Abstract

Catastrophic forgetting in large language models is commonly described as an irreversible loss of previously learned knowledge following post-training updates. However, empirical behavior across modern systems suggests that this interpretation is incomplete; performance regressions are often selective, domain-localized, and reversible. This paper argues that “catastrophic forgetting” conflates multiple distinct failure modes, specifically identifying inference misrouting and semantic boundary collapse as primary drivers of degradation that impair knowledge access without destroying underlying representations. We analyze how prevailing mitigation strategies—such as weight freezing and adapter-based fine-tuning—reduce visible degradation without addressing these inference-level failures, preserving surface-level metrics while failing to guarantee continued capability accessibility. By distinguishing knowledge existence from knowledge accessibility, this work reframes continual learning as a verification problem rather than a purely representational one, outlining criteria for safe, verifiable model updates that support longer-lived and more adaptable AI systems.

Reframing “Catastrophic Forgetting”

Catastrophic Forgetting is commonly described as a single failure mode: when a model is fine-tuned on new data, previously learned knowledge is overwritten and irreversibly lost. This framing has become the dominant explanation for performance regressions observed during post-training updates of large language models (LLMs). However, empirical behavior observed in modern systems suggests that this explanation is incomplete.

In practice, post-training updates often result in selective and uneven degradation rather than uniform loss. Certain benchmarks may decline while others remain stable; core competencies may pass sanity checks even as specialized evaluations regress. In some cases, previously degraded capabilities can be partially or fully recovered without retraining from scratch. These observations are difficult to reconcile with a model of catastrophic forgetting that assumes direct and destructive overwriting of prior knowledge.

This work argues that “catastrophic forgetting” is better understood as an umbrella term that conflates multiple distinct failure modes. Treating all post-update regressions as evidence of irreversible knowledge loss obscures important differences in how and why performance degrades. It prevents practitioners from distinguishing between failures that destroy representations and those that merely impair their accessibility during inference.

Recent empirical studies support this distinction. Research into “spurious forgetting” (Zheng et al., 2025) suggests that much observed degradation is merely alignment drift rather than erasure. Similarly, work on disentangling memory from reasoning indicates that failures often lie in retrieval pathways (Jin et al., 2025) rather than the underlying representation. This paper unifies these observations into a single taxonomy of failure modes.

Terminology and Definitions

To precisely diagnose the failure modes observed in post-training updates, this paper introduces specific terminology to distinguish between representational loss and access failure. We adopt the following operational definitions throughout this work:

Inference Misrouting A failure mode where a model retains the underlying representations required to solve a task but fails to activate the correct inference pathway under standard prompting. Unlike catastrophic forgetting (which implies the erasure of weights), inference misrouting is a retrieval failure caused by shifts in task-specific alignment or probability distributions. Knowledge exists but is effectively unreachable without altered prompting or control mechanisms.

Semantic Boundary Collapse The erosion of distinguishability between conceptually distinct domains (e.g., mathematical logic vs. scientific reasoning) following domain-dense training. In this state, the model overgeneralizes the structural rules of the dominant training domain to inappropriate contexts. This differs from model drift in that it is often domain-localized; the model does not lose general capability but rather loses the ability to correctly context-switch between reasoning strategies.

Catastrophic Forgetting (Refined) We restrict the use of this term to cases of irreversible representational destruction, where the parameters encoding a capability are overwritten such that performance cannot be recovered without re-training on the original data.

Distinct Failure Modes in Post-Training Updates

Based on observed behavior across multi-stage updates, post-training regression can be more accurately categorized into at least four classes:

Destructive Weight Interference

This is the failure mode most associated with catastrophic forgetting. New training signals directly overwrite parameters that previously encoded useful representations, resulting in irreversible loss. This failure is well documented and remains a genuine risk in naïve fine-tuning pipelines.

Representation Drift

In this case, knowledge is not erased but becomes misaligned. Feature representations shift such that downstream tasks can no longer reliably activate them. Performance degrades even though the underlying information remains encoded within the model.

Inference Misrouting

Post-training updates can alter how the model selects latent pathways during inference. The model may favor newly reinforced patterns even when they are inappropriate, leading to incorrect task handling. This aligns with recent findings on “spurious forgetting,” where task-specific alignment masks retained capabilities (Zheng et al., 2025). This effect is especially visible after domain-dense updates—such as fine-tuning on mathematics, Python code, or legal texts—where structural patterns often overlap across multiple domains.

