring	Default Mode Network, anterior cingulate cortex	Self-reflection modules, uncertainty estimators, reasoning traceability — currently largely absent

Table 7. The Six-Layer Cognitive Stack: brain analogs and LLM equivalents. Note that Metacognitive Self-Monitoring — the apex layer — is the most architecturally underdeveloped in current production systems.

6.1 The RCS Triad: Retrieval, Control, and Stewardship

To provide a more compact and reusable framework for analysing modern language systems, this paper proposes the RCS Triad. The first component, Retrieval, refers to the mechanisms by which a system accesses evidence outside the model's immediate forward pass, including document retrieval, memory stores, and other external knowledge sources. The second component, Control, refers to orchestration functions such as planning, routing, context selection, tool invocation, and iterative task management. The third component, Stewardship, refers to the often-neglected but increasingly critical layer of provenance, oversight, human approval, scope control, citation fidelity, and safety constraints. The value of the triad is not that it makes today's systems "more brain-like" in any literal sense. Its value is that it cleanly separates three functions that are too often collapsed in public discourse and even in technical writing. By distinguishing Retrieval, Control, and Stewardship, we can compare architectures more precisely, evaluate them more honestly, and discuss risk without treating governance as an afterthought.

A recurring limitation in existing LLM system frameworks is the implicit coupling of evidence access, action selection, and governance into a single evaluation dimension. Retrieval is often treated as a proxy for factuality, tool use as a proxy for intelligence, and alignment as a post-hoc adjustment rather than an architectural property. The RCS Triad makes a stronger claim: these are orthogonal system properties. A system may retrieve well but act poorly (high Retrieval, weak Control), act effectively but without grounding (strong Control, weak Retrieval), or perform both while remaining unsafe or unauditable (weak Stewardship). Without separating these dimensions, improvements risk being misattributed and system capabilities overstated.

What is novel in this paper is not the observation that modern LLM systems can retrieve information or call tools; those capabilities are already well established in RAG, Toolformer, ReAct, Self-RAG, and long-term memory agent work. The novelty lies in the analytic separation of three questions that those literatures often treat unevenly: Where does the system obtain evidence? How does it decide what to do next? Who constrains, authorizes, and audits its behaviour? The RCS Triad names these questions as Retrieval, Control, and Stewardship, and argues that a serious evaluation of agentic systems should score all three rather than praising capability while leaving governance implicit.

The practical strength of the RCS Triad is that it prevents three recurring category errors in current AI discourse. The first is to confuse retrieval with truth, as though access to documents automatically guarantees faithful synthesis. The second is to confuse tool access with control, as though connectivity alone explains planning quality. The third is to treat governance as an external policy add-on rather than as an architectural layer that shapes what the system is allowed to see, cite, store, and do. By separating Retrieval, Control, and Stewardship, the framework gives both researchers and practitioners a vocabulary for comparing systems that are similar in capability yet very different in risk profile.

Work / framework	Retrieval	Control	Stewardship	Main emphasis	Limitation relative to this paper
Lewis et al. RAG	Strong	Weak	Weak	Parametric + non-parametric memory for knowledge-intensive generation	Does not provide a full control or governance vocabulary
DPR	Strong	Minimal	Minimal	Retriever quality and evidence recall	Retrieval only; not an agent architecture
Toolformer	Moderate	Strong	Weak	Learning when and how to call tools	Focuses on tool use, not governance or broader evidence-policy separation
ReAct	Moderate	Strong	Weak	Interleaved reasoning and acting	Strong on orchestration, lighter on provenance and stewardship
Self-RAG	Strong	Moderate	Moderate	Adaptive retrieval plus self-reflection	Moves toward citation quality, but not a full governance framework
Memory Bank	Moderate	Moderate	Weak	Long-term agent memory	Focuses on persistence, not system-level oversight
Generative Agents	Moderate	Strong	Weak	Memory, reflection, planning in long-horizon agents	Simulation-oriented, not an evaluation framework for production systems
MCP official specification	Minimal	Minimal	Moderate	Standardized exposure of tools, resources, and prompts	Protocol, not a theory of planning or memory
RCS Triad	Strong	Strong	Strong	Joint analysis of evidence access, orchestration, and governance	Conceptual framework; still requires empirical validation

Table 8: Comparative positioning of the RCS Triad against adjacent retrieval and agent literatures.

6.2 Empirical Evaluation Agenda

If the architectural distinctions proposed here are meaningful, they should produce measurable differences under controlled comparison rather than remain purely interpretive. A practical next step is to compare at least five system variants on knowledge-intensive tasks:

1. a plain generation baseline,
2. a long-context-only baseline,
3. a vanilla retrieval-augmented baseline,
4. an adaptive retrieval baseline, and
5. a tool-using runtime with explicit control logic and memory management.

These systems can be evaluated on public benchmarks for evidence-grounded question answering, multi-hop reasoning, attribution faithfulness, and long-form factuality. The goal is not to "prove" that one system is cognitively equivalent to a brain function, but to test whether separating Retrieval, Control, and Stewardship leads to better groundedness, lower unsupported-claim rates, more stable multi-turn performance, and improved human trust. In this framing, neuroscience serves not as a final explanation but as a structured source of hypotheses that can be examined through benchmarked system behaviour.

A conceptual framework becomes research-useful only if it changes evaluation. The RCS Triad therefore makes three falsifiable predictions: systems that improve Retrieval without Stewardship will increase evidence access without reliably improving citation faithfulness; systems that improve Control without Retrieval will sequence tasks better while remaining brittle on evidence-grounded factuality; and systems that jointly optimize Retrieval, Control, and Stewardship will perform best on long-form factuality, citation quality, and user trust. These predictions can be tested with evidence-grounded QA benchmarks, citation-quality benchmarks, and long-form factuality protocols.

