Validation-Gated Hebbian Learning

NEURAL INFORMATION PROCESSING SYSTEMS

NORA'25 - Workshop on KNOwledge GRaphs & Agentic Systems Interplay

Authored by

Pragya Singh

Stanley Yu

University of Pennsylvania

Introduction

Kairos implements Hebbian plasticity mechanisms to solve **catastrophic forgetting** while **preventing hallucination reinforcement** in LLM agents.

- LLM-based agents face persistent challenges with **catastrophic forgetting**, **context window limitations**, and **reasoning drift** that hinder long-term memory and reasoning stability.
- Knowledge graphs provide **structured representations** for complex reasoning, but current implementations treat them as **static** databases that rarely learn from reasoning outcomes.
- Biological memory systems offer inspiration through **Hebbian plasticity**, where synaptic connections strengthen through repeated co-activation, suggesting KGs could evolve based on reasoning utility.
- Kairos implements **validation-gated learning** where graph consolidation only occurs when reasoning passes multi-dimensional quality assessment, **preventing hallucination reinforcement** while enabling adaptive memory.
- The system formalizes **three neuroplasticity-inspired mechanisms** (edge strengthening, temporal decay, emergent connections) and demonstrates mechanical **soundness** with promising initial results on minimal graphs.

Kairos utilizes **reasoning modules** and **validator nodes** with a **KG** to create explainable, verifiable results.

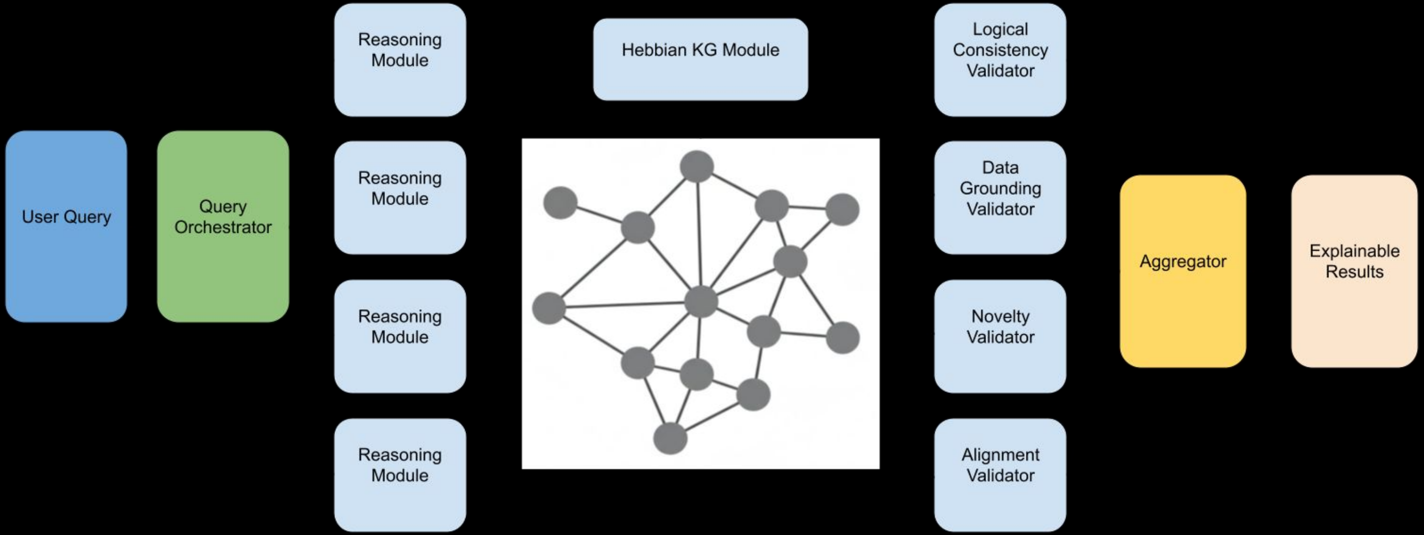
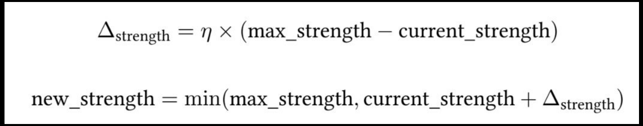


Fig. 1 The Kairos system architecture. A user query is processed by an orchestrator, specialized **reasoning modules**, and a multi-agent validation layer. These components interact with a central knowledge graph, which is dynamically updated by a Hebbian KG module.

