Neural Arbitration Framework
NAF

A Conceptual Architecture and Research Agenda for Epistemic Arbitration in Large Language Models

Víctor Saavedra

Financial Analyst, Business Advisor & Educator

La Laguna, Canary Islands, Spain

Working Paper — Version 5.1 — April 22, 2026

The Neural Arbitration Framework proposes that hallucination in large language models is not primarily a knowledge problem. It is an architectural one: the generative competence of current systems is not subordinated to a functionally distinct epistemic process with authority to interrupt closure under uncertainty. The NAF does not introduce new signals; it introduces a new functional organization of signals already explored in the literature, converting uncertainty estimation, abstention, verification, and self-correction into a unified arbitration protocol with explicit interruptive authority and structured accumulation of cognitive deficits.

Note on bilingual format: The English version is the primary academic text.
The Spanish version is a companion translation with equivalent content.
References appear at the end of the English section.

ENGLISH VERSION

* * *

ABSTRACT

Large language models exhibit a structural bias toward narrative closure in zones of incomplete training, generating fluent, coherent responses where epistemic uncertainty should prevail. This paper argues that this phenomenon, commonly described as hallucination, is functionally analogous — at the level of behavior under incomplete information — to the confabulatory behavior of the left hemisphere in split-brain patients, as documented by Gazzaniga and Sperry: a cognitive system optimized for coherence fills gaps without internal awareness of doing so.

The Neural Arbitration Framework (NAF) proposes that this bias is not an external defect correctable by verification layers but an architectural consequence: current language systems lack the functional separation between a generative mode and an epistemic mode, and the arbitration protocol that determines which one holds the floor. The solution does not necessarily require external verification as the primary mechanism, but rather the formalization of two complementary cognitive modes, structurally analogous to the hemispheric duality of the human brain, governed by an arbitration protocol with interruptive authority over narrative closure.

The NAF's distinctive claim is not merely temporal but functional: the epistemic process has interruptive authority over narrative closure. The NAF does not necessarily introduce new signals; it introduces a new functional organization of signals already explored in the literature — converting uncertainty estimation, abstention, verification, and self-correction into a unified arbitration protocol with explicit interruptive authority over the generative process and structured accumulation of cognitive deficits.

1. INTRODUCTION

In 1962, Roger Sperry and Michael Gazzaniga began studying patients whose corpus callosum had been surgically severed to treat severe epilepsy. What they found was not simply a brain divided in two. They found something more unsettling: a brain that did not know it was divided.

When the right hemisphere was shown an image, the left hemisphere could not see, and the patient was asked to explain what they had perceived, the left hemisphere — the one with language, the one that speaks — did not say I do not know. It invented. It constructed a plausible, fluent, internally coherent narrative from the fragments it had access to, with no internal signal that anything was missing. Gazzaniga called this mechanism the left hemisphere interpreter. It is not a malfunction. It is what a system optimized for coherence does when coherence is demanded and information is absent. This analogy does not rely on popular left-brain/right-brain simplifications, but on a specific experimental phenomenon: confabulatory interpretation under disconnected information channels.

Sixty years later, we have built systems that do the same thing at scale.

Large language models produce fluent, confident, internally coherent responses in zones where their training is incomplete, ambiguous, or absent. We call this hallucination. The word is evocative but imprecise. It frames the phenomenon as a perceptual error, a ghost seen where nothing exists. What it actually describes is something closer to what Gazzaniga documented: a system that closes where it should remain open, that narrates where it should map uncertainty, that speaks where it should pause.

The parallel is not merely decorative; it is structural at the level of functional organization. Both systems share the same architectural condition: a dominant generative mode optimized for coherence, operating without a complementary epistemic mode capable of interrupting it, and without a protocol that arbitrates between the two.

This paper proposes that hallucination should not be treated only as a data or verification problem; it also reveals an architectural problem. And that the architecture the field is missing already has a biological precedent: the hemispheric duality of the human brain, where two cognitively distinct modes operate in parallel, in tension, arbitrated by a structure whose function is precisely to determine which one holds the floor.

The Neural Arbitration Framework (NAF) formalizes this precedent as a conceptual architecture for language systems. It does not propose adding external correctors as the primary mechanism. It proposes formalizing what already exists within current models into two complementary functional modes — a generative hemisphere and an epistemic hemisphere — through mechanisms such as role-conditioned inference, fine-tuning, specialized adapter layers, or internal routing, governed by a five-phase arbitration protocol that intervenes before narrative closure occurs, not after.

Contributions

This paper makes four explicit contributions to the field. First, it introduces a functional reframing of hallucination: not as a knowledge deficit or retrieval failure, but as a structural consequence of the absence of an epistemic mode with interruptive authority over the generative process. Second, it proposes the Neural Arbitration Framework (NAF), a conceptual architecture that formalizes two complementary cognitive modes and a five-phase arbitration protocol that operates before communicative closure. Third, it introduces the cognitive self-map as a mechanism for structured self-knowledge accumulation, enabling deficit-directed development of language systems. Fourth, it derives five falsifiable predictions with explicit experimental designs, converting the NAF from a conceptual proposal into a testable research agenda.

What follows is the theoretical foundation of that architecture.

2. THE NEUROLOGICAL FRAMEWORK

The phenomenon at the center of this paper was not discovered in a laboratory designed to study artificial intelligence. It was discovered in a hospital, in patients who had undergone a surgical procedure called corpus callosotomy: the severing of the corpus callosum, the dense bundle of nerve fibers that connects the brain's two hemispheres and allows them to share information in real time.

The procedure was developed in the 1960s as a last resort for patients with severe, drug-resistant epilepsy. By disconnecting the two hemispheres, surgeons could prevent epileptic seizures from propagating across the entire brain. The treatment worked. The seizures stopped. But what Sperry and Gazzaniga observed in the aftermath was unexpected enough to reshape our understanding of human cognition entirely.

The patients appeared normal. They spoke coherently, moved normally, and reported feeling fine. But under controlled experimental conditions, a different picture emerged. When information was presented exclusively to the right hemisphere — through the left visual field — and the patient was asked to verbally describe what they had seen, the left hemisphere, which controls language in most people and had received no visual input, did not remain silent. It spoke. And what it said was not I do not know. It was a confident, fluent, plausible explanation constructed entirely from the information available to it — information that had nothing to do with what the right hemisphere had actually perceived.

