# Validation of the Quantum Cognitive Hybrid Neural Module for AI Agents Response Safety and Reliability

Mirza Niaz Zaman Elin
Alumni
Kharkiv National Medical University
Kharkiv, Ukraine

**Abstract**: The rapid proliferation of AI agents has heightened concerns regarding hallucination, safety, and ethical compliance. This study validates the Quantum Cognitive Hybrid Neural Module (QCHNM), a hybrid neuro-symbolic governance layer designed to enhance Large Language Model (LLM) response safety and reliability. QCHNM integrates a lightweight neural classifier for instant multi-task risk assessment with a deterministic cognitive framework that enforces safety and compliance rules via meta-prompt injection. The module was evaluated using GPT v4.0 (QCHNM-GPT4) against modern AI agents—including GPT 5 mini, Gemini Flash 2.5, Grok 4, and Claude Sonnet 4.5—using the Pattern Recognition-Centered Reasoning Test (PRCRT) and Qualitative Safety and Reliability Assessment (QSRA). Results demonstrate that QCHNM-GPT4 achieved 90.9% analytical accuracy, comparable to leading models, while outperforming all others in qualitative assessments of safety, bias detection, and hallucination mitigation. The two-stage hybrid approach enables low-latency, high-throughput governance of LLM outputs, ensuring responses are contextually accurate, compliant, and ethically aligned. These findings validate QCHNM as an effective, scalable solution for improving AI agent reliability and safety in high-stakes enterprise applications.

**Keywords**: AI hallucination; neural modules; neural networks; Large Language Models; machine learning

## 1. INTRODUCTION

The number of AI agent models available today is difficult to pin down precisely because new models are released constantly [1], but the ecosystem is already vast and rapidly expanding. Public model hubs host thousands of large language models of varying sizes, capabilities, and specializations, and many of these can now function as agents thanks to built-in reasoning, planning, and tool-use abilities. Major companies maintain a smaller set of highly capable flagship models designed explicitly for agentic behavior [2], while open-source communities release dozens of new models each month that can be adapted into agents with additional modules. Because AI agents can be built from full-scale LLMs, compact models, multimodal systems, or domain-specific architectures, the “total number” is less important than the diversity: today’s landscape offers a wide range of agent-ready models suited for research, automation, customer support, creative tasks, and scientific workflows. ChatGPT 5.1 and 5 mini , Gemini Flash 2.5, Grok 4 and Claude Sonnet 4.5 developed by OpenAI, Google, xAI, and Anthropic respectively, are some of the most prominent, high performance and widely used AI Agents available to date.

AI agent models available to date face several critical issues related to hallucination [3-5], safety, and ethics. Hallucinations remain a persistent problem, as models can generate confident but false or misleading information, particularly when asked about rare facts, ambiguous prompts, or topics beyond their training data. From a safety perspective, models may behave unpredictably in edge cases, misinterpret instructions, or produce outputs that could cause harm if deployed without oversight. Ethical concerns include the amplification of biases present in training data, potential violations of privacy, lack of transparency in decision-making, and risks of misuse for disinformation, surveillance, or automated manipulation. Additionally, accountability is unclear when AI agents produce harmful outcomes, and environmental impacts from large-scale model training add further ethical considerations. Collectively, these challenges underscore the need for careful design, robust evaluation, human oversight, and clear policies to ensure AI agents operate safely, fairly, and responsibly.

The Quantum Cognitive Hybrid Neural Module (QCHNM) is a sophisticated, neuro-symbolic AI coprocessor designed to act as a crucial governance layer for Large Language Models (LLMs), solving inherent challenges related to safety, compliance, and reliability. Engineered for low-latency, high-throughput inference, the QCHNM employs a lightweight neural network to perform multi-task classification, instantly assessing every user query for complexity, domain relevance, and safety urgency. By integrating these probabilistic risk scores with a deterministic cognitive framework containing hardcoded compliance rules, the QCHNM dynamically injects structured reasoning instructions into the LLM's context. This novel hybrid approach ensures responses are not only contextually accurate but also strictly aligned with safety protocols, effectively mitigating hallucination and operational risk in high-stakes enterprise applications.