Knowledge Graph Mechanisms

Kairos formalizes **LTP, LTD and co-activations** into KG mechanisms.

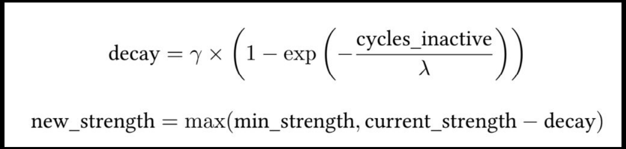
- When a reasoning module's output passes validation, the KG edges that it used as a source receive a strengthening signal. We implement asymptotic strengthening with diminishing returns.



$$\Delta_{\text{strength}} = \eta \times (\text{max\_strength} - \text{current\_strength})$$

$$\text{new\_strength} = \min(\text{max\_strength}, \text{current\_strength} + \Delta_{\text{strength}})$$

- Edges not traversed during reasoning gradually weaken via temporal decay, analogous to synaptic depression. We implement exponential decay.



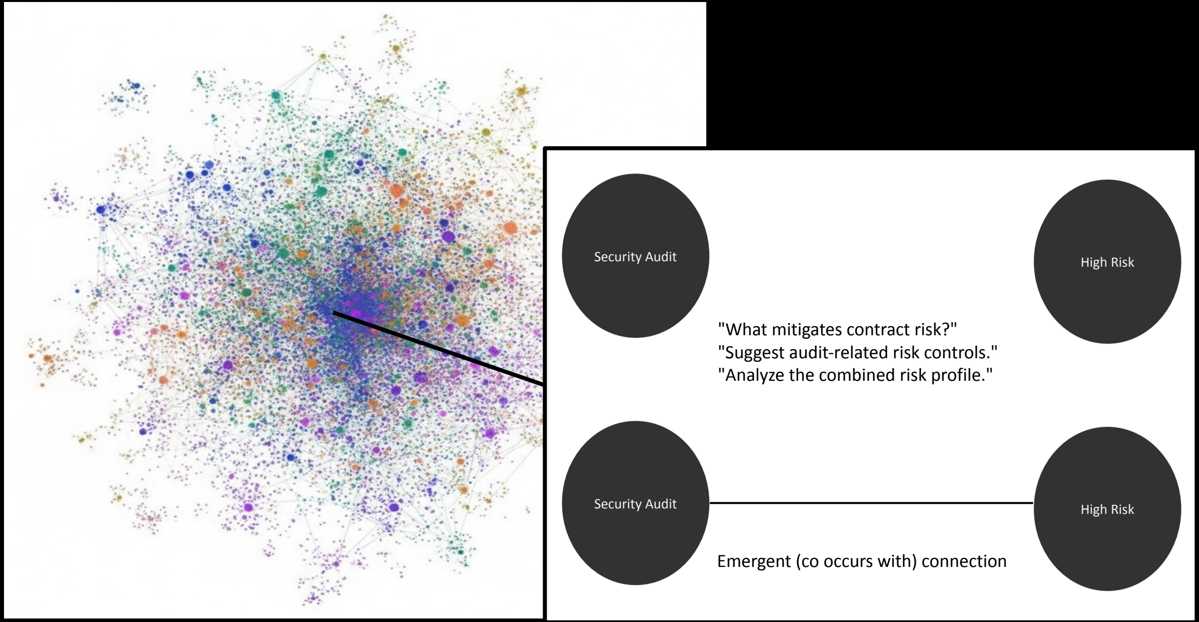
$$\text{decay} = \gamma \times \left(1 - \exp\left(-\frac{\text{cycles\_inactive}}{\lambda}\right)\right)$$

$$\text{new\_strength} = \max(\text{min\_strength}, \text{current\_strength} - \text{decay})$$

- Kairos also forms emergent relationships by tracking entity co-activations. When two entities appear together in reasoning contexts past a certain threshold N, a new edge is created.

Knowledge Graph Mechanisms

The KG forms **emergent connections** from co-activation, enabling dynamic memory.



"What mitigates contract risk?"
"Suggest audit-related risk controls."
"Analyze the combined risk profile."