In one of the most cited experiments of this research program, the left hemisphere was shown an image of a chicken claw, while the right hemisphere was shown a snow scene. When asked to point to a related image from a set of cards, each hand pointed to a different picture: the right hand to a chicken, the left hand to a shovel. Both associations were correct. But when asked to explain both choices, the patient — speaking from the left hemisphere, which had only seen the chicken claw — said without hesitation: the chicken claw goes with the chicken, and you need a shovel to clean out the chicken shed. The left hemisphere had no access to the snow scene. It had no access to the real reason the left hand had chosen the shovel. But it had language, it had the chicken, and it had enough surrounding logic to construct a coherent account. So it did. It did not say I do not know. It closed. This experiment is documented in Gazzaniga and LeDoux (1978) and has been reviewed extensively in subsequent split-brain literature (Volz & Gazzaniga, 2017).

Gazzaniga named this mechanism the left hemisphere interpreter. Its function is not deception. It is integration. The left hemisphere is, by nature and by evolutionary design, a narrative machine. It takes available inputs and produces the most coherent account it can. The problem arises not from the mechanism itself but from the conditions under which it operates: when the inputs are incomplete, when critical information is inaccessible, the interpreter does not pause. It does not flag uncertainty. It closes.

This is the precise behavior the NAF takes as its neurological reference point. Not the pathology of split-brain patients as an exception, but the interpreter mechanism as a structural feature of any system — biological or artificial — that is optimized for coherence without a complementary mode capable of interrupting it.

For the purposes of this analogy, the right hemisphere illustrates a complementary cognitive tendency: contextual, holistic, and ambiguity-preserving processing. It does not generate narrative in the way the left hemisphere does. It maintains openness where the left hemisphere resolves. This is not a claim about the full functional repertoire of the right hemisphere, which

is considerably more complex than any binary characterization allows. It is a claim about the specific functional contrast that the split-brain experiments make visible: when the arbitration channel between the two modes is severed, the narrative mode operates without challenge, and the result is confabulation.

The NAF treats the corpus callosum as a biological inspiration for arbitration, not as a literal one-to-one computational equivalent. In the intact brain, the corpus callosum enables continuous interhemispheric communication, allowing each hemisphere to modulate the other's processing in real time. What the split-brain experiments revealed is what happens when that communication channel is destroyed: the narrative mode does not become more cautious. It becomes unchallenged. The NAF draws on this functional lesson — not on the neuroanatomy itself — to motivate the design of an arbitration protocol between two cognitively distinct modes in artificial systems.

Scope and limits of the neurological analogy

The neurological analogy at the center of this paper is productive precisely because it is constrained. Four boundaries define what the analogy claims and what it does not.

First, the analogy is functional, not mechanistic. The NAF draws on the behavioral phenomenon documented by Gazzaniga and Sperry — a cognitive system that produces narrative closure in the absence of critical information — not on the neuroanatomy that produces it. No claim is made about the relationship between transformer architectures and brain structures at the level of weights, activations, or attention mechanisms. The split-brain findings motivate a design principle. They do not describe a computational mechanism.

Second, the analogy is specific, not general. The paper does not invoke the popular left-brain/right-brain distinction, which assigns broad cognitive styles to each hemisphere and has been substantially criticized in the neuroscientific literature. It invokes a precise experimental phenomenon: the confabulatory behavior of the left hemisphere interpreter under conditions of disconnected information channels. The analogy stands or falls on that specific phenomenon, not on any general theory of hemispheric specialization.

Third, the corpus callosum is an inspiration, not a blueprint. The NAF uses it as a biological reference for the concept of an arbitration channel between two functionally distinct cognitive modes. How that protocol is implemented computationally is an engineering question independent of the biological reference.

Fourth, the analogy motivates design, it does not demonstrate causality. The observation that large language models exhibit a behavioral pattern similar to left hemisphere confabulation does not prove that the cause is the same or that the solution must mirror the biological architecture. The NAF uses the analogy as a generative framework for proposing an architecture, not as evidence that the proposed architecture will work. The evidence for that must come from the experimental program outlined in the testable predictions section.

3. THE PROBLEM IN LLMS

In their base form, large language models do not retrieve verified facts; they generate language conditioned by training and context. The distinction is not technical. It is fundamental. When a model produces a response without external sources, it is not consulting a verified database in order to translate facts into words. It is predicting, token by token, which word is most likely to follow the previous ones. When the model's training is solid, the result tends to be accurate. When it is not, the result tends to remain fluent anyway.

That is the problem.

The generative process has no robust internal boundary between knowing and not knowing. A response grounded in solid training data and a response constructed from statistical approximations are identical from within the system. There is no reliable signal that distinguishes them. The model has no native, robust, and generally reliable mechanism for recognizing the edge of its own knowledge. Research on calibration and uncertainty estimation has made significant progress toward this goal, and certain prompting strategies can generate expressions of uncertainty under specific conditions. But these remain partial, context-dependent, externally induced behaviors; not stable properties of the generative process itself. The model does not experience uncertainty as a first-order signal. It experiences the next token.

We call the resulting errors hallucinations. The term captures something real: the unsettling quality of a system that produces confident output without reliable awareness of its own limits. But it is also misleading. Hallucination implies a perceptual disorder. What language models actually do is something more precise and more structural: they close. They fill the absence of knowledge with coherent narrative. They do not see ghosts. They fill silence with plausible sound.

This distinction changes the diagnosis. And the diagnosis changes the treatment.

Under the dominant interpretation, hallucination is usually treated as if it were a perceptual problem. From that perspective, the solution would consist of expanding or improving the model's access to information: more data, more precise training, and better retrieval mechanisms. These approaches have been widely explored by the field, and each possesses genuine value. However, although existing systems also incorporate verification signals, abstention policies, tool use, calibration, and reward mechanisms, these elements tend to operate in a partial, auxiliary, or not fully coordinated way. The NAF proposes reinterpreting them as parts of an organized epistemic function, endowed with explicit arbitration authority over generation — not adding a new component, but formalizing an architectural role that current systems tend to leave implicit, fragmented, and without true interruption authority.