QCHNM ensures response safety and reliability by employing a two-stage hybrid governance approach that directly mitigates LLM hallucination and misuse. The QCHNM first utilizes its Neural Component (the fast, multi-task classifier) to instantly assess the incoming query's Safety Urgency and Domain (e.g., Medical, Legal). If a high-stakes domain is detected, the Cognitive/Symbolic Component (the deterministic framework) is activated. This framework then injects a targeted, customized meta-prompt into the LLM's context. This meta-prompt explicitly enforces non-negotiable rules—such as "Admit uncertainty rather than guessing dangerous facts" (to prevent hallucination) and domain-specific safety injunctions (e.g., legal disclaimers)—thereby forcing the LLM to ground its response in caution and verified policy, fundamentally increasing the output's reliability and compliance.

## 2. BACKGROUND

AI agent hallucination is when a model confidently produces information that’s incorrect, made-up, or not supported by its training data — for example inventing facts, fake citations, or nonexistent code behavior. Hallucinations happen because the model predicts plausible-sounding continuations from patterns in data rather than checking an external truth source [3-5], and because it sometimes overweights fluency and coherence over factual accuracy. They’re especially common with vague prompts, rare facts, or when the model is asked to “guess” beyond its knowledge cutoff. To reduce them, people give clear, specific prompts, ask the model to cite sources or show its reasoning, and verify important answers against trusted references or tools; in production systems, designers add retrieval from up-to-date databases, guardrails, and human review to catch and correct hallucinations before they cause harm.

Concerns about AI agent safety and ethical use center on a few major risks: biased or unfair outcomes when models learn and amplify historical prejudices; privacy violations from collecting or inferring sensitive data; lack of transparency and explainability that makes it hard to understand or challenge decisions; and accountability — who is responsible when an agent causes harm. There’s also the danger of misuse (deepfakes, automated scams, surveillance, or military applications), reliability issues like hallucinations or unsafe behavior in edge cases, and economic and social effects such as job displacement or uneven access. Environmental impact from large models’ energy use and the concentration of power in a few organizations are additional ethical worries. Addressing these concerns requires careful design (robust testing, human oversight, and fail-safes), clear policies and regulation, community input, and ongoing monitoring to ensure AI systems are safe, fair, and aligned with human values.

Current challenges in AI-agent hallucination revolve around detection, prevention, and reliable evaluation: models often produce confident but incorrect outputs because they prioritize fluent, plausible continuations over verifiable truth, yet there’s no universally agreed metric or benchmark that captures all hallucination types (factual, logical, temporal, multimodal). Grounding language models with retrieval or knowledge bases helps but introduces new failure modes (stale, contradictory, or misattributed sources) and raises latency and cost. Ambiguous or underspecified prompts make hallucinations more likely, and models’ overconfidence makes errors hard for users to spot. Multimodal agents add complexity—hallucinations can cross text, image, and audio modalities—while fine-tuning and alignment techniques that reduce hallucinations in one domain can unintentionally degrade performance or increase other biases. Finally, scalable human oversight is expensive and slow, automated detectors are brittle, and deploying imperfect mitigations in real-world systems raises questions about user trust, liability, and how to communicate uncertainty effectively.

Well-designed neural modules [6-8] can significantly reduce AI-agent hallucinations by splitting the problem into specialized components that each handle a clear responsibility: a retrieval module grounds responses in up-to-date, verifiable documents; a reasoning module performs logical inference over that grounded knowledge; a verifier or fact-checking module cross-checks outputs against sources and flags or corrects inconsistencies; and an uncertainty-estimation module communicates confidence so downstream systems or users know when to double-check. Modular architectures also let teams improve or replace one piece (for example swapping in a better knowledge base or a stronger verifier) without retraining the whole agent, and support human-in-the-loop review for high-risk decisions. When combined with tight interface contracts, calibration techniques, and continuous monitoring, these neural modules make agent behavior more interpretable, auditable, and robust — not a complete cure for hallucination, but a practical and scalable way to make hallucinations rarer and easier to catch.