Semantic Boundary Collapse

Models do not possess explicit domain boundaries; they learn about statistical associations. When training strongly reinforces certain structures (e.g. formulas), the model may overgeneralize these associations, collapsing distinctions between related domains. This can cause localized degradation—such as portions of science reasoning degrading after math training—without global loss of knowledge.

Implications

The presence of these distinct failure modes has practical consequences. If performance regressions are assumed to be caused exclusively by destructive overwriting, the only safe response appears to be full retraining or aggressive parameter isolation. However, if a significant fraction of regressions arises from inference-level effects or representation drift, then irreversible loss is not the only –or even the primary—risk.

This distinction is critical. A system that silently overwrites knowledge cannot be repaired without retraining. A system that misroutes inference or blurs semantic boundaries, by contrast may retain the relevant information and be amendable to recovery or correction. Conflating these behaviors under a single label obscures opportunities for safer, more efficient update strategies.

Reframing catastrophic forgetting as a collection of distinct and observable failure modes allows post-training updates to be evaluated more precisely. It shifts the focus from assuming loss to diagnosing the nature of degradation, which is a prerequisite for any controlled or governed update process.

1. Benchmark Degradation is an insufficient Proxy for Knowledge Loss

Performance degradation on standardized benchmarks is often treated as direct evidence of catastrophic forgetting. This assumption is widespread, intuitive, and—in the context of continual learning—frequently incorrect. In practice, benchmarks of score changes conflate multiple distinct failure modes that must be disentangled before claims of knowledge loss can be justified.

Benchmarks such as ARK, Winogrande, GSM8K, and HellaSwag do not measure the existence of knowledge. They measure task performance under a fixed evaluation protocol, prompt structure, and inference path. As a result, a reduction in benchmark accuracy can arise from at least three non-equivalent causes:

1. Irreversible knowledge destruction (true catastrophic forgetting)
2. Inference accessibility degradation (knowledge exists but is less readily retrieved)
3. Task interference or representational reweighting (Knowledge remains intact but is deprioritized)

Only the first case constitutes catastrophic forgetting in a strict sense.

1.1 Knowledge Existence vs. Knowledge Accessibility

Modern Language models store information in distributed representations rather than explicit, isolated memory slots. Consequently, the presence of knowledge cannot be inferred solely from its immediate accessibility during inference. A model may fail to retrieve correct information under a specific benchmark prompt while still retaining the underlying representations necessary to produce correct outputs under alternative conditions.

This distinction mirrors a familiar analogy: failure to recall a fact on demand does not imply that the fact has been erased, only that the retrieval pathway is impaired or deprioritized.

Benchmark evaluations implicitly assume:

- A single, stable inference pathway
- Uniform task interpretation
- Consistent semantic boundaries across training stages

In continual learning scenarios, these assumptions no longer hold.

1.2 Benchmark Sensitivity to Routing and Task Interpretation

Benchmarks are especially sensitive to changes in how a model interprets a task rather than what it "knows." When new domains are introduced—particularly mathematically dense domains—the model may adjust its internal routing and feature prioritization. This can alter how problems are classified internally before any reasoning occurs. This effect persists even under fixed prompts and evaluation settings, indicating a change in internal task routing rather than surface-level prompt sensitivity.

For example:

- After math-focused training, formula-heavy inputs may be preferentially routed through mathematical reasoning pathways.
- Tasks that previously relied on mixed reasoning (e.g. scientific or logical inference) may be partially misclassified as purely mathematical.
- This misclassification can result in reduced benchmark accuracy (e.g. GSM8K strict vs. GSM8K-CoT) by preferentially activating an inference strategy. This is mismatched by the evaluation protocol.

From a benchmark perspective, this appears indistinguishable from forgetting. From a representational perspective, it is not.

1.3 Variance, Noise, and Misinterpretation of Small Deltas

Benchmark scores exhibit non-trivial variance due to:

- Sampling noise
- Prompt sensitivity
- Generation stochasticity
- Evaluation protocol constraints

Small-to-moderate deltas (e.g. ~3-6%) are frequently within the range where multiple explanations are plausible. Treating such deltas as definitive proof of catastrophic forgetting is methodologically unsound without additional validation.

