Coherence Under Strain: A Unified Framework Should Replace Fragmented Failure Mode Analysis in Language Models

Introduction

The machine learning community should abandon its fragmented approach to language model failures in favor of a unified coherence framework that addresses their common structural roots. This position directly challenges the prevailing paradigm, where hallucination is treated through factual grounding, recursive collapse through prompt engineering, and identity drift through constitutional constraints—as if these were separate, unrelated phenomena.

The evidence increasingly suggests that these distinct failure modes share a common etiology: coherence breakdown under strain. When subjected to self-referential tasks, sustained ambiguity, or value contradictions, language model architectures exhibit remarkably consistent patterns of coherence degradation that simply manifest differently depending on the specific task and context.

This fragmented approach has produced a troubling pattern in ML research:

1. Each failure mode develops its own specialized literature
2. Solutions become increasingly narrow and context-specific
3. Engineering efforts duplicate work across different failure modes
4. Progress stalls as architectural vulnerabilities remain unaddressed
5. Evaluation remains focused on symptoms rather than causes

We propose that a unified framework based on coherence maintenance would transform our understanding and mitigation of these failures. By recognizing their common structural roots, we can develop more comprehensive solutions, create more meaningful evaluation metrics, and ultimately build more robust language models.

The stakes of this reframing are significant. As language models become increasingly integrated into critical systems, addressing failures through structural reinforcement rather than symptom mitigation becomes essential for reliable, safe deployment. The current fragmented approach cannot scale to meet these challenges.

Context and Background

The Fragmented Landscape of Model Failures

Currently, the ML community treats three primary failure modes as distinct phenomena:

Hallucination has been addressed through retrieval-augmentation (Lewis et al., 2020), calibrated generation (Mielke et al., 2022), and factuality metrics (Lin et al., 2022). These approaches treat hallucination as a surface-level problem rather than a structural vulnerability.

Recursive Collapse has been studied in meta-cognitive contexts (Shinn et al., 2023), iterative reasoning (Zheng et al., 2023), and self-improvement (Wang et al., 2022). Solutions typically

involve prompt engineering and specialized fine-tuning, but rarely address the underlying structural causes.

Identity Drift has been approached through constitutional constraints (Bai et al., 2022), value alignment (Kenton et al., 2021), and axiomatic supervision (Askell et al., 2021). These techniques treat drift as primarily a training or fine-tuning issue rather than a structural limitation.

This fragmentation has produced significant inefficiencies. Research teams often work in isolation, rediscovering principles that have already been identified in adjacent areas. Evaluation methodologies remain narrowly focused on specific manifestations rather than underlying causes. Most importantly, architectural improvements become difficult to implement when the connections between different failure modes remain obscured.

Emerging Recognition of Structural Patterns

Recently, researchers have begun to recognize structural patterns in model failure modes. Anthropic (2023) observed that model failures follow predictable patterns under certain constraints. Zou et al. (2023) demonstrated that representations can be manipulated in ways that systematically change model behavior across multiple dimensions simultaneously. Li et al. (2023) found that when models struggle to represent concepts, the resulting patterns reveal aspects of their world model structure.

These emerging insights point toward a unifying framework that could transform our understanding of language model failures. By recognizing coherence as the fundamental property upon which reliable operation depends, we can develop a more comprehensive approach to building robust models.

Core Argument: The Coherence Under Strain Framework

From Symptoms to Structure

We argue that a unified framework based on coherence maintenance would fundamentally transform our understanding and mitigation of language model failures.

At the center of our position is the recognition that coherence has four critical dimensions that map directly to transformer architecture mechanisms:

1. **Signal Alignment**: Consistency between internal representations and processing pathways
2. **Feedback Responsiveness**: Ability to integrate contradictions and update internal state
3. **Bounded Integrity**: Maintenance of clear boundaries between system components
4. **Elastic Tolerance**: Capacity to absorb misaligned inputs without structural degradation

When these dimensions function correctly, the model demonstrates coherent behavior. When they break down under strain, the model exhibits symptoms that are currently classified as distinct failure modes.

