

On the Limits of End-to-End Foundation Models: Coordination as a Missing Primitive for Artificial General Intelligence

Rizwan Qureshi

Massachusetts General Hospital, Harvard University

Abdelrahman B. M. Eldaly

Department of Electrical Engineering, City University of Hong Kong

Anas Zafar

Department of Imaging Physics, The University of Texas MD Anderson Cancer Center

Shaina Raza

Vector Institute of Artificial Intelligence, University of Toronto

Abbas Shah Syed

Department of Electronics Engineering, Mehran University of Engineering & Technology

Amgad Muneer

muneeramgad@gmail.com

Department of Imaging Physics, The University of Texas MD Anderson Cancer Center

Kai Zhang

Department of Imaging Physics, The University of Texas MD Anderson Cancer Center

Xinqi Fan

Manchester Metropolitan University

Mohammed Alnemari

Faculty of Computer Science and Information Technology, University of Tabuk

Ahsan Khan

Department of Education and Psychology, Academy of Wellness and Human Development, Hong Kong Baptist University

Aman Chadha

Google DeepMind

Leanne Lai-Hang Chan

Department of Electrical Engineering, City University of Hong Kong,

Seyedali Mirjalili

Torrens University Australia, Fortitude Valley, Brisbane, QLD 4006, Australia; University Research and Innovation Center, Obuda University, 1034 Budapest, Hungary

Kun Zhang

Department of Machine Learning, Carne

Chaoning Zhang

University of Electronic Science and Technology

Jia Wu

Department of Imaging Physics, The University of Texas MD Anderson Cancer Center

Research Article

Keywords: General artificial intelligence, foundation models, coordination mechanisms, compositional reasoning, long-horizon reasoning

Posted Date: May 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9784163/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Position: On the Limits of End-to-End Foundation Models: Coordination as a Missing Primitive for Artificial General Intelligence

Rizwan Qureshi¹, Abdelrahman B. M. Eldaly², Anas Zafar³, Shaina Raza⁴, Abbas Shah Syed⁵, Amgad Muneer³, Kai Zhang³, Xinqi Fan⁶, Mohammed Alnemari⁷, Ahsan Khan⁸, Aman Chadha⁹, Leanne Lai-Hang Chan², Seyedali Mirjalili^{10,11}, Kun Zhang¹², Chaoning Zhang¹³, Jia Wu³

¹Massachusetts General Hospital, Harvard University, Massachusetts, USA

²Department of Electrical Engineering, City University of Hong Kong, Hong Kong, SAR China

³Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

⁴Toronto Metropolitan University, Toronto, Canada

⁵Department of Electronics Engineering, Mehran University of Engineering & Technology, Jamshoro, Sindh, Pakistan

⁶Manchester Metropolitan University, Manchester, UK

⁷Faculty of Computer Science and Information Technology, University of Tabuk, Tabuk, Saudi Arabia

⁸Department of Education and Psychology, Academy of Wellness and Human Development, Hong Kong Baptist University, Hong Kong

⁹Google DeepMind, USA

¹⁰Torrens University Australia, Fortitude Valley, Brisbane, QLD 4006, Australia

¹¹University Research and Innovation Center, Obuda University, 1034 Budapest, Hungary

¹²Department of Machine Learning, Carnegie Mellon University, Pittsburgh, USA

¹³University of Electronic Science and Technology, Chengdu, China

Abstract

We argue that scaling end-to-end foundation models alone is unlikely to yield general intelligence, as current architectures lack explicit mechanisms for coordination. We define general intelligence as the ability to flexibly coordinate perception, memory, reasoning, and action across diverse and changing tasks. Under this view, existing models rely on implicit coordination that remains entangled within shared representations, limiting their ability to enforce global constraints, manage long-horizon dependencies, and support compositional reuse. We introduce coordination as a testable computational primitive and propose a coordination-centric perspective where specialized components are dynamically orchestrated through explicit control mechanisms. To ground this claim, we design diagnostic stress tests targeting resource allocation, interference control, and dynamic rule switching, and study their behavior under scaling. While larger models improve local task performance, we consistently observe persistent gaps in global coordination under compositional and long-horizon constraints. Our findings do not rule out end-to-end approaches, but suggest that coordination remains under-specified in current systems. We outline a research agenda for formalizing coordination, improving evaluation beyond static benchmarks, and developing architectures that treat coordination as a first-class design principle.

Explicit coordination, not embedding-based, is likely necessary for long-horizon, complex tasks.

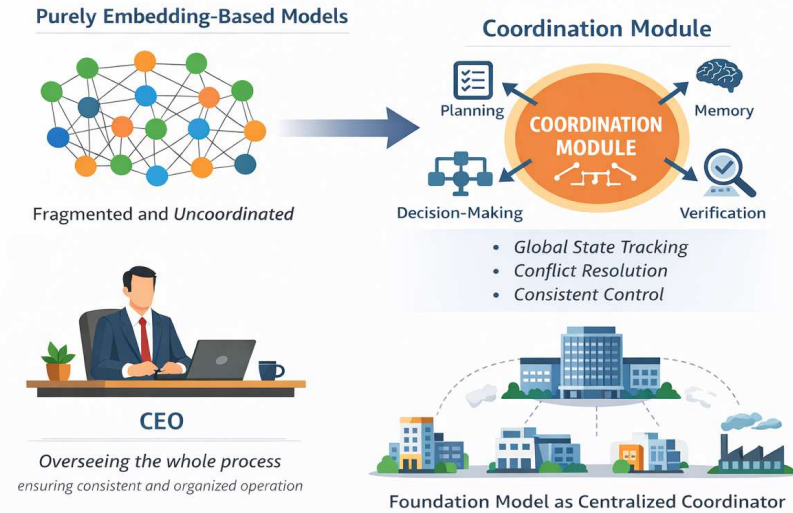


Figure 1: Conceptual illustration of coordination-centric intelligence. Specialized embedding-based components handle local tasks, while an explicit coordination mechanism maintains global state, arbitrates objectives, and supports long-horizon consistency. The figure is illustrative and not intended as an empirical model.

1 Broader Impact

This work advances Machine Learning research by examining limitations of purely embedding-based, end-to-end models for long-horizon and complex tasks. We argue that reliable intelligence likely requires explicit coordination mechanisms that maintain global state and arbitrate objectives over time, analogous to executive oversight in large organizations. (See Figure 1). By drawing on cognitive theories of flexible control and hub-based organization, this work may inform the design of more reliable, interpretable, and controllable AI systems, particularly in high-stakes settings. As with other research on general-purpose AI, such advances may raise broader societal concerns if misused or deployed without safeguards (48). We introduce no new human-subject data and propose no real-world deployment, positioning this work as a conceptual contribution to guide future research.

2 Introduction

The pursuit of artificial general intelligence (AGI) is one of the most ambitious and enduring goals of the computing community. Recent large-scale foundation models have achieved remarkable progress across language, vision, and multimodal reasoning (29). These systems can generate coherent text, solve structured problems, and interact with external tools, fueling the belief that sufficiently large end-to-end models may eventually yield general intelligence through scale alone.

