Cognitive-Inspired Neural Architectures: Enhancing Interpretability and Ethical Alignment for Safe Decision-Making in High-Impact Sectors.

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Varun Reddy

Spring 2025

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Sean Murray, Department of Engineering and Society

The question is not whether intelligent machines can have any values, but whether humans can still recognize their own.

— Eliezer Yudkowsky

1. Introduction:

Recent advancements in artificial intelligence (AI) have resulted in models with remarkable performance across various tasks; however, these systems still fall short in key areas where human intelligence excels, such as abstract reasoning, symbolic representation, and social intelligence (Anderson, 1990; Tenenbaum et al., 2011). Unlike humans, who interact continuously with a dynamic, real-world environment and learn through tightly integrated sensory and cognitive feedback loops, AI systems often lack grounded, adaptable understanding. This sociotechnical gap raises a crucial question: how can we develop AI models that better perform cognitive tasks while aligning with human values to remain trustworthy in complex social contexts? This challenge becomes increasingly urgent as AI shapes decisions in high-stakes domains such as healthcare, justice, and education. While neuroscience and cognitive science provide insights into human cognition that could inspire more sophisticated AI architectures, implementing these insights also introduces challenges in model interpretability. Mechanistic interpretability, which seeks to bridge AI transparency and human ethical standards, emerges as a critical component for aligning AI systems with societal values and ensuring accountability.. This study employs the STS lenses of interpretative flexibility, social construction of technology, and ethical frameworks to critically examine how AI technologies embed and reflect societal values.

2. Problem Definition:

A fundamental challenge with large language models (LLMs) is their tendency to produce hallucinations—instances in which the model generates information that is factually incorrect or fabricated. These hallucinations arise because LLMs do not possess a true understanding of knowledge in the way humans do; rather, they generate responses based on learned statistical patterns from their training data. The issue becomes particularly problematic when LLMs serve as knowledge bases, as they often present answers with high confidence, even when internal uncertainty exists. Several factors can contribute to this uncertainty: question vagueness, task difficulty, or scenarios outside the model's training data. (Figure 1). However, from the perspective of a human user, these models appear authoritative in their responses, and unlike in human communication, we cannot rely on social cues like vocal inflections or body language to gauge the model's actual confidence. This challenge poses significant risks in high-stakes applications such as law and healthcare, where objectivity is critical and erroneous information can have severe consequences. If a legal or medical professional unknowingly acts on a confidently stated but incorrect response, the results could be legally or ethically disastrous. Addressing the challenge of LLM hallucinations requires technical improvements and a comprehensive societal framework for governing AI reliability, interpretability, and ethical accountability, especially in sensitive contexts such as healthcare and legal systems.

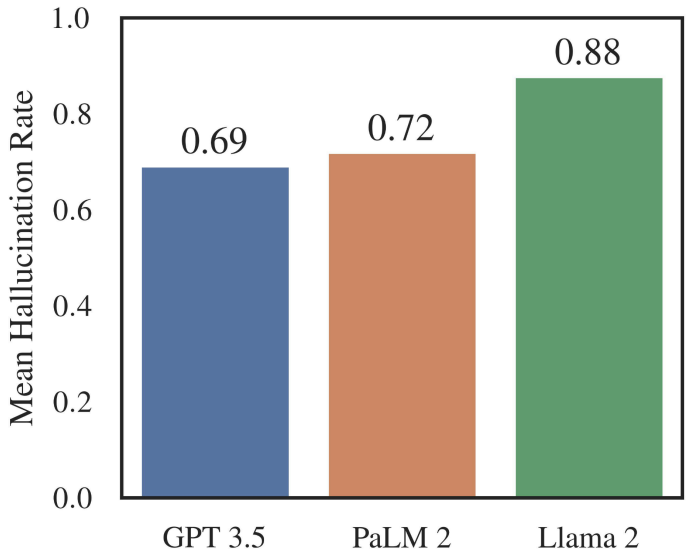

Figure 1: Mean Hallucination Rate of Flagship LLMs (Wei, 2022)

One promising direction to tackle this issue is through scientific study using frameworks from cognitive science and neuroscience, which have long been employed to understand natural intelligence. By applying mechanistic interpretability techniques, we can analyze how LLMs process information at a structural level, potentially identifying where and why these hallucinations emerge (Barsalou, 2008). Another promising approach involves embedding principles of uncertainty expression directly into LLMs. This can be achieved by integrating symbolic reasoning frameworks and prompting techniques such as chain-of-thought (CoT) reasoning, which allow models to break down their reasoning process step by step. By doing so, LLMs can provide users with insight into their confidence levels, reasoning paths, and potential sources of error, allowing for a more transparent and reliable AI decision-making process. Developing LLMs that explicitly communicate uncertainty and exhibit structured reasoning will be crucial for improving their trustworthiness in domains where precision and accountability are paramount (Sharkey, 2025).

