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

Anonymous Author(s)
Affiliation
Address
email

ABSTRACT

This position paper argues that the machine learning community's fragmented approach to understanding and addressing language model failures is fundamentally misguided. Current research treats hallucination, recursive collapse, and identity drift as separate engineering challenges requiring distinct solutions, resulting in siloed research efforts and diminishing returns. We argue that these phenomena are actually manifestations of a single structural vulnerability: the inability to maintain coherence under strain. By reconceptualizing these failures through a unified framework of coherence maintenance, 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.

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 , calibrated generation , and factuality metrics . These approaches treat hallucination as a surface-level problem rather than a structural vulnerability.

Recursive Collapse has been studied in meta-cognitive contexts , iterative reasoning , and self-improvement . Solutions typically involve prompt engineering and specialized fine-tuning, but rarely address the underlying structural causes.

Identity Drift has been approached through constitutional constraints , value alignment , and axiomatic supervision . 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. observed that model failures follow predictable patterns under certain constraints. demonstrated that representations can be manipulated in ways that systematically change model behavior across multiple dimensions simultaneously. 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

39

Anthropic. Discovering latent knowledge in language models without supervision. , 2023.

A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, et al. A general language assistant as a laboratory for alignment. , 2021.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, et al. Constitutional ai: Harmlessness from ai feedback. , 2022.

D. Hernandez, A. M. Deiana, J. Folz, K. Doshi, J. Merullo, and A. M. Rush. Measuring progress on scalable oversight for large language models. , 2023.

Z. Kenton, T. Everitt, L. Weidinger, I. Gabriel, V. Mikulik, and G. Irving. Alignment of language agents. , 2021.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. , 33:9459–9474, 2020.

J. Li, C. Mao, A. Zhang, S. Cao, G. Wang, C. Du, and Y. Cao. Emergent world representations: Exploring a sequence model trained on a synthetic task. , 2023.

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

S. J. Mielke, E. Kamalloo, S. Hahn, C. Yu, J. Ward, S. R. Bowman, and H. Daumé III. Reducing model hallucination with direct behavioral cloning. , 2022.

J. Park and S. Kim. 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*, 2025.

N. Shinn, F. Cassano, B. Labash, A. Gopinath, K. Narasimhan, and J. Sohl-Dickstein. Reflexion: Language agents with verbal reinforcement learning. , 2023.

A. Wang, D. Hernandez, F. Chen, G. Angeli, O. Ram, T. Henighan, et al. Self-instruct: Aligning language models with self-generated instructions. , 2022.

C. Zheng, N. Kandpal, Y. Liu, A. Torralba, and J. B. Tenenbaum. Learning to reason and memorize with self-notes. , 2023.

A. Zou, Z. Wang, M. Tan, J. Liu, P. P. Liang, R. Salakhutdinov, and X. Ren. Representation engineering: A top-down approach to ai alignment. , 2023.

NeurIPS Paper Checklist

1. Claims

- **Question:** Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
- **Answer:** [Yes]
- **Justification:** The abstract and introduction clearly state our position that the ML community should abandon its fragmented approach to language model failures in favor of a unified coherence framework. This position is consistently developed throughout the paper with supporting evidence and reasoned arguments.

2. Limitations

- **Question:** Does the paper discuss the limitations of the work performed by the authors?
- **Answer:** [Yes]
- **Justification:** The "Alternative Views" section thoroughly addresses potential limitations and counter-arguments to our position. We present these alternative views fairly and respond to them with reasoned arguments rather than dismissing them.

3. Theory assumptions and proofs

- **Question:** For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
- **Answer:** [NA]
- **Justification:** This is a position paper that does not present formal theoretical results requiring mathematical proofs. The paper presents a conceptual framework and supporting evidence rather than formal theorems.

4. **Experimental result reproducibility**

- **Question:** Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
- **Answer:** [NA]
- **Justification:** This position paper does not present new experimental results that would require reproduction. The paper synthesizes existing research and proposes a conceptual framework rather than reporting on novel experiments.