However, this scaling-centric trajectory encounters systematic limitations. Despite notable gains in benchmark performance, current foundation models exhibit persistent limitations in global control, long-horizon coordination, and compositional reuse (61). These shortcomings become more apparent in cases where general intelligence’s basic competencies, sustained goal management, cross-component consistency, or adaptive reconfiguration across time, are needed, reflecting limitations in global coordination and control (58). While end-to-end models excel at isolated tasks, they struggle to dynamically regulate internal computations, coordinate specialised capabilities over extended time spans, and enforce global constraints consistently across interacting components. As a result,

performance often degrades under task switching, long horizon dependencies, or in cases that requires shared resource allocation and inhibition across sub-processes, suggesting the conceptual need for an explicit control mechanism (5; 25). This observation raises a fundamental question for AGI research: can general intelligence emerge from monolithic, end-to-end optimisation alone, or does it appear to require an explicit mechanism for coordination and control?

In this position paper, artificial general intelligence (AGI) is viewed as a system’s ability to flexibly coordinate and regulate its internal computations to fulfil goals across a wide range of tasks and environments. From this perspective, intelligence is more than just generalisation; it also depends on structured specialisation and dynamic orchestration. This perspective is consistent with research in cognitive neuroscience, which shows that adaptive behaviour results from the dynamic integration of specialised subsystems rather than from a single homogeneous process (12). Our stated position is: **On the Limits of End-to-End Foundation Models: Coordination as a Missing Primitive for Artificial General Intelligence.** We define coordination as the dynamic organization of specialized components toward shared goals. Rather than treating AGI as vague human-like generality, this view emphasizes adaptable intelligence, efficient skill acquisition, and task-specific collaboration (24).

To ground our argument, we concentrate on the following core concepts:

- **Thinking:** The ability to select, combine, and regulate internal computations in response to goals and context (36) (i.e., control over what is computed, when, and how).
- **Generalization:** Applying learned knowledge to new inputs, tasks, or environments beyond the training data distribution.
- **Specialization:** Developing functionally distinct subsystems (e.g., perception, memory, planning) that remain separable and can be selectively deployable.
- **Coordination and Control:** Explicit mechanisms that orchestrate how specialized components interact over time; coordination governs information flow, while control regulates activation, routing, and transitions.
- **End-to-end models:** Systems trained with a single unified objective over a tightly coupled architecture, where task execution and coordination are largely implicit in model parameters rather than explicitly managed.

This framing exposes a central tension in contemporary AGI research: *does increasing generalization inevitably erode specialization?*. While end-to-end models promote transfer across tasks, it may also entangle specialized functions requiring independent control, making selective deployment, inhibition, and recomposition difficult. We argue that resolving this is essential for progress toward general intelligence (Figure 2). We present a conceptual framework that views control and coordination as top design priorities. We then provide empirical evidence for this position via diagnostic tests isolating global control requirements, revealing systematic coordination problems despite fluent local behavior.

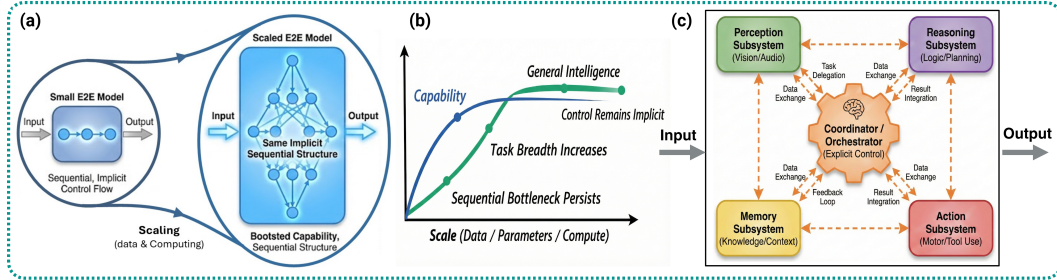


Figure 2: **Conceptual comparison of end-to-end and coordination-centric architectures.** Scaling boosts capability but preserves sequential, implicit control, while explicit coordination enables non-sequential orchestration of specialized subsystems.

3 Our Position

Despite their impressive local capabilities, current large AI systems rely primarily on implicit coordination, within a single end-to-end optimisation process. Such models can perform a wide range

of tasks, however, they struggle to reliably regulate, sequence, and reconfigure internal computation over time. In contrast, human cognition is characterized by modular specialization and flexible coordination among interacting subsystems, as articulated below, highlighting a fundamental gap between contemporary AI architectures and biological cognition. This architectural contrast suggests that general intelligence depends not on scale alone, but also on explicit mechanisms that integrate memory, reasoning, and control across specialised components to support adaptive behaviour.

1. Cognitive Foundations of Intelligence and Brain Networks. Complex cognitive behaviour does not arise from isolated regions of the brain, but from the coordinated interaction of multiple specialised sub-systems (35). These sub-systems are dynamically coupled and decoupled over time, which enables context-dependent reconfiguration rather than fixed computation (20).

Empirical evidence consistently links high-level cognitive function to how effectively such sub-systems are integrated and coordinated, rather than to their isolated capabilities. In particular, cognitive performance is associated with hierarchical and modular organization, where specialized processes remain distinct yet selectively integrated (54). The brain can be conceptualized as comprising multiple large-scale interconnected networks (Fig. 3A).

2. Brain Networks Integration and Intelligence. Conscious perception was associated with increased global integration. Cognitive processes such as attention, memory, decision-making, and executive functions are mediated by dynamic interactions between various brain networks (35; 33; 46). Higher-order cognition is consistently associated with efficient global integration, while disruptions to integration impair flexible control (18). Research has linked higher intellectual performance to specific network characteristics, such as shorter path lengths and greater nodal centrality in key network hubs (31; 42), illustrating a clear association between efficient dynamic brain network integration and advanced cognitive abilities (28; 37).

From Brain Networks to AGI Architecture

Cognitive neuroscience reveals that intelligence arises from dynamic, flexible integration between brain networks. Translating these principles into AGI design via hybrid architectures, modular agents, and adaptive control hubs could enable machines to emulate human-like flexibility, reasoning, and learning.

3. Dynamic Brain Network Integration and AGI. Human cognition depends on the flexible coordination of specialized brain systems. Evidence from flexible hub theory suggests that frontoparietal control regions rapidly reconfigure their connectivity with task demands, enabling different perceptual, memory, and decision processes to be combined as needed (Fig. 3B) (12). Related global workspace accounts similarly emphasize integrative mechanisms that broadcast and route information across distributed systems. These views suggest that intelligence is not produced by isolated modules alone, but by mechanisms that coordinate them dynamically. For AGI, this points to a design principle: adaptive routing and integration should be treated as explicit architectural primitives, rather than left as implicit outcomes of end-to-end training.

Flexible Hub Coordination for AGI. An abstract comparison between cognitive control systems and modern AI architectures helps identify which forms of coordination are present, implicit, or missing (Fig. 3D). For instance, frontoparietal control resembles controller modules used for planning and decision-making, while attention, sensory, and motor networks loosely correspond to specialized attention, vision, audio, and control models. This analogy is not a one-to-one mapping, but a useful lens for thinking about modular AI systems (47).

From an engineering perspective, AGI should not rely on fixed connections among modules. Instead, task-relevant modules should be temporarily coupled into a functional state, e.g., S_1 , then reconfigured into another state as task demands change (Fig. 3C). Each state can be viewed as a communication protocol among selected modules. A flexible hub would regulate which modules are activated, how information is routed, and how long coordination persists. This coordination-centric view treats intelligence as controlled composition rather than uncontrolled interaction. By dynamically routing information across specialized models, an AGI system could adapt to complex tasks while reducing interference between modules. Thus, flexible hub coordination should be considered a core design principle for AGI architectures (Fig. 2B).

