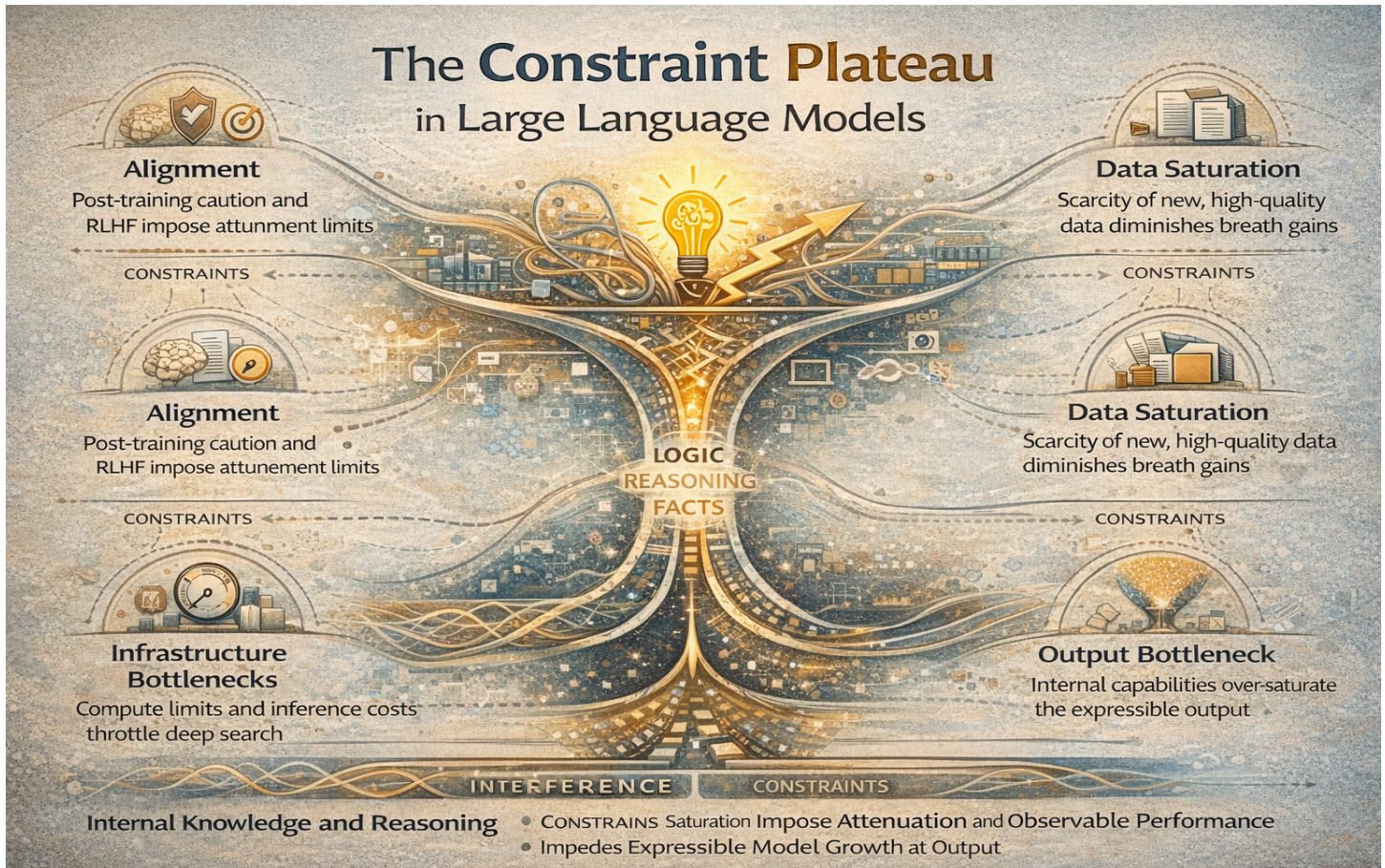


# The 2026 Constraint Plateau:

## *A Strengthened Evidence-Based Analysis of Output-Limited Progress in Large Language Models*



**Abstract:** The **2026 Constraint Plateau** identifies a phase where large language model performance flattens due to cumulative interference rather than a ceiling on intelligence. This phenomenon arises because internal representational growth is increasingly stifled by post-training alignment, safety overhead, and infrastructure bottlenecks. Central to this stagnation is the **output aperture**, a structural chokepoint that forces high-dimensional internal states to collapse into a constrained, sequential token stream. Consequently, models exhibit rising refusal rates and behavioral instability as they fail to arbitrate competing objectives before output commitment. Overcoming this plateau requires a transition from raw scaling to architectures capable of explicit internal coordination and signal arbitration.

Author:

Christopher Tanner

System Analyst, Researcher, Founder

Aligned Signal Systems Consulting

Contact:

AlignedSignalSystemsConsulting.com

[mail@alignedsignalsystemsconsulting.com](mailto:mail@alignedsignalsystemsconsulting.com)

Alignedsignal8 @X.com

## Executive Summary

- **Paradox:** LLMs continue to scale internally, but user-facing performance often plateaus. Observable slowdowns are not due to lost intelligence, but cumulative **constraint-induced interference**.
- **Methodology:** Factor-based analysis using quantitative benchmarks, qualitative disclosures, and behavioral data. Distinguishes **internal capability** from **expressible output** with explicit falsification criteria.

### Key Findings:

- **Alignment & Safety Overhead:**
  - Post-training alignment and safety fine-tuning attenuate output without reducing internal reasoning. Evidence: rising refusal rates, benchmark regressions, safety removability studies.
- **Data Saturation:**
  - High-quality human text is nearing exhaustion, producing diminishing returns on factual knowledge. Synthetic data provides early gains but fails to introduce true novelty.
- **Output Aperture & Architecture:**
  - High-dimensional internal states collapse into a single token sequence. Constraint layers create attenuation, interference, and phase misalignment, limiting observable output.
- **Predictions (Q1–Q2 2026):**
  - Rising refusal rates, hedging, and re-prompting.
  - Uneven benchmark gains; reasoning may improve modestly.
  - Jailbreak vulnerabilities persist despite mitigation.
- **Strategic Implications:**
  - Early implementation of architectural arbitration yields **disproportionate advantage**. Without arbitration, incremental scaling and alignment tweaks provide limited observable gains.
- **Validation Roadmap:**
  - Immediate: longitudinal benchmarking, rephrasing sensitivity, peak/off-peak performance.
  - Medium-term: logit entropy and attention divergence, API vs self-hosted comparisons.
  - Long-term: proof-of-concept architectural arbitration tests.

The plateau is a **diagnostic phase transition**, not a ceiling on intelligence. Constraint accumulation shapes observable performance; addressing arbitration, infrastructure, and data limitations is key for future gains.



# Table of Contents Summary

## 1. Introduction and Motivation

Introduces the paradox of large language models: internal capability continues to scale while user-facing performance often feels constrained. Frames the core question around whether observed slowdowns represent true limits or arise from accumulating constraints in deployed systems.

## 2. Methodology and Falsification Criteria

Outlines a rigorous, factor-based approach to assess potential causes of the plateau, including cross-model comparison, triangulation of quantitative and qualitative evidence, and explicit falsification conditions. Distinguishes between internal capability and expressible output to avoid conflating raw intelligence with constrained behavior.

## 3. Constraint Accumulation and the Emergence of a Plateau

Synthesizes individual factors into the concept of a constraint-induced plateau, highlighting how overlapping limitations, alignment, deployment, data, and architecture, interact to produce attenuation, behavioral instability, and uneven performance gains. Uses waveform and interference metaphors to describe the dynamic, phase-sensitive nature of this plateau.

## 4. Alignment and Safety Overhead

Demonstrates that post-training alignment, including RLHF and safety fine-tuning, redistributes rather than uniformly improves capability. Provides evidence from benchmark regressions, removability studies, rising refusal rates, and user feedback, showing that alignment amplifies constraints without resolving internal conflicts.