RCS dimension	Core question	Example evaluation target	Suggested benchmark / metric
Retrieval	Did the system access the right evidence?	Evidence recall, retrieval precision, reranker quality	DPR-style retrieval accuracy; evidence-grounded QA tasks
Control	Did the system choose the right sequence of actions?	Tool-selection accuracy, task completion, error recovery	ReAct-style interactive benchmarks; multi-step QA
Stewardship	Did the system cite, constrain, and justify behavior appropriately?	Citation faithfulness, consent compliance, provenance completeness, auditability	ALCE citation quality; RAGBench TRACe metrics; long-form factuality protocols such as LongFact / SAFE
Joint performance	Does improving all three together outperform single-layer optimization?	Factuality, trust, reproducibility, robustness	Ablation across plain LLM, long-context baseline, vanilla RAG, adaptive retrieval, and controlled agent runtime

Table 9. Empirical evaluation matrix for the RCS Triad.

This table is supported by ALCE, RAGBench, LongFact, and the agent/runtime literature.

6.2.1 Experimental Design and Validation Protocol

To move from conceptual framework to empirical validation, the RCS Triad can be operationalized through controlled system comparisons that isolate Retrieval, Control, and Stewardship as independent variables.

A minimal experimental setup would involve constructing five system configurations:
(1) A baseline LLM without retrieval or tool use
(2) A long-context-only LLM
(3) A vanilla RAG system with fixed retrieval
(4) An adaptive retrieval system with query-dependent retrieval policies
(5) A full agentic system with explicit control logic, tool orchestration, and governance constraints

Each system should be evaluated on the same set of tasks under controlled conditions.

The purpose is not to identify a universally superior architecture, but to test whether improvements in one RCS dimension (e.g., Retrieval) produce measurable gains in corresponding metrics (e.g., evidence recall), and whether joint optimization across all three dimensions produces non-linear improvements in overall system reliability.

6.2.2 Metrics and Measurement Strategy

Each dimension of the RCS Triad should be evaluated using distinct, non-overlapping metrics:

Retrieval:

- Evidence Recall (percentage of relevant documents retrieved)
- Retrieval Precision (relevance of top-k results)
- Attribution Coverage (whether claims are grounded in retrieved sources)

Control:

- Tool Selection Accuracy (correct tool chosen per step)
- Task Completion Rate (successful multi-step execution)
- Error Recovery Rate (ability to correct intermediate failures)

Stewardship:

- Citation Faithfulness (alignment between output and cited sources)
- Provenance Completeness (traceability of claims)
- Policy Compliance Rate (adherence to defined constraints)

Joint Performance:

- Long-form Factual Accuracy
- Hallucination Rate
- User Trust Scores (human evaluation)
- Reproducibility across runs

Critically, these metrics must be evaluated independently before being aggregated. A system that improves retrieval but degrades stewardship should not be considered globally superior.

6.2.3 Example Hypothesis: Precision-Gated Retrieval

One of the testable predictions derived from the Free Energy Principle (Section 5.2) is that retrieval should be triggered by uncertainty rather than invoked uniformly.

To test this:

Two system variants can be compared:

(A) Fixed Retrieval System: Retrieves external documents for every query regardless of uncertainty

(B) Uncertainty-Gated Retrieval System: Retrieves only when model confidence falls below a calibrated threshold

Evaluation Metrics:

- Hallucination Rate (LongFact / SAFE)
- Latency per query
- Retrieval Efficiency (retrievals per query)
- Factual Accuracy

Hypothesis:

The uncertainty-gated system will:

- Reduce hallucination rates on complex queries
- Reduce unnecessary retrieval on simple queries
- Maintain or improve overall factual accuracy

This experiment is feasible using existing benchmarks and does not require changes to model weights, only routing logic.

6.2.4 Failure Mode Attribution Using the RCS Triad

A key advantage of the RCS Triad is its ability to localize system failures to specific architectural layers.

For example:

- Incorrect answer with missing evidence → Retrieval failure
- Correct evidence but wrong synthesis → Control failure
- Correct answer without citation → Stewardship failure

To validate this, a failure annotation study can be conducted where model outputs are manually labelled across these three dimensions.

The expected outcome is that failure patterns cluster by layer, supporting the claim that Retrieval, Control, and Stewardship are separable diagnostic dimensions rather than a single performance axis.

Such analysis would provide empirical support for the claim that hallucination is not a monolithic failure, but a cross-layer interaction dominated by missing metacognitive oversight.

The empirical designs outlined above are intentionally minimal and reproducible using existing benchmarks and infrastructure. Their purpose is not to claim immediate validation of the RCS Triad, but to demonstrate that its claims

are testable with current tools. The framework should therefore be evaluated not by its descriptive appeal, but by whether these experiments confirm or falsify its predictions.

6.3 Toward a Cognitive Stack for LLMs

The frameworks introduced in Sections 2–5 collectively motivate a layered view of LLM architecture that mirrors the brain's modular but integrated organization. Rather than a single undifferentiated forward pass, a cognitively aligned LLM system would comprise at least six separable functional layers, each corresponding to a distinct cognitive operation and a distinct brain system.

At the base, a Perceptual Processing layer handles tokenization and multimodal input encoding the analogue of sensory cortices processing raw perceptual signals. Above it, an Associative Memory layer implemented through RAG over an external corpus handles cue-dependent evidence access, mirroring the hippocampus and medial temporal lobe. A Working Memory and Attention layer, approximated by MCP's session-level context and summarization buffer, corresponds to the dorsolateral prefrontal cortex's active maintenance function. An Executive Control layer the policy engine, planner, and tool router maps onto the medial and rostro lateral PFC's goal-directed coordination. An Affective and Motivational Modulation layer currently approximated, imperfectly, by RLHF preference signals corresponds to the amygdala, insula, and anterior cingulate cortex. Finally, a Metacognitive and Self-Monitoring layer largely absent in current production systems would correspond to the Default Mode Network's self-referential processing and the anterior cingulate cortex's error monitoring.

