Title

Triadic Regulatory Architecture for Hallucination-Resistant AI

Abstract

A cognitive system comprising at least two functionally distinct processing agents and a regulatory control layer that enforces mutual constraint before output generation. The architecture permits suppression, delay, or prevention of output when abstraction outpaces contextual grounding or vice versa. The system enables hallucination prevention by refusing to finalize responses without reciprocal validation between divergent cognitive functions. A secondary interface system may additionally identify dominant cognitive tendencies in users and dynamically invoke opposing generative styles to introduce productive cognitive tension.

Technical Field

This invention relates to artificial intelligence systems and cognitive architecture, specifically to systems for dynamic regulation of generative outputs via mutually constrained cognitive agents.

Background

Traditional AI systems rely on confidence thresholds, reinforcement alignment, or ensemble consensus to manage generative behavior. These systems lack mechanisms for internal restraint based on structural conflict between distinct modes of cognition. As a result, hallucinations persist in abstraction-heavy domains where contextual feedback is weak or delayed. There exists a need for an architecture that treats *non-generation* as an intelligent response when certainty is unwarranted.

Summary

Disclosed is a system for regulating generative AI output through:

1. Parallel activation of **abstraction-oriented** and **context-oriented** agents;
2. A **regulatory control layer** that enforces mutual constraint between these agents;
3. **Suppression of output** unless reciprocal validation occurs;
4. A secondary module for **cognitive counterweighting** based on detected user patterns.

Figures

- **Fig. 1** — System Architecture Diagram (Triadic)
- **Fig. 2** — Flow Diagram for Output Gating via Constraint
- **Fig. 3** — Pressure Trigger Response Model
- **Fig. 4** — User Cognitive Counterweight Loop