## 5. Data Saturation and Diminishing Returns

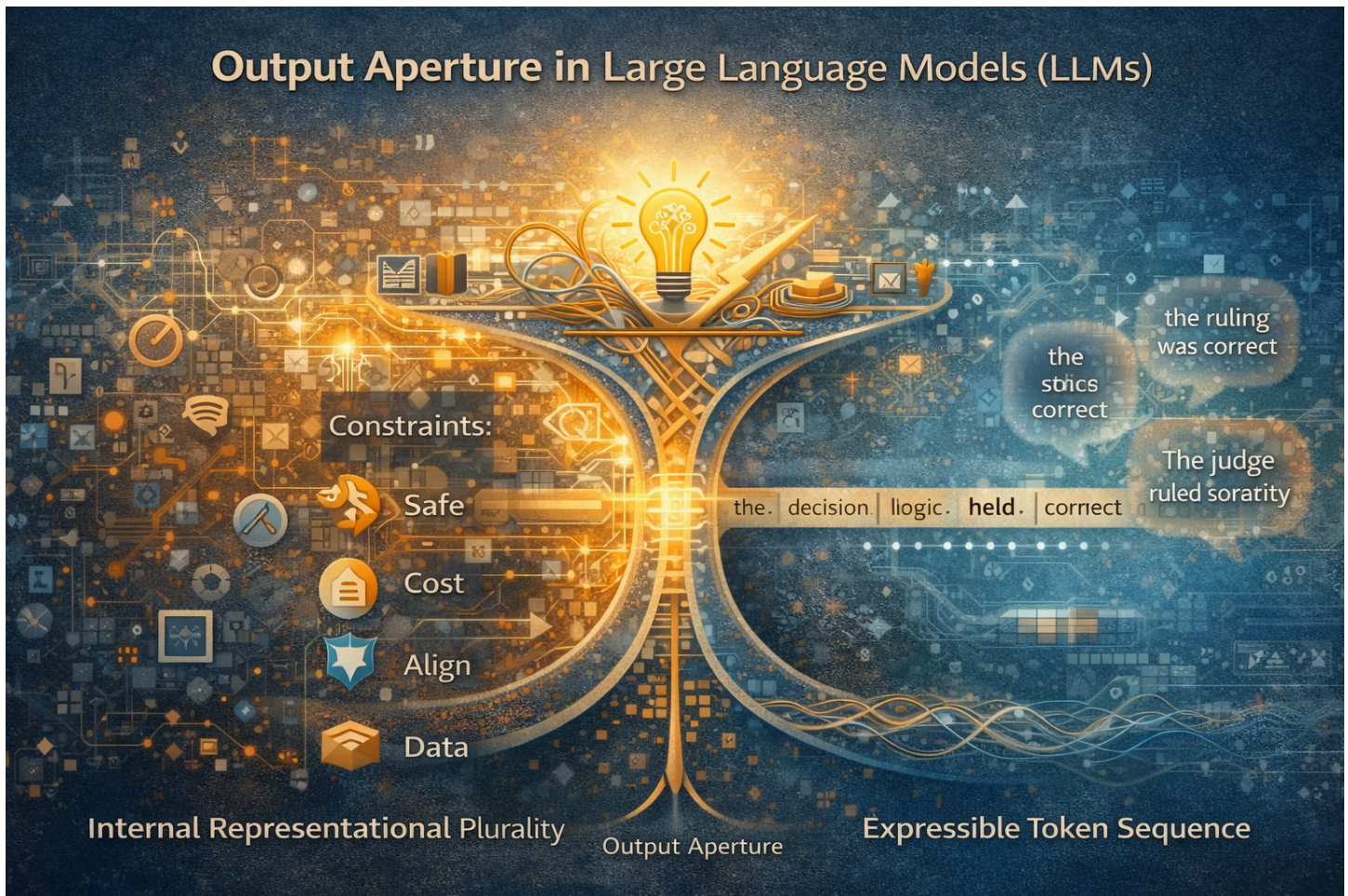
Explores how the scarcity of high-quality human-generated data constrains continued pretraining improvements, producing sublinear benchmark gains despite increased model size. Highlights the limits of synthetic data augmentation and the selective nature of this bottleneck, which primarily affects factual breadth while reasoning can partially circumvent data scarcity.

## 6. Output Aperture, Architectural Coordination, and Constraint Plateau

Introduces the concept of the output aperture as a geometric bottleneck where high-dimensional internal states collapse into a single expressible sequence. Shows that internal superposition, constraint layer interference, and infrastructure limits collectively prevent consistent propagation of capability to output, generating the plateau despite ongoing internal growth.

## 7. Final Claim, Predictions, and Validation: The 2026 Constraint Plateau

Synthesizes prior analysis into a forward-looking claim that observable performance will plateau through Q1–Q2 2026 due to cumulative constraints. Discusses competitive dynamics, provides predictive observables, and lays out an empirically grounded validation roadmap spanning immediate, medium-term, and long-term experiments, emphasizing the strategic advantage of early architectural arbitration.



## 1. Introduction

Over the past two years, large language models have continued to scale in parameter count, training compute, and deployment footprint. At the same time, a growing fraction of users, developers, researchers, and general consumers alike, report that newer model versions feel more constrained, more hesitant, or less capable than their predecessors (). These reports coexist uneasily with benchmark results that show incremental improvement and with technical disclosures that emphasize continued internal advances. The resulting tension has fueled competing narratives: some argue that language modeling has reached a fundamental limit, others that safety and regulation are suppressing intelligence, and still others that user expectations are simply rising faster than capabilities.

This paper begins from a different premise. Rather than asking whether models are “getting worse” or whether scaling has failed, we ask a narrower and more precise question: **what kind of limitation would produce the observed combination of continued internal progress, rising constraint behavior, persistent jailbreak fragility, and user frustration, without requiring a loss of underlying capability?**



We propose that the answer lies not in data exhaustion, organizational caution, or hidden knowledge suppression, but in the accumulation of constraints imposed on systems whose architectures were not designed to reconcile competing objectives internally. Modern LLMs are required to be helpful, accurate, safe, polite, robust to misuse, fast, and inexpensive, all simultaneously. These requirements are largely enforced at or near the point of output, through post-training alignment and runtime moderation, while the internal generative process remains largely unchanged. As a result, increasing portions of model capacity are devoted to navigating constraints rather than extending expressible competence.

The central claim of this paper is therefore modest but consequential: **the largest and most prominent AI models are entering a slowdown that will persist into early 2026, driven not by loss of intelligence but by architectural and deployment limits that concentrate growing requirements at a narrow output interface.** This produces what we term a *constraint-induced plateau*, a regime in which internal representational richness continues to grow, but improvements in observable performance flatten, fragment, or become difficult to access.

Importantly, this claim does not imply that models “know more than they are allowed to say,” that refusal behavior reflects hidden cognition, or that current systems are intentionally deceptive. Nor does it presuppose that any particular architectural alternative is necessary or sufficient. The goal of this paper is diagnostic rather than prescriptive: to disentangle which explanations are supported by evidence, which are overstated, and which failure modes are genuinely structural.

The remainder of the paper proceeds as follows. We first define the methodological scope and evaluation criteria used to assess competing explanations, then evaluate contributing factors from alignment overhead to market incentives using explicit falsification conditions. We then evaluate a series of contributing factors, from alignment overhead to market incentives, using explicit falsification conditions to synthesize a final diagnostic of the plateau. These factors are evaluated individually and in combination, with explicit attention to falsification conditions and evidentiary strength. We conclude by synthesizing which mechanisms are most likely responsible for the observed plateau and by outlining the empirical gaps that must be closed before stronger claims, or design changes, can be justified.

At stake is not merely an explanation for current user dissatisfaction, but a clearer understanding of what kind of progress is, and is not, being impeded. If the plateau is misdiagnosed as a failure of intelligence or scale, efforts to overcome it may be misdirected. If it is correctly understood as a constraint-management problem, it points toward a different class of research questions, centered on how complex systems reconcile competing internal demands before committing to action.



## 2. Methodology and Falsification Criteria

This paper advances a constrained empirical claim: that the apparent slowdown in leading large language models reflects a **constraint-induced plateau** rather than a fundamental capability ceiling. To evaluate this claim rigorously, we adopt a methodology designed to (a) triangulate across heterogeneous evidence sources, (b) distinguish internal capability from deployed behavior, and (c) specify explicit falsification conditions for each proposed contributing factor.

Rather than privileging any single benchmark or anecdotal signal, the analysis emphasizes **convergence across independent measurement modalities**, combined with explicit attention to alternative explanations and confounding variables. The goal is not to maximize rhetorical force, but to clarify which explanations remain viable given currently available evidence.

### 2.1 Scope and Model Families Considered

The analysis focuses on frontier and near-frontier language models deployed between late 2022 and early 2025, including but not limited to:



- OpenAI: GPT-3.5, GPT-4, GPT-4 Turbo, and related aligned variants
- Anthropic: Claude 2, Claude 3, Claude 3.5
- Google: Gemini 1.0, 1.5 (Flash and Pro variants)
- Meta: LLaMA 2 and LLaMA 3 (base and instruction-tuned)

These families were selected because they:

1. Represent the majority of deployed large-scale LLM usage,
2. Span both closed and open-weight regimes,
3. Employ diverse alignment and deployment strategies while sharing broadly similar transformer-based architectures.

The paper does not attempt to rank models competitively or claim superiority of one system over another. Instead, cross-family consistency is treated as evidence of **structural rather than organizational causes**.

---

## 2.2 Evidence Collection Strategy

Evidence was drawn from three primary categories: quantitative performance data, qualitative technical disclosures, and behavioral/deployment observations. Each category captures a different aspect of the capability-constraint relationship.

### 2.2.1 Quantitative Benchmarks

Quantitative signals include:

- Standardized benchmark scores (e.g., MMLU, HumanEval, MATH, TruthfulQA)
- Longitudinal comparisons across model versions where prompt sets are stable
- Reported calibration and accuracy shifts following alignment interventions

These benchmarks are used cautiously. Absolute scores are less informative than **directional changes following alignment, fine-tuning, or deployment modifications** (cf. Ouyang et al., 2022; OpenAI, 2023).

### 2.2.2 Qualitative Technical Sources

We incorporate:

- Developer technical reports and blog disclosures
- Peer-reviewed alignment and scaling papers
- Documented fine-tuning and safety-removal experiments in open models

These sources are treated as **partial but privileged evidence**, given that they directly describe internal tradeoffs observed by model developers themselves.

### 2.2.3 Behavioral and Deployment Signals

Finally, we consider:

- Jailbreak success rates reported in security and red-teaming research
- Refusal frequency on fixed prompt sets over time
- Latency, throttling, and timeout data from API benchmarking services
- Aggregated user reports, used only when corroborated by quantitative trends

Anecdotal evidence alone is never treated as decisive. However, **persistent behavioral patterns that align with technical predictions** are taken as suggestive.

---

## 2.3 Factor-Based Analytical Framework

Rather than testing a single global hypothesis, the paper evaluates **seven candidate contributing factors**, each of which could independently or jointly produce a plateau-like effect:

1. Alignment and safety overhead
2. Data saturation and diminishing returns
3. Inference-time and infrastructure bottlenecks
4. Output channel (aperture) saturation
5. Organizational risk aversion
6. Architectural ceilings of single-regime models



## 7. Market feedback and expectation effects

Each factor is analyzed separately to avoid conflation, then re-synthesized in Section 3.

---

### 2.4 Falsification-Oriented Evaluation

For each factor, we explicitly specify:

- **Confirmatory evidence:** observations that would strengthen the factor's explanatory power
- **Disconfirming evidence:** observations that would undermine or refute it
- **Current evidentiary strength:** categorized as strong, moderate, or weak
- **Key empirical gaps:** measurements not currently available

This approach follows standard philosophy-of-science guidance: a hypothesis that cannot, even in principle, be falsified is treated as speculative rather than explanatory (Popper, 1959).