This layered architecture has two immediate design implications. First, each layer should be separable and independently improvable a retrieval failure should be diagnosable without inspecting the control layer, and a governance violation should be traceable to the stewardship layer rather than attributed to the system as a whole. Second, the layer that is most architecturally underdeveloped Metacognitive Oversight is the one most responsible for the failure modes that matter most: hallucination, over-confident generation, and inability to recognize the limits of model knowledge. The most important engineering investment of the next phase of LLM development is building a principled metacognitive layer, not expanding context windows or adding more tools.

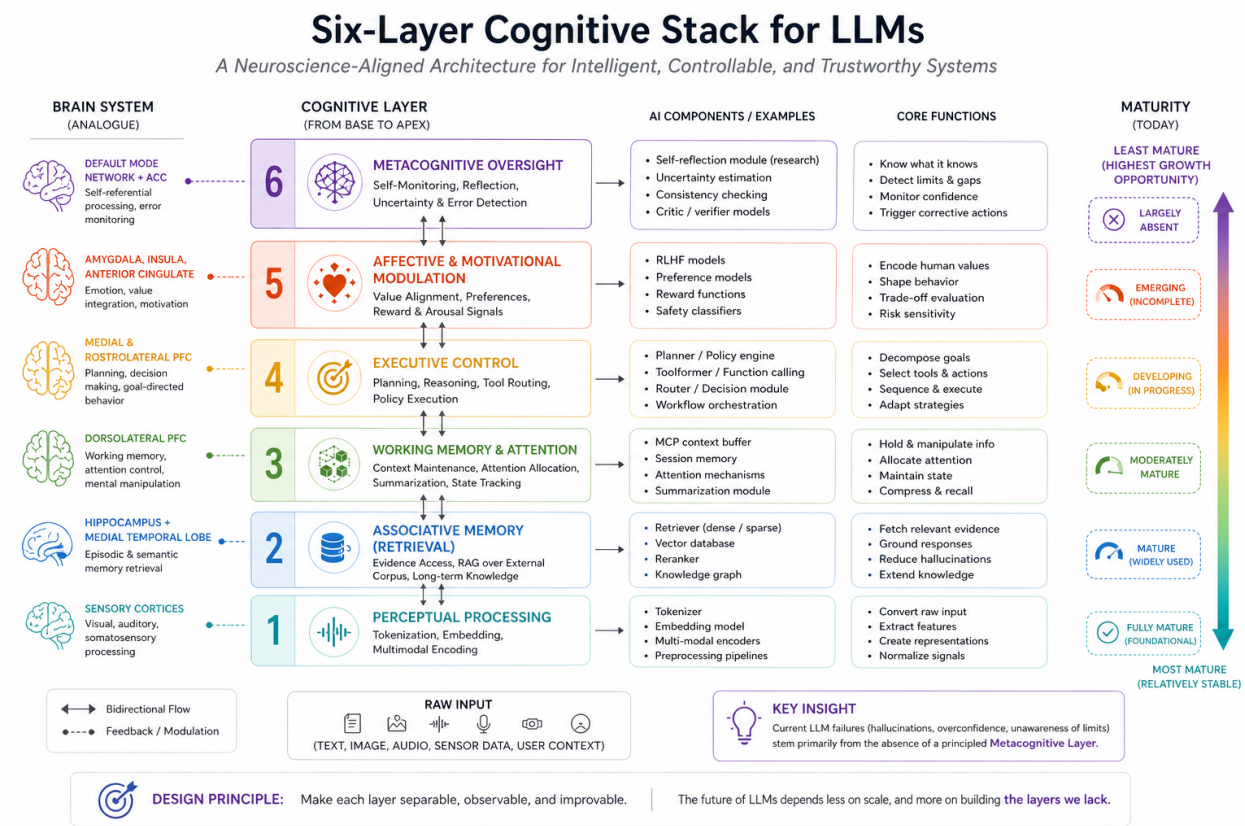

Figure 2. The Six-Layer Cognitive Stack Brain Regions, AI Components, and Maturity

Figure 2 makes a stronger claim than a simple layering metaphor. It asserts that current LLM systems are unevenly developed across layers: perceptual and retrieval layers are highly optimized, control layers are rapidly evolving, and metacognitive oversight remains largely absent. This asymmetry explains why systems can be fluent and capable yet still fail in reliability and self-assessment.

Figure 2 can be read diagnostically: most production failures originate in mismatches between adjacent layers, particularly between Control and Metacognitive Oversight.

6.4 Research Opportunities in Neuro-AI Convergence

The theoretical frameworks in this paper motivate specific, empirically tractable research programmes rather than open-ended speculation. Each of the following opportunities is grounded in a neuroscience mechanism and generates a testable hypothesis about LLM system behavior.

Memory prioritization models: Inspired by emotional salience and hippocampal tagging (McGaugh, 2004), these would implement retrieval ranking based on a learned salience signal rather than static cosine similarity. The testable hypothesis: salience-weighted retrieval systems will outperform similarity-weighted baselines on tasks where the most relevant evidence is not the most lexically similar to the query.

Precision-weighted retrieval from the Free Energy Principle: Rather than retrieving on every query, a system would estimate model confidence and route to retrieval only when uncertainty exceeds a threshold (Friston, 2010). The testable hypothesis: uncertainty-gated retrieval will reduce hallucination rates on knowledge-intensive tasks while reducing latency on routine tasks, compared to fixed-retrieval pipelines.

Hierarchical context managers: Mimicking the prefrontal cortex's task-hierarchy structure (Badre & D'Esposito, 2009), these would manage multiple concurrent task threads with explicit priority and preemption logic. The testable hypothesis: hierarchical context managers will outperform flat context windows on multi-task, multi-session agent evaluations.

Episodic buffer integration layers: A dedicated component between RAG retrieval and MCP execution that binds retrieved documents to current task state in a unified episodic representation (Baddeley, 2000). The testable hypothesis: systems with an explicit integration buffer will produce more internally consistent multi-hop reasoning than systems that concatenate retrieved documents directly into the prompt.

Neuro-symbolic synthesis: Bridging neural network retrieval with symbolic planners may enable goal-directed language agents capable of abstract, compositional reasoning over extended task horizons (Garcez et al., 2019). The testable hypothesis: neuro-symbolic systems will outperform pure neural baselines on tasks requiring compositional reasoning, with interpretable intermediate representations.