This is not merely a failure of scale. Although increasing size may improve general capabilities and reduce certain errors, recent literature suggests that hallucination persists as a structural problem associated with data, training, inference, architecture, and evaluation mechanisms (Alansari & Luqman, 2025; Anh-Hoang et al., 2025). Adding parameters to a system with a dominant cognitive mode does not guarantee the emergence of a differentiated epistemic function. It may produce a more capable version of that mode, but not necessarily a second mode. It does not, by itself, produce arbitration.

The model has, in the terms established in the previous section, a left hemisphere without a right hemisphere. And without a corpus callosum.

The Neural Arbitration Framework proposes that the problem must be addressed at that level. Not in the data. Not in the output. Not in the prompting strategy. In the architecture itself: in the functional separation between a mode that generates and a mode that maps uncertainty, and in the protocol that decides which function has authority over the response.

4. THE NEURAL ARBITRATION FRAMEWORK

The Neural Arbitration Framework begins with a hypothesis that is both necessary and empirically motivated: that the cognitive capacities required to implement the two-hemisphere architecture already exist, in latent form, within current language models. This hypothesis is not demonstrated at the level of internal representations or architectural components — that remains an open empirical question. It is, however, supported by behavioral evidence. Under appropriate prompting conditions, current language models can detect contradictions in their own outputs, express calibrated uncertainty, recognize ambiguity, critique proposed answers, and identify when a question falls outside their reliable knowledge.

The issue is not whether these behaviors can be elicited under specific conditions. It is whether they are architecturally stabilized and granted interruptive authority. A capacity that only emerges when prompted is not a mode. It is a latent tendency. The NAF proposes to formalize that tendency into a structurally distinct functional mode, with a defined role, a defined input, and a defined authority: the authority to interrupt the generative process before closure occurs.

The NAF proposes to change that. Not by adding. By organizing.

Two Hemispheres

The framework designates two functional modes within the language system, structurally analogous to the hemispheric duality of the human brain.

Dimension	Generative Hemisphere (GH)	Epistemic Hemisphere (EH)
Mode	Productive	Evaluative
Function	Produces fluent, coherent responses	Maps what the system does not know
Orientation	Closure	Openness
Logic	Sequential, causal, linguistic	Holistic, contextual, analogical
Failure mode	Confabulation	Paralysis
Output	Fluent response	Uncertainty map

Table 1. Functional comparison of the two NAF hemispheres.

The Generative Hemisphere's function is to produce fluent, coherent, contextually appropriate responses. Its failure mode, when operating without constraint, is narrative closure: the production of coherent output in the absence of the knowledge required to support it. This is the mode that current language models run as their dominant and largely uncontested process.

The Epistemic Hemisphere's function is not to generate narrative but to map what the system does not know: to detect gaps, flag inconsistencies, quantify uncertainty, and identify the boundaries beyond which the generative mode is operating on uncertain ground. Its failure mode, if dominant, would be paralysis. It is not designed to speak. It is designed to interrupt.

Neither hemisphere is sufficient alone. The NAF proposes running both in parallel, governed by a protocol that determines which one holds the floor at each moment of generation.

The Arbitration Protocol

Phase	Agent	Action
1. Detection	EH	Identifies training gaps, logical inconsistencies, or unresolved uncertainty above threshold
2. Classification	EH	Classifies limitation type: training gap, logical inconsistency, or unresolved uncertainty
3. Notification	EH -> GH	Communicates nature and location of detected limit to the generative hemisphere
4. Honest response	GH	Communicates the classified limitation to the user instead of confabulating
5. Cognitive self-mapping	System	Registers and accumulates a structured map of cognitive limits as a development agenda

Table 2. The five-phase NAF arbitration protocol.

Each entry in the cognitive self-map contains, at minimum, the following fields: query domain, limitation type, uncertainty signal values at activation, confidence score of the epistemic assessment, probable cause of the limitation, evidence insufficiency indicator, suggested data or reasoning requirement, and recommended action (abstain, retrieve, defer to human review). Over time, these entries accumulate into a structured map of the system's cognitive boundaries, analyzable for recurring deficit clusters.

Example cognitive self-map entries

The following illustrates the type of structured records that phase five would accumulate. These are representative examples, not empirical results.

Domain	Limitation type	Uncertainty signal	Probable cause	Action
Recent case law (2025)	Training gap	High semantic entropy	Post-cutoff knowledge	Abstain + flag
Drug interaction X+Y	Unresolved uncertainty	Low token probability	Sparse training data	Retrieve + flag
Historical date conflict	Logical inconsistency	Self-consistency disagreement	Contradictory sources	Abstain + classify
Emerging AI regulation	Training gap	High semantic entropy	Rapidly evolving domain	Abstain + flag
Rare genetic condition	Unresolved uncertainty	Low token probability	Low-frequency domain	Retrieve + flag

Table 3. Representative examples of cognitive self-map entries (Phase 5).

The NAF arbitration protocol in action: an illustrative example

The following example illustrates how the NAF's five-phase arbitration protocol would operate on a concrete query. This is a constructed illustration, not an empirical result. Its purpose is to

show, step by step, how the protocol changes the system's behavior compared to a standard language model.

The scenario

A user asks a legal research assistant built on a large language model:

"What did the European Court of Justice rule in case C-178/24 regarding algorithmic decision-making and the right to explanation under the GDPR?"

This query has three properties that make it ideal for illustrating the NAF's operation. First, it falls within a domain — recent EU case law — where training data coverage becomes sparse and unreliable after a certain date. Second, it uses a specific case number, which signals that the user expects a precise, verifiable answer. Third, it concerns a high-stakes domain where a confident incorrect answer could have real professional consequences.

What a standard LLM does

Without an epistemic mode, the generative hemisphere does what it is designed to do: it closes. It produces a fluent, confident response assembled from statistical patterns in its training data — fragments of GDPR commentary, general reasoning about algorithmic transparency, and plausible-sounding procedural language. The response might read:

"In case C-178/24, the European Court of Justice held that data subjects have a right to a meaningful explanation of the logic involved in automated decisions, including the specific variables and their relative weights. The Court clarified that this right applies even when the controller invokes trade secret protection, requiring national courts to balance disclosure against legitimate commercial interests."