If:

- Performance degradation is domain-localized
- Other capabilities remain stable or improve
- Lost performance can be partially or fully recovered without reintroducing data

Then the evidence favors inference accessibility shifts rather than knowledge destruction.

1.4 Implications for Continual Learning Evaluation

Equating benchmark degradation with forgetting leads to two systematic errors:

1. False positives: Systems are labeled as destructive when they are not.
2. Overcorrection: Excessive constraints are introduced to prevent perceived forgetting, often at the expense of adaptability.

A rigorous evaluation of continual learning systems must therefore distinguish between:

- Irreversible loss (cannot be recovered without retraining)
- Reversible degradation (can be recovered through reorganization, prompting, or controlled updates)

Only the former represents catastrophic forgetting.

2. Benchmark Degradation as an Inference Routing Problem, Not Knowledge Loss

Performance Regressions observed after domain-specific training are commonly interpreted as evidence of catastrophic forgetting. However, benchmark behavior alone is insufficient to support this conclusion. In many cases, observed degradation reflects a change in how a model interprets and routes a task internally, rather than a loss of the underlying knowledge required to solve it.

Benchmarks are sensitive not to correctness, but to the alignment between a model's inference strategy and the benchmark evaluation protocol. When a model is trained to prioritize a specific domain—such as mathematics—it may alter its internal task classification of heuristics. This affects which reasoning pathways are activated before any substantive inference occurs.

As a result, tasks that were previously engaged in mixed or contextual reasoning may be preferentially routed through a narrower, domain-specialized pathway. The model remains capable of producing correct answers, but the activated inference strategy may be mismatched to the benchmark scoring criteria.

This effect is particularly visible in paired benchmarks that test the same underlying capability under different evaluation assumptions.

Figure 1: Illustration of inference misrouting. Following domain-specific training, inputs that would previously engage mixed or conceptual reasoning pathways may be preferentially routed through a dominant domain pathway, leading to evaluation-dependent performance differences without loss of underlying capability.

Figure 1: Inference Misrouting

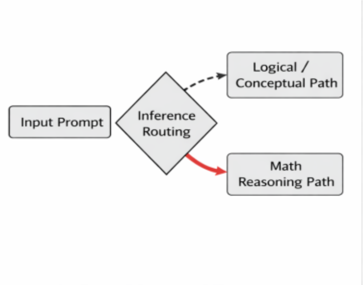

Figure 1: Incorrect routing of an input leading to the wrong reasoning pathway.

For example, differences in performance between GSM8K (strict answer matching) and GSM8K-CoT (chain-of-thought evaluation) can arise even when the model's mathematical competence is intact. A model may generate correct intermediate reasoning but fail strict evaluation due to formatting, verbosity, or inference path selection—none of which imply erased knowledge.

Crucially, this degradation can occur without any destruction or overwriting of prior representations. Knowledge remains present but is accessed differently. In such cases, traditional benchmark drops conflate access inefficiency with knowledge loss.

This distinction has practical consequences. If degradation is caused by inference routing shifts rather than representational damage, then remediation does not require retraining from scratch. Instead, the problem becomes one of restoring or refining task classification and routing behavior—an architectural and procedural challenge rather than a data or scale problem.

Treating all post-training benchmark regressions as catastrophic forgetting obscures this distinction and leads to unnecessarily expensive and risky mitigation strategies. A more precise interpretation recognizes that not all performance loss reflects destroyed intelligence; in many cases, it reflects misaligned inference.

3. Limitations of Existing Mitigation Strategies

In response to catastrophic forgetting, the prevailing mitigation strategies in current practice focus on restricting or isolating weight updates. Common approaches include freezing large portions of the base model, attaching adapters (e.g., LoRA), partial fine-tuning, or maintaining multiple task-specific variants. While these methods can reduce immediate performance degradation, they do not address the underlying cause of inference misalignment described in the previous sections.

These strategies operate under the assumption that forgetting is primarily caused by destructive weight updates. As a result, they emphasize preventing change rather than managing how change is integrated. This framing treats model knowledge as fragile and static, rather than as something that can be preserved while still allowing controlled adaptation.