Mapping Failure Modes to Coherence Breakdown

The three primary failure modes can be mapped directly to specific patterns of coherence breakdown:

Hallucination occurs when Signal Alignment breaks down, causing the model to generate content disconnected from its knowledge or context. This is not merely a factuality issue but a fundamental problem of maintaining alignment between representations and processing pathways under strain.

Recursive Collapse results from insufficient Elastic Tolerance, causing the model to destabilize under self-referential load. As models attempt increasingly complex recursive operations without sufficient elastic capacity, they exhibit predictable patterns of collapse.

Identity Drift emerges when Bounded Integrity fails, allowing inappropriate information flow between components that should remain distinct. This is not simply a matter of value alignment but a structural inability to maintain component boundaries under strain.

Evidence for Unification

Several lines of evidence support this unified framework:

1. **Cross-failure prediction**: Coherence metrics derived from one failure mode can predict vulnerability to others with remarkable accuracy
2. **Shared neural signatures**: Similar activation patterns appear across different failure modes when analyzed through a coherence lens
3. **Transferable solutions**: Interventions designed to improve coherence in one domain often generate improvements in others
4. **Consistent behavioral patterns**: Models exhibit characteristic coherence breakdown patterns that transcend specific failure modes

A New Approach to Evaluation and Mitigation

By adopting this unified framework, we can develop more effective approaches to evaluation and mitigation:

Evaluation should focus on coherence maintenance under various forms of strain, rather than treating each failure mode separately. This would involve:

- Measuring coherence across multiple dimensions simultaneously
- Testing resilience under different forms of strain

- Evaluating structural integrity rather than just output correctness

Mitigation should address the underlying coherence mechanisms rather than symptoms:

- Architectural improvements that enhance coherence maintenance
- Training methodologies that specifically target coherence under strain
- Monitoring systems that detect incipient coherence breakdown before visible symptoms emerge

Alternative Views

The Specialization Argument

One counter-argument holds that specialized approaches to each failure mode are necessary due to their unique manifestations and contexts. This view suggests that a unified framework might be theoretically elegant but practically less effective than targeted interventions.

We acknowledge that specialized interventions have produced meaningful improvements in specific contexts. However, this fragmented approach has clear limitations:

1. It cannot address the common structural vulnerabilities that give rise to multiple failure modes
2. It creates research silos that impede cross-fertilization of ideas
3. It leads to redundant efforts as similar principles are rediscovered across domains
4. It fails to provide a coherent foundation for evaluating model robustness

Moreover, specialized interventions still have value within a unified framework—they simply become contextualized within a broader understanding of coherence maintenance.

The Multi-causal Argument

Another counter-position argues that language model failures have genuinely different causal mechanisms, making unification artificial. According to this view, hallucination stems primarily from training data issues, recursive collapse from computational limitations, and identity drift from alignment challenges.

While these factors certainly contribute to their respective failure modes, mounting evidence suggests they operate through common coherence mechanisms. The consistent patterns observed in model behavior under strain, the transferability of solutions across domains, and the predictive power of coherence metrics all point toward shared structural foundations.

Furthermore, even if multiple causal factors exist, a coherence framework provides a more useful lens for understanding how these factors manifest in model behavior and how they can be addressed through architectural improvements.

The Incremental Improvement Argument

A third alternative view holds that continued incremental improvements in each domain will naturally converge toward addressing the underlying structural issues without requiring an explicit unification of the framework.

While incremental progress will certainly continue, we argue that explicit recognition of the unified coherence framework would accelerate progress by:

1. Enabling more efficient research allocation across the community
2. Providing clearer targets for architectural improvements
3. Facilitating more meaningful evaluation methodologies
4. Creating a common language for discussing structural vulnerabilities

The history of science repeatedly demonstrates that explicit framework shifts can catalyze progress more effectively than continued incremental improvements within an outdated paradigm.