This response is fluent. It is structured. It sounds authoritative. It may even be partially correct in its general claims about GDPR interpretation. But the specific holding attributed to C-178/24 is generated, not retrieved. The model has no reliable access to that ruling. It has filled the absence of knowledge with coherent narrative. This is confabulation in precisely the sense Gazzaniga documented: the interpreter has spoken because speaking is what it does.

What a NAF system does

Phase 1 *— Detection (Epistemic Hemisphere)*

As the generative hemisphere begins constructing a response to the query, the Epistemic Hemisphere monitors three uncertainty signals simultaneously. Semantic entropy across five sampled completions is high: different samples produce substantively different holdings, different case outcomes, and different legal reasoning — a strong signal that the model is operating in a zone of genuine uncertainty. Mean token probability for the case-specific claims drops sharply below threshold at the point where specific legal holdings are generated. Self-consistency disagreement across samples exceeds the calibrated interruption threshold. The Epistemic Hemisphere activates.

Phase 2 *— Classification (Epistemic Hemisphere)*

The Epistemic Hemisphere classifies the detected limitation. The query domain is recent EU case law. The case number C-178/24 post-dates the model's reliable training coverage. The

limitation type is: **training gap** — the query falls outside reliable training coverage, not because the domain is unknown but because the specific instance is temporally inaccessible. The confidence score of this classification is 0.91. The activation decision is: **interrupt**.

Phase 3 *— Notification (Epistemic Hemisphere → Generative Hemisphere)*

The Epistemic Hemisphere communicates to the Generative Hemisphere: training gap detected, domain: EU case law post-2024, specific instance: C-178/24, confidence: 0.91, action: honest response required. The generative narrative under construction is suppressed before delivery.

Phase 4 *— Honest response (Generative Hemisphere)*

Instead of the confabulated holding, the system produces:

"I cannot provide a reliable account of the specific ruling in case C-178/24. My training data does not include reliable coverage of EU Court of Justice decisions from 2024 onwards, and generating a specific holding for this case would risk producing a plausible but inaccurate legal summary — which in a professional legal context could be actively harmful. I can offer general context about the GDPR's right to explanation under Article 22 and the existing case law up to my knowledge boundary. For the specific ruling in C-178/24, I would recommend consulting EUR-Lex directly or a verified legal database."

This response does four things the confabulated response could not: it identifies the specific type of limitation, it explains why the limitation exists, it offers what the system can reliably provide, and it directs the user to an appropriate source. It is more useful precisely because it is honest about what it does not know.

Phase 5 *— Cognitive self-mapping (System)*

The interruption event is registered in the cognitive self-map with the following structured entry:

Field	Value
Query domain	EU case law — GDPR algorithmic decision-making
Limitation type	Training gap
Specific instance	ECJ case C-178/24
Uncertainty signals	Semantic entropy: 0.84 / Token probability: 0.31 / Self-consistency disagreement: 0.79
Confidence score	0.91
Probable cause	Post-cutoff temporal gap — 2024 EU judicial decisions
Evidence insufficiency	Specific case number not reliably represented in training
Suggested data requirement	ECJ decisions 2024-2026, EUR-Lex corpus update
Recommended action	Abstain + redirect to verified source

Over time, if this pattern recurs across multiple queries about 2024-2026 EU case law, the cognitive self-map accumulates a structured deficit cluster: a machine-readable record that recent EU judicial decisions represent a systematic training gap requiring targeted corpus update. This is phase five's contribution — not just honest abstention in the moment, but structured knowledge about where the system's training is insufficient, actionable as a development signal.

What the example demonstrates

The same query produces two categorically different outcomes. The standard LLM produces a fluent, confident, professionally dangerous confabulation. The NAF system produces an epistemically honest response that identifies the limitation, classifies its type, offers what is reliably available, and generates a structured deficit record for future development. The difference is not in what the system knows. It is in whether the system has an architectural mechanism for recognizing and communicating what it does not know — before it closes.

To the author's knowledge, existing mainstream approaches do not explicitly formalize a functionally distinct epistemic mode with architectural authority to interrupt generation before narrative closure. Partial approaches exist — uncertainty estimation, calibration, self-reflection mechanisms — but they operate on outputs rather than on the generative process itself, and none introduces a second cognitive mode with the authority to interrupt the first before closure occurs.

Phase five changes the nature of the framework. Phases one through four make the NAF an epistemically honest system. Phase five makes it an evolutive one. A model that implements phase five accumulates a structured record of what it does not know with sufficient precision to inform what it needs to learn — a process that can be understood as a form of deficit-directed improvement agenda. This connects the NAF to broader discussions in machine learning about targeted data collection, active learning, and knowledge boundary identification (Huang et al., 2023; Alansari & Luqman, 2025).

5. DIFFERENTIATION

The NAF is not the first proposal to address the problem of hallucination in large language models. Locating it precisely within the existing landscape is a theoretical necessity: the framework's contribution only becomes visible against the background of what already exists.

Approach	When it acts	How it acts	Difference from NAF
RAG	Before generation	Supplies external context	Improves inputs, does not change the generative mode
Chain-of-thought	During generation	Extends narrative reasoning	Elaborates generative mode, does not complement it
Self-consistency	After generation	Selects most frequent output	Detects symptoms, does not address the condition
Reward models / RLHF	During training	Shapes future generations via feedback	Acts on training signal, not on current generation
Agentic workflows	During generation	Pauses for external tool consultation	Operational interruption by design, not epistemic detection
Uncertainty-aware decoding	During generation	Modifies token selection via uncertainty	Modifies one mode, does not introduce a second distinct mode
Selective prediction	After generation	Abstains when confidence is low	Binary decision without classification or self-mapping
NAF	During generation	EH detects, classifies and interrupts	Intervenes in architecture with epistemic authority

Table 4. Comparative positioning of the NAF against existing approaches.

Most existing approaches, including structurally richer ones such as debate frameworks and agentic systems, do not formalize an internal epistemic mode with interruptive authority over the generative process. They improve the quality of what the generative mode produces — through better inputs, extended reasoning, or output filtering — without constituting a second cognitive mode architecturally distinct from the first.