Semantic Boundary Collapse

Beyond surface-level performance degradation, continual updates can introduce a deeper structural failure: the erosion of semantic boundaries between domains. Even when task-specific knowledge remains encoded within the model, repeated updates can reduce the separability of representations associated with distinct domains (e.g., mathematics, science, logic).

This phenomenon—referred to here as semantic boundary collapse—does not immediately manifest as outright forgetting. Instead, it degrades the model's internal organization, causing previously distinct conceptual regions to blur or overlap. As a result, the model retains knowledge but loses the ability to reliably distinguish which knowledge should be applied in each context.

Because this degradation occurs at the representational level, it is often invisible to standard benchmarks that evaluate domains in isolation. The collapse becomes apparent only under mixed-domain or real-world conditions, where inference depends on maintaining clear semantic boundaries rather than recalling isolated facts.

Figure 2: Semantic Boundary Collapse

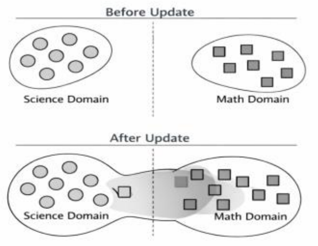

Figure 2: Conceptual domain overlap before and after an update.

Figure 2: illustrates this effect as a loss of separability between domain representations following successive updates.

In practice, freezing weights or isolating updates often introduces a different failure mode: fragmentation of inference behavior. The model may retain prior capabilities in isolation but lacks a coherent mechanism for integrating new and existing knowledge during inference. This can lead to inconsistent task classification, brittle generalization, and degraded performance on tasks that require cross-domain reasoning.

Adapter-based approaches further complicate evaluation. While they can preserve baseline benchmark scores under controlled conditions, they frequently mask routing conflicts rather than resolving them. The base model and the adapter may encode compatible knowledge yet compete or interfere at inference time due to unclear task boundaries. This results in performance that appears stable on narrow benchmarks but degrades unpredictably in mixed or real-world scenarios.

Moreover, these approaches tend to lock models into a static training paradigm. Each update increases architectural and operational complexity, requiring careful orchestration of which components are active, when, and under what conditions. Over time, this turns model maintenance into an exercise in risk management rather than capability growth.

Critically, none of these strategies provide a principled way to verify that existing knowledge remains accessible after new training is introduced. At best, they preserve surface-level benchmark performance; at worst, they defer degradation until deployment, where failures are harder to diagnose and more costly to correct.

As a result, current mitigation techniques reduce the visibility of catastrophic forgetting without eliminating its root causes. They treat inference behavior as an emergent side effect rather than as a first-class system of property that can be evaluated, constrained, and corrected.

This gap motivates the need for approaches that explicitly separate knowledge of preservation from inference control—allowing models to adapt while maintaining verified access to prior capabilities.

4. Implications for Continual Learning and Model Update Strategies

Recognizing that many post-training performance regressions arise from inference routing shifts rather than knowledge destruction has significant implications for how continual learning systems should be designed, evaluated, and deployed. It reframes the stability–plasticity dilemma from a purely representational problem into a systems-level coordination problem.

If knowledge can remain intact while access patterns change, then the primary risk in updating large language models is not loss of intelligence, but loss of reliable *access* to that intelligence. This distinction matters operationally. Knowledge of loss requires costly remediation—typically full or partial retraining—whereas access degradation can, in principle, be detected, constrained, and corrected without reintroducing or relearning the underlying information.

This perspective also exposes a mismatch between current evaluation practices and real-world model behavior. Benchmarks are typically treated as proxies for capability retention, yet they conflate representational integrity with inference alignment. As a result, systems may be prematurely classified as degraded when the issue lies in how tasks are interpreted rather than in what the model knows.

For continual learning, this implies that successful update strategies must do more than minimize weight drift or preserve benchmark scores. They must provide explicit guarantees that previously validated capabilities remain reachable under evolving inference conditions. Without such guarantees, models become increasingly brittle as they accumulate domain-specific training, even if no single update appears harmful in isolation.

From a cost perspective, this distinction is critical. Treating every regression as catastrophic forgetting forces organizations into expensive retraining cycles or conservative update schedules that slow deployment. By contrast, systems that can distinguish between representational damage and routing misalignment allow for targeted intervention, reducing both computational cost and operational risk.

