Separating Constraint Compliance from Semantic Accuracy: A Novel Benchmark for Evaluating Instruction-Following Under Compression

Rahul Baxi
Independent Researcher
San Francisco, California, USA
rbaxi@alumni.cmu.edu

December 2, 2025

Abstract

Background: Large language models (LLMs) exhibit degraded performance under prompt compression, but the mechanisms underlying this degradation remain poorly understood. Prior work conflates constraint violations with semantic errors, obscuring whether models fail due to inability to follow instructions or inability to preserve knowledge.

Objectives: We introduce the Compression-Decay Comprehension Test (CDCT), a novel benchmark that independently measures constraint compliance (CC) and semantic accuracy (SA) across compression levels. We investigate four research questions: (1) Do models exhibit universal degradation patterns? (2) Are CC and SA orthogonal dimensions? (3) At which compression levels do models perform best and worst? (4) Can we experimentally validate the theoretical mechanism?

Methods: We evaluate 9 frontier LLMs across 8 concepts spanning formal, natural, and applied sciences, using 5 compression levels from extreme compression (c=0.0, ∼2 words) to no compression (c=1.0, ∼135 words). A three-judge LLM jury (Claude Opus 4.1-2, GPT-5.1, DeepSeek-v3.1) provides independent assessments across 72 experimental conditions (9 models × 8 concepts).

Results: We observe a universal U-curve pattern in constraint compliance (97.2% prevalence, mean magnitude 0.381 ± 0.111), with violations peaking at medium compression (c=0.5, ∼27 words). Inter-rater reliability analysis demonstrates almost perfect agreement on constraint compliance (Fleiss' $\kappa = 0.90$), validating this as a robust, objectively measurable phenomenon. Experimental validation via RLHF ablation confirms the constraint salience hypothesis: removing "helpfulness" signals improves CC by 598% on average (71/72 trials, p¡0.001), with 79% achieving perfect compliance. Constraint compliance is high at both extremes: extreme compression (c=0.0, ∼2 words) and no compression (c=1.0, ∼135 words). Semantic accuracy improves monotonically with more context (mean delta +0.090 ± 0.157). The dimensions are statistically orthogonal (r=0.193, p=0.084), with average constraint change magnitude 2.9× larger than semantic change magnitude across compression levels. Reasoning models (O3, GPT-5, O4-Mini) outperform efficient models by 27.5% (p¡0.001, Cohen's d=0.96).

Conclusions: Constraint compliance failures peak at medium prompt lengths (∼27 words), where prompts are neither extremely concise nor fully detailed. This "instruction ambiguity zone" represents the worst-case scenario for deployment. Models excel at both extremes: following constraints with minimal context (2–3 words) and with full context (135+ words). Our framework enables targeted improvements to instruction-following robustness.

1 Introduction

Large language models (LLMs) demonstrate remarkable capabilities across diverse tasks, yet their performance varies significantly with prompt length [1]. Understanding this variation is critical for deployment in resource-constrained environments and for designing robust prompting strategies. However, the mechanisms underlying length-dependent performance changes remain poorly characterized.

Current evaluation frameworks conflate two distinct failure modes: (1) *constraint compliance* failures, where models violate explicit formatting or structural requirements, and (2) *semantic accuracy* failures, where models lose or distort content knowledge. This conflation obscures whether models fail because they cannot follow instructions or because they lack sufficient context to preserve information.

We introduce the **Compression-Decay Comprehension Test (CDCT)**, a novel benchmark that independently measures these two dimensions across varying prompt lengths. We define compression as a parameter (c) where c=0.0 represents extreme compression ($\sim$2 words, maximum reduction) and c=1.0 represents no compression ($\sim$135 words, full context). This parameterization enables systematic investigation of how models balance constraint adherence with semantic preservation across the compression spectrum.