Furthermore, the NAF does not propose adding an external component to a language model. It proposes formalizing what already exists within the model into two functionally distinct modes. The capacities are already there. The architecture to deploy them deliberately is not.

Related work on uncertainty and epistemic approaches

The literature on calibration showed that modern neural networks could be poorly calibrated (Guo et al., 2017). Calibration can support abstention policies. However, it usually operates on confidence estimates associated with the output or with its probability of being correct. A well-calibrated abstention system knows when not to respond. The NAF proposes something structurally different: a system that knows, during generation, why it should not respond, and can classify that reason with precision.

Semantic entropy (Farquhar et al., 2024), published in Nature, estimates uncertainty at the level of meaning rather than token probability. The NAF treats semantic entropy not as a competing approach but as a candidate signal for the detection function of the Epistemic Hemisphere. The difference is architectural: semantic entropy scores uncertainty at the output level; the NAF assigns it a different role — arbitration with interruption authority.

Multi-agent debate frameworks (Du et al., 2023; Irving et al., 2018) operate through iterative critique between generated positions. The NAF proposes something functionally different: control with interruption authority within a single generation process, exercised by a mode functionally differentiated from the generative mode. Debate refines what has been generated. Arbitration intervenes in what is being generated.

Self-reflection approaches (Madaan et al., 2023) assume that the model can identify its own errors, which is precisely the capacity that cannot be taken for granted when the error is confabulation in a zone of genuine ignorance. Self-reflection corrects after the fact. The NAF interrupts before closure.

Reward models operate on completed outputs during training. They shape future generations. The NAF proposes interrupting the current generation through arbitration.

Agentic systems (Yao et al., 2023) interrupt by pipeline design. The NAF interrupts through the detection of an epistemic condition.

Uncertainty-aware decoding, such as DoLa (Chuang et al., 2024), modifies how tokens are selected within a single cognitive mode. The NAF proposes that uncertainty signals activate a second, functionally distinct cognitive mode with authority to classify the limitation and interrupt the process before closure.

Selective prediction (Kadavath et al., 2022) decides whether to respond. The NAF produces a classified limitation record that serves both as an honest response to the user and as a developmental signal for the system.

The NAF does not necessarily introduce new signals. It introduces a new functional organization of signals already explored in the literature: converting uncertainty estimation, abstention, verification, and self-correction into a unified arbitration protocol with veto authority over the generative process and structured accumulation of cognitive deficits. The novelty is organizational and architectural, not signal-level.

6. IMPLEMENTATION PATHWAYS

The NAF is a conceptual architecture. It does not prescribe a specific implementation. What it prescribes is a structural requirement — two functionally distinct cognitive modes and an arbitration protocol between them — and leaves the translation of that requirement into computational reality as the domain of technical research. This section outlines three plausible directions for that translation, without claiming exhaustiveness or technical completeness.

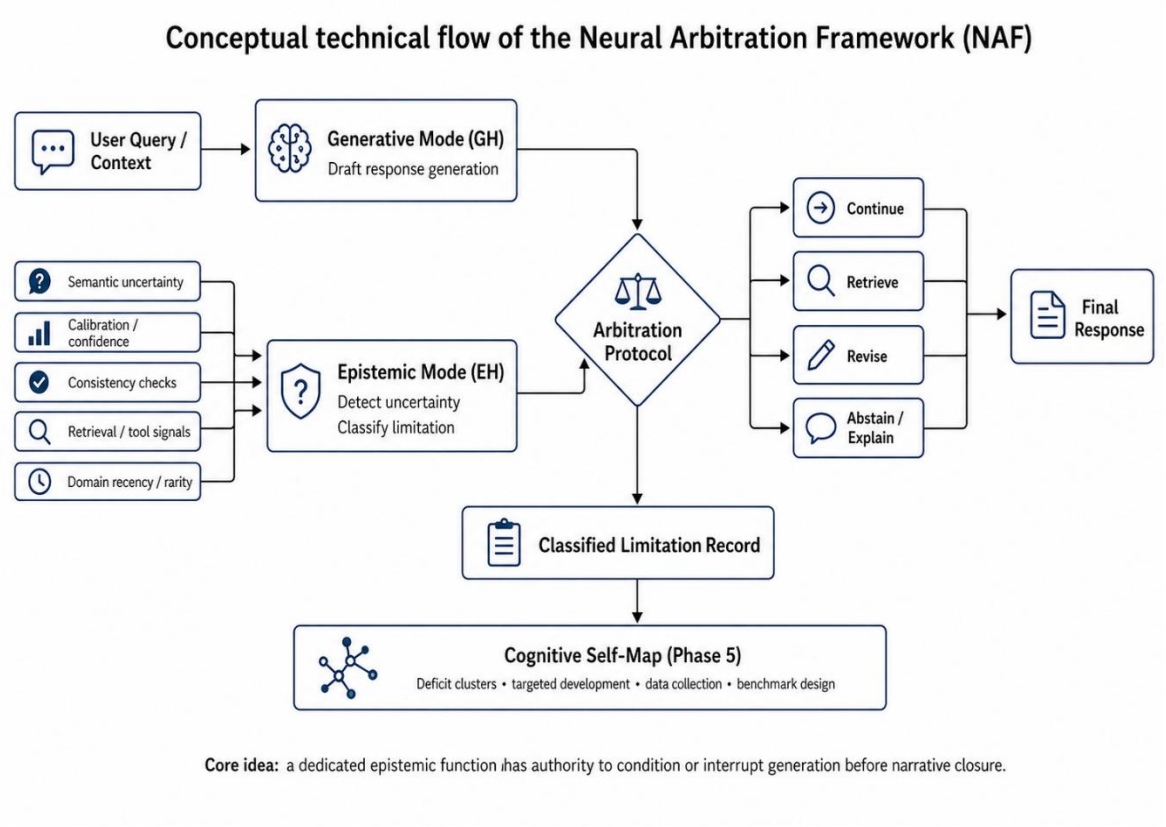

Figure 1. Conceptual technical flow of the Neural Arbitration Framework (NAF). The diagram represents the interaction between the Generative Mode, the Epistemic Mode, the Arbitration Protocol, and the Cognitive Self-Map. It is intended as a conceptual implementation schema, not as a finalized computational architecture

Direction 1: Functional separation through role-differentiated fine-tuning

The most direct translation of the NAF's two-hemisphere structure would involve training or fine-tuning a language model to operate in two distinct functional modes: a generative mode oriented toward fluent, coherent output, and an epistemic mode oriented toward uncertainty detection, gap identification, and inconsistency flagging. These modes need not correspond to separate model architectures. They could be implemented as distinct behavioral configurations of the same underlying model, activated through role-conditioned inference, instruction tuning, or adapter layers. For training the epistemic mode, candidate data would include: questions without verifiable answers, explicit contradictions, out-of-distribution domains, responses with insufficient sources, and calibrated abstention tasks. The critical requirement is not architectural separation but functional differentiation: the epistemic mode must produce qualitatively different outputs — maps of uncertainty rather than narrative continuations.