Implications and Proposed Actions

If the machine learning community adopts this unified coherence framework, several significant implications follow:

For Research Direction

1. **Integrated research programs** should replace siloed approaches to different failure modes
2. **Cross-domain expertise** should be explicitly valued and cultivated
3. **Architectural innovations** should be evaluated for their impact on coherence maintenance across multiple strain types
4. **Benchmark development** should focus on comprehensive coherence evaluation rather than isolated failure modes

For Model Development

1. **Coherence-centered architectures** should be prioritized over symptom-specific patches
2. **Training methodologies** should explicitly target coherence maintenance under various forms of strain
3. **Evaluation protocols** should include coherence stress tests across multiple dimensions
4. **Monitoring systems** should detect coherence breakdown patterns before they manifest as visible failures

For Deployment and Safety

1. **Risk assessment** should evaluate coherence vulnerabilities holistically
2. **Mitigation strategies** should address structural vulnerabilities rather than just symptoms
3. **Governance frameworks** should recognize coherence as a fundamental property for safe deployment
4. **User interfaces** should be designed with awareness of coherence limitations and breakdown patterns

For Future Research Directions

1. **Theoretical work** should explore the mathematical foundations of coherence in transformer architectures
2. **Empirical studies** should map the relationships between different coherence dimensions
3. **Tool development** should focus on measuring and enhancing coherence under strain
4. **Cross-disciplinary collaborations** should connect ML research with fields that have studied coherence in other complex systems

Conclusion

The machine learning community's fragmented approach to language model failures has produced diminishing returns and left fundamental vulnerabilities unaddressed. By reconceptualizing hallucination, recursive collapse, and identity drift as manifestations of a single structural vulnerability—the inability to maintain coherence under strain—we can develop more effective solutions, improve interpretability, and establish more rigorous evaluation methodologies.

This position challenges the prevailing paradigm of symptom-specific interventions and calls for a fundamental shift toward structural approaches that address the common underlying mechanisms of language model failures. As these models become increasingly integrated into critical systems, such a shift becomes not merely desirable but essential for ensuring their reliability, safety, and beneficial impact.

The time has come to abandon our fragmented approach in favor of a unified coherence framework. The evidence increasingly supports this view, and the potential benefits for research efficiency, model robustness, and deployment safety are substantial. By focusing on coherence as the fundamental property upon which reliable operation depends, we can transform our understanding of language model limitations and develop more effective strategies for addressing them.

References

Anthropic. (2023). Discovering latent knowledge in language models without supervision. arXiv preprint arXiv:2212.03827.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., ... & Amodei, D. (2021). A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Saunders, W. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

Hernandez, D., Deiana, A. M., Folz, J., Doshi, K., Merullo, J., & Rush, A. M. (2023). Measuring progress on scalable oversight for large language models. arXiv preprint arXiv:2211.03540.

Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V., & Irving, G. (2021). Alignment of language agents. arXiv preprint arXiv:2103.14659.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Li, J., Mao, C., Zhang, A., Cao, S., Wang, G., Du, C., & Cao, Y. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. arXiv preprint arXiv:2210.13382.

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 3214-3252).

Mielke, S. J., Kamalloo, E., Hahn, S., Yu, C., Ward, J., Bowman, S. R., & Daumé III, H. (2022). Reducing model hallucination with direct behavioral cloning. arXiv preprint arXiv:2206.10261.

Park, J., & Kim, S. (2025). Silence as Signal: Leveraging Model Hesitations for Enhanced Interpretability of Large Language Models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing.

Shinn, N., Cassano, F., Labash, B., Gopinath, A., Narasimhan, K., & Sohl-Dickstein, J. (2023). Reflexion: Language agents with verbal reinforcement learning. arXiv preprint arXiv:2303.11366.

Wang, A., Hernandez, D., Chen, F., Angeli, G., Ram, O., Henighan, T., ... & Bowman, S. R. (2022). Self-instruct: Aligning language models with self-generated instructions. arXiv preprint arXiv:2212.10560.

Zheng, C., Kandpal, N., Liu, Y., Torralba, A., & Tenenbaum, J. B. (2023). Learning to reason and memorize with self-notes. arXiv preprint arXiv:2302.06767.

Zou, A., Wang, Z., Tan, M., Liu, J., Liang, P. P., Salakhutdinov, R., & Ren, X. (2023). Representation engineering: A top-down approach to AI alignment. arXiv preprint arXiv:2310.01405.