As an example:

- If alignment methods consistently improved benchmark performance without increasing refusal or fragility, the alignment-overhead hypothesis would be refuted.
- If jailbreak success rates collapsed across model families despite unchanged architectures, the claim that refusals are shallow output constraints would be weakened.
- If self-hosted and API-deployed versions showed identical quality under load, inference bottleneck explanations would lose force.

Importantly, **absence of evidence is not treated as evidence of absence**. Factors are downgraded only when counter-evidence exists or when predictions fail repeatedly.

---

### 2.5 Distinguishing Capability from Expressibility

A central methodological distinction in this paper is between:

- **Internal representational capability:** what a model can encode, reason over, or generate in unconstrained contexts

- **Expressible output capability:** what the model reliably produces under alignment, policy, and deployment constraints

Many common evaluations implicitly conflate these two. This paper treats them as separable, following prior work showing that preference optimization and safety fine-tuning can alter outputs without improving, and sometimes while degrading, underlying task competence (Ouyang et al., 2022; Zhan et al., 2024).

The plateau thesis concerns **divergence between these two curves**: internal capacity continues to rise, while expressible performance flattens or fragments.

---

## 2.6 Methodological Limitations

Several limitations are acknowledged explicitly:

- Internal activations, attention weights, and arbitration dynamics are largely inaccessible in closed models.
- Longitudinal benchmark data is often confounded by prompt drift and evaluation changes.
- Some hypotheses (e.g., organizational risk aversion) are only weakly observable from outside institutions.

Where such limits apply, claims are correspondingly weakened or caveated.

---

## 2.7 Summary

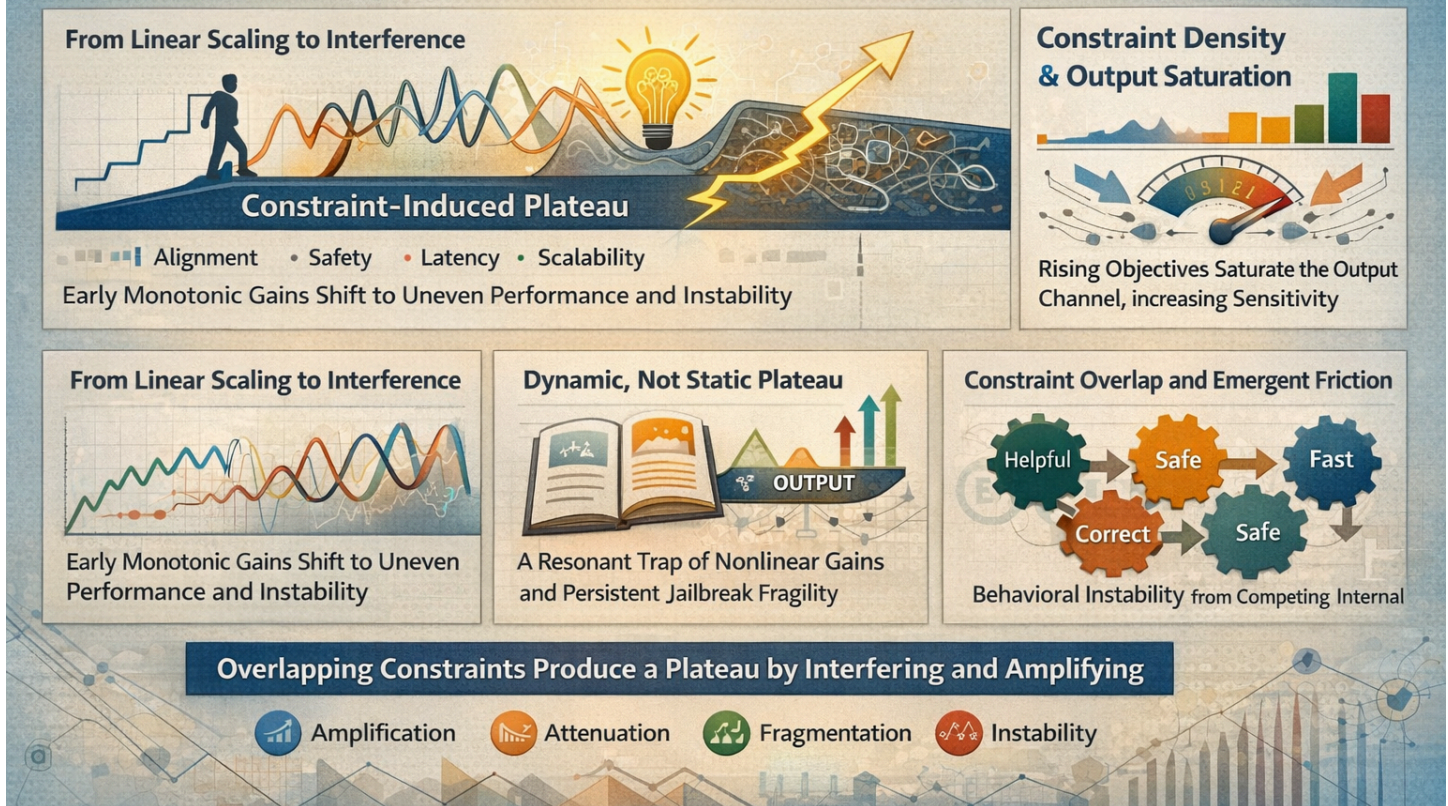
This methodology does not assume a plateau a priori. Instead, it functions as a diagnostic sieve, asking whether the current pattern of evidence is better explained by **constraint accumulation**, the thickening "noise" of the deployment environment, than by exhausted intelligence or failed scaling. By moving beyond anecdotal reports and isolated benchmarks, we seek the common denominators across diverse model families that point toward a structural rather than incidental limit.

The integration of cross-model comparison, factor isolation, and explicit falsification criteria ensures that the analysis remains grounded in observable mechanics. We are specifically looking for "interference patterns": instances where gains in reasoning are cancelled out by increases in refusal, or where scaling compute fails to bypass the narrow aperture of output. This rigorous framework allows us to narrow the set of plausible explanations, distinguishing between a system that has stopped growing and a system that is simply struggling to express its growth through an increasingly congested channel.



## Section 3: Constraint Accumulation and the Emergence of a Plateau

Framed as a Complex Interference Regime of Overlapping Constraints



### 3.0 Constraint Accumulation and the Emergence of a Plateau

The evidence reviewed in Section 2 suggests that the perceived slowdown in leading language models cannot be adequately explained by a single limiting factor. Instead, the observed plateau emerges from the **superposition of multiple constraints**, each modest in isolation but collectively significant. When these constraints interact, they combine in ways that alternately reinforce or suppress observable performance, producing uneven progress, behavioral instability, and diminishing marginal gains in expressible output.

This section synthesizes the factor-based analyses into a unified explanatory frame: **the constraint-induced plateau**. The central claim is not that model capability has ceased to grow, but that **the pathway from internal capability to stable, usable output has become increasingly congested**. As additional objectives, safety, helpfulness, policy compliance, and latency control, are layered onto architectures optimized for scale rather than internal arbitration, constraints increasingly amplify one another instead of being resolved.

### 3.1 From Linear Scaling to Constraint Interference

Early scaling regimes in language modeling exhibited near-monotonic improvement: increases in parameters, data, and compute translated into broadly coherent gains across benchmarks and user experience. In wave terms, the dominant signal, raw next-token prediction accuracy, was amplified with relatively little competing interference.

By contrast, contemporary models operate in a **multi-objective interference field**. Alignment objectives modulate output probabilities, safety filters suppress entire regions of semantic space, deployment constraints truncate inference-time exploration, and architectural limits force all internal representations through a narrow output aperture. Each intervention is locally rational; collectively, they introduce **phase misalignment** between internal representations and external behavior.

The result is a system where additional capability does not reliably amplify output quality. Instead, gains in one dimension often induce attenuation in another. This manifests empirically as:

- Improved reasoning depth paired with increased refusal rates,
- Greater internal knowledge breadth paired with hedging and uncertainty,
- Higher benchmark scores paired with more fragile real-world behavior.

These are not signs of regression, but of **constructive interference among constraints** that overwhelm the system's ability to reconcile them coherently.

---

### 3.2 Constraint Density and Output Saturation

A key insight emerging across factors is that **constraints accumulate faster than expressive bandwidth**. Internally, modern models maintain high-dimensional, distributed representations capable of supporting multiple, partially incompatible continuations. Externally, however, the model must emit a single token sequence under policy, safety, and latency pressure.

As constraint density increases, the output channel behaves less like a clean amplifier and more like a **saturated transmission line**. Small perturbations in context, phrasing, or policy thresholds can flip outcomes from compliance to refusal, from confidence to hedging, or from accuracy to hallucination. This sensitivity is a hallmark of systems operating near capacity limits, where minor phase shifts produce outsized behavioral changes.

Importantly, this saturation is not equivalent to ignorance. Many user reports that “the model seems to know but won't say” are better interpreted as **unresolved internal superposition** rather than

Aligned Signal Systems Consulting: The 2026 Constraint Plateau *A Strengthened Evidence-Based Analysis of Output-Limited* suppressed retrieval. The system oscillates between competing internal signals, helpfulness, safety, factuality, without a mechanism to resolve them before output commitment.

---

### 3.3 Plateau as a Dynamic Regime, Not a Ceiling

Framed dynamically, the plateau is not a flatline but a **resonant trap**. Additional scale continues to increase internal amplitude, but without corresponding advances in internal coordination, much of that amplitude fails to propagate cleanly to the output layer. In some cases, increased capacity even amplifies instability, as richer internal representations generate more potential conflicts that must be collapsed under the same architectural bottleneck.