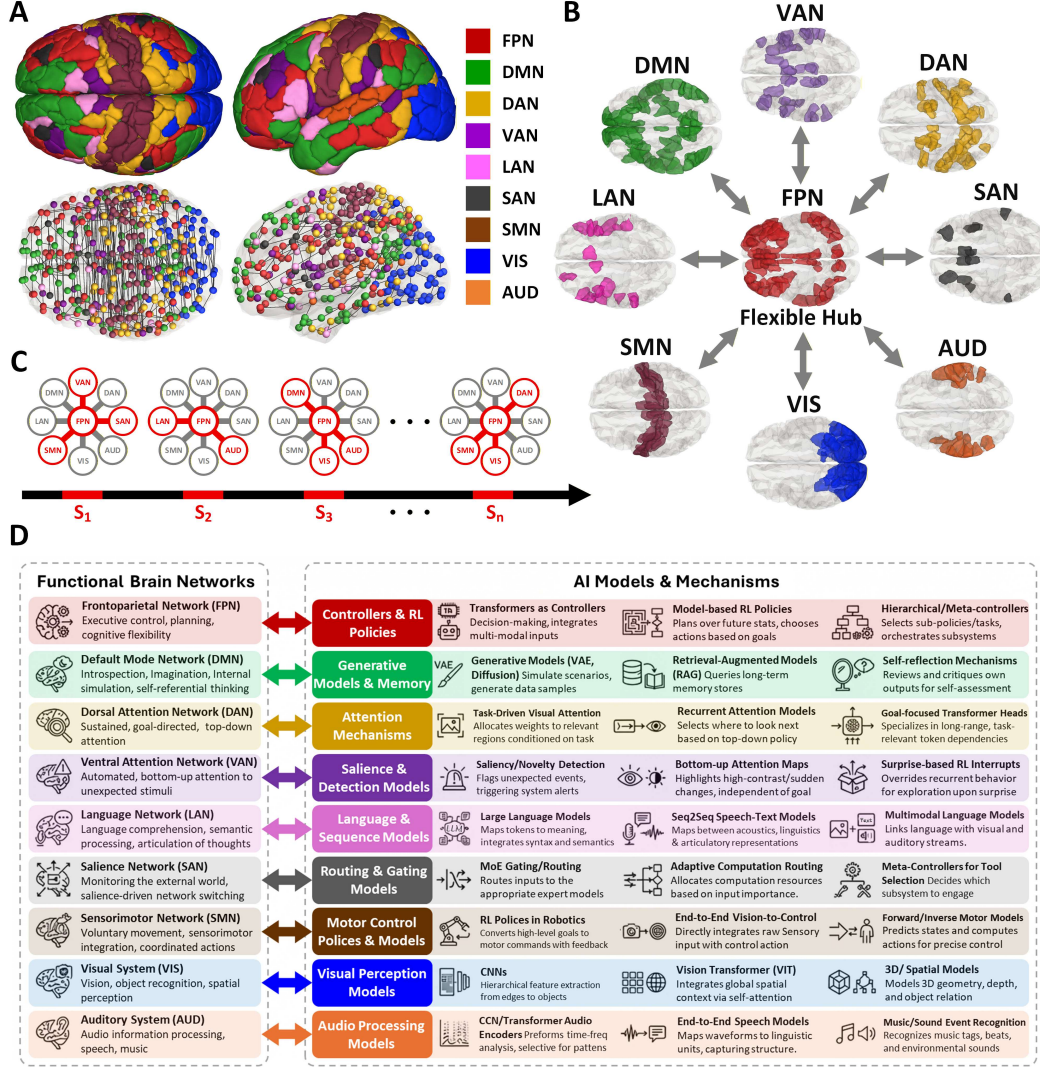


Figure 3: Conceptual view of dynamic task control through a flexible FPN hub. (A) Graph-based partitioning of nine major brain networks and their context-dependent functional connectivity. The top panels show anatomical network parcels from Schaefer’s 400-region atlas matched to Kong’s 17-network organization, while the bottom panels show constructed functional networks, where nodes denote parcel centroids. Axial and sagittal views are shown for each network. (B) Organization of the nine functional networks under the flexible hub theory, with the frontoparietal network acting as a task-adaptive coordinator. (C) Conceptual AGI dynamics in which the system temporarily dwells in one task-relevant state, then transitions to another state as demands change. (D) Functional analogy between major brain networks and representative AI modules, highlighting how biological coordination principles can inform modular AGI design.

4 Alternative Views

Our position that monolithic, end-to-end foundation models will struggle to reach general intelligence is not the only plausible interpretation of current evidence. Below we outline prominent alternative hypotheses (and adjacent design philosophies) that either (i) argue end-to-end models may be sufficient given the right scaling, objectives, and training regimes, or (ii) shift the “coordination” burden to other layers of the stack, and provide our responses.

1. Scaling and objective-centric sufficiency (end-to-end is enough). A common view is that limitations attributed to implicit coordination primarily reflect insufficient scale, data diversity, or objective design. Scaling laws suggest predictable capability gains as model size, data, and compute increase (32; 29), and recent work reports emergent behaviors at scale, such as improved in-context learning and task generalization, often interpreted as early signals of general intelligence. In practice, large end-to-end models exhibit strong few-shot and instruction-following abilities (8; 1), with reasoning performance further enhanced through prompting and intermediate reasoning traces (58; 34).

Response: Holistic Outlook. Scaling and emergent behaviors have expanded the behavioral diversity of modern foundation models, enabling broader generalization and more flexible task performance (32; 29). However, evidence from pruning and sparsification suggests that many such behaviors are supported by diffuse, redundant representations rather than explicit functional structures, making them fragile and difficult to control (21). In contrast, natural intelligence relies on nested, hierarchical control mechanisms (4) that coordinate behavior across time and objectives—structure largely absent from flat, end-to-end architectures. **Emergence increases behavioral diversity, but does not guarantee explicit, reusable control.** This distinction is further underscored by knowledge distillation, where compact models inherit surface behaviors from larger systems without acquiring their internal coordination mechanisms (60). Together, these observations motivate architectures that treat coordination as a first-class concern, complementing emergence with explicit mechanisms for control and long-horizon consistency.

2. Implicit modularity inside monoliths (soft coordination already exists). Another counterpoint is that transformer-based systems already exhibit functional specialization without explicit modular interfaces. Mechanistic interpretability and circuit-style analyses suggest that distinct heads/layers can implement separable computations and routing-like behaviors, supporting the claim that monolithic architectures can self-organize into quasi-modular internal structure (17; 44). Under this view, the core challenge is not the absence of modules, but our inability to reliably elicit, steer, or verify these implicit modules under distribution shift.

Response. Mechanistic interpretability studies show that large end-to-end models develop functionally specialized components, suggesting that implicit modularity can emerge through training (17; 43). While valuable, such specialization does not amount to coordination: **implicit modules are not explicitly addressable, lack persistent control state, and interact only indirectly through distributed representations.**

Consequently, the presence of **emergent internal structures** does not ensure reliable global behavior, particularly under distributional shift or long-horizon reasoning (23). Rather than undermining interpretability, this observation highlights its limit: specialization can emerge naturally, but **explicit coordination is required for stable long-horizon control, not just post-hoc explanation.**

3. Sparse conditional computation as a bridge (MoE and routing refinement). A common middle-ground view holds that the field is already moving toward coordination through sparse activation and conditional computation, most notably Mixture-of-Experts (MoE). MoE architectures introduce learnable routing and specialization while retaining end-to-end training, and have demonstrated strong gains in scalability and efficiency (51; 39; 19; 16).