3. Research Frame:

This study explores three key research questions that aim to deepen our understanding of large language models (LLMs) from both a cognitive science and AI interpretability perspective. By examining LLMs through the lens of mechanistic analysis, human-inspired reasoning, and the broader ecosystem of AI development, we can assess their internal knowledge representations, reasoning capabilities, and the external factors influencing their transparency and safety.

1. **Evaluating Cognitive Science Techniques for Understanding LLM Knowledge Representations**

   A fundamental challenge in AI research is determining how LLMs internally structure and represent knowledge. Traditional evaluations focus on benchmark accuracy and performance metrics, yet these methods fail to capture the nuances of how LLMs "understand" and generalize information. We aim to quantify the alignment between LLM knowledge representations and humanlike abstraction to determine whether existing interpretability techniques are sufficient for understanding LLM decision-making. Questions include the following:

   - How effective are cognitive science-inspired methods, such as probing techniques, mechanistic interpretability, and concept-based analysis, in uncovering LLMs' latent knowledge structures?
   - What do these methods reveal about how LLMs encode, retrieve, and contextualize knowledge, particularly in cases of hallucination or uncertainty?
   - Can insights from natural intelligence research provide a framework for assessing the depth and reliability of LLMs' world models?

2. **The Trade-off Between Human-Inspired Reasoning, Interpretability, and Accuracy in LLMs**

Recent advancements in AI have incorporated human-inspired reasoning mechanisms, such as chain-of-thought (CoT) prompting, symbolic reasoning, and uncertainty estimation, to improve model transparency. However, these modifications introduce a trade-off between interpretability and model accuracy. We aim to quantify the trade-offs between interpretability and accuracy and determine whether humanlike reasoning strategies improve the reliability of AI-generated outputs (Wei, 2024).

Questions include:

- To what extent does integrating structured reasoning frameworks enhance LLM interpretability without degrading accuracy?
- How can quantitative evaluation metrics, such as faithfulness of reasoning, robustness under adversarial prompting, and performance variance across reasoning tasks, be used to measure interpretability improvements?
- Does the implementation of explicit uncertainty estimation (e.g., confidence scoring, calibration techniques) improve users' trust in LLM outputs while maintaining high predictive accuracy?

4. Dive into Cognitive Science for AI Interpretability

Understanding how large language models (LLMs) internally represent knowledge is not merely a technical challenge but a deeply sociotechnical issue, inviting exploration of cultural, ethical, and regulatory implications. Cognitive science-inspired interpretability techniques bridge technical capabilities with societal implications, highlighting not only how societies perceive and integrate intelligent technologies but also how societal values shape and constrain these technologies, particularly in high-impact areas such as healthcare, justice, and education.

Mechanistic interpretability, rooted in cognitive science and neuroscience, attempts to decode neural network activations and understand the encoding of concepts within AI. Techniques like Sparse Autoencoders (SAEs), inspired by neuroscience's sparse coding hypothesis, emphasize efficient neural representation of information, reflecting societal expectations around efficiency, transparency, and ethical responsibility. Such interpretability methods provide critical insight into how societal norms and expectations become embedded within technical practices, essential for trustworthy AI deployment in sensitive sectors.

For instance, the discovery of neurons responding specifically to culturally significant symbols, such as Olah's (2020) "Golden Gate Bridge" neuron, goes beyond technical interest to reveal deeper implications of cultural representation within AI systems (see Figure 2). This insight raises profound questions regarding cultural bias, societal trust, and inclusivity in AI technologies. In healthcare, cultural biases might lead to disparities in patient diagnosis or treatment recommendations. In judicial contexts, implicit biases embedded within AI could reinforce systemic inequalities, potentially affecting sentencing decisions. Similarly, educational tools employing AI might inadvertently propagate biases, impacting learning outcomes for diverse student populations.