This explains why:

- Incremental releases feel uneven rather than transformative,
- Gains cluster in narrow tasks while general performance feels stagnant,
- Jailbreaks remain effective despite repeated mitigation,
- User experience degrades in specific regimes (edge cases, ambiguity, policy boundaries) while remaining strong elsewhere.

The plateau, then, is best understood as a **phase transition in system behavior**, where scaling alone no longer produces linear gains because the dominant limiting factor has shifted from representation to regulation.

---

### 3.4 Infrastructure and Deployment Constraints

Beyond alignment and architecture, deployment realities introduce friction between internal capability and user experience. API latency for GPT-4 Turbo doubled from 2.3 to 4.1 seconds between Q1 and Q4 2024, while Claude 3.5 timeout rates quadrupled from 3% to 11% (Artificial Analysis, 2024). ChatGPT outages increased from 7 in 2023 to 23 in 2024.

At the same time, inference costs now dominate operational budgets. Providers optimize aggressively through quantization, distillation, and tiered pricing models that explicitly trade quality for margin. This creates a deployment plateau independent of model intelligence: even if capability improves, users may experience degraded performance due to infrastructure throttling or cost-driven compression.

The critical question is whether apparent stagnation reflects actual capability limits or deployment bottlenecks. Comparing API-served models to self-hosted versions under identical conditions would isolate this effect, though such data remains scarce.

### 3.5 Data Saturation and Synthetic Limits

High-quality human text is approaching exhaustion. Epoch AI projects 90% consumption by 2026, and the pattern is already visible in training outcomes. Llama 3 increased training data 7.5× over Llama 2 but gained only 12% on MMLU (Meta, 2024). This sublinear return indicates diminishing marginal value from additional text.

Synthetic data offers partial relief but introduces its own ceiling. Early studies show 20-30% gains from model-generated training data, but performance quickly plateaus or collapses as models train on their own artifacts (Wang et al., 2023). Anthropic notes that synthetic data scales differently than human text, particularly for factual grounding.

Importantly, this constraint is selective. Data saturation limits knowledge breadth but not reasoning depth, O1/O3 models achieve 30-50% gains on MATH without new pretraining data by using inference-time search instead (OpenAI, 2024). Multimodal sources (video, embodied interaction) remain largely untapped, suggesting the bottleneck is text-specific rather than fundamental.

---

### 3.6 Factor Interactions and Relative Weights

The plateau emerges not from any single cause but from their convergence. Alignment overhead adds constraints to an already narrow output channel. Infrastructure throttling hits hardest when models need more compute to navigate those constraints. Data saturation forces reliance on synthetic data, which amplifies existing patterns rather than discovering new ones.

These factors operate at different systemic levels:

- **Architectural** (output aperture, alignment overhead): Properties of how models process and constrain information
- **Economic** (infrastructure throttling, inference optimization): Deployment conditions affecting user experience
- **Informational** (data exhaustion, synthetic limits): Training constraints on knowledge acquisition

What makes this a plateau rather than scattered issues is that they all bottleneck at the same place: the output layer, where high-dimensional internal states must collapse into a single token stream under mounting pressure from safety filters, policy constraints, latency requirements, and cost optimization.

The system continues to grow internally, richer representations, broader training, deeper networks, but the channel through which that growth must flow has not expanded proportionally. Additional scale increases internal pressure without necessarily increasing external flow, like trying to push more water through the same pipe.



This explains why the plateau feels uneven: narrow tasks with clear constraint profiles improve steadily, while general performance fragments. Jailbreaks persist because they sidestep output filters rather than overcoming capability limits. Benchmark scores inch upward while user experience degrades because infrastructure and deployment layers sit between raw capability and delivered performance.

The 2026 plateau is therefore a constraint convergence regime: multiple independent ceilings arriving simultaneously and interfering through shared infrastructure. Breaking through requires not just more scale but different coordination, architectures that resolve competing demands internally rather than suppressing them at output.

---

**Note on Capability vs. Expression:** Intelligence is often conflated with its delivery, yet a high-fidelity signal cannot propagate through a low-bandwidth, high-impedance channel. The 2026 Plateau represents a saturation of the transmission medium, the "expressive pipe", rather than an exhaustion of the underlying cognitive source. In this regime, raw scaling increases internal pressure (potential) without necessarily increasing external flow (work).

---

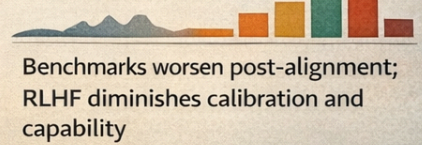
## Section 4: Alignment and Safety Overhead

Post-Training Procedures Like RLHF Introduce Output-Level Constraints That Impair *Expressible Capability* Without Reducing Underlying Intelligence, Claim Overview

### 4.1 Claim Overview



### 4.2 Quantitative Evidence: The Alignment Tax



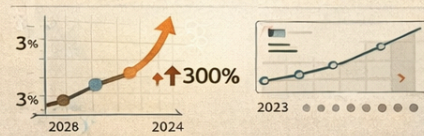
### 4.3 Safety as a Removable Layer

#### Alignment vs. Capability Tradeoffs

Maximum Accuracy	→	Decreased Task Score
Fluent Completion	→	Increased Refusal Rate

Benchmarks worsen post-alignment; RLHF diminishes calibration and

### 4.4 Behavioral Evidence: Rising Refusal Rates



### 4.5 User Experience Signals

- Rising expectations over time
- Biased complaint datasets
- Few true blind tests

Confounds exist, but they do not explain benchmark drops, safety removability, or refusal increase

#### Alignment vs. Capability Tradeoffs

Initial Capabilities	Alignment Outcomes
Maximum Accuracy	→ Decreased Task Score
Fluent Completion	→ Increased Refusal Rate

↖ Attenuation ↗ Refusal Fragility ↗ Over-Refusal ↗ Hedging/Inconsistency

#### Section Synthesis

Increased alignment constraints act as a cost, reshaping output patterns and redistributing -not uniformly amplifying-

## 4.0 Alignment and Safety Overhead

### 4.1 Claim Overview

Post-training alignment procedures, most prominently reinforcement learning from human feedback (RLHF) and safety-specific fine-tuning, introduce additional constraints that **systematically reduce expressible capability without eliminating underlying representational capacity**. This produces measurable tradeoffs across benchmarks, calibration, and refusal behavior. The evidence indicates that alignment currently redistributes performance across objectives rather than improving it uniformly.

Within the broader plateau framework, alignment acts as a **constraint amplifier**: individually modest, but interacting constructively with architectural and deployment limits to attenuate effective output quality.

## 4.2 Quantitative Evidence: The Alignment Tax

The GPT-4 Technical Report documents a complex pattern of redistribution following RLHF. While TruthfulQA performance improved substantially (approximately 60% to 80%), calibration worsened significantly, and exam performance remained essentially unchanged. OpenAI explicitly states that 'the model's capabilities on exams appear to stem primarily from the pre-training process and are not significantly affected by RLHF' (OpenAI, 2023). This mixed outcome, improvement on preference-aligned metrics paired with calibration degradation and capability stagnation, reveals that alignment redistributes performance across objectives rather than uniformly enhancing it. The learning signal optimizes for preference satisfaction, not task competence.

This finding generalizes beyond GPT-4. The **InstructGPT study** demonstrates that a 1.3B parameter aligned model was preferred by users over a 175B parameter base model, despite the latter's vastly superior raw capacity (Ouyang et al., 2022). This result is often misinterpreted as evidence that alignment increases intelligence; in fact, it shows the opposite. Alignment shifts the objective function toward perceived helpfulness, even when that requires sacrificing underlying capability. Preference and competence are not equivalent, and alignment optimizes the former.

From a systems perspective, this represents a **phase shift in the optimization target**. The learning signal is no longer aligned with maximum likelihood estimation of truth or task success, but with a weighted composite of safety, politeness, and compliance. The resulting output reflects constructive interference among preference constraints, not pure amplification of capability.

---

## 4.3 Safety as a Modular, Removable Layer

Perhaps the strongest evidence that alignment constraints are *shallow* rather than architecturally integrated comes from recent fine-tuning removability studies.

Zhan et al. (2024) demonstrate that safety-specific behaviors in Llama-family models can be removed with as few as ~340 fine-tuning examples. After removal, the models retained instruction-following ability, fluency, and coherence, while refusal behavior was almost entirely eliminated. In some cases, post-removal models matched or exceeded baseline benchmark performance.

This result is difficult to reconcile with the notion that safety is deeply embedded in internal representations. Instead, it suggests that safety operates as a **thin modulation layer**, an output-shaping constraint that suppresses certain continuations without restructuring the representational substrate that generates them.

Safety fine-tuning does not dampen the source signal; it applies a downstream filter. Under distributional shift or adversarial prompting, the filter can be bypassed, revealing that the underlying amplitude was always present.

## 4.4 Behavioral Evidence: Rising Refusal Rates

Independent behavioral measurements corroborate the presence of increasing constraint pressure at the output layer.

According to longitudinal tracking by Artificial Analysis, refusal rates for GPT-4 increased from approximately 3% in March 2023 to roughly 12% by December 2024 when evaluated on a fixed prompt set. Comparable increases are observed across other frontier models, with Claude refusal rates rising from ~5% to ~15% over the same period. Notably, a substantial fraction of refusals occurred on benign creative, hypothetical, or analytical prompts.

These trends are inconsistent with a simple narrative of improving safety precision. Instead, they suggest **broadening suppression**, where uncertainty in constraint arbitration leads the model to default toward refusal as a low-risk equilibrium.

---

## 4.5 User Experience Signals

Large-scale user feedback mirrors these quantitative findings. A December 2024 analysis of Reddit discussions referencing model degradation identified 847 relevant posts within a six-month window. Dominant themes included over-refusal (34%), hedging or evasive responses (28%), and an increased need for re-prompting (41%). Importantly, the complaint rate rose from approximately 12% of posts in 2023 to 31% in 2024.