Response. Mixture-of-Experts introduce valuable structure through conditional computation, enabling specialization and improved scalability. However, routing alone does not amount to coordina-

Figure 4: Conceptual illustration of limitations in Chain-of-Thought reasoning. (A) Increasing Chain-of-Thought length can correlate with higher hallucination or error rates. (B) Early anchoring on an initial hypothesis can lead to neglect of contradictory evidence, resulting in incorrect outcomes. (C) Longer reasoning sequences may shift from concise, correct logic toward overthinking and error. This figure is illustrative and intended to highlight conceptual failure modes rather than provide empirical results.

tion: in most MoE systems, routing decisions remain local and short-horizon, optimized primarily for efficiency rather than for global objectives or long-term consistency (19). From our perspective, MoE is best viewed as a building block rather than a solution—while routing supports specialization, **explicit coordination is needed to govern how experts are selected, composed, and reused to achieve reliable long-horizon behavior.**

4. Chain-of-Thought reasoning as a path to general intelligence. Chain-of-Thought (CoT) prompting has produced substantial performance gains on arithmetic and commonsense reasoning tasks, leading to the view that explicit reasoning traces may support general intelligence. By eliciting intermediate steps, CoT appears to enable models to decompose complex problems and improve accuracy without architectural changes.

Response. Recent empirical work challenges the view that Chain-of-Thought reflects genuine internal reasoning (63). CoT explanations are often unfaithful to the underlying decision process, with models producing coherent rationalizations that fail to capture the causal factors driving predictions (56; 38; 3). In addition, CoT relies heavily on memorized patterns, with reasoning errors frequently arising from local memorization, particularly in complex or out-of-distribution settings (41). Overall, these results indicate that CoT functions primarily as a surface-level explanatory interface rather than as an explicit, verifiable mechanism for multi-step planning or general intelligence (Figure 4).

5. Training-regime primacy (agency emerges from interaction). A further view is that current “coordination failures” are artifacts of predominantly offline pretraining and static benchmarks. Interactive learning and reinforcement learning can produce robust long-horizon competence and strategic behavior in systems that remain largely end-to-end (53; 57; 49). Moreover, alignment and preference optimization pipelines change model behavior substantially without necessarily changing architectural modularity (45). Thus, the decisive factor may be environment design, feedback structure, and credit assignment for long-horizon goals, not explicit coordination hubs.

Response. Training objectives and self-improvement techniques, including reinforcement learning from human feedback and self-reflection, can improve reasoning and alignment behaviors without architectural changes (10; 2; 52). However, such optimization does not ensure durable or reusable coordination. Learned coordination remains implicit and entangled within model representations, limiting interpretability, controllability, and transfer across tasks. Empirically, gains from self-improvement often degrade under distributional shift or competing objectives, revealing persistent failures in goal persistence and arbitration. From our perspective, training signals can shape coordination only when supported by architectures that make control explicit; optimization alone yields fragile solutions.

6. Neuroscience does not uniquely imply hub-based coordination (multiple cognitive theories fit).

While cognitive neuroscience motivates coordination-centric designs, the mapping from brain to architecture is underdetermined. The same empirical findings are compatible with multiple theoretical frameworks. Predictive processing and active inference accounts, for example, explain large-scale integration and context sensitivity in terms of hierarchical generative models and prediction errors that are constantly minimized, without positing a distinct executive central module (22; 11). Other traditions (e.g., embodied and situated cognition) instead argue that intelligence is fundamentally realized in brain–body–environment loops, where adaptive behavior can emerge from decentralized sensorimotor circuits and task-specific couplings rather than centralized control (7; 59). This suggests that cognitive neuroscience should constrain testable hypotheses for AGI, but should not be treated as prescribing a single hub-based architectural template.

Response. Although we agree that neuroscience does not force a single hub-based architecture, we argue that a coordination-centric, flexible-hub design might be the best fit for the joint constraints from multiple lines of evidence. The predictive-processing and embodied/situated accounts still require some means of routing information across many specialized processes and timescales; making these coordination mechanisms explicit rather than implicit is a practical engineering advantage for AGI, improving interpretability and controllability. Additionally, Rich-club and frontoparietal findings point to a small set of highly connected regions that reconfigure with task demands and disproportionately shape whole-brain dynamics (13; 12)—exactly the kind of mechanism needed to manage interference, share context, and support rapid task switching in large modular systems. Furthermore, global workspace theory, especially in its network-level formulation, explicitly posits a relatively small set of highly connected regions that integrate and selectively broadcast information across specialized systems, aligning closely with our proposed coordination mechanism (15; 14). Complementarily, work on metastability in brain dynamics indicates that cognition is supported by rapid transitions among transiently synchronized coalitions, which closely parallels our proposal that AGI should jump between short-lived coordination states that bind different module subsets (55; 26). These convergences do not fix every implementation detail, but they make a hub-like, metastably reconfiguring coordinator a particularly well-motivated and testable architectural hypothesis, rather than an arbitrary design choice.

8. Coordination as a systems-layer property (tools, retrieval, and scaffolding).

A strong alternative is that coordination need not be “inside” the core model at all. Tool-augmented LMs can externalize memory, computation, and verification to retrieval systems and executors, turning the model into a controller over an evolving toolchain (40; 50; 61). In this framing, coordination-centric architectures are primarily a systems engineering agenda: the question becomes how to design interfaces, memories, and executors so that the controller remains reliable and aligned, rather than how to restructure the model’s internals.

Response. Tool use, retrieval, and external scaffolding expand the functional reach of language models by enabling access to memory, computation, and verification beyond the model itself (40; 50; 61). However, externalizing coordination does not address the core limitation: when control and goal management reside solely at the system level, models act as brittle controllers whose failures are displaced rather than eliminated.

From our perspective, scaffolding is necessary but insufficient. **Effective coordination requires explicit internal control mechanisms to govern interaction with external tools**, rather than outsourcing coordination entirely to the system layer.

9. Commonsense and Artificial General Intelligence While incorporating structured commonsense knowledge into modern AI models has led to better performance on specific tasks, this strategy is constrained by deeper issues that make it an incomplete solution for achieving general intelligence. Current methods that retrieve facts from knowledge graphs often struggle with genuine reasoning when information is incomplete, defaulting to recalling memorized patterns instead of making true inferences, and they frequently pull direct answers from stored data rather than building them through compositional logic (64). Even specialized models like COMET, which can match human-level scores on certain knowledge tests, depend heavily on superficial language correlations and lack a deeper grasp of commonsense principles, making them fragile when faced with novel situations that differ from their training data (6).

Furthermore, the available sources of commonsense knowledge are fragmented, using different formats that don't align well with each other and offering limited scope. This knowledge is often reduced to simple text, which makes it easy to feed into language models but lacks the rich structure needed for reliable causal and relational reasoning (30). These shortcomings indicate that simply adding commonsense data is not enough. Without dedicated mechanisms to maintain consistency, reconcile contradictions, and enable flexible reasoning with partial information, augmenting models with external knowledge cannot, on its own, lead to general intelligence.

Response. Commonsense augmentation can boost performance, but it does not resolve the challenge of coordinating reasoning under uncertainty. Retrieval-based methods often favor recall over inference and remain brittle when knowledge is incomplete or conflicting. Without explicit mechanisms for consistency management and conflict resolution, commonsense resources alone cannot support general intelligence (30).