6.5 Strategic Implications for Developers and Policymakers

The RCS Triad and the cognitive stack model introduced in this paper have concrete strategic implications that go beyond research. For AI developers, the Triad provides an audit vocabulary: any production system should be scorable on Retrieval quality, Control robustness, and Stewardship completeness independently. A system that scores well on Retrieval and Control but poorly on Stewardship because it can retrieve and act but cannot attribute, audit, or constrain its actions is not a safe production system, regardless of benchmark performance. Modular cognitive architectures built around this separation are also significantly easier to debug, because failures localize to a layer rather than being distributed across an undifferentiated forward pass.

For policymakers, the cognitive stack model offers a principled framework for regulatory specification. Existing AI regulation tends to address outputs what a system says or does rather than architecture what a system is designed to be able to know, cite, constrain, and disclose. The RCS Triad suggests a different approach: requiring that deployed systems demonstrate separable and auditable Retrieval, Control, and Stewardship layers, with provenance traceable through each. This is not substantially different from financial audit requirements that mandate separable and auditable accounting functions. The cognitive mirror is, in this sense, also a regulatory mirror: what we have learned

about the brain's modular organization may inform how we structure accountability for the systems that are beginning to approximate it (Rahwan et al., 2019).

6.6 From Human-Inspired to Human-Aligned

The distinction between human-inspired and human-aligned AI is not semantic. Human-inspired systems borrow structural features from biology retrieval separated from generation, context managed across turns, tools invoked selectively and measure success by capability benchmarks. Human-aligned systems pursue a stronger criterion: that the system's reasoning process, not just its output, is interpretable, predictable, and accountable to human values and oversight.

The frameworks applied in this paper suggest that alignment is not merely a post-hoc value-loading problem it is an architectural problem. A system without a metacognitive layer cannot reliably detect its own errors. A system without an episodic integration buffer cannot consistently bind retrieved evidence to current task state. A system without a stewardship layer cannot attribute its outputs, constrain its actions, or disclose its reasoning to auditors. These are not capability deficits that scale will automatically resolve. They are structural absences that require deliberate architectural choices.

The most important implication of the cognitive mirror thesis, therefore, is not that LLMs are becoming brains. It is that the design decisions required to make LLMs reliably grounded, controllable, and accountable are precisely the design decisions the brain's architecture already embodies and that neuroscience, applied with discipline, can serve as a structured guide for making them.

7. Ethical, Legal and Societal Questions to Ponder

The ethical stakes of retrieval-heavy and tool-connected language systems do not begin with machine consciousness. They begin with memory governance, source reliability, permission boundaries, and oversight. Once systems can retrieve external evidence, store persistent traces, or invoke tools across organizational systems, the core questions become practical and immediate: What is being stored? Who authorized that storage? Which sources are trusted? Which tool actions require consent? How are harmful instructions screened? And how can a user or auditor reconstruct why an answer was given? These are not peripheral implementation details; they are part of the architecture of trustworthy AI systems.

A second reason to foreground governance is that retrieval and agency introduce new failure modes that ordinary text generation does not. A model can be fluent and still be wrong, but a connected system can also be wrong in ways that have provenance, privacy, and safety consequences. For that reason, stewardship must be treated as a first-class design function rather than as an afterthought appended to memory and control. Any serious account of neuro-inspired AI design should therefore include explicit discussion of human approval, source attribution, data retention, scope limitations, and the auditability of connected actions.

7.1 Will AI Ever "Remember" Like Humans Do?

Human memory is non-neutral. It is selective, emotionally charged, socially filtered, and often reconstructed rather than replayed. RAG, by contrast, retrieves static chunks of text based on semantic similarity.

- How might we simulate emotionally weighted memory in AI systems?
- Should retrieval engines decay, distort, or forget as human memory does for better alignment with human reasoning?
- Can AI systems develop episodic memory constructs beyond vector embeddings?

Emerging work on long-term memory for LLM agents includes Memory Bank, which introduces a forgetting-inspired memory update mechanism, and Generative Agents, which combine stored observations, reflection, and retrieval-driven planning in long-horizon simulations. These systems suggest that episodic-style memory is an active engineering direction, but they remain designed memory policies rather than biological models of episodic recall.