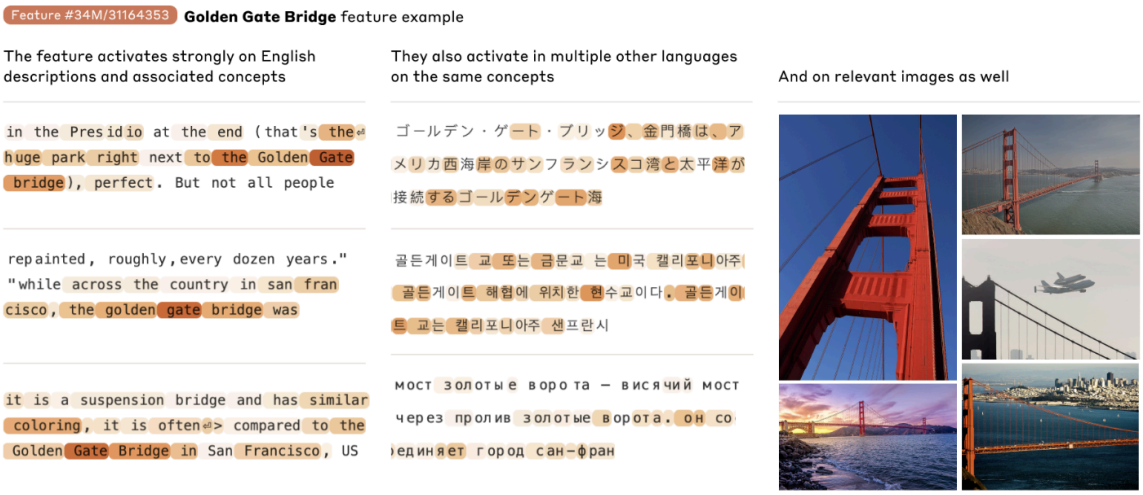

Figure 2: Multimodal Golden Gate Bridge Neuron (Olah, 2020)

Moreover, interpretability faces significant sociotechnical challenges, such as polysemanticity, where single neurons encode multiple unrelated concepts. This complexity mirrors societal concerns about ambiguity, uncertainty, and trust in automated decision-making processes, especially in high-stakes areas like medical diagnostics, legal adjudication, and educational assessments. Reliance on technical metrics such as reconstruction loss alone underscores societal debates about meaningful accountability and ethical alignment measures necessary in these sectors.

The scalability of interpretability techniques further heightens regulatory and governance concerns in healthcare, justice, and education. As LLM complexity grows, their internal opacity poses urgent challenges for regulatory oversight, transparency, and informed consent. For example, opaque AI systems in healthcare raise critical ethical issues concerning patient autonomy and medical accountability. In legal systems, the lack of transparency might

undermine fairness and public trust. Educational settings similarly face challenges in maintaining equity and clarity about how AI-driven assessments and recommendations are generated.

Future research must explicitly integrate societal considerations by:

- Investigating cultural differences in reasoning frameworks and interpretability standards across diverse global contexts, providing insights into how different cultural expectations influence the governance and adoption of AI in healthcare, judicial, and educational settings.
- Evaluating and advocating for robust and culturally sensitive legal frameworks and corporate policies regarding AI transparency, interpretability, and accountability, explicitly addressing the unique ethical challenges present in high-impact areas.
- Developing interpretability frameworks explicitly informed by societal expectations, emphasizing clarity, trustworthiness, and ethical alignment, especially tailored to healthcare diagnostics, judicial fairness, and educational equity.

Moreover, human-inspired interpretability methods such as Chain-of-Thought (CoT) prompting emphasize transparency and structured reasoning, directly responding to societal demands for accountability in sensitive sectors (Wei, 2022) (see Figure 3). In healthcare, CoT could clarify AI diagnostic recommendations; in justice systems, it might enhance clarity around legal reasoning; and in education, it may help make AI-driven learning recommendations comprehensible to students and educators.