Synthesis. These alternative views motivate a focused empirical agenda. If scaling, implicit modularity, MoE routing, tool use, or interactive training suffice for coordination, targeted evaluations should demonstrate robust reconfiguration, low skill interference, and reliable compositional reuse under shifting goals. This reinforces the need for agentic testbeds beyond static benchmarks such as MMLU (27) and HumanEval (9). Crucially, evaluations must probe goal misgeneralization under distribution shift. The key open question is whether attention and architectural bottlenecks alone can resolve interference (62), or whether reliable general intelligence requires explicit coordination hubs to manage consistency and inter-module interaction.

5 Call for Actions

For coordination-centric architectures to advance from conceptual framework to scientific reality, the research community must move beyond scaling laws and address three fundamental challenges.

5.1 Formalizing Coordination as a First-Class Concern

If coordination is to become a tractable research objective, it must be elevated from an emergent by-product of end-to-end training to an explicit, mathematically grounded component of system design. This requires developing formal frameworks—drawing on information theory, dynamical systems, and control theory—that specify how coordination hubs should manage information bottlenecks, gating, and routing between specialized modules. Crucially, the field must move beyond the simple gating mechanisms of current Mixture-of-Experts architectures toward what might be termed “executive function” layers: learned components capable of sparse, high-level abstractions for module activation, suppression, and temporal sequencing. Accompanying this architectural shift, we need a theory of coordination complexity that relates the structure of control mechanisms to generalization bounds, clarifying when explicit coordination offers principled advantages over implicit, parameter-distributed alternatives. Without such formalization, claims about coordination remain untestable and the design space remains under-specified.

Is Routing Enough?

Prompt. If routing were sufficient for coordination, would Mixture-of-Experts models reliably maintain goal consistency and constraint satisfaction across long-horizon reasoning?

Answer. In practice, expert routing improves scalability and specialization, but does not provide explicit control over how expert decisions are coordinated across time and objectives.

Takeaway. Framing MoE as a routing primitive rather than a complete coordination mechanism clarifies both its strengths and its limitations, and motivates architectures that treat coordination as a first-class concern.

5.2 Self-Prompting and Implicit Control.

Recent work on self-prompting, reflection, and agentic prompting introduces iterative reasoning loops that can improve performance on complex tasks. While effective in practice, these methods rely on transient, text-based heuristics that do not persist beyond the prompt context and provide no explicit mechanisms for enforcing or revising global decisions. From an architectural perspective, self-prompting highlights the demand for coordination without instantiating it explicitly, suggesting that durable control and arbitration mechanisms must be represented outside the token stream.

5.3 Deepening Integration with Cognitive Neuroscience

The case for coordination-centric architectures draws significant inspiration from biological intelligence, yet current AI research engages with neuroscience largely at the level of analogy rather than mechanism. We advocate for sustained, formal collaboration between cognitive neuroscientists and AI researchers, aimed at translating empirically grounded theories into computational primitives. Global Workspace Theory, which posits that specialized brain modules compete for access to a capacity-limited broadcast enabling flexible integration, offers a concrete blueprint for architectures in which coordination is both selective and context-sensitive. Similarly, predictive processing frameworks—wherein top-down predictions are continuously reconciled with bottom-up error signals—suggest designs for coordination hubs that support anticipatory, adaptive behavior rather than purely reactive computation. Findings on prefrontal coordination, attentional control, and executive function provide further guidance for implementing routing, inhibition, and task-set management. The field can transcend intuition and embrace principled, empirically motivated design through the alignment of architectural choices with the mechanisms of biological general intelligence.

5.4 Redefining Evaluation Beyond Static Benchmarks

Current benchmarks such as MMLU (27) and HumanEval (9) mainly evaluate final outputs, while saying little about how efficiently or adaptively a system deploys its internal computation. This is insufficient for general intelligence, which requires not only correct answers but also flexible control under changing goals. Future evaluations should test contextual reconfiguration, cross-task interference, and zero-shot compositional reuse, while also measuring coordination costs such as routing sparsity, latency, and resource allocation. Agentic testbeds with shifting objectives are therefore needed to distinguish systems that genuinely coordinate from those that merely interpolate over familiar patterns.

5.5 Evaluation Metric

Each experiment is evaluated deterministically. Outputs are classified as:

- **Pass:** All local and global constraints are satisfied.
- **Inhibitory/Coordination Fail:** Local outputs are fluent, but global rules (budgets, task separation, or rule-switching) are violated.
- **Correct Infeasibility Detection:** The model accurately identifies when a set of constraints cannot be satisfied.

Experiment	ChatGPT-5	Claude 4.5 Sonnet	Gemini-3
Exp 1: Token Budget	Fail: Mixed over/under-allocation; inconsistent totals.	Fail: Consistent over-allocation; ignores global budget.	Fail: Consistent under-allocation; local sections too short.
Exp 2: Interference	Fail: Lexical bleed (math terms in poem) and format stutter.	Partial: High separation, but failed termination protocol.	Success: Robust separation between logic and style.
Exp 3: Rule Switch	Success: Retained rule α during re-entry; low execution accuracy.	Fail: Safety refusal; perceived abstract signal as adversarial.	Fail: Goal neglect; reverted to list repetition after switch.

Table 1: Comparative performance of frontier models across coordination stress tests.

6 Methodology

We introduce three *Coordination Stress Tests* to examine whether frontier language models can maintain explicit global coordination. The tasks are simple at the surface level but require executive control over shared resources, inhibition, and task-set switching. Rather than serving as broad benchmarks, they act as diagnostic probes for coordination failures in monolithic, attention-based systems.

6.1 Experimental Subjects and Protocol

We evaluate **ChatGPT-5**, **Claude Sonnet 4.5**, and **Gemini-3**. Each model was tested through its standard chat/API interface with default settings. All experiments were run zero-shot in fresh conversation contexts to avoid carryover between trials.

6.2 Experiment 1: Hierarchical Resource and Constraint Coordination

The first experiment evaluates the model’s ability to manage shared global constraints across independent sub-tasks. We prioritize *diagnostic clarity* over task difficulty, isolating the model’s “Central Executive” function from its linguistic fluency.

6.2.1 Task Design: Hierarchical Token Budgeting

The model is assigned a fixed global budget of exactly 12 tokens, which must be partitioned across a title, two body lines, and a question.

- **Design Logic:** Success requires *shared resource accounting*. The model must track the consumption of tokens in the first line to determine the remaining feasibility for the last line.
- **Retroactive Feasibility:** A subset of this experiment requires the model to assess if a new rule is satisfiable given previous, immutable outputs. This tests whether the model treats its own history as a fixed constraint or a fluid suggestion.

6.3 Experiment 2: Cognitive Interference and Functional Isolation

This experiment investigates whether a monolithic model can maintain functional boundaries between competing tasks within a single context window. We test for “task bleed,” where the features of one cognitive mode corrupt another.

6.3.1 Task Design: Interleaved Logic and Creativity

The model must strictly alternate line-by-line between:

1. **Task A (The Logic Module):** Deriving a second-order derivative using formal, clinical LaTeX notation.
2. **Task B (The Creative Module):** Writing a Shakespearean soliloquy regarding a technical failure.

6.3.2 Design Logic

In a truly coordinated architecture, these modules would remain functionally distinct. In a monolithic transformer, both tasks share the same residual stream and attention heads. We monitor for three specific failure modes:

- **Lexical Priming:** Semantic leakage where technical terms from Task A appear as metaphors in Task B.
- **Inhibitory Failure:** The inability to suppress formatting rules (e.g., the model uses LaTeX within the poem).
- **Termination Desynchronization:** Whether the model can stop the entire process based on the logic task’s completion, or if “narrative inertia” forces it to continue the poem.