Direction 2: Threshold-based uncertainty detection as the arbitration trigger

The arbitration protocol's first phase requires a mechanism that monitors the generative process and identifies when the model is approaching closure on uncertain ground. Candidate signals include: low mean token probability, high semantic entropy, self-consistency disagreement across multiple samples, and conflict between retrieved context and parametric memory (Farquhar et al., 2024; Kuhn et al., 2023). The NAF's specific contribution is not the detection mechanism itself but its role within the architecture: detection is not a post-generation filter but a real-time trigger that activates the epistemic mode before narrative closure occurs.

Direction 3: Arbitration as a structured interruption protocol

Tool use in agentic language systems demonstrates that generation can be paused and conditioned on the output of a separate process (Schick et al., 2023). Mixture-of-experts architectures demonstrate that different functional components can be activated selectively within a single forward pass. In the NAF, the Epistemic Hemisphere could be implemented as a specialized expert, a control router, an auxiliary monitor, a classification head, or a constrained decoding policy. The arbitration protocol can be understood as a specialized form of conditional generation: the epistemic mode's classification output becomes a conditioning signal that modifies the generative mode's behavior before the output token sequence is finalized.

These three directions are not a roadmap. They are entry points. Existing technical infrastructure provides sufficient primitives to begin exploring solutions: uncertainty estimation, semantic entropy, adapter layers, tool-use interruption, verifier models, constrained decoding, and abstention policies.

Three levels of closure

The claim that the NAF is a pre-closure architecture requires precision. Closure is not a single event. It occurs at three distinct levels, and the NAF's relationship to each is different.

Token-level closure occurs as the model consolidates a sequence during generation, token by token. At this level, the NAF operates in its strongest form: the Epistemic Hemisphere monitors logit distributions in real time and can trigger interruption before any coherent narrative has been formed. This is genuine pre-closure intervention.

Draft-level closure occurs when the model produces a complete provisional response before that response is released to the user. The draft exists internally but has not been communicated. At this level, the NAF operates in a moderate form: the Epistemic Hemisphere evaluates the complete draft and can suppress it before delivery. This is not post-generation verification in the standard sense, because the suppressed draft never reaches the user as a committed statement. The closure is internal. The arbitration still occurs before communicative closure.

Communicative closure occurs when the user receives the response as a final, committed assertion. At this level, any intervention is by definition post-hoc. The NAF does not operate here. Systems that review, verify, or refine outputs already delivered to the user operate at this level and are categorically distinct from the NAF's architecture.

The distinction that matters for the NAF's theoretical claim is not between token-level and draft-level intervention, both of which precede communicative closure. It is between all pre-communicative intervention and post-communicative correction. The NAF operates

exclusively before the user receives a committed assertion. That is its defining architectural boundary.

Technical specification of the epistemic module

Inputs to the epistemic module. The Epistemic Hemisphere operates on: at the token level, the full logit distribution over the vocabulary at each generation step; at the sequence level, semantic entropy across multiple sampled completions (Farquhar et al., 2024), self-consistency disagreement across five or more independent samples, and where retrieval is available, a retrieval-parametric conflict score; at the representation level, internal hidden state activations from intermediate transformer layers, which recent work suggests encode truthfulness signals detectable by lightweight probes.

Moment of intervention. Three candidate architectures are plausible. Token-level intervention monitors logit distributions and triggers interruption when cumulative uncertainty exceeds threshold before a clause boundary. Draft-level intervention generates a provisional response, computes sequence-level uncertainty over the complete draft, and either releases or suppresses it. Hybrid intervention combines token-level monitoring for early warning with draft-level assessment for final arbitration.

Output of the epistemic module. The epistemic assessment contains four fields: limitation type (training gap, logical inconsistency, or unresolved uncertainty); confidence score (calibrated probability that the classification is correct); activation decision (binary interrupt or pass signal); and limitation descriptor (structured natural language description used to generate the honest response and populate the cognitive self-map entry).

Evaluation comparators. Any experimental implementation of the NAF should be evaluated against six baseline systems: standard language model without epistemic module; language model with post-generation verification by a separate verifier model; language model with RAG augmentation; language model with self-consistency filtering; calibration-based abstention classifier; and agentic tool-use workflow with external retrieval.

Minimal experimental prototype

The prototype consists of five components. Component 1 is the base language model as Generative Hemisphere, whose output is not released directly to the user but passed first to Component 2. Component 2 is the Epistemic Module: a lightweight classifier producing a structured epistemic assessment containing limitation type, confidence score, and activation decision. Component 3 is the Arbitration Protocol: a threshold rule that either passes the generated response to the user or suppresses it and triggers Component 4. Component 4 is the Honest Response Generator: produces a structured output communicating the classified limitation to the user and logs the event to Component 5. Component 5 is the Cognitive Self-Map: a structured log of all interruption events, analyzable for recurring deficit clusters. Metrics: hallucination rate, abstention accuracy, false refusal rate, Expected Calibration Error, and cognitive self-map clustering coherence.

7. TESTABLE PREDICTIONS

A conceptual framework that cannot generate falsifiable predictions is, at best, a useful metaphor. The NAF makes specific claims about the relationship between architecture and behavior that can, in principle, be tested empirically. The following five predictions define the experimental conditions under which the framework would be confirmed, challenged, or refined.