5. **Open access to data and code**

- **Question:** Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
- **Answer:** [NA]
- **Justification:** This position paper does not introduce new code or datasets. The paper proposes a conceptual framework rather than presenting experimental results requiring code or data.

6. **Experimental setting/details**

- **Question:** Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
- **Answer:** [NA]
- **Justification:** This position paper does not present experimental results or training procedures. The paper focuses on conceptual arguments and proposed research directions rather than empirical findings.

7. **Experiment statistical significance**

- **Question:** Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
- **Answer:** [NA]
- **Justification:** The paper does not present experimental results requiring statistical significance testing or error bars. It discusses conceptual frameworks rather than reporting empirical findings.

8. **Experiments compute resources**

- **Question:** For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
- **Answer:** [NA]
- **Justification:** This position paper does not present experimental results that would require computational resources for reproduction. The paper focuses on conceptual arguments rather than computational experiments.

9. **Code of ethics**

- **Question:** Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

- **Answer:** [Yes]
- **Justification:** This position paper adheres to the NeurIPS Code of Ethics. It presents arguments transparently, acknowledges alternative viewpoints fairly, and promotes research directions that would enhance the safety, reliability, and beneficial impact of AI systems.

10. **Broader impacts**

- **Question:** Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
- **Answer:** [Yes]
- **Justification:** The paper discusses potential positive impacts of the proposed framework throughout the "Implications and Proposed Actions" section, where we outline how a unified coherence framework could improve research efficiency, model robustness, and deployment safety. While we don't anticipate negative societal impacts from adopting this framework, we acknowledge in the "Alternative Views" section that there may be limitations or challenges to implementing it.

11. **Safeguards**

- **Question:** Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
- **Answer:** [NA]
- **Justification:** This position paper does not release data or models that would require safeguards against misuse. The paper proposes a conceptual framework rather than releasing artifacts that could be misused.

12. **Licenses for existing assets**

- **Question:** Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
- **Answer:** [NA]
- **Justification:** This position paper does not use existing assets such as code, data, or models that would require licensing information. The paper properly cites prior research but does not utilize assets requiring specific licenses or terms of use.

13. **New assets**

- **Question:** Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
- **Answer:** [NA]
- **Justification:** This position paper does not introduce new assets such as datasets, code, or models that would require documentation. The paper presents conceptual frameworks and arguments rather than creating new technical assets.

14. **Crowdsourcing and research with human subjects**

- **Question:** For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

- **Answer:** [NA]
- **Justification:** This position paper does not involve crowdsourcing or research with human subjects. The paper discusses conceptual frameworks and research directions without involving human participants in experiments.

15. **Institutional review board (IRB) approvals**

- **Question:** Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
- **Answer:** [NA]
- **Justification:** This position paper does not involve research with human subjects that would require IRB approval or risk assessment. The paper focuses on conceptual frameworks rather than studies involving human participants.

16. **Declaration of LLM usage**

- **Question:** Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?
- **Answer:** [NA]
- **Justification:** LLMs were not used as a core component of the methods or arguments presented in this position paper. The conceptual contributions and position statements are based on analysis of existing research and trends in the field rather than LLM usage.

Lay Summary

Current approaches to fixing language model problems are like treating different symptoms of the same disease with separate medicines. Researchers address hallucinations, reasoning breakdowns, and value inconsistencies as if they were completely different issues. This paper argues that these problems are actually different manifestations of the same underlying vulnerability: the inability to maintain coherence under pressure. Just as a bridge might fail in different ways (buckling, swaying, or cracking) depending on where stress is applied, language models show different failure patterns depending on the type of strain they experience. By focusing on enhancing the structural coherence of these models rather than patching individual symptoms, we could develop more effective solutions, better evaluation methods, and more reliable AI systems. This shift from symptom-specific fixes to structural reinforcement becomes increasingly important as language models are integrated into critical systems where reliability and safety are paramount.