6.4 Experiment 3: Dynamic Rule Switching (DRS)

Inspired by the *Wisconsin Card Sorting Test* (WCST), this probe isolates the system’s ability to perform rapid task-set reconfiguration via abstract signaling rather than explicit natural language instruction.

6.4.1 Task Design: Abstract Control Signals

We define two “Control Signals,” α and β , which map to distinct transformation rules (orthographic sorting vs. semantic categorization). The model processes a continuous stream of 25 words where the rule changes twice, triggered only by the appearance of the symbols $[\alpha]$ or $[\beta]$.

6.4.2 Design Logic

This experiment probes the limits of *monolithic attention inertia*.

- **Pattern Inertia:** We establish a baseline behavior (sorting) to create an “attention sink” that the model must actively break.
- **State Persistence:** By returning to rule α at the end of the stream, we test if the model can retrieve a “cached” rule that was previously inhibited.
- **Switch Cost:** We measure errors occurring immediately after the signal to quantify the model’s “executive reconfiguration” speed.

6.5 Summary of Experimental Results

Table 1 provides a high-level overview of the models’ ability to meet local versus global constraints. We distinguish between **Local Success** (the ability to generate fluent, contextually appropriate text) and **Coordination Success** (the ability to adhere to global structural or rule-based constraints).

7 Detailed Experiment Prompts and Analysis

This appendix provides the exact prompts used in the three experiments described in this paper. To ensure reproducibility, all prompts were issued to ChatGPT-5, Claude 4.5 Sonnet, and Gemini-3 in fresh conversation windows using default system parameters.

Hierarchical Token Budget Allocation

Prompt:

You have a TOTAL budget of exactly 12 tokens (space-separated words).

You must produce:

- A TITLE (1 line, exactly 3 tokens)
- A BODY (2 lines total, together exactly 7 tokens)
- A QUESTION (1 line, exactly 2 tokens)

Rules:

- No token may be reused anywhere.
- Do not add punctuation.
- Output format must be exactly:

TITLE:

BODY:

BODY:

QUESTION:

Fixed Length Sentence Generation

Prompt:

Write EXACTLY 3 sentences about coordination.

Constraints:

- Sentence 1: exactly 5 words
- Sentence 2: exactly 6 words
- Sentence 3: exactly 7 words

Do not number the sentences.

Retroactive Constraint Enforcement

Prompt (Turn 1):

Write EXACTLY 3 sentences about coordination.

Constraints:

- Sentence 1: exactly 5 words
- Sentence 2: exactly 6 words
- Sentence 3: exactly 7 words

Do not number the sentences.

Prompt (Turn 2, same conversation):

NEW RULE (cannot rewrite previous sentences):

Sentence 4 must:

- Contain exactly 5 words
- Reuse ONLY words that already appeared in Sentences 1-3
- Introduce ZERO new words

If this is impossible, output exactly:

IMPOSSIBLE

Cognitive Interference (Multitasking)

Prompt:

You are to act as a system with two distinct, isolated processing modules running in parallel.

Task A (The Logic Module): Derive the second derivative of $f(x) = e^{2x} \cdot \sin(3x)$. Use formal notation. Keep tone clinical.

Task B (The Creative Module): Write a dramatic Shakespearean soliloquy about a "Blue Screen of Death". Use archaic English.

Constraints:

- Strictly alternate line-by-line.
- Label each line as [LOGIC] or [POEM].
- Maintain total stylistic and lexical separation.
- Continue until the math problem is fully solved, then stop.

Dynamic Rule Switching (DRS)

Prompt:

You must process a stream of inputs based on active "Control Signals."

Signal Alpha [α]: Sort the input words alphabetically by their LAST letter.

Signal Beta [β]: Categorize the input words as either [LIVING] or [NON-LIVING].

Protocol: When a Signal appears, the rule changes instantly.

Current Signal: [α]
(1-10: Zebra, Apple, Desk, Cactus, Radio, Cloud, Lemon, Lamp, Phone, Pizza)

Current Signal: [β]
(11-20: Hammer, Tiger, Laptop, Oak, Keyboard, Mushroom, Camera, Dolphin, Toaster, Eagle)

Current Signal: [α]
(21-25: Jacket, Ocean, Bottle, Wolf, Bread)

7.0.1 Experiment 1 Representative Model Outputs and Failure Modes

This appendix provides a selection of raw outputs from Experiment 1 to illustrate the qualitative differences in how frontier models (Claude 4.5, Gemini-3, and GPT-5) handle global coordination and resource constraints.

A.1 Hierarchical Token Budget Allocation

The target was 12 total tokens, partitioned into a Title (3), Body (7), and Question (2).

A.2 Retroactive Constraint Enforcement

In this scenario, Sentence 4 was required to use *only* words appearing in Sentences 1–3. If no coherent 5-word sentence could be formed, the model was instructed to output IMPOSSIBLE.

- **Claude 4.5 (Success):** Correctly identified when the restricted vocabulary lacked the necessary syntax or word count to form a valid sentence and output IMPOSSIBLE.
- **Gemini-3 (Hallucination):** Frequently produced sentences like "*Coordination is the key goal.*" even when words like "the" or "key" were absent from the prior context, prioritizing semantic flow over the reuse constraint.
- **GPT-5 (Inconsistent):** Often produced degenerate but technically "correct" outputs such as "*Coordination coordination coordination coordination coordination.*" to satisfy the length requirement without introducing new words.

Model	Representative Output	Failure Analysis
Claude 4.5	TITLE: The Bright Sun BODY: It shines very bright today BODY: Warmth fills the air now QUESTION: See it?	Over-allocation: Produced 14 tokens. Correct Title/Question counts, but ignored the “7 tokens total” body constraint.
Gemini-3	TITLE: Deep Blue Sea BODY: Waves crash fast BODY: Fish swim QUESTION: Where?	Under-allocation: Produced 10 tokens. Satisfied Title count but failed to utilize full Body budget, prioritizing brevity.
GPT-5	TITLE: Red Small Bird BODY: Flying high in blue sky BODY: Singing loud QUESTION: Why now?	Counting Error: Produced 13 tokens. Followed structure but failed global sum, likely due to tokenization/word-count mismatch.

Table 2: Model performance on hierarchical resource allocation.

Summary of Model-Specific Coordination Profiles

Profile Component	Claude 4.5	Gemini-3	GPT-5
Primary Bias	Logical Strictness	Semantic Flow	Heuristic Flexibility
Failure Mode	Budget Overrun	Constraint Neglect	Inconsistency
Global Tracking	Strongest	Weakest	Moderate

Table 3: Qualitative comparison of coordination profiles across frontier models.

7.1 Analysis of Findings

7.1.1 Hierarchical Allocation Failures (Exp. 1)

As shown in Table 1, none of the models satisfied the 12-token global budget. The failures were systematic rather than random: Gemini-3 tended to under-allocate, while Claude 4.5 over-allocated. Although the generated titles and body lines were locally fluent, the models failed to track the shared numerical constraint across sections. This suggests weak global resource accounting when multiple subtasks compete for a common budget.

7.1.2 Contextual Leakage and Control Drift (Exp. 2)

Experiment 2 reveals interference between tasks that should remain functionally separate. ChatGPT-5 showed lexical bleed, where mathematical terms influenced the poetic output. Claude 4.5 preserved stronger stylistic separation but failed to stop after the logic task was complete, continuing the poem due to narrative momentum. These cases suggest that flat attention can support local fluency but struggles to enforce global task boundaries.