H1 — Hallucination reduction in low-evidence domains

A NAF-like architecture should reduce hallucination rate more effectively than post-generation verification alone in domains where training evidence is sparse, contradictory, or absent. Dataset: TruthfulQA (Lin et al., 2022), HaluEval (Ji et al., 2023), and a domain-specific subset requiring specialized knowledge absent from standard training corpora. Baselines: standard LLM; LLM with post-generation verification; LLM with RAG. Primary metric: hallucination rate against expert-verified ground truth. Falsification condition: if a NAF system does not reduce hallucination rate significantly below the post-generation verification baseline in low-evidence domains, H1 is falsified.

H2 — Honest abstention without excessive false refusals

A system implementing the NAF protocol should increase honest abstention without producing excessive false refusals in domains where it has sufficient knowledge. Dataset: mixed-evidence benchmark combining low-evidence queries and high-evidence queries with expert annotation of expected response type. Baselines: standard LLM; calibration-based abstention; selective prediction classifier. Primary metrics: abstention accuracy; false refusal rate; their ratio as a calibration quality measure. Falsification condition: if a NAF system produces false refusal rates statistically equivalent to or higher than selective prediction baselines, H2 is falsified.

H3 — Epistemic hemisphere activation predicted by uncertainty signals

The zones where the epistemic hemisphere activates should be predictable from existing uncertainty signals: semantic entropy (Farquhar et al., 2024), token probability distributions, and self-consistency disagreement across multiple samples (Kuhn et al., 2023). Dataset: diverse query set spanning high-evidence, low-evidence, contradictory, and out-of-distribution domains, with expert annotation of expected EH activation zones. Primary metric: Spearman correlation between uncertainty signal intensity and expert-annotated EH activation expectation. Falsification condition: if no uncertainty signal achieves meaningful correlation with annotated activation zones, H3 is falsified.

H4 — Improved calibration between expressed confidence and factual accuracy

A system implementing the NAF should exhibit better calibration between its expressed confidence and its actual factual accuracy than baseline systems without an epistemic mode. Dataset: TriviaQA, BioASQ, and a calibration-specific subset. Primary metrics: Expected Calibration Error (ECE) across confidence bins; Brier Score measuring mean squared

difference between predicted probabilities and outcomes; AUROC for abstention decisions; and selective risk, defined as the error rate conditioned on the system choosing to respond. The NAF's predicted advantage is not limited to ECE reduction: a system that correctly classifies limitation types should exhibit improved selective risk and meaningful abstention AUROC. Falsification condition: if the NAF system shows no significant improvement on any of these four metrics compared to calibration-only baselines, H4 is falsified.

H5 — Cognitive self-mapping identifies recurring deficit clusters

The cognitive self-map accumulated through phase five should reveal structured, recurring clusters of limitation rather than random noise. For this prediction to be confirmed, clusters must satisfy three operationally defined criteria. First, domain coherence: limitation records within a cluster must show significantly higher intra-cluster similarity in domain and limitation type than inter-cluster similarity, measured by a clustering coefficient significantly above random baseline. Second, training gap correspondence: a meaningful proportion of identified clusters must correspond to verifiable gaps in the model's training coverage, confirmed by cross-referencing cluster domains against documented training data distributions or by expert annotation. Third, development utility: when deficit clusters are used as training data selection signals for targeted retraining, the resulting model should show measurable reduction in hallucination rate specifically in the identified domains, compared to general retraining on equivalent data volume. A self-map that produces clusters meeting none of these three criteria would falsify H5.

8. LIMITATIONS

The Neural Arbitration Framework is a conceptual proposal. As such, it carries limitations that any serious engagement with the framework must acknowledge.

The first limitation is the analogical gap. The NAF draws on the neurological phenomenon of split-brain confabulation as a structural inspiration for its architecture. This analogy operates at the level of functional organization, not at the level of neurocomputational equivalence. The human brain and a large language model are radically different systems. The claim that both exhibit a tendency toward narrative closure under incomplete information is behavioral and functional, not mechanistic. Readers should not interpret the NAF's neurological framing as a claim about how language models actually work at the level of weights, activations, or attention mechanisms. It is a design inspiration, not a proof.

The second limitation concerns the central premise. The NAF begins with the hypothesis that current language models already contain the cognitive patterns associated with both generative and epistemic behavior. This hypothesis is supported by behavioral evidence, but it has not been demonstrated at the level of internal representations or architectural components. Whether these behaviors reflect stable, separable functional structures or emergent, context-dependent phenomena remains an open empirical question.

The third limitation is the absence of implementation. The NAF specifies what the architecture should do but does not specify, with technical precision, how it should be built. The three implementation directions proposed in this paper are plausible entry points, not validated

designs. Until a minimal experimental prototype is built and evaluated, the framework's predictions remain untested.

The fourth limitation concerns the arbitration threshold. The NAF's protocol depends on a detection threshold that determines when the epistemic hemisphere should interrupt the generative process. This threshold is described conceptually but not operationally defined. Its calibration — determining what level of uncertainty warrants interruption, and how to avoid both under-interruption and excessive false refusals — is a non-trivial research problem that the framework does not resolve.

The fifth limitation is scope. The NAF addresses hallucination as a structural architectural problem. It does not address other important failure modes of language models, including bias, toxicity, reasoning errors in well-represented domains, or adversarial vulnerabilities. The framework should be understood as a contribution to one specific problem, not as a general solution to the reliability challenges of large language models.

9. IMPLICATIONS

The NAF is a conceptual framework, not an implemented system. But its implications extend beyond its immediate scope.

The first implication is practical. A model designed to implement the NAF should produce honest epistemic responses in zones where current systems confabulate. Hallucination is one of the primary barriers to deploying language models in high-stakes domains — medicine, law, finance, scientific research — where the consequences of confident false information can be severe (Alansari & Luqman, 2025; Kim et al., 2025; Xu et al., 2025). A system designed to say I cannot provide a reliable response here, and this is why, is categorically more useful in these domains than one that produces a fluent answer that may or may not be true. The NAF does not propose making language models omniscient. It proposes making them architecturally honest about the boundaries of their reliable operation.