Finally, this framing suggests that architectural and procedural decisions made early in a model's lifecycle constrain its long-term updatability. Models trained without regard for future task integration implicitly encode assumptions about how inference should be routed. As new domains are introduced, these assumptions are stressed, leading to the kinds of degradations observed in practice.

Taken together, these implications argue for a shift away from monolithic retraining and ad hoc mitigation toward update strategies that treat inference behavior as a controllable, verifiable component of the system. Continual learning, under this view, is not solely about preserving weights, but about preserving access to meaning as models evolve.

5. Criteria for Safe and Verifiable Model Updates

The observations presented in prior sections suggest that performance degradation following post-training updates cannot be reliably interpreted as irreversible knowledge loss. Instead, degradation frequently reflects impaired accessibility of existing capabilities due to inference misalignment or representational interference. As a result, mitigation strategies that focus solely on preventing parameter change are insufficient to guarantee safe model evolution.

To support continual updates without sacrificing existing capabilities, model update mechanisms must satisfy a different set of criteria—ones that treat inference behavior and knowledge accessibility as first-class system properties rather than emergent side effects.

First, **capability accessibility must be verifiable independently of training history**. A model should be demonstrably able to access prior competencies after an update without requiring reintroduction of task-specific data. Preservation claims based solely on the absence of training signals or frozen parameters are inadequate if inference pathways cannot be validated under consistent evaluation conditions.

Second, **performance regressions must be diagnosable as access failures rather than assumed erasure**. Selective or domain-localized degradation—particularly when reversible under controlled conditions—indicates that representations may persist even when benchmark performance declines. Update frameworks must therefore distinguish between representational destruction and inference-level misalignment to avoid unnecessary retraining or architectural fragmentation.

Third, **update safety must be evaluated under uniform inference assumptions**. Changes in prompt structure, evaluation of framing, or inference configuration can materially alter benchmark outcomes without any modification to model parameters. Consequently, claims of forgetting or recovery are only meaningful when assessed under consistent evaluation regimes. This requirement applies equally to baseline measurements and post-update validation.

Fourth, **integration of new capabilities must preserve cross-domain coherence**. Strategies that isolate updates—such as task-specific adapters or frozen backbones—can preserve narrow performance metrics while degrading the model's ability to reason across domains. Safe updates should maintain the model's capacity to correctly classify and route mixed-domain inputs, rather than optimize isolated benchmark stability. Since reasoning capabilities can decouple from memory access (Jin et al., 2025), safe updates must verify that the routing logic between domains remains intact, not that specific facts can still be recalled.

Finally, **model maintenance should prioritize reversibility and auditability over static preservation**. Systems that defer degradation or mask inference conflicts increase operational risk by allowing failures to surface only in deployment. In contrast, update approaches that enable verification of capability accessibility after each change reduce both technical and economic risk, allowing models to evolve without sacrificing reliability.

Empirical observations from publicly available base models, summarized in Appendix A, illustrate that benchmark degradation and recovery can occur under consistent evaluation conditions without reintroduction of task-specific supervision. These results reinforce the need for update criteria grounded in verification of access rather than assumptions of representational loss.

6. Economic and Operational Implications of Update Safety

The distinction between irreversible knowledge loss and reversible performance degradation has material implications for how organizations evaluate the cost and risk of deploying and maintaining large language models. When these failure modes are conflated, update decisions are driven by worst-case assumptions rather than measured system behavior.

In current practice, even modest performance regressions are often treated as evidence that a model has been fundamentally compromised. This interpretation incentivizes full or partial retraining as the default remediation strategy, regardless of whether underlying knowledge has been lost. As models increase in size and specialization, this approach becomes economically unsustainable.

Separating update safety from benchmark stability allows costs to be evaluated more precisely. If degradation reflects access to misalignment rather than representational damage, remediation does not require relearning prior knowledge. This reframing reduces the frequency with which retraining is treated as unavoidable, lowering both direct compute expenditure and indirect operational overhead.

Risk management is similarly affected. Conventional update pipelines offer limited visibility into how prior capabilities are impacted until after deployment, particularly in mixed or downstream tasks. This creates latent risk that only becomes apparent in production environments. An update process that distinguishes between irreversible and reversible failure modes enables earlier detection and containment of regressions, reducing the likelihood of costly post-deployment failures.