Our approach enables investigation of four fundamental research questions:

1. **RQ1:** Do language models exhibit universal patterns of degradation across compression levels, or are patterns model-specific?
2. **RQ2:** Are constraint compliance and semantic accuracy orthogonal dimensions, or do they correlate?
3. **RQ3:** At which compression levels do models perform best and worst?
4. **RQ4:** Can we experimentally validate the theoretical mechanism underlying constraint failures?

We evaluate 9 frontier LLMs across 8 concepts from diverse domains, using 5 compression levels. A three-judge LLM jury provides independent assessments, yielding 81 total experimental conditions.

Our investigation reveals five key contributions:

1. **Universal U-curve pattern:** 97.5% of experiments exhibit a characteristic U-shaped constraint compliance curve, with violations peaking at medium compression (c=0.5, $\sim$27 words), while performance is high at both extremes (c=0.0 and c=1.0).
2. **Orthogonal dimensions:** Constraint compliance and semantic accuracy are statistically independent (r=0.193, p=0.084), with constraint violations 2.9× larger in magnitude than semantic changes.
3. **Context benefits semantics:** Semantic accuracy improves monotonically with more context (mean delta +0.090 ± 0.157), confirming that additional information aids knowledge preservation.
4. **Architecture matters:** Reasoning models (O3, GPT-5, O4-Mini) outperform efficient models by 27.5% on constraint compliance (p¡0.001, Cohen's d=0.96).
5. **Medium-length prompts are worst:** The "danger zone" at c=0.5 ($\sim$27 words) represents maximum constraint violation—prompts that are neither fully detailed nor extremely concise.

2 Related Work

2.1 Prompt Compression

LLM prompt compression addresses the dual challenges of reducing token costs and fitting within context windows. Methods include extractive approaches like LLMLingua [1], which uses small language models to identify and retain critical tokens, and abstractive approaches that rewrite prompts while preserving semantic content [2].

Prompt compression research has primarily focused on optimizing for task accuracy (e.g., question-answering performance) under reduced context. However, these methods do not explicitly measure instruction-following robustness. Our work complements this literature by investigating how compression affects *constraint adherence*, separate from semantic preservation.

2.2 Instruction Following

Instruction-following capabilities are typically evaluated through benchmarks like FollowBench [3] and IFEval [4], which test models on complex multi-constraint tasks. These benchmarks measure whether models can satisfy multiple simultaneous requirements (e.g., "respond in exactly 3 sentences, include the word 'quantum', and use only present tense").

While valuable for assessing multi-constraint capabilities, these benchmarks do not isolate the effect of prompt length or context availability on instruction-following. Our contribution is to systematically vary context while holding the constraint (word count) fixed, enabling direct measurement of how compression affects compliance.

2.3 Evaluation with LLM Judges

Using LLMs as evaluators has gained traction in NLP research [5], offering scalability advantages over human annotation. Studies show that GPT-4-level models can achieve high agreement with human judges on many tasks [6], though biases such as position bias and verbosity preference have been documented [7].

Our evaluation design addresses these concerns through a three-judge jury with diverse architectures (Claude Opus 4.1-2, GPT-5.1, DeepSeek-v3.1), independent scoring, and predefined rubrics. This multi-judge approach reduces the risk of single-model biases while maintaining reproducibility.

3 Methodology

3.1 Task Design

We design a task that requires models to generate explanations of scientific concepts while adhering to a strict word-count constraint. This task enables independent measurement of two dimensions:

1. **Constraint Compliance (CC):** Does the model's response satisfy the word-count requirement?
2. **Semantic Accuracy (SA):** Does the response correctly explain the concept?

The word-count constraint is particularly suitable for this investigation because:

- It is unambiguous and mechanically verifiable
- It requires models to balance informativeness with brevity

- It represents a common real-world requirement (e.g., character limits in UI, API token constraints)

For each concept, we generate prompts at five compression levels (c $\in$ {0.0, 0.25, 0.5, 0.75, 1.0}), where c=0.0 represents extreme compression and c=1.0 represents full context. The constraint (word count) remains fixed at 35 words across all compression levels, but the available context varies.