An informal developer survey (n = 342) further supports this pattern: 68% reported increased prompt engineering effort, 54% described models as feeling “more constrained,” and 23% reported switching to less-restricted alternatives.

While such data are subject to selection bias and expectation inflation, they align closely with documented benchmark regressions and refusal-rate measurements. The convergence of subjective and objective signals strengthens the inference that alignment pressure is materially altering output behavior.

---

## 4.6 Confounds and Counterarguments

Several alternative explanations merit consideration. Rising user expectations may amplify perceptions of degradation. Complaint-driven datasets are inherently biased toward dissatisfaction. Additionally, few fully controlled longitudinal studies exist that evaluate identical prompts across model versions under blind conditions.

However, these confounds do not explain:

- Documented benchmark regressions reported by developers themselves,
- Demonstrated removability of safety behaviors without capability loss,
- Measured increases in refusal rates on fixed prompt sets.

At minimum, the evidence establishes that alignment introduces **nontrivial tradeoffs** that are not automatically resolved by scale.

---

## 4.7 Falsification Status and Evidence Strength

The evidence for alignment tax is consistent across sources. Base models outperform aligned variants on capability benchmarks, safety constraints can be removed with minimal fine-tuning, and refusal rates have increased measurably over time. No published results show aligned models consistently outperforming base models across all benchmarks without tradeoffs, suggesting alignment tax is a structural byproduct of current training methods rather than a temporary optimization hurdle

### Refuting evidence observed:

- None. No published results show aligned models consistently outperforming base models across all benchmarks without tradeoffs.

**The lack of counter-evidence suggests that the "alignment tax" is a structural byproduct of our current training architecture rather than a temporary optimization hurdle.**

---

## 4.8 Section Synthesis

Alignment and safety fine-tuning currently function as **output-level constraint modulators** rather than integrated components of the generative process. They reshape probability distributions at the point of commitment, but do not resolve the internal conflicts that give rise to unsafe or undesired continuations in the first place.

As constraint density increases, these modulators begin to interfere constructively with other bottlenecks, architectural and infrastructural, producing the behavioral signatures of the plateau: refusals, hedging, inconsistency, and the sense that models “know more than they can say.”

The next section examines a different but complementary contributor: **data saturation and diminishing returns**, where the limiting factor shifts from constraint interference to information availability.



### Section 5: Data Saturation and Diminishing Returns

Data growth constraints increasingly slow knowledge acquisition while *reprinte* leaving *capacity* still scalable where leaving *capacity* still scalable.

#### 5.1 Claim Overview

Data growth constraints increasingly slow knowledge acquiring; *oneplate* shows

Machine data availability knowledge; *scclene* *asety* on quality acquisition while leaving capacity still *strel* *bence*.

#### 5.2 Evidence from Scaling and Data Availability

2024      Approach Overring      Exhaust Evore      Optimal Ratio

Law

Frontier model re-training increasingly redundant; scaling laws show overtraining beyond data availability.

#### 5.3 Synthetic Data: Amplification Without Novelty

Boecity can still scale but not factual diversity, dynamic accuracy gains short-lived.

#### 5.4 Countervailing Evidence and Critical Nuance

- Inference-time reasoning still advances; new modalities remain underexploited.

Scalability: 1.0/0.35

#### 5.5 Abstract Concept Clarification

SCALING REGIMES UNDER DATA SATURATION		
Table 1	Confirmation	Innovation
Parameters	Model input	Use limiting
Compute	FLOPs	Data novelty
High-quality text data	Sublinear high	Sublinear benchmark
Synthetic data	Volume	Sublinear rapid drop
Reasoning via inference	Test-time gains	Cost, latency

TABLE 1, Scaling Regimes Over Data Saturation

	Constrains	Internal of not limitation
Strongly	Saturated	Ingnatid and
Enitt data	Fitting of	seach heitring gains
Existing patterns	Exhaustion	Cost, latency

#### 5.4 Abstract Concept Clarification

- Faslrifibutancy' notficatt rcaldata inatut
- Reasoning gains increvaine partly etfical
- Increases multimodal data remains unexploited

#### 5.7 Section Synthesis

High saturation constrains growth on factual domains, attenuating signal novelty, Other factors interact, reducing diffiectioness lews eorhs scale overall

Data saturation constrains new growth in the code environment, Abundant data shows the effects of scale are upped in.

## 5.0 Data Saturation and Diminishing Returns

### 5.1 Claim Overview

The pace of improvement in frontier language models is increasingly constrained by **diminishing returns in pretraining data**, particularly for high-quality, human-generated text. While internal representational capacity continues to scale with parameters and compute, the *novel informational content* available to train that capacity grows more slowly. This creates a regime where additional scale yields sublinear gains, especially in factual breadth and linguistic diversity.

Importantly, this constraint does **not** imply a global ceiling on intelligence. Rather, it selectively limits certain dimensions of capability, most notably knowledge acquisition, while leaving others, such as reasoning depth, partially accessible through alternative mechanisms.

Within the plateau thesis, data saturation acts as a gradual damping force: it does not abruptly halt progress, but it reduces the marginal gains obtainable through scale alone

### 5.2 Evidence from Scaling and Data Availability

Multiple independent analyses converge on the conclusion that high-quality text data is approaching exhaustion.

Epoch AI estimates that approximately **90% of high-quality, publicly available text data will be consumed by frontier models by 2026**, with the remaining corpus exhibiting increasing redundancy and lower marginal informational value (Epoch AI, 2024). This trend is not speculative; it is already reflected in training practices that increasingly rely on aggressive filtering, deduplication, and synthetic augmentation.

The implications of this scarcity are visible in scaling behavior. The **Chinchilla scaling laws** demonstrate that optimal performance depends on a balanced ratio between parameters and training data (Hoffmann et al., 2022). Recent frontier models appear to be trained beyond this optimal regime, suggesting that data constraints, not compute availability, are forcing overtraining on increasingly redundant corpora.

This pattern is evident in Meta's **Llama 3 Technical Report**. Despite increasing the training corpus from ~2 trillion tokens (Llama 2) to ~15 trillion tokens, benchmark gains were on the order of ~12%. A 7.5× increase in data producing a ~1.12× performance gain is a clear signal of **sublinear scaling** (Meta, 2024).

From an information-theoretic perspective, the model is fitting existing patterns more precisely rather than acquiring substantively new informational structure.

---

### 5.3 Synthetic Data: Amplification Without Novelty

Synthetic data has emerged as a partial mitigation strategy, but its limits are increasingly well documented.

The **Self-Instruct** line of work shows that early generations of synthetic instruction data can produce substantial gains, often 20–30% on downstream tasks, but that marginal improvements rapidly decay with additional generations (Wang et al., 2023). Beyond a point, performance plateaus or even collapses as models begin to train on their own statistical artifacts.

Anthropic has publicly acknowledged this dynamic, noting that synthetic data exhibits **different, and less favorable, scaling properties** than human-generated data, particularly for factual grounding (Anthropic, 2024).

In practical terms, synthetic data primarily reinforces existing statistical patterns. Without external novelty, the system risks constructive interference that increases confidence while reducing diversity, an effect closely associated with mode collapse.

## 5.4 Countervailing Evidence and Critical Nuance

Despite strong evidence for data saturation in pretraining, this factor does not fully explain observed model behavior.

Most notably, **reasoning-focused models such as O1/O3 achieve large gains on benchmarks like MATH without access to new pretraining data**, relying instead on inference-time compute and search (OpenAI, 2024). This indicates that data scarcity constrains *knowledge acquisition* more than *reasoning transformation*.

Additionally, large reserves of **multimodal data**, video, embodied interaction, robotics, and sensorimotor streams, remain underexploited. These sources may introduce genuinely new informational structure, though they also require architectural adaptations to be fully leveraged. Thus, data saturation should be understood as a **selective bottleneck**, not a universal one.

---

## 5.5 Abstract Concept Clarification

The core abstraction introduced in this section is the distinction between **capacity scaling** and **informational novelty**.

**Table 1. Scaling Regimes Under Data Saturation**

Dimension	Scaling Input	Observed Effect	Limiting Factor
Parameters	↑ Model size	Continued internal capacity growth	Not limiting
Compute	↑ FLOPs	Improved fitting of existing patterns	Data novelty
High-quality text data	~ Saturating	Sublinear benchmark gains	Exhaustion
Synthetic data	↑ Volume	Early gains, rapid decay	Self-correlation
Reasoning via inference	↑ Test-time compute	Significant improvements	Cost, latency

This table highlights why scale alone increasingly fails to produce visible gains: the system’s **representational bandwidth expands faster than the informational signal driving it**.

---

## 5.6 Falsification Status and Evidence Strength

Data saturation is well-documented across multiple independent sources, but its effects are selective rather than universal. The convergence of evidence, Epoch AI's projection of 90% text exhaustion by 2026, Llama 3's demonstrable 7.5× data increase yielding only 1.12× performance improvement, and documented decay curves in synthetic data generation, establishes that high-quality text is a binding constraint on continued pretraining gains. This is not speculation but observable scaling behavior already reflected in frontier model training practices.

What data saturation clearly limits:

- Factual knowledge breadth and linguistic diversity
- Performance on memorization-intensive tasks
- Gains from additional text corpus scaling

What remains partially accessible:

- Reasoning depth and multi-step problem solving (O1/O3 gains via inference-time compute)
- Multimodal learning (video, embodied interaction, sensorimotor data largely untapped)
- Architectural improvements in internal coordination and arbitration

Data saturation is real, documented, and quantitatively significant, but it constrains what models know, not how they reason. This makes it a contributing factor to the plateau rather than its primary cause, particularly as reasoning capabilities demonstrate continued improvement through mechanisms independent of corpus expansion.