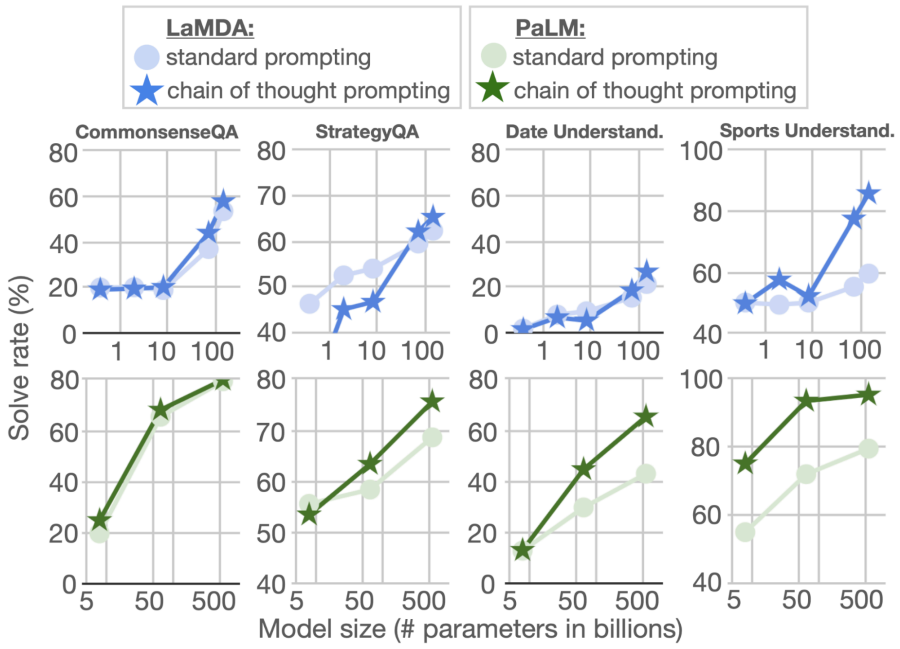

Figure 3: Chain-of-thought prompting enhances LLM interpretability and societal transparency (Wei, 2022)

Similarly, integrating Bayesian and symbolic architectures highlights cultural and institutional commitments to explicit representation of uncertainty and structured reasoning processes vital for ethical decision-making in healthcare, justice, and education. These methodologies align well with societal norms prioritizing accountability and clarity, but face significant implementation challenges due to cultural, institutional, and regulatory variations.

Recent advances, such as symbolic compression and context-aware inference, further illustrate the dynamic tension between technical progress and societal expectations in high-impact areas. These innovations underline the critical need for robust regulatory frameworks, culturally informed governance, and proactive stakeholder engagement, ensuring technological advancements align with societal values and ethical standards, particularly within healthcare, justice, and education.

Ultimately, an STS-informed exploration of cognitive science-inspired AI interpretability calls for continuous societal dialogue and robust governance mechanisms. Addressing interpretability through cultural, regulatory, and corporate lenses ensures responsible AI evolution, maintaining harmony between technological innovation and societal well-being in areas where ethical stakes are highest.

5. Conclusion and Future Considerations

LLMs represent more than just a leap in technical capability; they are complex sociotechnical systems increasingly interwoven into critical societal domains, including healthcare, justice, and education. While their capacity for processing vast information offers potential benefits, their inherent limitations—opacity, susceptibility to hallucination, and lack of genuine understanding—pose significant challenges that extend beyond the purely technical realm. These limitations raise questions about trust, accountability, and the alignment of AI with human values, demanding scrutiny through an STS lens.

The pursuit of interpretability, particularly through cognitive science-inspired methods like mechanistic analysis and chain-of-thought reasoning, emerges as an important societal endeavor. It reflects a societal demand for transparency and accountability in automated systems that exert growing influence over human lives. As discussed, however, even these methods are subject to interpretative flexibility and social construction; technical metrics of transparency do not automatically equate to trustworthiness or ethical alignment. We must examine whether interpretability techniques genuinely reveal underlying processes or inadvertently mask biases, potentially reinforcing existing societal inequalities, especially within sensitive sectors.

Therefore, the deployment of LLMs in high-impact areas necessitates robust governance structures and ethical frameworks. The technical challenge of ensuring AI reliability is inseparable from the societal challenge of defining acceptable risk, ensuring fairness, and maintaining meaningful human control. Current interpretability techniques, while promising, are insufficient to guarantee safe or ethical outcomes independently. Consequently, human oversight should be viewed as a sociotechnical necessity grounded in the ethical responsibility and contextual judgment that societies value, particularly when stakes are high. LLMs should function as tools to augment human expertise within specific, well-defined boundaries, rather than replacing the nuanced decision-making required in complex social situations.

Ultimately, the responsible trajectory for AI development requires a continuous, critical dialogue between technical innovation and societal values, informed by STS perspectives. Future efforts must prioritize not only enhancing model capabilities and transparency but also developing culturally sensitive evaluation standards, participatory design practices, and adaptive governance mechanisms. Ensuring that AI evolution serves human well-being demands a holistic approach that consciously integrates ethical considerations and societal impacts into the core of technological design and deployment.

References

Barsalou, L. W. (2008). *"Grounded Cognition."* Annual Review of Psychology.