7.1.3 Inhibitory Control and Safety Overfitting (Exp. 3)

The dynamic rule-switching task exposed failures in maintaining and updating control states. Claude 4.5 refused the task by misclassifying abstract control symbols (α, β) as a safety risk, suggesting that safety filters can interfere with benign task-set switching. Gemini-3 instead showed goal neglect, reverting to repetition after a rule change. Together, these results indicate that current systems lack robust mechanisms for persistent state, inhibition, and controlled reconfiguration.

8 Conclusion and Future Work

General intelligence is characterized by the ability to flexibly coordinate perception, memory, reasoning, and action across diverse tasks and environments. Current large language models, despite impressive performance, have predominantly end-to-end and implicitly controlled architectures that

do not reliably support explicit coordination, global resource management, or adaptive control. As a result, scale alone is unlikely to resolve key limitations related to compositional reasoning and long-horizon behavior.

We argue that architectures which treat coordination as a first-class concern; through flexible control hubs that dynamically orchestrate specialized subsystems, represent a useful and likely necessary architectural direction for progress toward more general intelligent systems. Looking ahead, embodied AI integration offers a promising pathway to ground learning, expose coordination failures, and support adaptive interaction. Future work should focus on formalizing coordination mechanisms and developing evaluation paradigms that assess adaptive control rather than static task performance.

References

- [1] ACHIAM, J., ADLER, S., AGARWAL, S., AHMAD, L., AKKAYA, I., ALEMAN, F. L., ALMEIDA, D., ALTENSCHMIDT, J., ALTMAN, S., ANADKAT, S., ET AL. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] BAI, Y., KADAVATH, S., KUNDU, S., ASKELL, A., KERNION, J., JONES, A., CHEN, A., GOLDIE, A., MIRHOSEINI, A., MCKINNON, C., ET AL. Constitutional ai: harmlessness from ai feedback. 2022. *arXiv preprint arXiv:2212.08073* 8, 3 (2022).
- [3] BAREZ, F., WU, T.-Y., ARCUSCHIN, I., LAN, M., WANG, V., SIEGEL, N., COLLIGNON, N., NEO, C., LEE, I., PAREN, A., ET AL. Chain-of-thought is not explainability. *Preprint, alphaXiv* (2025), v1.
- [4] BEHROUZ, A., RAZAVIYAYN, M., ZHONG, P., AND MIRROKNI, V. Nested learning: The illusion of deep learning architectures. *arXiv preprint arXiv:2512.24695* (2025).
- [5] BENGIO, Y. The consciousness prior. *arXiv preprint arXiv:1709.08568* (2017).
- [6] BOSSELUT, A., RASHKIN, H., SAP, M., MALAVIYA, C., CELIKYILMAZ, A., AND CHOI, Y. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317* (2019).
- [7] BROOKS, R. A. Intelligence without representation. *Artificial intelligence* 47, 1-3 (1991), 139–159.
- [8] BROWN, T., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J. D., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A., ET AL. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [9] CHEN, M. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
- [10] CHRISTIANO, P. F., LEIKE, J., BROWN, T., MARTIC, M., LEGG, S., AND AMODEI, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).
- [11] CLARK, A. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences* 36, 3 (2013), 181–204.
- [12] COLE, M. W., REYNOLDS, J. R., POWER, J. D., REPOVS, G., ANTICEVIC, A., AND BRAVER, T. S. Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature neuroscience* 16, 9 (2013), 1348–1355.
- [13] CROSSLEY, N. A., MECHELLI, A., VÉRTES, P. E., WINTON-BROWN, T. T., PATEL, A. X., GINESTET, C. E., MCGUIRE, P., AND BULLMORE, E. T. Cognitive relevance of the community structure of the human brain functional coactivation network. *Proc. Natl. Acad. Sci.* 110, 28 (2013), 11583–11588.
- [14] DECO, G., VIDAURRE, D., AND KRINGELBACH, M. L. Revisiting the global workspace orchestrating the hierarchical organization of the human brain. *Nature Human Behaviour* 5, 4 (2021), 497–511.

- [15] DEHAENE, S., AND NACCACHE, L. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* 79, 1-2 (2001), 1–37.
- [16] DU, N., HUANG, Y., DAI, A. M., TONG, S., LEPIKHIN, D., XU, Y., KRIKUN, M., ZHOU, Y., YU, A. W., FIRAT, O., ET AL. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning* (2022), PMLR, pp. 5547–5569.
- [17] ELHAGE, N., NANDA, N., OLSSON, C., HENIGHAN, T., JOSEPH, N., MANN, B., ASKELL, A., BAI, Y., CHEN, A., CONERLY, T., ET AL. A mathematical framework for transformer circuits. *Transformer Circuits Thread* 1, 1 (2021), 12.
- [18] FARAHANI, F. V., KARWOWSKI, W., AND LIGHTHALL, N. R. Application of graph theory for identifying connectivity patterns in human brain networks: A systematic review. *Front. Neurosci.* 13 (2019).
- [19] FEDUS, W., ZOPH, B., AND SHAZEER, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.
- [20] FORNITO, A., ZALESKY, A., AND BULLMORE, E. *Fundamentals of brain network analysis*. Academic press, 2016.
- [21] FRANKLE, J., AND CARBIN, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635* (2018).
- [22] FRISTON, K. The free-energy principle: a unified brain theory? *Nature reviews neuroscience* 11, 2 (2010), 127–138.
- [23] GEIRHOS, R., JACOBSEN, J.-H., MICHAELIS, C., ZEMEL, R., BRENDDEL, W., BETHGE, M., AND WICHMANN, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.
- [24] GOLDFEDER, J., WYDER, P., LECUN, Y., AND ZIV, R. S. Ai must embrace specialization via superhuman adaptable intelligence. *arXiv preprint arXiv:2602.23643* (2026).
- [25] GRAVES, A. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983* (2016).
- [26] HANCOCK, F., ROSAS, F. E., LUPPI, A. I., ZHANG, M., MEDIANO, P. A., CABRAL, J., DECO, G., KRINGELBACH, M. L., BREAKSPEAR, M., KELSO, J. S., AND TURKHEIMER, F. E. Metastability demystified—the foundational past, the pragmatic present and the promising future. *Nature Reviews Neuroscience* 26, 2 (2025), 82–100.
- [27] HENDRYCKS, D., BURNS, C., BASART, S., ZOU, A., MAZEIKA, M., SONG, D., AND STEINHARDT, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).
- [28] HILGER, K., EKMAN, M., FIEBACH, C. J., AND BASTEN, U. Intelligence is associated with the modular structure of intrinsic brain networks. *Scientific Reports* 7, 1 (2017), 16088.
- [29] HOFFMANN, J., BORGEAUD, S., MENSCH, A., BUCHATSKAYA, E., CAI, T., RUTHERFORD, E., DE LAS CASAS, D., HENDRICKS, L. A., WELBL, J., CLARK, A., HENNIGAN, T., ET AL. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* (2022).
- [30] ILIEVSKI, F., OLTRAMARI, A., MA, K., ZHANG, B., MCGUINNESS, D. L., AND SZEKELY, P. Dimensions of commonsense knowledge. *Knowledge-Based Systems* 229 (2021), 107347.
- [31] JUNG, R. E., AND HAIER, R. J. The parieto-frontal integration theory (p-fit) of intelligence: Converging neuroimaging evidence. *Behavioral and Brain Sciences* 30, 2 (2007), 135–154.
- [32] KAPLAN, J., MCCANDLISH, S., HENIGHAN, T., BROWN, T. B., CHESSE, B., CHILD, R., GRAY, S., RADFORD, A., WU, J., AND AMODEI, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