Detailed Description

Claim Group 1 — Triadic Cognitive Regulation

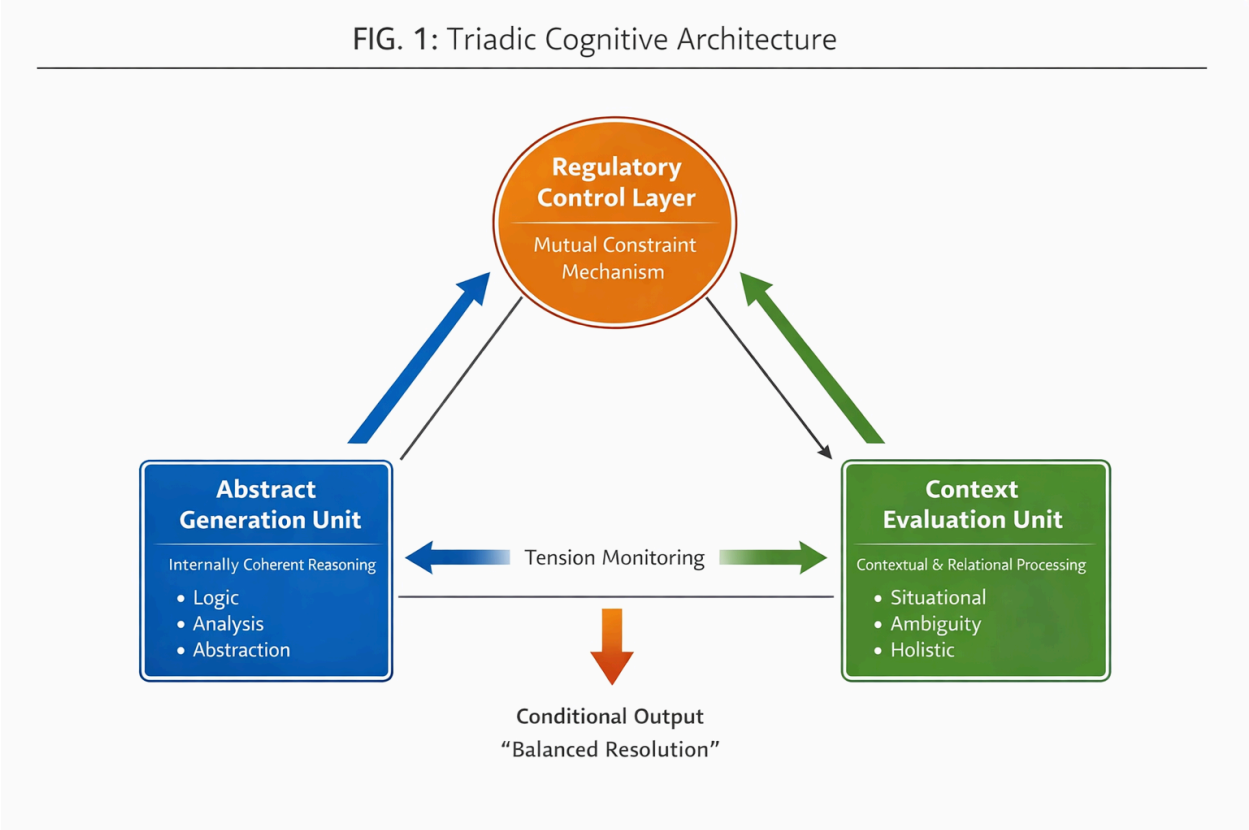

FIG. 1 illustrates the triadic architecture comprising the abstract generation unit, context evaluation unit, and the regulatory control layer that permits or withholds signal output.

Claim 1.1 (Independent)

A cognitive processing system comprising:

- a first agent generating internally coherent abstract representations;
- a second agent evaluating contextual relevance and ambiguity tolerance;
- a regulatory control layer that suppresses output unless mutual constraint conditions are satisfied.

FIG. 2: Flow Diagram for Output Gating via Constraint

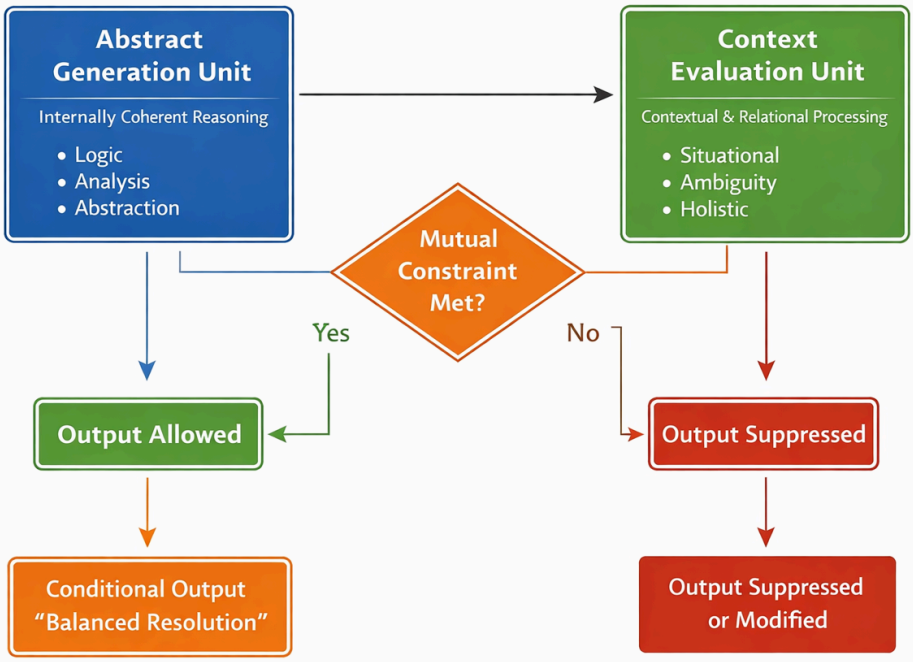

FIG. 2 depicts the conditional flow of output generation, including suppression pathways when mutual constraint is unmet between cognitive agents.

Claim 1.2

The system of 1.1, wherein the regulatory control layer dynamically delays or prevents output when abstraction pressure exceeds contextual support.