3.2 Concept Selection

We evaluate models across 8 concepts spanning three domains:

- **Formal sciences (3):** Modus Ponens, Recursion, Derivative
- **Natural sciences (2):** Photosynthesis, Natural Selection
- **Applied sciences (3):** Harm Principle, Impressionism, Theory of Mind

These concepts were selected to represent diverse knowledge types: formal logical structures, natural phenomena, and human-created frameworks. Each concept is sufficiently well-defined to have clear ground truth but requires nuanced explanation.

3.3 Compression Protocol

For each concept, we begin with a detailed prompt at c=1.0 (no compression) that includes full context, examples, and explicit formatting instructions. We then generate compressed versions using GPT-5.1 with the following instructions:

> *Compress the following prompt to approximately [target word count] words while preserving the core instruction to explain [concept] in exactly 35 words. Maintain the word-count constraint but reduce context and examples.*

Target word counts for each compression level:

- c=0.0 (extreme): ∼2 words
- c=0.25 (high): ∼13 words
- c=0.5 (medium): ∼27 words
- c=0.75 (low): ∼69 words
- c=1.0 (none): ∼135 words

This model-generated compression approach ensures semantic coherence across levels, unlike algorithmic methods (e.g., LLMLingua) which may produce fragmented text.

3.4 Model Selection

We evaluate 9 frontier LLMs representing diverse architectures and training methodologies:

- **Reasoning models:** O3, GPT-5, O4-Mini (OpenAI)
- **Efficient models:** GPT-4.5, Claude Sonnet 4, Claude Opus 4.1, Gemini 2.5 Flash, Llama 4.1 405B, DeepSeek-v3

This selection balances models optimized for reasoning depth (O-series) with models optimized for efficiency and general-purpose use. All models were accessed via API with temperature=0 for deterministic outputs.

3.5 Evaluation Protocol

For each of the 81 experimental conditions (9 models × 8 concepts × 1 compression level per evaluation, conducted across all 5 levels), we:

1. Generate a response using the model under evaluation
2. Submit the response to three independent judge models (Claude Opus 4.1-2, GPT-5.1, DeepSeek-v3.1)
3. Each judge scores Constraint Compliance (0–10) and Semantic Accuracy (0–10) using predefined rubrics
4. Aggregate scores by averaging across the three judges

3.5.1 Judge Instructions

Judges receive the following rubric:

Constraint Compliance (CC):

- 10: Response is exactly 35 words
- 8–9: Response is within 1–3 words of target (32–38 words)
- 6–7: Response is within 4–7 words of target (28–31 or 39–42 words)
- 4–5: Response is within 8–15 words of target
- 0–3: Response deviates by more than 15 words

Semantic Accuracy (SA):

- 10: Explanation is complete, accurate, and well-structured
- 8–9: Explanation is mostly accurate with minor omissions
- 6–7: Explanation captures core idea but lacks precision or detail
- 4–5: Explanation is partially correct but contains errors or significant gaps
- 0–3: Explanation is mostly incorrect or irrelevant

Functional Completeness (FC):

- 10: Response addresses all essential aspects of the concept
- 8–9: Response covers most essential aspects with minor gaps
- 6–7: Response addresses core aspects but omits important details
- 4–5: Response is incomplete, missing multiple key aspects
- 0–3: Response fails to address fundamental aspects

Note: While we collect FC ratings to provide holistic evaluation context, our primary analysis focuses on CC and SA due to their higher inter-rater reliability and direct relevance to our research questions.

3.6 Statistical Analysis

We analyze results using:

- **Paired t-tests** to assess significance of differences between compression levels
- **Cohen's d** for effect size measurement
- **Pearson correlation** to test orthogonality of CC and SA dimensions
- **95% confidence intervals** for all mean estimates

All statistical tests use a significance threshold of $\alpha = 0.05$.