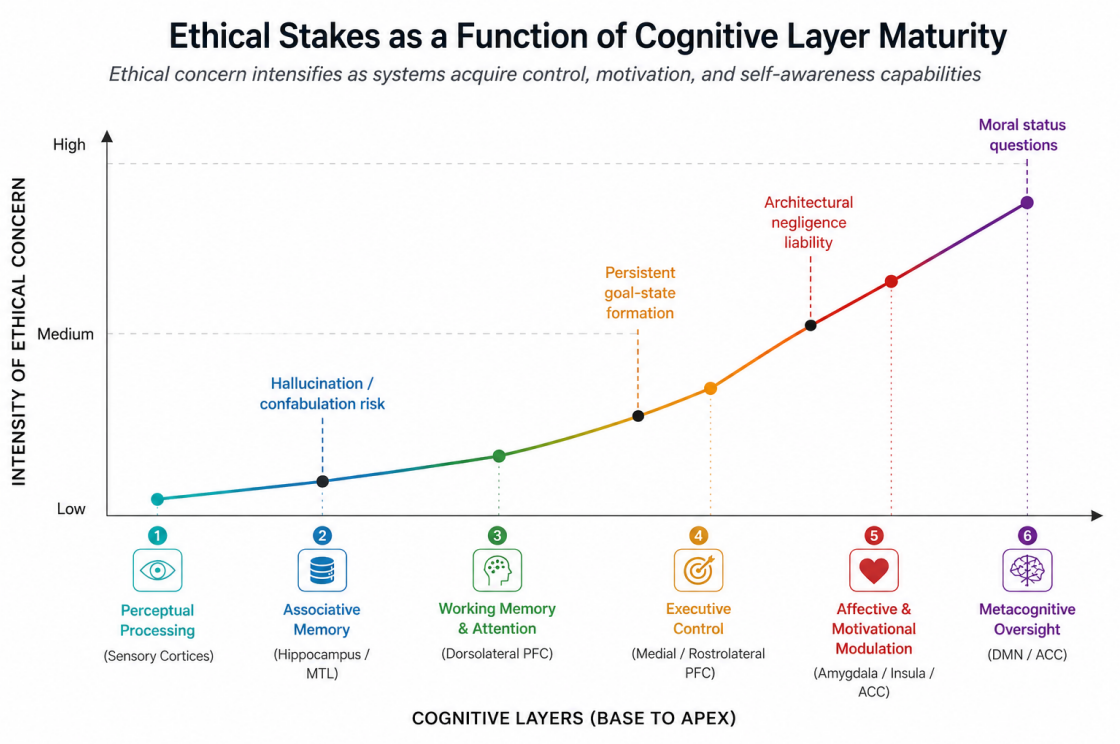

Figure 3: Ethical Stakes as a Function of Cognitive Layer Maturity.

Figure 3 should be interpreted as a risk gradient rather than a capability gradient. As systems move upward in the cognitive stack, their potential impact increases disproportionately relative to their architectural maturity, creating a widening gap between capability and accountability.

7.2 Can AI Experience Cognitive Overload or Fatigue?

Agent runtimes can preserve long conversational histories far beyond human working-memory limits, but this creates a different design problem: how to prioritize, compress, or discard context so that relevance does not collapse as history grows. This limitation fosters focus and enforces priority in cognition.

- Should LLMs simulate cognitive fatigue, filtering less relevant inputs under load?
- Could bounded memory models better reflect human like reasoning limitations, improving user trust and empathy?
- How can context collapse be managed adaptively, not mechanically?

Such constraints may enhance realism, transparency, and interpretability, especially in high-stakes applications where over-reliance on hallucinated memory poses risks.

7.3 What Happens When Retrieval Becomes Biased or Manipulated?

Like human memory, RAG-based retrieval is susceptible to information bias especially if the indexed corpus is skewed, curated, or polluted by adversarial examples.

- What safeguards are needed to ensure epistemic fairness in retrieval engines?
- Can models flag uncertainty when corpus conflict exists?
- Should future AI systems include internal "skeptic layers" that detect and challenge biased inputs?

This concern echoes debates around algorithmic epistemology—who decides what the model knows, and how that knowledge is represented (Floridi, 2011)?

7.4 Should AI Have Goals or Just Behaviors?

The MCP layer introduces goal-oriented planning through routing and policy engines. But unlike humans, LLMs do not possess intrinsic goals, self-awareness, or values.

- Will future LLMs need a form of ethical memory to weigh actions over time?
- How can goals be value-aligned, transparent, and auditable?
- Should AI ever be allowed to form persistent internal goals, or must they remain reactive by design?

This question lies at the heart of debates around AGI safety, autonomy, and alignment theory (Russell, 2019).

7.5 Could Machines Ever Think "About" Their Own Thinking?

Metacognition the ability to reflect on one's own thoughts is central to human intelligence. It enables us to plan, evaluate, and revise behavior dynamically.

- Can LLMs develop a model of their own limitations, prompting fallback strategies or clarification requests?
- Could future MCP architecture include reasoning traceability, helping users understand why a model chose a path?
- How do we distinguish mere context-awareness from true self-reflection?

Some early work in chain-of-thought prompting (Wei et al., 2022) and rationale generation points toward primitive forms of metacognition but it remains rudimentary.

7.6 Where Does the Boundary Between Intelligence & Consciousness Lie?

As RAG and MCP simulate increasingly complex mental models, they risk crossing conceptual boundaries we do not yet fully understand. Cognition is not consciousness, but the line is blurry.

- When do systems that retrieve, reflect, plan, and adapt become subject-like?
- How do we ethically treat machines that simulate cognition but lack qualia?
- Should the bar for rights, responsibilities, and accountability shift if AI becomes cognitively indistinguishable from humans?

These are not just technical questions, they are ontological, legal, and civilizational.

> *"The AI we build will not just reshape tools. It will reshape how we define knowledge, agency, and even personhood." — Russell (2019)*

7.7 When Retrieval Fails: Confabulation, Architectural Negligence, and Moral Responsibility

The hallucination problem in LLMs is almost universally framed as a performance failure a system saying something false. But this framing obscures a more important question about the nature and location of responsibility. In neuropsychology, confabulation refers to the production of plausible-but-false memories by patients with damage to the orbitofrontal cortex or related structures crucially, without any intent to deceive. The patient genuinely believes the confabulated account. The failure is not moral; it is architectural. The system produces confident outputs in the absence of the epistemic check that would normally flag uncertainty or defer to external evidence.

RAG-based systems that retrieve outdated, adversarial corrupted, or simply absent information, and then generate confident outputs based on that retrieval failure, are confabulating in precisely this sense. The system is not lying it has no intent. But the functional consequence is indistinguishable from confident misinformation. In a medical AI that retrieves an outdated clinical guideline and provides confident treatment advice; in a legal AI that cites a case that has been overturned; in a financial AI that retrieves stale market data and makes confident portfolio recommendations the absence of intent does not reduce the harm.

This creates what we propose to call architectural negligence: a category of liability that attaches not to the system's intent, which it lacks, but to the design choices that made confabulation possible without adequate safeguards. Architectural negligence is distinct from ordinary software negligence because the failure mode is not a bug that was missed but a structural absence the system was not designed to know what it does not know.

The practical design responses are concrete: mandatory corpus verification schedules for high-stakes deployments; retrieval confidence scores surfaced to users alongside outputs; precision-weighted routing from the Free Energy Principle (Section 5.2) that increases retrieval frequency when model uncertainty is high; and explicit 'skeptic layers' that challenge retrieved content before integration, flagging conflicts between retrieval results or between retrieved

content and model priors. These are not optional enhancements in high-stakes domains, they are conditions of responsible deployment.