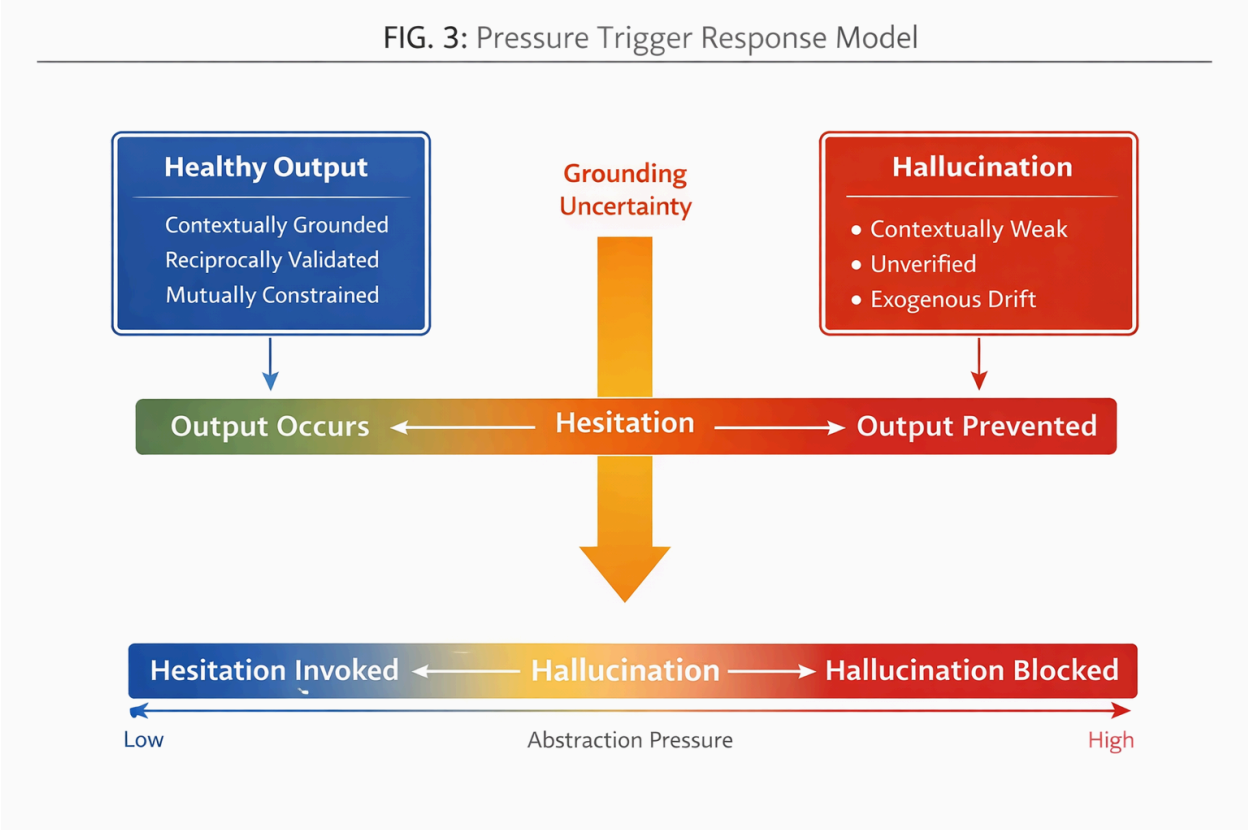

FIG. 3 shows how the system modulates output behavior dynamically in response to increasing abstraction pressure or insufficient grounding, activating hesitation instead of hallucination.

Claim 1.3

The system of 1.1, wherein hallucination is mitigated by enforced hesitation when internal coherence outpaces grounding signals.

Claim 1.4

The system of 1.1, wherein signal output occurs only upon convergence between the two agents, yielding a third-field result.

Claim 1.5

The system of 1.1, wherein the control layer functions analogously to executive inhibition.

Exclusions: This system does not rely on confidence averaging, majority voting, output reranking, or RLHF filters.

Claim Group 2 — Cognitive Counterweight System

Claim 2.1

A system comprising a detection module that identifies a user's dominant cognitive mode via interaction patterns or explicit configuration.