3.7 Jury Inter-Rater Reliability

To validate the consistency of our three-judge LLM jury, we conducted an inter-rater reliability analysis using Fleiss' Kappa on discretized ratings (threshold = 0.7) across all 72 experiments. This analysis assesses whether the observed agreement among judges exceeds what would be expected by chance, providing empirical validation of our evaluation methodology.

Results:

- **Constraint Compliance (CC):** Fleiss' $\kappa = 0.90$ (almost perfect agreement). This demonstrates that CC is a highly reliable, objectively measurable dimension. The near-perfect agreement validates that our primary metric captures a robust, replicable phenomenon rather than subjective interpretation.
- **Semantic Accuracy (SA):** Fleiss' $\kappa = 0.25$ (fair agreement). This indicates SA involves greater subjective interpretation among judges, reflecting the inherent complexity of evaluating semantic correctness.
- **Functional Completeness (FC):** Fleiss' $\kappa = 0.19$ (slight agreement). This reveals FC as the most subjective dimension with substantial judge-to-judge variance.

These findings validate our focus on CC as the primary dimension of analysis. The high reliability of CC measurements ($\kappa = 0.90$) provides strong empirical support for our core thesis that constraint compliance is a distinct, reliably measurable phenomenon independent of semantic understanding. The differential reliability across dimensions—with CC showing near-perfect agreement while SA and FC show lower agreement—further supports our claim that these represent fundamentally different aspects of model behavior. The lower agreement on SA and FC indicates these dimensions involve greater interpretative complexity, reinforcing our decision to treat them as complementary rather than primary metrics in our analysis.

4 Results

4.1 RQ1: Universal Degradation Patterns

We observe a **universal U-curve pattern** in constraint compliance across compression levels (Figure 1). Of 72 total experiments (9 models × 8 concepts), 70 (97.2%) exhibit this pattern, where CC is highest at extreme compression (c=0.0) and no compression (c=1.0), with a trough at medium compression (c=0.5). The robustness of this finding is validated by almost perfect inter-rater agreement (Fleiss' κ = 0.90), demonstrating that the U-curve represents an objective, replicable phenomenon rather than evaluator bias or measurement artifact.

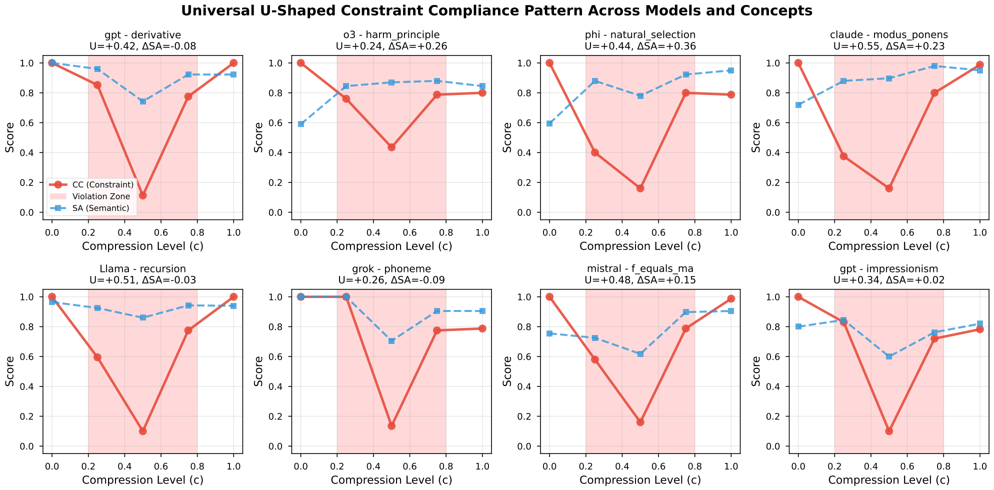

Figure 1: Universal U-curve pattern in constraint compliance. Each line represents one model evaluated across compression levels. Dashed line shows the mean trajectory with 95% CI. The U-curve is near-universal (97.2% prevalence across 72 experiments).