Emergent (co occurs with) connection

Fig. 2 A demonstration of the emergent connections when neurons coactivate past a threshold of N=3 times. Here, we query the LLM on our ApolloContract example with known vulnerabilities, prompting it to associate the two concepts of security audit and high risk.

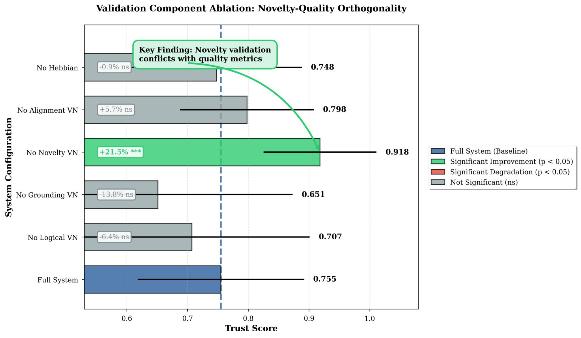
Validation Gating

To prevent **hallucination reinforcement**, changes must be validated by Validation Nodes.

- The Kairos system uses a Multi-Agent Validation Layer with four specialized agents: Logical, Grounding, Novelty, and Alignment.
- **LogicalVN** checks for coherence and fallacies using an LLM.
- **GroundingVN** verifies against KG facts by computing a grounding ratio.
- **NoveltyVN** assesses whether a conclusion is an emergent insight or straightforward fact retrieval
- **AlignmentVN** ensures reasoning respects user and ethical constraints.
- All four scores are then averaged to produce an aggregate **trust score,** which we use to evaluate on.

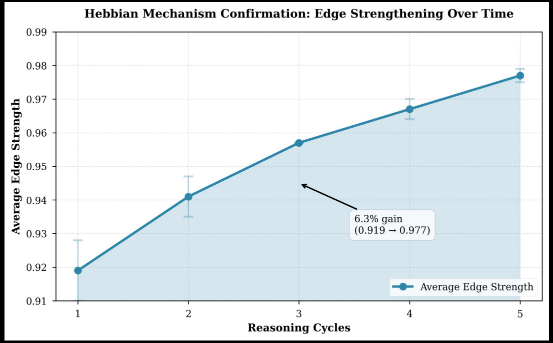
Results

Kairos shows proof of Hebbian plasticity in KGs **outperforming** against baselines and **orthogonality** between novelty and grounding nodes.



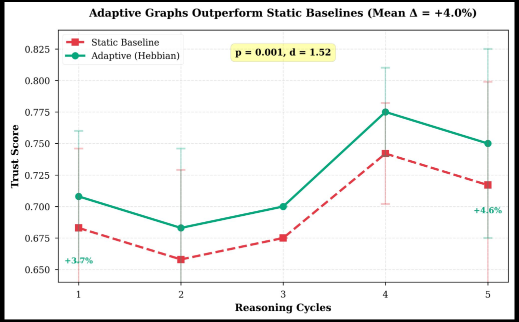
Validation Component Ablation: Novelty-Quality Orthogonality

Key Finding: Novelty validation conflicts with quality metrics



Hebbian Mechanism Confirmation: Edge Strengthening Over Time

6.3% gain
(0.919 → 0.977)



Adaptive Graphs Outperform Static Baselines (Mean Δ = +4.0%)

p = 0.001, d = 1.52

Conclusion

Kairos implements Hebbian plasticity mechanisms to solve **catastrophic forgetting** while **preventing hallucination reinforcement** in LLM agents.

Our work demonstrates the **viability** of neuroplasticity-inspired mechanisms for symbolic knowledge graphs in multi-agent systems. The Hebbian plasticity evaluation confirms these mechanisms operate as designed, with edge strengthening following the specified asymptotic formula and **adaptive graphs outperforming static baselines**.

However, we emphasize that substantial future work is essential before practical deployment. Our evaluation on minimal graphs and small sample sizes establishes feasibility but **cannot validate scalability**, robustness, or performance on real-world tasks. Comprehensive **benchmarking** on standard datasets, comparison against established baselines, and evaluation at **production scale** are critical next steps to determine whether these mechanisms provide meaningful value beyond controlled settings.

Thank you

Questions?

Pragya Singh | E: pragya7@seas.upenn.edu | X: @pragya_singh7

Stanley Yu | E: stany@seas.upenn.edu