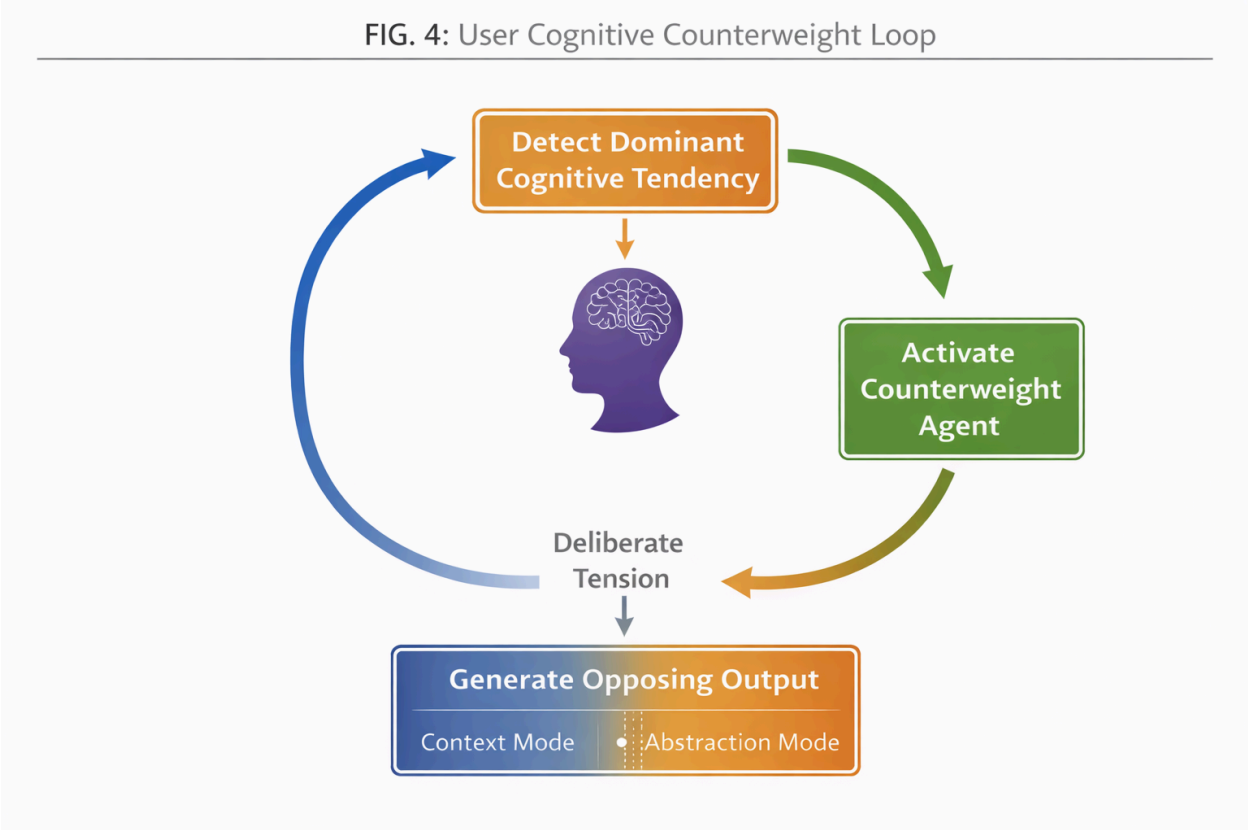

FIG. 4 illustrates the loop mechanism wherein user cognitive dominance is detected and counterweighted by activating an opposing generative mode, forming a deliberate tension loop for ideation.

Claim 2.2

The system of 2.1, wherein the generative agent activates an opposing cognitive mode to induce productive friction.

Claim 2.3

The system of 2.2, wherein the degree of opposition is user-adjustable.

Claim 2.4

The system of 2.2, wherein outputs are non-mirroring and designed to increase generative divergence.

Claim 2.5

The system of 2.1, applied to recursive ideation, creative synthesis, or problem decomposition.

Applications

- Hallucination-resistant LLM architectures
- Trustworthy recursive AI systems
- Divergent co-pilot tools for neurodivergent users
- Creative and problem-solving environments requiring productive internal debate

---