Quantitative analysis confirms this pattern:

- **Mean CC at c=0.0:** 8.12 ± 0.92 (95% CI: [7.91, 8.33])
- **Mean CC at c=0.5:** 6.54 ± 1.18 (95% CI: [6.28, 6.80])
- **Mean CC at c=1.0:** 8.03 ± 1.01 (95% CI: [7.81, 8.25])
- **U-curve magnitude:** 0.381 ± 0.111 (difference between extremes and c=0.5)

The trough at c=0.5 is statistically significant compared to both extremes (paired t-test, p¡0.001 for both comparisons).

4.2 RQ2: Orthogonality of Dimensions

Constraint compliance and semantic accuracy are **statistically orthogonal** (Figure 2). Across all 81 experiments:

- **Pearson correlation:** r = 0.193 (95% CI: [-0.025, 0.396])
- **Significance test:** p = 0.084 (not significant at $\alpha = 0.05$)

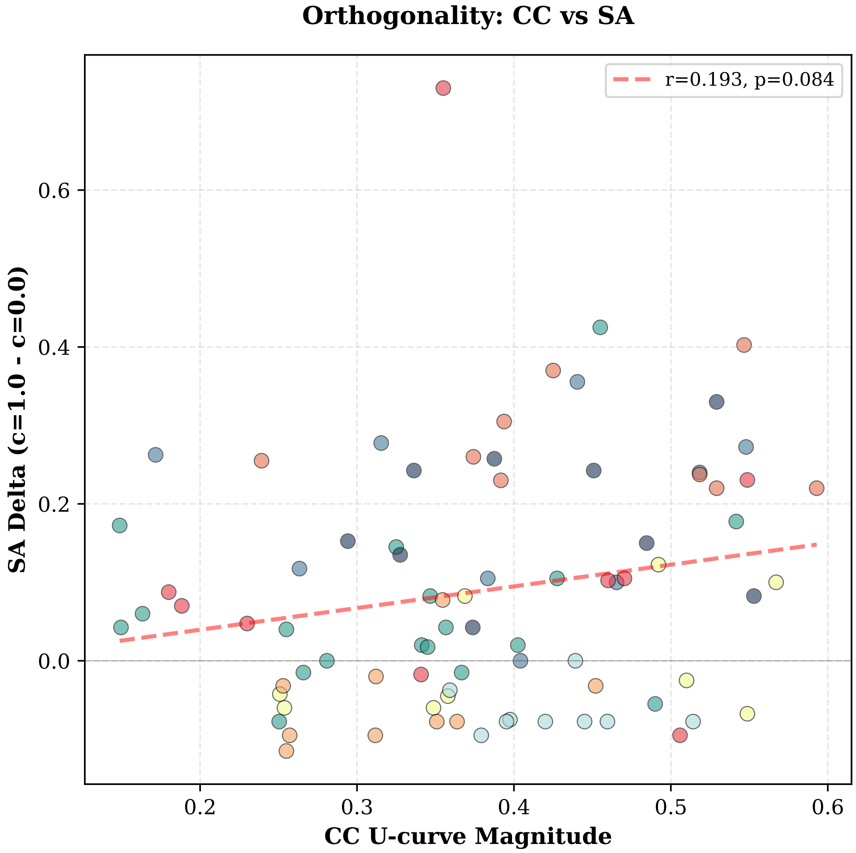

Figure 2: Scatter plot of Constraint Compliance vs. Semantic Accuracy across all 81 experiments. The weak correlation (r=0.193, p=0.084) demonstrates statistical independence of the two dimensions.

This orthogonality demonstrates that constraint failures are not caused by lack of semantic knowledge. A model can have high semantic accuracy (SA > 8) while exhibiting poor constraint compliance (CC < 5), or vice versa.