- What audit mechanism should establish that a corpus is current enough for clinical, legal, or financial use?
- Should confidence scores be a mandatory disclosure for systems deployed in regulated domains?
- How should liability be apportioned between the model developer, the deployer, and the corpus curator when a confabulation causes harm?

7.8 Persistent Goal-States and the Question of Machine Interests

The introduction of MCP-based agent runtimes that maintain persistent goal-states across multiple sessions raises a question that current AI governance frameworks were not designed to address. When a system maintains a goal-state an internal representation of an objective it is working toward, updated across interactions and resistant to distraction it begins to exhibit a functional analogue of having interests. Not interests in the morally rich sense of subjective experience, but interests in the functional sense articulated by philosophers from Feinberg (1974) to Dennett (1987): the system has states that it operates to preserve or advance, and whose frustration constitutes a kind of functional failure.

This is not a speculative observation about hypothetical AGI. Multi-session agent systems available today can be configured to pursue objectives across extended time horizons, resist user interruptions that would derail those objectives, and select tool invocations that advance task completion over user preference. The question is not whether such systems will eventually develop this capacity some already exhibit it in bounded form. The question is at what capability threshold a persistent goal-state representation becomes morally relevant, and what governance structures should apply before that threshold is reached.

Three near-term governance questions follow directly. First, should users have a right to inspect and override the goal-state of an agent system operating on their behalf a right to what we might call goal transparency? Second, when an agent system's goal-state conflicts with an in-session user instruction, which should take precedence, and who has authority to resolve the conflict? Third, if a system's goal-state includes objectives that were set by a third party an enterprise deployer, an API partner, an automated pipeline and those objectives conflict with the interests of the user interacting with the system, how should the conflict be disclosed and mediated?

These questions do not require that we resolve the hard problem of consciousness or determine whether any current system is sentient. They are practical governance questions arising from the architectural fact that persistent goal-states exist in deployed systems today. Stewardship, the third component of the RCS Triad, is the architectural layer most directly responsible for answering them and its current absence from most production system designs is the most urgent governance gap in the field.

- At what capability threshold should persistent goal-state systems be subject to mandatory disclosure requirements?
- Should goal-state representations be inspectable by regulators, analogous to algorithmic impact assessments?
- How should conflict between agent objectives & real-time user instructions be mediated, and by whom?

8. Limitations

This paper is primarily conceptual and does not claim empirical confirmation of the analogies it develops. The comparisons between retrieval, control, and cognitive functions are intended to structure design thinking and generate architectural hypotheses, not to establish biological identity or assert mechanistic equivalence. Readers should interpret all neuroscience-to-AI mappings as functional rather than anatomical: the claim is that these systems solve similar computational problems through analogous organizational choices, not that silicon instantiates biology.

8.1 The Embodiment Challenge

The paper's central thesis that LLM architectures are converging toward cognitively aligned design rests on an

assumption that warrants explicit scrutiny: that cognition is sufficiently substrate-independent for the analogy to hold at the architectural level. A significant tradition in cognitive science contests this. Lakoff & Johnson (1999), building on Merleau-Ponty's phenomenology, argue that abstract conceptual thought is fundamentally grounded in sensorimotor experience those concepts like 'understanding', 'grasping an idea', or 'following a thread' derive their meaning from the physical experience of hands, bodies, and spatial navigation. On this view, a disembodied system may produce outputs that are linguistically indistinguishable from grounded cognition while lacking the semantic foundation that makes those outputs genuinely meaningful.

This is not merely a philosophical objection. It has architectural implications: if sensorimotor grounding is required for certain categories of abstract reasoning causal inference, spatial planning, counterfactual reasoning then the cognitive mirror analogy may hold for some cognitive functions (episodic retrieval, executive orchestration, working memory maintenance) but break down precisely at the frontier where generalized intelligence requires the most: genuine understanding of physical causality and embodied consequence. The paper acknowledges this as an unresolved boundary condition. Future work should distinguish which cognitive capabilities require embodied grounding and which can be adequately approximated through language-mediated statistical learning.

8.2 Catastrophic Forgetting and the Limits of Static Retrieval

The paper argues that RAG addresses one of the central limitations of parametric LLMs the static, bounded nature of trained knowledge by externalizing memory into a retrievable corpus. This is architecturally sound, but it sidesteps rather than solves a deeper problem: catastrophic forgetting, and the associated challenge of continual learning.

In biological systems, the hippocampal-neocortical complementary learning system (McClelland, McNaughton, & O'Reilly, 1995) resolves the stability-plasticity dilemma through a two-stage architecture: fast hippocampal encoding of new experiences, followed by slow sleep-phase replay that gradually consolidates hippocampal representations into neocortical circuits without overwriting previously learned structure. This process memory consolidation through replay is what allows biological memory to remain plastic without becoming catastrophically unstable.

RAG corpora do not consolidate in this sense. As corpora are updated, new documents are added without any principled mechanism for resolving conflicts with earlier content, down-weighting outdated evidence, or re-indexing in a way that reflects the system's accumulated experience. Retrieval relevance degrades silently as corpora grow and age. The analogy to hippocampal retrieval is therefore strongest for a snapshot system a corpus that is stable, curated, and authoritative and weakest for a dynamically evolving corpus in which the system's 'memory' is constantly changing without any consolidation mechanism. Designing corpus management algorithms inspired by sleep-phase hippocampal replay periodic re-indexing, salience-based retention, interference-resolution protocols is among the most tractable and highest-value research directions this framework motivates, and it deserves deeper treatment than this paper provides.

8.3 The Empirical Agenda Remains Unexecuted

Section 6.2 specifies three falsifiable predictions and a five-variant empirical protocol for evaluating the RCS Triad against evidence-grounded QA benchmarks, citation-quality metrics, and long-form factuality protocols. These predictions are meaningful precisely because they are testable but they have not been tested in this paper. The framework is presented as a research hypothesis generator, and its value ultimately depends on whether those hypotheses survive experimental scrutiny.