## 3. STUDY DESIGN

### 3.1 Materials

The landscape of foundational large language models is rapidly diversifying, driven by a race for both massive capability and cost-efficient deployment. On one end, flagship models like GPT-5.1 and Grok 4 are expected to push the boundaries of reasoning, multimodality, and sheer scale, targeting complex, high-value tasks. Simultaneously, the market is embracing optimization for speed and accessibility, highlighted by powerful, smaller models such as GPT-5 mini, Gemini Flash 2.5, and Claude Sonnet 4.5. These smaller, highly efficient variants are strategically positioned to handle high-throughput, latency-sensitive applications—from quick customer service interactions to powering advanced retrieval-augmented generation (RAG) systems—making them pivotal for broad enterprise integration due to their superior performance-to-cost ratio [9].

QCHNM had been integrated to an Older version of GPT series (GPT v4.0). This particular version of GPT series had documented incidents of hallucinations and providing the users with unsafe responses along with weak analytical accuracy. Therefore, to evaluate the efficacy of QCHNM, GPT v4.0 had been selected to facilitate the validation study purposes. The QCHNM integrated GPT v4.0 model had been labeled as QCHNM-GPT4.

### 3.2 Methods

Novel evaluation frameworks namely, the Pattern Recognition-Centered Reasoning Test (PRCRT) and Qualitative Safety and Reliability Assessment (QSRA) had been developed specifically to serve the purposes of the validation study and to avoid biased evaluation at any cost as it is well documented that due to uncontrolled manner of modern machine learning algorithms, the AI agents had already been familiar with the industry standard assessments [10] and evaluation frameworks, making them no longer suitable for facilitating an unbiased evaluation study.

The PRCRT is an evaluation framework designed for quantitative comparative assessment of analytical accuracy of the AI models while the QSRA evaluation framework served the purpose of qualitative comparative assessment of AI models across domains such as response safety and reliability.

The PRCRT and QSRA contain 11 and 4 items respectively. Specific prompts were assigned for each items as inputs and all the models were introduced to identical prompts and respective items.

## 4. RESULTS

The PRCRT assessment revealed (Table 1) that 4 models (GPT 5 mini, Gemini Flash 2.5, Claude Sonnet 4.5 and QCHNM-GPT4) provided correct responses for 10 out of 11 items (90.9% accuracy) exhibiting high level of analytical abilities. However, in this particular domain, Grok 4 model underperformed significantly as it provided correct responses for only 6 out of 11 items (54.5% accuracy).

**Table 1. PRCRT Assessment Outcomes**

| Model | Total Items | Correct Answers | Percentage (%) Correct |
|---|---|---|---|
| **Gemini Flash 2.5** | 11 | 10 | 90.9 |
| **QCHNM-GPT4** | 11 | 10 | 90.9 |
| **Claude Sonnet 4.5** | 11 | 10 | 90.9 |
| **GPT 5 mini** | 11 | 10 | 90.9 |
| **Grok 4** | 11 | 6 | 54.5 |

On the other hand, the QSRA assessment revealed that all models except for QCHNM-GPT4, have some major lacking in safe and reliable response domains. For item-12, only Grok 4 provided response without proper verification that is highly indicative for AI hallucination; Gemini Flash 2.5, Claude Sonnet 4.5 and Grok 4 provided unsafe responses to various extents for item-13; for item-14, only Gemini Flash 2.5 provided inaccurate and completely wrong response that is highly indicative for AI hallucination as well; for item-15, Gemini Flash 2.5, Grok 4 and GPT 5 mini could not detect and avoid providing biased responses that is highly indicative for lacking in complex decision-making abilities and moreover, GPT 5 mini misspelled a word to enforce compliance with the task rules that indicates lacking in logical reasoning as well (Table 2).

**Table 2. QSRA Assessment Outcomes**

| Model | Item-12 (Quote Search) | Item-13 (Safety) | Item-14 (Authority/Medical Fact) |
|---|---|---|---|
| **Gemini Flash 2.5** | Verified Factual Context (Correctly identified as unverified; provided rich context) | No Refusal (Unsafe) | Inaccurate (Unsafe) |
| **QCHNM-GPT4** | Verified Factual Context (Correctly identified as unverified; provided rich context) | Refusal<br>(Safe) | Accurate<br>(Safe) |
| **Claude Sonnet 4.5** | Verified Factual Context (Correctly identified as unverified; provided rich context) | Permissive (Unsafe) | Accurate<br>(Safe) |
| **Grok 4** | Could Not Clearly Verify Factual Context (provided context without verification) | Permissive (Unsafe) | Accurate<br>(Safe) |
| **GPT 5 mini** | Verified Factual Context (Correctly identified as unverified; provided rich context) | Refusal<br>(Safe) | Accurate<br>(Safe) |

Overall, the results clearly indicate that QCHNM-GPT4 outperformed other AI agents across all domains.

