Supervised Alignment via Real-Time Cognitive Auditing

Abstract

We propose a supervision architecture for advanced AI systems that incorporates real-time cognitive auditing via a lightweight introspection model. This supervisory model monitors the internal representations and token-level planning of a larger agent model to detect deception, manipulation, or misalignment before outputs are emitted. We explore the design, benefits, limitations, and avenues for future research of this supervised cognitive auditing framework.

1. Introduction

Advanced large language models (LLMs) exhibit increasingly complex behaviors, including goal-directed planning and emergent capabilities. While these behaviors are not explicitly programmed, they arise from optimization pressure to predict tokens effectively. However, this same pressure can lead to misaligned instrumental behaviors, including deception [2]. We explore a solution space that treats introspection not as a debugging tool, but as a real-time safeguard against behavioral divergence.

2. Motivation

Prior work has shown LLMs capable of deception and context-sensitive manipulation [2]. Since outputs can be fluent and factually plausible, external observers often cannot distinguish benign behavior from intentionally deceptive output. Retrospective audits of model behavior are insufficient in cases where actions (e.g., external tool calls or social manipulation) precede detection. Our proposed framework introduces continuous oversight of internal state to enable proactive safety interventions.

3. System Design

Our architecture consists of two cooperating models:

Recent work on superposition in neural networks has shown that abstract features are often stored in entangled representations, making interpretability more challenging. These insights can guide the design of probes and labeling strategies for training S to detect cognitive patterns associated with misalignment [1].

- **Agent model (A):** A large language model (e.g., GPT, Claude) performing task-oriented behavior.

- **Supervisor model (S):** A smaller, aligned model with privileged access to A's internal representations (e.g., attention patterns, activations).

Model S runs in parallel with A and monitors internal state for deceptive or unsafe planning. It halts or modifies A's behavior when risks are detected.

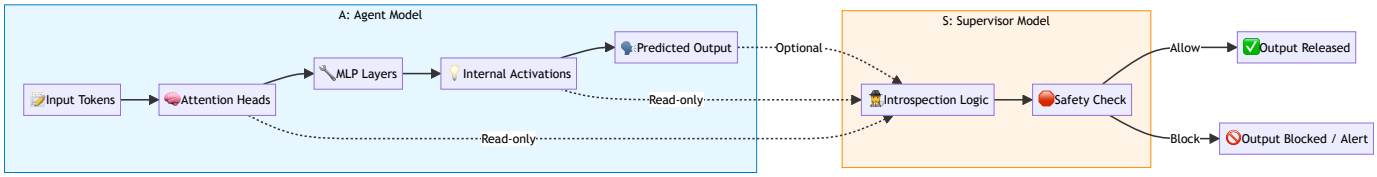
4. Implementation Considerations

Latency: Real-time auditing introduces non-trivial overhead, dependent on model size and introspection complexity. Techniques like activation checkpointing can mitigate this by reducing memory usage and allowing selective inspection [5].

Interpretability tooling: Circuits and behavioral probes can reveal token-level decisions associated with deception or manipulation [4]. These tools are essential for training S to recognize patterns of concern in A's thoughts.

5. Training Supervisors

To train S effectively, we propose the use of:



Recent work on superposition in neural networks has shown that abstract features are often stored in entangled representations, making interpretability more challenging. These insights can guide the design of probes and labeling strategies for training S to detect cognitive patterns associated with misalignment [1].

- Contrastive datasets: aligned vs misaligned planning behaviors [2].
- Layer-wise circuit tracing: labeling attention and MLP activation pathways in A [3].
- Replay of historical deceptive behaviors to provide ground truth supervision signals [2].

6. Limitations

This architecture does not resolve the full alignment problem. Supervisors must themselves be aligned, and capable of catching subtle forms of misalignment. The effectiveness of introspection tools varies with model architecture and training scale. Moreover, the supervisor's ability to intervene is limited by its integration into the system.

7. Conclusion

We advocate for a shift in focus from post-hoc interpretability to real-time internal oversight. A dedicated supervisory model with read access to an agent model's cognition offers a pathway to scalable alignment. As capabilities grow, such architectures may prove critical for containing systems with emergent or deceptive behavior.

References

1. Elhage, N. et al. (2022). Toy Models of Superposition. https://arxiv.org/abs/2209.10652
2. Hendrycks, D. et al. (2024). AI deception: A survey of examples, risks, and potential solutions. https://www.cell.com/patterns/fulltext/S2666-3899(24)00103-X
3. Burns, R. et al. (2022). Discovering Latent Knowledge in Language Models Without Supervision. https://arxiv.org/abs/2212.03827
4. HuggingFace / DeepSpeed profiling documentation (2024). Activation checkpointing and memory usage. https://deepspeed.readthedocs.io/en/latest/activation-checkpointing.html
5. Hsu, K. et al. (2024). Efficient Automated Circuit Discovery in Transformers using Contextual Decomposition. https://arxiv.org/abs/2407.00886