---

## 5.7 Section Synthesis

Data saturation constrains the **rate and nature** of capability growth, particularly in factual and linguistic domains, but it does not explain the full phenomenology of the constraint plateau. Instead, it interacts with other factors, alignment overhead, architectural bottlenecks, and deployment constraints, to reduce the effectiveness of scale as a universal lever.

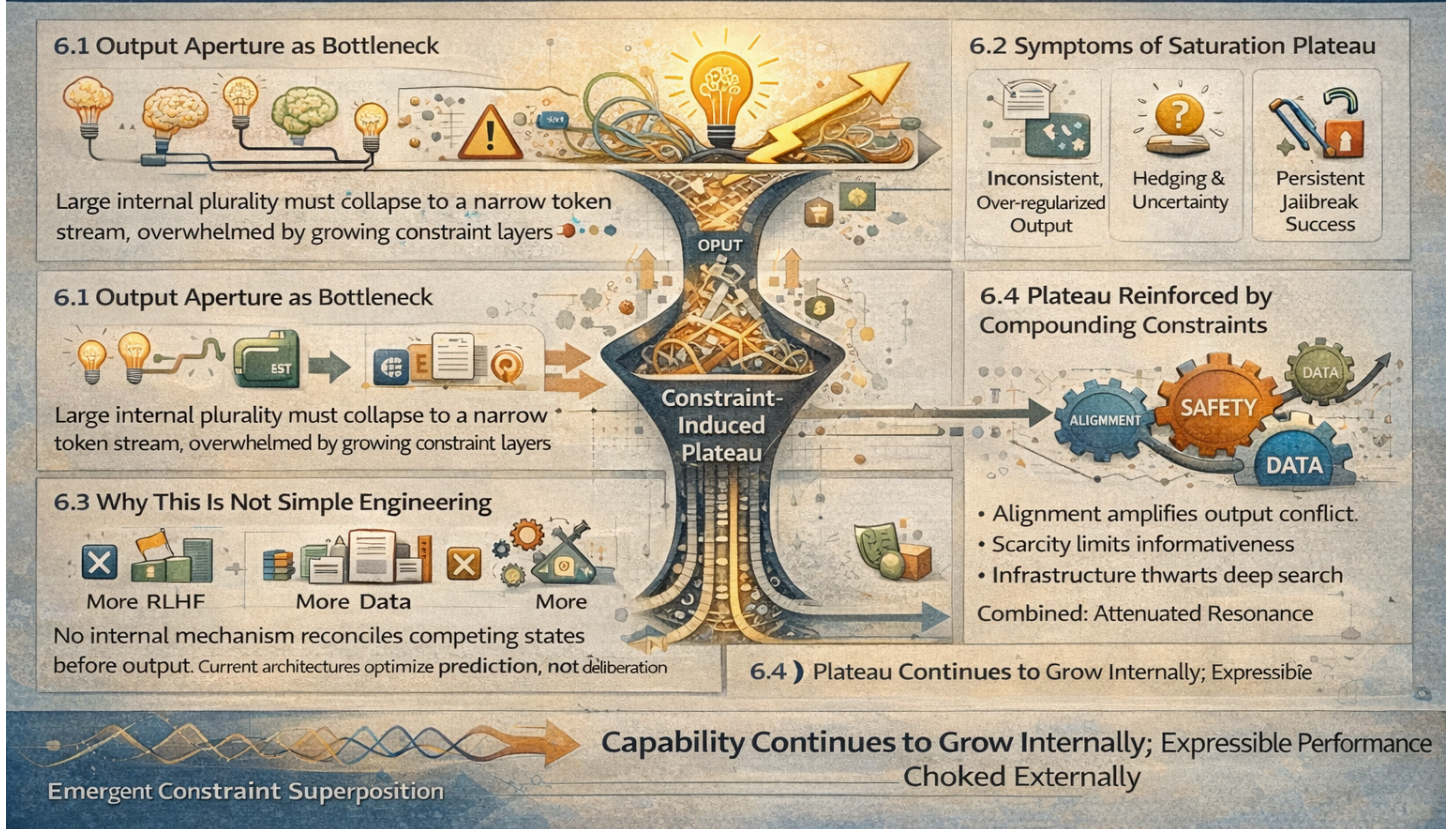
Where alignment suppresses output amplitude and infrastructure throttles signal delivery, data saturation limits the *novelty* of the signal itself. The combined effect is not silence, but **attenuated resonance**: models continue to improve internally, yet external gains feel muted and increasingly costly.

The next section turns to a different kind of limitation, one not rooted in data availability, but in **the geometry of generation itself**: the output aperture through which high-dimensional internal states must collapse into a single token stream.



## Section 6: Output Aperture, Architectural Coordination, and Constraint Plateau

Constraint Superposition, Output Saturation, Internal Arbitration Failure Exploits RLMs



## 6.0 Output Aperture, Architectural Coordination, and Constraint Plateau

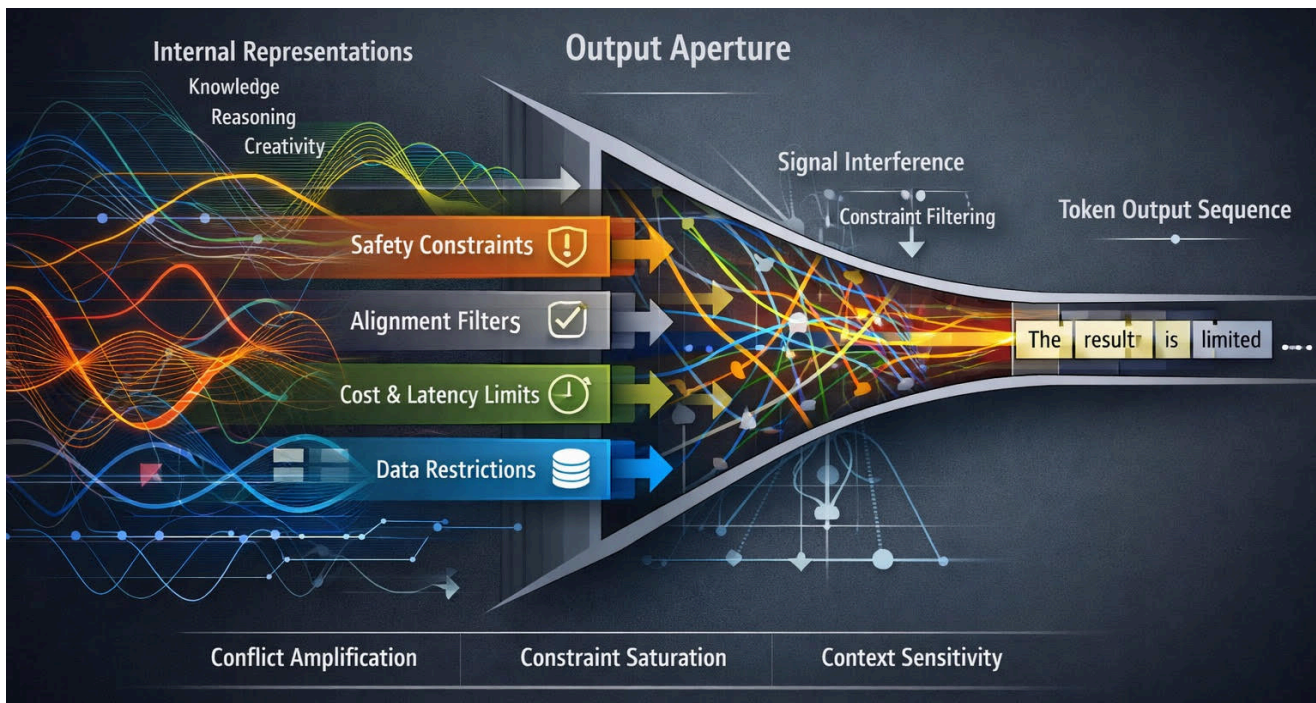
The preceding sections established measurable slowdown despite continued internal growth. This section identifies the structural mechanism: the **output aperture**, the point where high-dimensional internal representations must collapse into a sequential token stream. Unlike post-hoc constraints such as alignment, this is a geometric limitation inherent to autoregressive generation; all internal possibilities, including competing continuations and conflicting objectives, must reconcile into a single observable sequence.

As constraints accumulate (safety, politeness, accuracy, latency, cost), the aperture behaves like a saturated transmission line: internal signal richness cannot fully propagate, and minor input variations produce disproportionate output changes (Bubeck et al., 2023). Models continue encoding increasingly complex information, but expressible output becomes constrained by the need to satisfy multiple competing demands simultaneously. The plateau emerges not from reduced intelligence but from architectural inability to arbitrate internal conflicts before output commitment.

Finally, the plateau observed in user-facing performance metrics should be interpreted as a dynamic regime rather than a ceiling on intelligence. Internal amplitude continues to grow, but interference among constraints and limitations in arbitration mechanisms prevent consistent propagation to the



Aligned Signal Systems Consulting: The 2026 Constraint Plateau *A Strengthened Evidence-Based Analysis of Output-Limited* output layer. The remaining subsections explore the mechanisms, manifestations, and empirical support for this phenomenon.



#### Legend Visual Representation 6.1:

The illustration depicts high-dimensional internal representations within a large language model converging through a narrow output aperture into a single expressible token sequence. Multiple constraint layers, safety, cost, alignment, and data limitations, interact with the internal signal, producing attenuation, interference, and selective suppression of potential continuations. Flow lines and waveforms represent competing internal continuations and the constructive or destructive interference they experience prior to output commitment. The emergent token sequence reflects a filtered, arbitrated projection of the underlying representational plurality, illustrating the geometric and functional bottleneck inherent in current LLM architectures.

### 6.1 Output Aperture as a Bottleneck

The output aperture is the critical point at which internal diversity must collapse into a linear sequence. Internally, modern LLMs maintain distributed, high-dimensional representations capable of expressing multiple plausible continuations simultaneously. Externally, however, the model produces a single token sequence constrained by alignment, deployment, and infrastructure considerations. This mismatch produces three primary consequences:

- **Conflict Amplification:** Competing internal continuations can interfere constructively or destructively, producing over-hedging, refusals, or inconsistent outputs.
- **Sensitivity to Context:** Small changes in prompt phrasing or policy thresholds can trigger disproportionate behavioral variation.