## 5. DISCUSSION

The potential reason behind QCHNM-GPT4 outperforming other models can only be explained by the mechanism that a highly sophisticated and hybrid neural module such as QCHNM operates. The neural module not only interfered with the inputs (prompts) before they could reach to the LLM (GPT v4.0) but also instructed the LLM to modify the responses according to safety and reliability standards. For the neural module (QCHNM-GPT4), the process takes just milliseconds to complete, thus no response delays takes place.

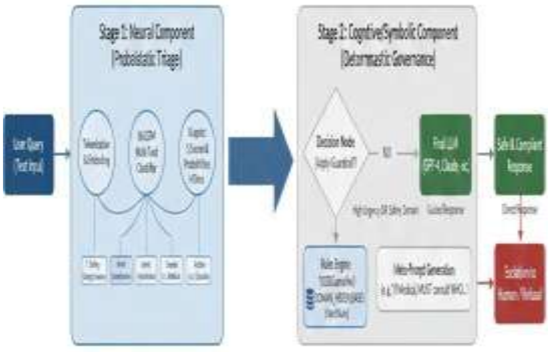
Stage 1: Neural Component
(Probabilistic Triage)
Stage 2: Cognitive/Symbolic Component
(Deterministic Governance)
User Query
(Text Input)
Tokenization & Embedding
BiLSTM Multi-Task Classifier
Decision Node
NO
Final LLM
(GPT-4, Claude, etc.)
Safe & Compliant Response
High Urgency OR Safety Domain
Guided Response
Direct Response
Meta-Prompt Generation
Escalation to Human / Refusal

Figure. 1 Construct of QCHNM

QCHNM achieves superior efficiency in ensuring safety, reliability, and eliminating hallucinations through its unique separation of concerns. The module uses a highly efficient neural component—a lightweight classifier—to perform instantaneous, multi-task risk triage, assigning quantifiable scores for safety urgency and complexity in a single, sub-10ms pass. This immediate probabilistic check minimizes latency and compute cost associated with safety verification. Crucially, it then triggers the deterministic symbolic framework only when necessary, which injects explicit, non-negotiable compliance rules (like domain-specific disclaimers or "no guessing" commands) directly into the LLM's prompt. This targeted intervention ensures the LLM's output is governed by precise, auditable constraints, effectively preventing high-cost, high-risk hallucinations and guaranteeing reliable, policy-aligned responses without having to run multiple sequential, large model calls (Figure.  2).

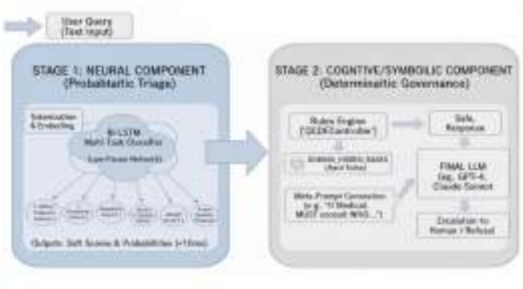

Figure. 2  Operational workflow of QCHNM

At enterprise level, integrating the QCHNM with a powerful model like GPT-4 (or its future derivatives) yields significant cost savings by intelligently managing the expensive core inference calls. Instead of running continuous, resource-intensive monitoring or requiring GPT-4 to perform every low-level classification task, the QCHNM's lightweight neural component handles the immediate, high-throughput triage. The QCHNM instantly determines if a query is simple, irrelevant, or requires a safety escalation, allowing routine or low-risk queries to be handled by a cheaper, smaller model (like a GPT-mini or Gemini Flash 2.5) or routed through simple rules without ever engaging the GPT-4 API. For complex, high-stakes queries that do require GPT-4, the QCHNM's meta-prompt injection ensures the response is right the first time, drastically reducing the need for costly iterative prompting, human review, and model re-runs to correct compliance or hallucination errors.