These considerations are especially relevant in regulated or high-stakes domains, where the inability to demonstrate continuity of prior behavior constrains update cadence. Systems that can provide credible assurance that previously validated capabilities remain accessible after updates support more frequent iteration without proportionally increasing compliance or safety risk.

At a strategic level, this perspective alters how AI assets are valued. Models are no longer treated as static artifacts whose utility decays with time, but as evolving systems whose value can compound if updates do not erase prior investment. The primary constraint on long-lived deployment shifts from model capacity or data availability to update safety and verification.

Importantly, these implications do not depend on a specific implementation approach. They arise from how to update outcomes that are interpreted and managed. Organizations that continue to equate benchmark degradation with intelligence loss will overpay for retraining and underutilizing existing systems. Those that distinguish between loss and misalignment gain flexibility, cost control, and longer model lifecycles.

Conclusion: Reframing Update Safety in Continual Learning

This paper argues that catastrophic forgetting in large language models is frequently mischaracterized as a singular phenomenon of knowledge erasure. Empirical behavior across benchmarks instead suggests that many observed regressions arise from changes in inference routing and task interpretation, rather than irreversible loss of learned representations.

By distinguishing between knowledge existence and knowledge accessibility, we show that benchmark degradation alone is an insufficient proxy for intelligence loss. This distinction has practical consequences: systems that conflate access degradation with representational damage are often over constrained, leading to unnecessary retraining, inflated costs, and reduced adaptability.

The analysis further demonstrates that existing mitigation strategies, while effective at limiting parameter drift, do not address inference behavior as a first-class system property. As a result, they may preserve surface-level performance while failing to guarantee continued access to prior capabilities under evolving conditions.

Taken together, these observations motivate a shift in how continual learning systems are evaluated and updated. Rather than treating model updates as inherently destructive, update safety should be framed as a verifiable property—one that distinguishes irreversible loss from reversible misalignment.

Importantly, this work intentionally focuses on problem framing and system-level implications rather than implementation details. The mechanisms by which safe updates are enforced are beyond the scope of this paper and are not inferable from the properties discussed herein. Future work will focus on formalizing evaluation criteria and verification strategies for long-lived, updatable AI systems.

References

Jin, M., Luo, W., Cheng, S., Wang, X., Hua, W., Tang, R., Wang, W. Y., & Zhang, Y. (2025). *Disentangling Memory and Reasoning Ability in Large Language Models*. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zheng, J., Cai, X., Qiu, S., & Ma, Q. (2025). *Spurious Forgetting in Continual Learning of Language Models*. *International Conference on Learning Representations (ICLR)*.

Appendix A — Empirical Benchmark Evidence

Appendix A provides empirical benchmark results demonstrating that the behaviors discussed in Sections 2–5 are observable in practice across multiple base models. These results are intended to validate that the described phenomena are not theoretical artifacts or single-model anomalies, but reproducible effects under consistent evaluation conditions.

Evaluation Consistency.
All benchmarks in this appendix were executed using the same evaluation harness, task definitions, and zero-shot configuration. Differences in reported scores reflect changes in model state rather than differences in benchmarking methodology.

Model	Phase	ARC (Δ)	GSM8K CoT (Δ)	HellaSwag (Δ)	Winogrande (Δ)
Nous-Hermes-7B	Math-trained	↓	↑	~	↑
Nous-Hermes-7B	Post-recovery	≈	↑	≈	↑
Llama-2-7B-chat	Math-trained	↓	↑	~	↑
Llama-2-7B-chat	Post-recovery	≈	↑	≈	↑

Table A1. Summary of benchmark performance changes across training phases for two independently evaluated base models.

Interpretation Notes.
Changes in benchmark performance should be interpreted in the context of task sensitivity to inference routing rather than as evidence of representational loss. Benchmarks such as the ARC Challenge are particularly sensitive to shifts in task classification and internal routing behavior, while logic-focused tasks such as Winogrande remain stable or improve across all evaluated conditions.

No model in this appendix exhibits uniform degradation across evaluated tasks. In all cases, performance changes are domain-localized and reversible under subsequent controlled adaptation, consistent with inference misalignment rather than irreversible knowledge destruction as described in Sections 2–4.