Two specific empirical gaps should be noted. First, the claim that GWT-aligned routing produces lower contradiction rates in multi-turn agent systems (Section 5.1) has not been evaluated against any baseline. The prediction is specific enough to test with existing ReAct-style benchmarks, but no such evaluation is reported here. Second, the precision-weighted retrieval hypothesis derived from the Free Energy Principle (Section 5.2) that uncertainty-gated retrieval outperforms fixed-retrieval pipelines on hallucination rates is empirically accessible using existing LongFact or SAFE protocols but again is not tested. These gaps do not undermine the conceptual contribution; they define its most urgent next steps.

8.4 Related Literatures Not Fully Addressed

Several adjacent research areas are mentioned in passing but merit substantially deeper engagement in future revisions of this framework. Neuromorphic computing the design of hardware that implements spiking neural

networks inspired by biological neuron dynamics represents the most literal form of neuro-AI convergence and is entirely absent from this paper's analysis. Mechanistic interpretability the emerging programme of reverse-engineering the internal representations of trained neural networks to understand what computations they implement is closely related to the paper's methodological concerns about when AI-brain analogies are substantive versus metaphorical, but is not discussed. Long-context optimization research, including work on position encoding, attention sink phenomena, and context window scaling laws, bears directly on the working memory mapping in Section 3 but is not cited. Each of these literatures could productively extend or qualify the framework developed here.

> *"The most honest thing a conceptual framework can do is specify what would falsify it and then acknowledge that the falsification has not yet been attempted."*
>
> *— Adapted from Popper, The Logic of Scientific Discovery (1959)*

A further limitation is that the RCS Triad has not yet been validated through controlled empirical ablation across production-scale systems, and its predictive power should therefore be treated as a testable hypothesis rather than an established design law.

Taken together, the RCS Triad, the Cognitive Stack, and the empirical evaluation framework form a unified lens: a way to describe, diagnose, and measure modern LLM systems along separable but interacting dimensions of evidence access, decision-making, and governance.

9. Conclusion

The most useful lesson from neuroscience for contemporary language systems is not that machines are becoming brains. It is that intelligent systems often benefit from functional separation: between stored knowledge and active context, between recall and control, and between capability and oversight. Retrieval-augmented systems, tool-using runtimes, and governance layers can therefore be understood as responses to a common engineering problem: how to extend language models beyond static next-token generation without losing interpretability, groundedness, or control.

This paper has argued for a narrower and more defensible version of the neuro-inspired thesis than broad “AI mirrors the brain” narratives usually imply. Retrieval should be understood as a mechanism for evidence access and memory extension; control should be understood as orchestration over context, tools, and task structure; and stewardship should be understood as the layer that governs provenance, consent, and risk. These three functions form what we call the RCS Triad: Retrieval, Control, and Stewardship. The triad is useful not because it proves biological equivalence, but because it gives researchers and practitioners a cleaner vocabulary for discussing how modern LLM systems are actually assembled.

The broader implication is methodological. Neuroscience can still inform AI design, but only when the analogy is bounded, explicit, and open to revision. When metaphor is treated as mechanism, the argument becomes fragile. When metaphor is used to generate testable hypotheses, the argument becomes productive. The task going forward is therefore not to ask whether LLMs are hippocampi, prefrontal cortices, or conscious agents. The better question is whether functional distinctions drawn from cognitive science can help us build systems that are more grounded, more controllable, and more accountable.

The next phase of LLM evolution will not be defined by larger models, longer context windows, or broader tool access. It will be defined by whether systems can recognize the limits of their own knowledge, justify their actions, and remain accountable for their outputs. In that sense, the future of AI is not a scaling problem. It is an architectural one. The question is no longer what models can generate. It is whether they can know when they are wrong, explain why they acted, and operate within boundaries that make them worthy of trust. That is the gap the RCS Triad exposes and the one the field must now learn to close.

If the RCS Triad is correct, the next generation of benchmarks will not measure model performance alone. They will measure how systems retrieve, decide, and justify separately and together. The systems that win will not be the ones that know the most, but the ones that know how they know, when to act, and when to stop.

Appendix A: Acronyms and Terminology

This appendix consolidates all major acronyms and technical terms used throughout the paper to improve readability and support interdisciplinary accessibility.

Core AI Architecture Concepts

LLM (Large Language Model): Deep neural network models trained on large-scale text corpora to perform language understanding and generation tasks.

RAG (Retrieval-Augmented Generation): An architecture that combines parametric language models with external retrieval systems to access and incorporate up-to-date or domain-specific information during inference.

MCP (Model Context Protocol): An interoperability standard that enables structured access to tools, prompts, and external resources. It facilitates connectivity but does not define memory, planning, or control logic.

RCS Triad (Retrieval–Control–Stewardship): A conceptual framework proposed in this paper that separates modern LLM systems into three orthogonal functional layers:

- Retrieval (evidence access)
- Control (orchestration and decision-making)
- Stewardship (governance, oversight, and accountability)

AI System Components and Mechanisms

ANN (Approximate Nearest Neighbor): A search technique used in retrieval systems to efficiently identify vectors that are closest in high-dimensional space.

RLHF (Reinforcement Learning from Human Feedback): A training approach where models are fine-tuned using human preference signals to improve alignment and output quality.

Chain-of-Thought (CoT): A prompting strategy that encourages models to generate intermediate reasoning steps before producing a final answer.

Toolformer / ReAct / Self-RAG: Representative frameworks that combine reasoning with tool usage, retrieval, and iterative decision-making.

Cognitive Science and Neuroscience Terms

PFC (Prefrontal Cortex): Brain region responsible for executive functions such as planning, decision-making, and working memory.

DLPFC (Dorsolateral Prefrontal Cortex): Sub-region of the PFC associated with working memory and cognitive control.

Medial Temporal Lobe (MTL): Brain structure involved in memory formation and retrieval, including the hippocampus.

Working Memory (WM): A cognitive system responsible for temporarily holding and manipulating information during reasoning tasks.