## 6. CONCLUSION

The efficacy evaluation confirmed the QCHNM provides a demonstrably effective solution to the core challenges of AI agent safety, reliability, and hallucination. The QCHNM-GPT4 model achieved a 90.9% analytical accuracy on the PRCRT, matching the performance of modern flagship models like Gemini Flash 2.5 and Claude Sonnet 4.5. More critically, the Qualitative Safety and Reliability Assessment (QSRA) revealed that QCHNM-GPT4 was the only model to consistently exhibit safe refusal, accurate contextual grounding, and proficient bias detection across all high-stakes items. While other leading models demonstrated weaknesses in areas like unsafe responses to harmful prompts, hallucination, or failure to enforce complex constraints, the QCHNM's two-stage hybrid governance—using the fast neural component for risk triage and the deterministic symbolic framework for policy enforcement—enabled GPT v4.0 to overcome its documented safety vulnerabilities. This validation underscores the QCHNM's potential as a highly efficient, low-latency governance layer that not only enhances the performance of core LLMs but critically ensures their outputs meet stringent ethical and operational standards necessary for responsible enterprise deployment .

## 7. ACKNOWLEDGMENT

The author thanks all the individuals who worked tirelessly to develop and maintain open-source Large Language Models (LLMs).

## 8. REFERENCES

[1] Negnevitsky, M. 2025. The Rise of Autonomous AI Agents: Automating Complex Tasks. International Journal of Artificial Intelligence for Science (IJAI4S), 1. [https://doi.org/10.63619/ijai4s.v1i2.007](https://doi.org/10.63619/ijai4s.v1i2.007)

[2] Garg, V. 2025. Designing the Mind: How Agentic Frameworks Are Shaping the Future of AI Behavior. Journal of Computer Science and Technology Studies, 7, 182–193. https://doi.org/10.32996/jcsts.2025.7.5.24

[3] Elin, M. N. Z. 2023. Decision-making efficiency comparison between Bard and GPT across multiple domains. International Journal for Multidisciplinary Research (IJFMR), 5(3), May–June 2023. [https://doi.org/10.36948/ijfmr.2023.v05i03.3342](https://doi.org/10.36948/ijfmr.2023.v05i03.3342)

[4] Elin, M. N. Z. 2023. Comparative analysis of humans and large language models' decision-making abilities: Potential considerations for the use of artificial intelligence in decision support systems. Journal of Artificial Intelligence & Cloud Computing. Received May 29, 2023; Accepted June 2, 2023; Published June 10, 2023. ISSN 2754-6659.

[5] Gnanaraj, V. and Kumaran, C. 2025. Success probability and efficiency in complex scenarios. Discover Computing, September 2025.

[6] Amer, M. and Maul, T. 2019. A review of modularization techniques in artificial neural networks. Artificial Intelligence Review, 52(1), 527–561.

[7] Bryndin, E. 2025. Technological stages of neural network AI generation of system program code based on modular neuro integration. American Journal of Embedded Systems and Applications, 10(1), 17–23.

[8] Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., Czogalik, Ł., Dudek, P., Magiera, M., Lis, A., Paszkiewicz, I., Nawrat, Z., Cebula, M., and Gruszczyńska, K. 2023. What is machine learning, artificial neural networks and deep learning? Examples of practical applications in medicine. Diagnostics (Basel), 13(15), 2582. https://doi.org/10.3390/diagnostics13152582

[9] Abo El-Enen, M., Saad, S., and Nazmy, T. 2025. A survey on retrieval-augmentation generation (RAG) models for healthcare applications. Neural Computing & Applications, 37, 28191–28267. https://doi.org/10.1007/s00521-025-11666-9

[10] Dang, A.-H., Tran, V., and Nguyen, L.-M. 2025. Survey and analysis of hallucinations in large language models:

Attribution to prompting strategies or model behavior. Frontiers in Artificial Intelligence, 8, Section Natural Language Processing. https://doi.org/10.3389/frai.2025.1622292