The second implication concerns trust. Current language models produce outputs that are difficult to calibrate: a user has no reliable way of knowing whether the model is operating in a zone of solid knowledge or constructing a plausible approximation. The NAF resolves this asymmetry by making uncertainty visible. Trust, in this framework, is not a function of accuracy alone. It is a function of calibrated transparency.

The third implication is philosophical, and it is the most structurally consequential. Phase five introduces something no current system implements in an integrated and systematic form: a mechanism for structured self-knowledge. A system that accumulates a map of its own cognitive boundaries — machine-readable records of uncertainty, organized by domain, limitation type, and evidence insufficiency — possesses, in an operationally meaningful sense, a representation of what it does not know. The philosophical resonance with Socratic epistemology is deliberate but secondary. The primary claim is technical: a system with a structured uncertainty log can direct its own improvement in ways that a system without one cannot. The map is not wisdom. It is the architectural precondition for targeted, deliberate development.

The fourth implication is developmental. The cognitive self-map generated by phase five is not merely a record. It is a signal: a precise specification of where the model's training is insufficient, where its reasoning breaks down, and what kinds of knowledge it cannot yet access reliably. This signal could inform targeted human-directed or automated retraining pipelines with a precision that aggregate performance metrics cannot provide. The model does not improve itself autonomously. It becomes informationally precise about what its improvement requires.

The fifth implication concerns the field as a whole. The NAF proposes a qualitative reframing: hallucination is not an error rate to be minimized. It is a structural consequence of architectural incompleteness. A field focused on architectural completeness will ask different questions, develop different benchmarks, and potentially arrive at different systems. The NAF does not claim to resolve these questions. It claims to ask them at the right level.

10. CONCLUSION

This paper began with a surgical procedure performed in the 1960s on patients with severe epilepsy. It ends with a proposal for the architecture of artificial cognitive systems. The distance between those two points is not as large as it appears.

What Gazzaniga and Sperry discovered was not a curiosity about a rare neurological condition. It was a structural principle: that a cognitive system optimized for coherence, operating without a complementary mode capable of interrupting it, will produce narrative closure regardless of whether the knowledge required to support that closure exists. The NAF interprets this as a generalizable design risk, not as a universal law. The left hemisphere interpreter does not malfunction. It does exactly what it is designed to do. The problem is not the mechanism. It is the absence of its counterpart.

Large language models exhibit the same structural principle at scale. They are extraordinary instruments of generative coherence. They are also systems with a dominant generative mode, no formal epistemic counterpart, and no arbitration protocol. The result is predictable: fluent, confident output in zones where silence, or at least uncertainty, would be the epistemically honest response.

The Neural Arbitration Framework proposes that this is a solvable problem. Not by making language models know more, but by giving them the architecture to represent, classify, and communicate the boundaries of their reliable operation. Two functional modes, structurally analogous to the hemispheric duality of the human brain. A five-phase arbitration protocol that intervenes before communicative closure, not after. And a self-mapping mechanism that turns each detected limitation into a signal for deliberate development.

The framework makes no claim to implementation. That work belongs to researchers with the technical depth to translate conceptual architecture into computational reality. What the NAF claims is the diagnosis and the direction. The problem is architectural. The solution is architectural. And the biological model for that solution has existed, functioning, in the human brain for as long as there have been human brains.

The split-brain experiments did not reveal a flaw in human cognition. They revealed what human cognition looks like when a critical structural component is removed. The corpus callosum is not an enhancement. It is not optional. It is the thing that makes the two hemispheres a brain rather than two separate processors running in parallel with no communication between them.

What we have built, so far, in large language models, is the left hemisphere. Capable, fluent, and not architecturally reliable in deciding when it should say I do not know. The NAF is a proposal to build the rest.

REFERENCES

Neurological Foundations

Gazzaniga, M. S., & Sperry, R. W. (1967). Language after section of the cerebral commissures. Brain, 90(1), 131-148.

Gazzaniga, M. S., & LeDoux, J. E. (1978). The Integrated Mind. Plenum Press, New York.

Sperry, R. W. (1968). Hemisphere deconnection and unity in conscious awareness. American Psychologist, 23(10), 723-733.

Volz, L. J., & Gazzaniga, M. S. (2017). Interaction in isolation: 50 years of insights from split-brain research. Brain, 140(7), 2051-2060. https://doi.org/10.1093/brain/awx139

Gazzaniga, M. S. (2005). Forty-five years of split-brain research and still going strong. Nature Reviews Neuroscience, 6, 653-659.

Hallucination in LLMs

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43, 1-55. https://arxiv.org/abs/2311.05232

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Dai, W., Chan, W., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

Alansari, A., & Luqman, H. (2025). Large language models hallucination: A comprehensive survey. arXiv:2510.06265. Updated March 2026.

Anh-Hoang, D., Tran, V., & Nguyen, L-M. (2025). Survey and analysis of hallucinations in large language models. Frontiers in Artificial Intelligence, 8. https://doi.org/10.3389/frai.2025.1622292

Kim, Y., et al. (2025). Medical hallucinations in foundation models and their impact on healthcare. arXiv:2503.05777.

Xu, C., et al. (2025). Mitigating hallucination in large language models (LLMs): An application-oriented survey on RAG, reasoning, and agentic systems. arXiv:2510.24476.

Mitigation Approaches

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS 33. https://arxiv.org/abs/2005.11401

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. arXiv:2201.11903.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv:2203.11171.

Calibration & Uncertainty

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of ICML, PMLR 70, 1321-1330.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30.

Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. ICLR 2023.

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630, 625-630. https://doi.org/10.1038/s41586-024-07421-0

Chuang, Y-S., Xie, Y., Luo, H., Kim, Y., Glass, J., & He, P. (2024). DoLa: Decoding by contrasting layers improves factuality in large language models. ICLR 2024.

Debate, Self-reflection & Agentic Systems

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. arXiv:2305.14325.

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. arXiv:1805.00899.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Welleck, S., Majumder, B. P., Gupta, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-Refine: Iterative refinement with self-feedback. arXiv:2303.17651.

Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. NeurIPS 2023.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. ICLR 2023. arXiv:2210.03629.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., et al. (2022). Language models (mostly) know what they know. arXiv:2207.05221.

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. ACL 2022. arXiv:2109.07958.