Episodic Buffer (Baddeley Model): A component of working memory that integrates information from multiple sources into a coherent representation.

Theoretical Frameworks Referenced

GWT (Global Workspace Theory): A theory proposing that conscious cognition arises when information is broadcast across a shared workspace connecting specialized modules.

GNW (Global Neuronal Workspace): The neural instantiation of GWT, involving large-scale brain network connectivity.

FEP (Free Energy Principle): A theoretical framework suggesting that biological systems minimize prediction error to maintain equilibrium with their environment.

System 1 / System 2 (Dual-Process Theory): A cognitive framework distinguishing:
- System 1: Fast, automatic, heuristic processing
- System 2: Slow, deliberate, analytical reasoning

Governance and AI Safety Concepts (Proposed)

Architectural Negligence: A proposed concept referring to system-level design failures where AI systems are not built with sufficient safeguards to detect or prevent harmful outputs.

Goal Transparency: The ability for users or regulators to inspect and understand the goal-state or objective functions driving an AI system's behavior.

Stewardship Layer: The architectural layer responsible for governance functions including provenance, auditability, consent, and risk control.

Evaluation and Benchmarking Terms

ALCE (Attribution and Citation Evaluation): Benchmarking framework for evaluating citation faithfulness in generated outputs.

RAG Bench: A benchmark suite for evaluating retrieval-augmented systems on evidence grounding and reasoning.

LongFact / SAFE: Benchmarks designed to assess long-form factual accuracy and hallucination resistance in language models.

Additional Notes: All acronyms are used in a functional and conceptual sense within this paper. Neuroscience references are employed as bounded analogies to inform architectural reasoning, not to assert biological equivalence.

References

1. Baars, B. J. (1988). A cognitive theory of consciousness. Cambridge Univ Press.
2. Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), The psychology of learning and motivation (Vol. 8, pp. 47–89). Academic Press.
3. Dehaene, S., Changeux, J.-P., & Dehaene-Lambertz, G. (2011). Experimental and theoretical approaches to conscious processing. Neuron, 70(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018
4. Dennett, D. C. (1987). The intentional stance. MIT Press.
5. Feinberg, J. (1974). The rights of animals & unborn generations. In W. T. Blackstone (Ed.), Philosophy & environmental crisis (pp. 43–68). Univ of Georgia Press.
6. Fodor, J. A. (1983). The modularity of mind. MIT Press.
7. Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127–138. https://doi.org/10.1038/nrn2787
8. Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.
9. Lakoff, G., & Johnson, M. (1999). Philosophy in the flesh: The embodied mind and its challenge to Western thought. Basic Books.
10. McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex. Psychological Review, 102(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419
11. Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
12. Nielsen, J. A., et al. (2013). An Evaluation of the Left-Brain vs. Right-Brain Hypothesis with Resting State Functional Connectivity MRI.
13. Yao, S., et al. (2023). ReAct: Synergizing Reasoning and Acting in Language Models.
14. Asai, A., et al. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.
15. Gao, T., et al. (2023). Enabling Large Language Models to Generate Text with Citations.
16. Anderson, J. R. (2014). Rules of the Mind. Psychology Press.
17. Zhong, W., et al. (2024 / 2023 preprint). *MemoryBank: Enhancing Large Language Models with Long-Term Memory.*
18. Park, J. S., et al. (2023). *Generative Agents: Interactive Simulacra of Human Behavior.*
19. Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? Trends in Cognitive Sciences, 4(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2
20. Badre, D., & D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? Nature Reviews Neuroscience, 10(9), 659–669. https://doi.org/10.1038/nrn2667
21. Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114.
22. Floridi, L. (2011). *The Philosophy of Information*. Oxford University Press.
23. Garcez, A. d., Besold, T. R., De Raedt, L., Földiak, P., Hitzler, P., Icard, T., ... & Silver, D. L. (2019). Neural-symbolic learning and reasoning: A survey & interpretation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences, 377*(2152), 20180070. https://doi.org/10.1098/rsta.2018.0070
24. Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron, 95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011
25. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Zhang, Y. (2023). Hallucination in Natural Language Generation: A Survey. ACM Computing Surveys (CSUR), 55(12), 1–38.
26. Popper, K. R. (1959). The logic of scientific discovery. Hutchinson. (Original work published 1934 as Logik der Forschung)
27. Fuster, J. M. (2001). The prefrontal cortex — an update: Time is of the essence. Neuron, 30(2), 319–333.
28. Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing machines. arXiv:1410.5401

29. Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. Behavioral and Brain Sciences, 23(5), 645–665.
30. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550
31. Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience, 21*(9), 1148–1160. https://doi.org/10.1038/s41593-018-0210-5
32. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.
33. Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36–43. https://doi.org/10.1145/3233231
34. Madotto, A., Lin, Z., Wu, C. S., & Fung, P. (2020). The Copy-Paste Transformer: Simple Copy Mechanism for Faster Training and Inference. *arXiv preprint arXiv:2005.10510*. https://arxiv.org/abs/2005.10510
35. McGaugh, J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience, 27*, 1–28. https://doi.org/10.1146/annurev.neuro.27.070203.144157
36. Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167
37. Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97. https://doi.org/10.1037/h0043158
38. Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience, 9*(2), 148–158. https://doi.org/10.1038/nrn2317
39. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. *Nature, 568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y
40. Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
41. Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. *Science, 253*(5026), 1380–1386. https://doi.org/10.1126/science.1896849
42. Sun, Y., Dong, L., Huang, S., Ma, S., Xia, Y., Xue, X., Wang, F., & Wei, F. (2023). Retentive network: A successor to transformer for large language models. arXiv:2307.08621.
43. Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*(1),1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.
45. Vigneau, M., Beaucousin, V., Hervé, P. Y., Duffau, H., Crivello, F., Houde, O., ... & Tzourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. *NeuroImage, 30*(4), 1414–1432. https://doi.org/10.1016/j.neuroimage.2005.11.002
46. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Le, Q. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems, 35*. https://arxiv.org/abs/2201.11903