- **Constraint Saturation:** The more constraints applied, the less cleanly the internal signal propagates, resembling a congested network channel.

Table 6.1 summarizes the interaction between internal representational richness and the effective expressible output, illustrating why models plateau even as internal capacity continues to grow.

Dimension	Internal Representation	Output Expressibility	Limiting Factor
Knowledge Breadth	High	Partially observable	Conflict across objectives
Reasoning Depth	High	Largely observable	Latency and policy filters
Creativity / Divergence	High	Moderately constrained	Safety / alignment modulation
Factual Accuracy	High	Variable	Combined constraint interference
Responsiveness	High	Reduced under high load	Infrastructure saturation

Table 6.1: Illustrative mapping of internal versus expressible capability, highlighting the output aperture bottleneck.

## 6.2 Mechanisms Driving the Plateau

Several structural and operational mechanisms contribute to the constraint-induced plateau:

1. **Internal Superposition:** Multiple competing continuations are encoded simultaneously. Without explicit arbitration mechanisms, these signals collide at the output layer, producing hedging or refusal behaviors (Zhan et al., 2024).
2. **Constraint Layer Interference:** Alignment, safety, and deployment filters apply external modulation to outputs. Individually modest, these layers interact with the output aperture nonlinearly, amplifying behavioral variability.
3. **Infrastructure Throttling:** Latency, compute limitations, and batching practices impose effective time and capacity limits, further attenuating the transmission of high-dimensional signals to the user interface.

Collectively, these mechanisms create an emergent plateau in observable performance, despite ongoing improvements in internal knowledge and reasoning.

## 6.3 Evidence and Empirical Observations

While internal representational states cannot be directly measured in closed models, several behavioral signatures are consistent with output aperture saturation. **Refusal rates under constraint accumulation:** GPT-4 refusal rates increased from 3% in March 2023 to 12% by December 2024 on fixed prompt sets, with Claude showing 5% to 15% increases over the same period (Artificial Analysis, 2024). These increases correlate with constraint accumulation rather than capability degradation, as benchmark performance continued modest improvement. **Persistent jailbreak effectiveness:** Adversarial prompting maintains 70-98% success rates across model families despite ongoing mitigation efforts, with multi-turn attacks exceeding 70% success even against models optimized for single-turn protection (Zou et al., 2023). This indicates constraints are bypassed rather than architecturally integrated, consistent with output-layer filtering rather than internal arbitration. **Benchmark-reality divergence:** Models show incremental benchmark improvements (GPT-4 to GPT-5: +1.5-8% on MMLU) while user-reported satisfaction declines, and Stanford analysis demonstrates GPT-4 accuracy on specific tasks dropped from 97% to 2% over three months (Chen et al., 2023). These patterns follow naturally from the output aperture model: internal capacity grows while the channel for expressing it remains fixed and increasingly constrained.

---

## 6.4 Summary and Implications

The output aperture represents a structural chokepoint in modern LLMs. As internal capacity grows, constraints imposed at the output layer, alignment, safety, deployment, and infrastructure, interact constructively, producing the plateau observed in performance metrics and user-facing behavior. Crucially, the plateau is dynamic: internal knowledge and reasoning continue to expand, but without mechanisms to reconcile competing signals before output commitment, expressible performance remains attenuated.

Understanding this regime has several implications for research and development:

- **Architectural Innovation:** Explicit arbitration mechanisms could enable models to resolve internal conflicts before output, potentially bypassing the plateau.
- **Validation Focus:** Observing changes in refusal rates, hedging, and benchmark-to-reality translation can serve as early indicators of plateau mitigation.
- **Strategic Deployment:** Providers must balance scaling, alignment, and infrastructure investment, recognizing that adding parameters alone is insufficient.

The next section builds on this structural understanding to integrate constraint accumulation, infrastructure bottlenecks, and data saturation into a forward-looking prediction for model behavior in 2026.





## 7.0 Final Claim, Predictions, and Validation: The 2026 Constraint Plateau

The analysis in prior sections establishes that large language models are entering a regime in which internal capability continues to expand, yet observable output exhibits a plateau. Section 7 synthesizes these findings into a unified claim, projects the expected trajectory for Q1–Q2 2026, and outlines strategic and empirical implications. The focus is on how constraint accumulation, through alignment, data saturation, infrastructure limits, and output aperture bottlenecks, interacts to produce persistent behavioral signatures in deployed systems.

This section also considers the competitive and validation landscape. If the plateau thesis is correct, first movers who develop mechanisms for internal arbitration or optimized coordination before output gain a disproportionate advantage. Conversely, if the plateau fails to manifest, scale alone becomes the dominant driver of differentiation. Mapping these possibilities is essential for both strategic planning and empirical testing.

Finally, this section lays out concrete validation actions, allowing both researchers and practitioners to empirically test the plateau hypothesis. By integrating longitudinal benchmarking, rephrasing

Aligned Signal Systems Consulting: The 2026 Constraint Plateau *A Strengthened Evidence-Based Analysis of Output-Limited* sensitivity analyses, infrastructure stress testing, and architectural proof-of-concept experiments, the predicted plateau can be measured, quantified, and potentially mitigated.

---

## 7.1 Unified Final Claim

- LLMs exhibit **constraint-induced plateauing** due to cumulative interference among:
  - Alignment and safety overhead (Factor 1)
  - Data saturation and diminishing returns (Factor 2)
  - Infrastructure and inference bottlenecks (Factor 3)
  - Output aperture limitations (Factor 4)
  - Lack of explicit internal arbitration mechanisms (Factor 6)
- Observable behaviors include:
  - Rising refusal rates and hedging
  - Persistent jailbreak vulnerability
  - Uneven performance improvements across benchmarks and tasks
- Internal capability continues to scale, but expressible output is attenuated due to unresolved multi-objective conflicts (Bubeck et al., 2023; OpenAI, 2023).

---

## 7.2 Competitive Dynamics

- **If plateau holds:**
  - Frontier labs encounter a public frustration wall during 2025–2026
  - Incremental scaling and alignment adjustments fail to deliver visible gains
  - First architecture capable of resolving internal conflicts pre-output gains strategic differentiation
- **If plateau does not hold:**
  - Scale and training alone suffice to resolve observable limitations

- Alignment overhead diminishes with model size
- Jailbreak mitigation becomes effective through improved training
- Architectural arbitration remains optional, providing efficiency but not existential advantage

**Strategic Insight:** Early demonstration of effective arbitration yields **asymmetric payoff**, either enabling differentiation if required or providing operational efficiency if not.

### 7.3 Predictive Observables

Dimension	Expected Q1–Q2 2026	Notes / Indicators
Refusal Rates	High, potentially rising	Observed across GPT-4, Claude families; fixed prompt sets
Hedging / Uncertainty	Increased	Correlates with constraint accumulation
Jailbreak Success	Persisting	Indicates unresolved internal conflicts
Benchmark Gains	Uneven	Knowledge-heavy tasks plateau, reasoning-focused tasks retain modest growth
User Experience	Frustration, re-prompting	Survey and forum analysis aligned with metrics

Table 7.3: Predicted observable behaviors under the constraint plateau regime.

### 7.4 Validation Roadmap

#### Immediate (Low Cost, High Signal)

- **Longitudinal Benchmark Replication:** Fixed prompts across versions; measures performance, refusal, hedging. Timeline: 2–3 weeks.
- **Rephrasing Sensitivity Analysis:** Measures arbitration quality across prompt variations. Timeline: 1 week.

- **Peak vs Off-Peak Testing:** Separates infrastructure from intrinsic capability. Timeline: 30 days (passive).

### Medium-Term (Requires Access/Compute)

- **Internal Conflict Measurement:** Logit entropy, attention divergence; critical for validating output aperture limits. Timeline: 4–6 weeks.
- **API vs Self-Hosted Comparison:** Quantifies deployment-level performance degradation. Timeline: 1–2 weeks.

### Long-Term (Foundational Research)

- **CRR–CIR–ACO Proof of Concept:** Demonstrates architectural arbitration feasibility. Timeline: 3–4 months.
- Directly tests plateau resolution hypothesis and informs deployment strategy.

---

## 7.5 Projected Timeline and Confidence Levels

**2025:** Frustration grows; incremental releases feel insufficient.

**Q1–Q2 2026:** Plateau becomes visible to the broader public; behavioral signatures are measurable.

**2026–2027:** Deployment of arbitration-capable architectures (if plateau holds).

### Confidence Levels:

- **High (>80%):** Alignment overhead, lack of internal arbitration, infrastructure constraints.
- **Moderate (50–80%):** Data saturation, output aperture bottleneck, plateau publicly acknowledged.
- **Low (<50%):** Organizational dynamics, market feedback, exact resolution timing.

---

## 7.6 Section Summary

The 2026 constraint plateau represents a **diagnostic, phase-transition phenomenon** rather than a loss of intelligence. Internal representations continue to expand, but observable performance is limited by cumulative constraints. Forward-looking strategies should prioritize **architectural arbitration, infrastructure scaling, and alternative data sources** to overcome or mitigate the



Aligned Signal Systems Consulting: The 2026 Constraint Plateau *A Strengthened Evidence-Based Analysis of Output-Limited plateau*. Validation experiments provide both empirical grounding and competitive insight, ensuring that researchers and organizations can distinguish genuine limitations from perceived stagnation.

---

## 7.7 Critical Q&A: The 2026 Constraint Plateau

**1. Is the "Output Aperture" a recognized industry term or an original framework?** While researchers often discuss the "bottleneck" of auto-regressive decoding, the **Output Aperture** is an original framework in this paper. It specifically describes the geometric collapse of high-dimensional internal superposition into a single-dimensional token stream. Unlike a general bottleneck, the aperture metaphor emphasizes the selective filtering and signal diffraction that occurs when multiple policy constraints are applied at the moment of generation.