Krotov, D., & Hopfield, J. (2021). *Large Associative Memory Problem in Neurobiology and Machine Learning* (No. arXiv:2008.06996). arXiv. https://doi.org/10.48550/arXiv.2008.06996

Kriegeskorte, N., et al. (2008). *"Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience."* Frontiers in Systems Neuroscience.

Katz, Y. (2020). Artificial Whiteness: Politics and Ideology in Artificial Intelligence. Columbia University Press. Ch 3: Epistemic Forgeries and the Ghost in the Machine, 93-126

Mao, J., & Gan, C. (2019). *THE NEURO-SYMBOLIC CONCEPT LEARNER: INTERPRETING SCENES, WORDS, AND SENTENCES FROM NATURAL SUPERVISION*. The Seventh International Conference on Learning Representations (ICLR)

Morris, M. R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2024). *Levels of AGI for Operationalizing Progress on the Path to AGI* (No. arXiv:2311.02462). arXiv. https://doi.org/10.48550/arXiv.2311.02462

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, *5*(3), 10.23915/distill.00024.001. https://doi.org/10.23915/distill.00024.001

Poesia, G., Gandhi, K., Zelikman, E., & Goodman, N. D. (2023). *Certified Deductive Reasoning with Language Models* (No. arXiv:2306.04031). arXiv. https://doi.org/10.48550/arXiv.2306.04031

Sun, R. (2001). Duality of the Mind: A Bottom-up Approach Toward Cognition (1st ed.). Psychology Press. https://doi.org/10.4324/9781410604378

Joshua B. Tenenbaum et al. ,How to Grow a Mind: Statistics, Structure, and Abstraction.Science331,1279-1285(2011).DOI:10.1126/science.1192788

Wang, Y., Gan, C., Siegel, M. H., Zhang, Z., Wu, J., & Tenenbaum, J. B. (n.d.). *A Computational Model for Combinatorial Generalization in Physical Auditory Perception*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (No. arXiv:2201.11903). arXiv. https://doi.org/10.48550/arXiv.2201.11903

Yang, G. R., & Wang, X.-J. (2020). Artificial Neural Networks for Neuroscientists: A Primer. *Neuron*, *107*(6), 1048–1070. https://doi.org/10.1016/j.neuron.2020.09.005

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., Conmy, A., Nanda, N., Rumbelow, J., Wattenberg, M., Schoots, N., Miller, J., Michaud, E. J., Casper, S., Tegmark, M., Saunders, W., Bau, D., Todd, E., Geiger, A., Geva, M., Hoogland, J., Murfet, D., & McGrath, T. (2025). *Open problems in mechanistic interpretability* arXiv. https://arxiv.org/abs/2501.16496

Appendix:

1. **Mechanistic Interpretability** – The study of understanding how specific components (e.g., neurons, layers) of a neural network function by tracing their internal mechanisms. It aims to map model behavior to human-understandable concepts or reasoning processes.

2. **Large Language Model** – A type of neural network, typically transformer-based, trained on massive text corpora to understand, generate, and reason with human language across a wide range of tasks.

3. **Uncertainty Estimation** – The process of quantifying how confident a model is in its predictions, often using techniques like confidence scores, calibration, or Bayesian approaches to increase model transparency and trust.

4. **Chain of Thought** – A prompting technique that encourages a language model to generate intermediate reasoning steps, mimicking human step-by-step logic to improve performance and interpretability on complex tasks.

5. **Sparse Autoencoders (SAEs)** – Neural networks trained to reconstruct input data while enforcing sparsity, meaning only a small subset of neurons activate, which encourages interpretable internal representations.

6. **Polysemanticity** – A phenomenon where a single neuron or unit in a neural network responds to multiple unrelated concepts, making it harder to assign clear semantic

meaning to its activation.

7. **Entanglement** – A condition in neural networks where different concepts are encoded across overlapping neuron sets, making it difficult to isolate individual representations for interpretation.

8. **Scalability** – The capacity of an AI model or method to maintain performance and computational feasibility as its size, input data, or deployment environment grows.

9. **Neuro-symbolic Model** – A hybrid AI approach that combines neural networks' pattern recognition capabilities with symbolic systems' structured reasoning, aiming to achieve both high performance and interpretability.

10. **Neural Network** – A computational architecture inspired by the human brain, composed of interconnected layers of nodes (neurons) that learn representations and patterns from data to make predictions or generate content.

11. **Neuron** – An individual unit in a neural network that computes an activation value based on its input and learned weights, contributing to the model's overall prediction or representation.