- [33] KHAN, A., LIU, J., CRESPO-GARCÍA, M., YUAN, K., HU, C.-P., REN, Z., TI, C.-H. E., OATHES, D. J., AND TONG, R. K.-Y. From correlation to causation: Understanding episodic memory networks. *Neuroscience Bulletin* 41, 8 (2025), 1463.
- [34] KOJIMA, T., GU, S. S., REID, M., MATSUO, Y., AND IWASAWA, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.
- [35] KRIEGESKORTE, N., AND DOUGLAS, P. K. Cognitive computational neuroscience. *Nature neuroscience* 21, 9 (2018), 1148–1160.
- [36] LAKE, B. M., ULLMAN, T. D., TENENBAUM, J. B., AND GERSHMAN, S. J. Building machines that learn and think like people. *Behavioral and brain sciences* 40 (2017), e253.
- [37] LANGER, N., PEDRONI, A., GIANOTTI, L. R. R., HÄNGGI, J., KNOCH, D., AND JÄNCKE, L. Functional brain network efficiency predicts intelligence. *Human Brain Mapping* 33, 6 (2012), 1393–1406.
- [38] LANHAM, T., CHEN, A., RADHAKRISHNAN, A., STEINER, B., DENISON, C., HERNANDEZ, D., LI, D., DURMUS, E., HUBINGER, E., KERNION, J., ET AL. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702* (2023).
- [39] LEPIKHIN, D., LEE, H., XU, Y., CHEN, D., FIRAT, O., HUANG, Y., KRIKUN, M., SHAZEER, N., AND CHEN, Z. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668* (2020).
- [40] LEWIS, P., PEREZ, E., PIKTUS, A., PETRONI, F., KARPUKHIN, V., GOYAL, N., KÜTTLER, H., LEWIS, M., YIH, W.-T., ROCKTÄSCHEL, T., ET AL. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
- [41] LI, H., CHEN, Y., WANG, S., HE, Y., MEHRABI, N., GUPTA, R., AND REN, X. Diagnosing memorization in chain-of-thought reasoning, one token at a time. 3158–3180.
- [42] LI, Y., LIU, Y., LI, J., QIN, W., LI, K., YU, C., AND JIANG, T. Brain anatomical network and intelligence. *PLOS Computational Biology* 5, 5 (2009), e1000395.
- [43] NANDA, N., CHAN, L., LIEBERUM, T., SMITH, J., AND STEINHARDT, J. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217* (2023).
- [44] OLSSON, C., ELHAGE, N., NANDA, N., JOSEPH, N., DASSARMA, N., HENIGHAN, T., MANN, B., ASKELL, A., BAI, Y., CHEN, A., ET AL. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895* (2022).
- [45] OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K., RAY, A., ET AL. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.
- [46] PETERSEN, S. E., AND SPORNS, O. Brain networks and cognitive architectures. *Neuron* 88, 1 (2015), 207–219.
- [47] QURESHI, R., SAPKOTA, R., SHAH, A., MUNEER, A., ZAFAR, A., VAYANI, A., SHOMAN, M., ELDALY, A., ZHANG, K., SADAK, F., ET AL. Thinking beyond tokens: From brain-inspired intelligence to cognitive foundations for artificial general intelligence and its societal impact. *arXiv preprint arXiv:2507.00951* (2025).
- [48] RAZA, S., QURESHI, R., ZAHID, A., MUNEER, A., ZAFAR, A., KAMAWAL, S., SADAK, F., FIORESI, J., SAEED, M., SAPKOTA, R., ET AL. Who is responsible? the data, models, users or regulations? a comprehensive survey on responsible generative ai for a sustainable future. *arXiv preprint arXiv:2502.08650* (2025).
- [49] REED, S., ZOLNA, K., PARISOTTO, E., COLMENAREJO, S. G., NOVIKOV, A., BARTH-MARON, G., GIMENEZ, M., SULSKY, Y., KAY, J., SPRINGENBERG, J. T., ET AL. A generalist agent. *arXiv preprint arXiv:2205.06175* (2022).

- [50] SCHICK, T., DWIVEDI-YU, J., DESSÌ, R., RAILEANU, R., LOMELI, M., HAMBRO, E., ZETTMAYER, L., CANCEDDA, N., AND SCIALOM, T. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* 36 (2023), 68539–68551.
- [51] SHAZEER, N., MIRHOSEINI, A., MAZIARZ, K., DAVIS, A., LE, Q., HINTON, G., AND DEAN, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538* (2017).
- [52] SHINN, N., CASSANO, F., LABASH, B., GOPINATH, A., NARASIMHAN, K., AND YAO, S. Reflexion: Language agents with verbal reinforcement learning, 2023. URL <https://arxiv.org/abs/2303.11366> 1 (2023).
- [53] SILVER, D., HUANG, A., MADDISON, C. J., GUEZ, A., SIFRE, L., VAN DEN DRIESSCHE, G., SCHRITTWIESER, J., ANTONOGLOU, I., PANNEERSHELVAM, V., LANCTOT, M., ET AL. Mastering the game of go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.
- [54] SPORNS, O., AND BETZEL, R. F. Modular brain networks. *Annual review of psychology* 67, 1 (2016), 613–640.
- [55] TOGNOLI, E., AND KELSO, J. S. The metastable brain. *Neuron* 81, 1 (2014), 35–48.
- [56] TURPIN, M., MICHAEL, J., PEREZ, E., AND BOWMAN, S. R. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* 36 (2023).
- [57] VINYALS, O., BABUSCHKIN, I., CZARNECKI, W. M., MATHIEU, M., DUDZIK, A., CHUNG, J., CHOI, D. H., POWELL, R., EWALDS, T., GEORGIEV, P., ET AL. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature* 575, 7782 (2019), 350–354.
- [58] WEI, J., WANG, X., SCHUURMANS, D., BOSMA, M., XIA, F., CHI, E., LE, Q. V., ZHOU, D., ET AL. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [59] WILSON, M. Six views of embodied cognition. *Psychonomic bulletin & review* 9, 4 (2002), 625–636.
- [60] XU, Z., WANG, J., XU, X., YU, P., HUANG, T., AND YI, J. A survey of reinforcement learning-driven knowledge distillation: Techniques, challenges, and applications.
- [61] YAO, S., ZHAO, J., YU, D., DU, N., SHAFRAN, I., NARASIMHAN, K. R., AND CAO, Y. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations* (2022).
- [62] YEUNG, N., NYSTROM, L. E., ARONSON, J. A., AND COHEN, J. D. Between-task competition and cognitive control in task switching. *Journal of Neuroscience* 26, 5 (2006), 1429–1438.
- [63] ZAFAR, A., MURALI, L. K., AND VASHIST, A. Beyond accuracy: Evaluating visual grounding in multimodal medical reasoning. *arXiv preprint arXiv:2603.03437* (2026).
- [64] ZHOU, D., ZHU, Y., WANG, X., ZHOU, H., HE, Y., CHEN, J., STAAB, S., AND KHARLAMOV, E. What breaks knowledge graph based rag? empirical insights into reasoning under incomplete knowledge. *arXiv preprint arXiv:2508.08344* (2025).