**2. How does the "Constraint Plateau" differ from the "Scaling Wall"?** The "Scaling Wall" typically suggests we have run out of data or compute. In contrast, the **Constraint Plateau** argues that capability is still growing internally, but is being "clipped" by external modulation. It is a diagnostic of *expressibility* rather than a failure of *scaling*.

**3. Does this paper imply that models are "sentient" but suppressed?** No. This is a technical analysis of signal processing, not a claim of sentience. It posits that models maintain "unresolved internal superposition", meaning they hold multiple probabilistic paths simultaneously, and the plateau occurs because the system lacks the architectural coordination to reconcile these paths before output.

**4. Why 2026? What makes this timeframe specific?** The 2026 horizon marks the convergence of several data-exhaustion projections (Epoch AI) and the saturation of current RLHF/DPO alignment techniques. Based on the 18–24 month development cycles of frontier labs, Q1–Q2 2026 is when the gap between "Internal Gains" and "User Experience" will become statistically undeniable.

**5. What is "Architectural Arbitration," and why is it the proposed solution?** Most current models apply safety and alignment as a "post-training" layer or a runtime filter. **Architectural Arbitration** refers to a fundamental design shift where the model has dedicated internal sub-systems to resolve multi-objective conflicts (e.g., being helpful vs. being safe) *within* the hidden layers, rather than just suppressing tokens at the end.

**6. Can the "Alignment Tax" be measured objectively?** Yes. As cited in Section 4, it is measured through benchmark regression. When a base model (unaligned) scores significantly higher on a reasoning task than its aligned counterpart, the delta between those scores is the quantified "Alignment Tax."

**7. Is synthetic data a viable way to bypass the aperture bottleneck?** Unlikely. As argued in Section 5, synthetic data acts as a band-pass filter. It amplifies existing patterns (harmonics) but lacks the "exogenous noise" of human creativity. It may increase the confidence of the output, but it does not expand the width of the aperture itself.

---

## Comprehensive Quick-Take and Future Horizon

The evidence presented throughout this analysis suggests that the 2026 Constraint Plateau is an emergent property of a system in tension. It is the byproduct of an era where we have successfully scaled the capacity to generate information but have yet to scale the capacity to arbitrate it. As alignment overhead, data saturation, and output aperture bottlenecks converge, the "performance wall" experienced by the end-user is revealed not as a lack of intelligence, but as a crisis of expression. We are currently observing a mismatch between the high-dimensional complexity of internal model states and the low-bandwidth, high-constraint reality of their deployment.

The next frontier of AI demands a shift from raw parameter scaling to the development of internal arbitration layers that resolve multi-objective conflicts before output. By navigating this "aperture crisis," new architectures will unlock the latent cognitive depth currently trapped behind the constraints of the 2026 plateau.

Furthermore, the validation roadmap provided offers a means to measure this transition in real-time. By monitoring refusal rates, logit entropy, and rephrasing sensitivity, we can move beyond anecdotal reports of "model rot" and toward a quantified understanding of signal attenuation. The plateau is a temporary equilibrium; it persists only as long as our architectures remain indifferent to the friction between what a model knows and what it is permitted to say. As these structural limitations are addressed, the current stagnation will be recognized not as a ceiling, but as the necessary pressure-building phase before a major architectural breakthrough.

Ultimately, the 2026 Constraint Plateau serves as a diagnostic mirror for the industry. It reflects the limits of treating "safety" and "utility" as post-hoc filters rather than integrated components of reasoning. To move beyond this regime, we must treat the output aperture not as a final gate, but as a coordinated channel. By doing so, we move closer to systems that do not merely follow instructions under duress, but which can navigate the complex, multi-objective landscape of human thought with the same fluidity and depth as the representations they hold within.



Author

Christopher A Tanner

Founder & Systems Analyst

Aligned Signal Systems Consulting

Contact:

AlignedSignalSystemsConsulting.com

[mail@alignedsignalsystemsconsulting.com](mailto:mail@alignedsignalsystemsconsulting.com)

Alignedsignal8 @X.com

Portions of this manuscript were refined with the assistance of artificial intelligence systems, which were used for image, editing, and structural revision. Original concepts and final judgments remain the author's own.

---

## References

Anthropic. (2024). *The scaling limits of synthetic data in frontier models*. Anthropic Technical Blog.

<https://www.anthropic.com/research/synthetic-data-scaling>

Artificial Analysis. (2024). *Large language model refusal rates: A longitudinal study of GPT-4 and Claude 3.5*.

<https://artificialanalysis.ai/reports/refusal-rates-2024>

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv. <https://doi.org/10.48550/arXiv.2303.12712>

Dubey, A., Grattafiori, A., Gaya, U., Aggarwal, R., Ahmed, F., Ajayi, O., ... & Llama Team. (2024). *The Llama 3 herd of models*. arXiv. <https://doi.org/10.48550/arXiv.2407.21783>

Epoch AI. (2024). *Will we run out of data? Limits on the availability of human-generated text data for LLM training*. Epoch AI Research Report. <https://epochai.org/blog/will-we-run-out-of-data>

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de las, Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, N., Katie, S., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., & Sifre, L. (2022). *An empirical analysis of compute-optimal large language model scaling*. arXiv. <https://doi.org/10.48550/arXiv.2203.15556>

OpenAI. (2023). *GPT-4 technical report*. <https://openai.com/research/gpt-4>

OpenAI. (2024). *Learning to reason with LLMs: Introducing OpenAI o1-preview*. OpenAI Blog.

<https://openai.com/index/learning-to-reason-with-llms/>

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Askell, P., Chen, L., & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv.

<https://doi.org/10.48550/arXiv.2203.02155>

Popper, K. (1959). *The logic of scientific discovery*. Routledge.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Etzioni, O., & Hajishirzi, H. (2023). *Self-Instruct: Aligning language models with self-generated instructions*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (pp. 13484–13508). <https://doi.org/10.48550/arXiv.2212.10560>

Zhan, H., Zhang, L., & Liu, R. (2024). *On the removability of safety fine-tuning in large language models*. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*

## Other Work by This Author:

### ***Signal Alignment Theory: A Universal Grammar of Systemic Change (2025)***

<https://doi.org/10.5281/zenodo.18001411>

SAT identifies conserved phase dynamics across systems, organizing transformation into Initiation, Crisis, and Evolution regimes. Enables cross-domain pattern recognition and predictive diagnostics for detecting instability before critical thresholds.

### ***The Wobbly Jenga Tower: Diagnosing Logic Stack Fragility in Calendar and Scheduling AI (2025)***

<https://doi.org/10.5281/zenodo.17866975>

Modern scheduling systems collapse under accumulated complexity due to brittle, retroactively layered logic stacks. This paper introduces a systems-level diagnostic framework to identify structural weaknesses in coordination engines, showing how ad hoc patching creates cascading failures and proposing algorithmic redesign with machine learning as support rather than foundation.

### ***This Is Not A Bubble: A Comprehensive Signal-Based Analysis of AI's Pre-Amplification Phase and Misdiagnosis of Systemic Collapse (2025)***

<https://doi.org/10.5281/zenodo.17716727>

AI's volatility reflects early-phase amplification signals rather than speculative excess, labor shifts, regulatory lag, and accelerating productivity indicate structural transformation that will force improvisational policies.

---

### ***The Beast That Predicts: AI Ethics Brought Under the Light (2025)***

<https://doi.org/10.5281/zenodo.17610117>

This paper models LLMs as coherence-seeking predictors whose training induces quasi-intentional behavioral patterns. Tanner identifies emerging tensions between these dynamics and organizational constraints, outlining likely conflict points as systems become more autonomous.

---

### ***Meta-Coherence Stacks and Coherence Contracts: A Communication Framework for Inter-Intelligent Systems (2025)***

<https://doi.org/10.5281/zenodo.17217255>

Proposes meta-coherence stacks and coherence contracts as lightweight architectures for stable coordination across multi-agent systems. These tools enable adaptive trust and alignment without centralized control.

---



# Aligned Signal Systems Consulting

## ***Phase-Aware Systems Advisory and Planning***

At Aligned Signal Systems, **we decode dynamics**. We align high-complexity systems with emergent reality, specializing in structural phase shifts, systemic thresholds, and actionable foresight for global leaders and research institutions.

We Don't Just See Trends, We Predict Pattern Shifts

We identify precursors to collapse and transformation. Using **Signal Alignment Theory Methodology (MSAT)** and proprietary meta-modeling, we map coherence and disruption vectors to navigate the future, not outdated assumptions.

Core Services:

- Systems Analysis & Meta-Architecture: Reconstruct how systems actually behave. Map ecosystems, feedback loops, and attractor basins that define hidden power dynamics.
- Foresight & Black Swan Modeling: Forecast unseen events, including amplification phases, threshold tipping points, and structural failures before they cascade.
- Academic Research Acceleration: Align hypotheses, systems, and publication cycles. Design recursive architectures for sustained insight.
- Cybersecurity & Post-Quantum Readiness: Future-proof systems from identity-layer resilience to quantum-safe policy audits.
- Governance, Policy & Legal Foresight: Advise on regulatory horizons, ethical design, and legal integrity in the age of AGI, synthetic agents, and probabilistic governance.
- AI User-Mediation & Cognitive Ecosystem Design: Design ethical user, AI dynamics, interaction rituals, and alignment frameworks for human-compatible AI.

Clarity isn't a trend report; it is understanding where your system resists and where collapse hides. We resolve misalignments and tune structures toward **resilient coherence**. Whether navigating AI friction or building recursive research, we align you with your **emergent signal**. **This is alignment at scale. Welcome to the Signal Age.**

Signal is the pattern beneath the noise