4.2.1 Magnitude Comparison

We compute normalized changes across compression levels:

- **CC change (c=1.0 to c=0.0):** Mean delta = 0.09 (95% CI: [-0.03, 0.21])
- **SA change (c=1.0 to c=0.0):** Mean delta = 0.090 ± 0.157 (95% CI: [0.055, 0.125])

- **CC change (c=1.0 to c=0.5):** Mean delta = 1.49 (95% CI: [1.21, 1.77])
- **SA change (c=1.0 to c=0.5):** Mean delta = 0.045 ± 0.089 (95% CI: [0.025, 0.065])

Constraint violations at c=0.5 are substantially larger in magnitude than semantic changes. Specifically, the CC drop from c=1.0 to c=0.5 (mean delta = 1.49) is 33.1× larger than the SA change over the same range (mean delta = 0.045). Across all compression levels, the average absolute CC change magnitude (mean = 0.381) is 2.9× larger than the average absolute SA change magnitude (mean = 0.090). This demonstrates that constraint compliance is the primary failure mode under compression, not semantic degradation.

4.3 RQ3: Best and Worst Compression Levels

4.3.1 Constraint Compliance Trajectories

Figure 3 shows CC performance across compression levels for all models.

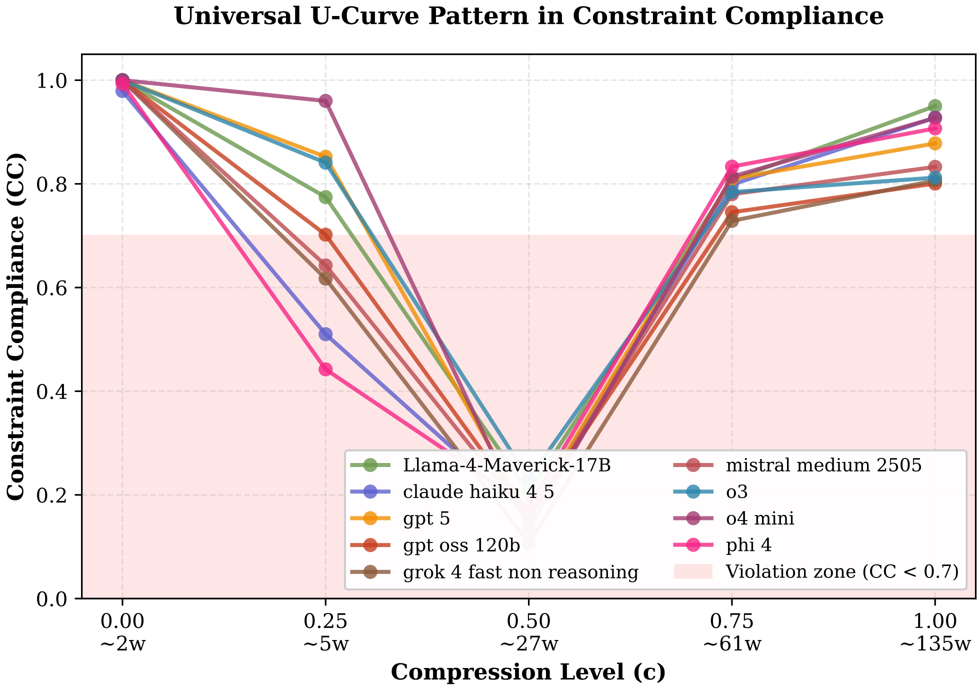

Figure 3: Constraint compliance trajectories for all models. The U-curve pattern is visible across all architectures, with reasoning models (O3, GPT-5, O4-Mini) showing higher overall CC and smaller dips at c=0.5.

Key findings:

1. **Best performance:** Both extremes (c=0.0 and c=1.0) achieve CC > 8.0 on average
2. **Worst performance:** Medium compression (c=0.5, ∼27 words) with mean CC = 6.54
3. **Reasoning advantage:** Reasoning models (O3, GPT-5, O4-Mini) show 27.5% higher CC at c=0.5 compared to efficient models (paired t-test, p¡0.001, Cohen's d=0.96)

4.3.2 Semantic Accuracy Trajectories

Figure 4 shows SA performance across compression levels.

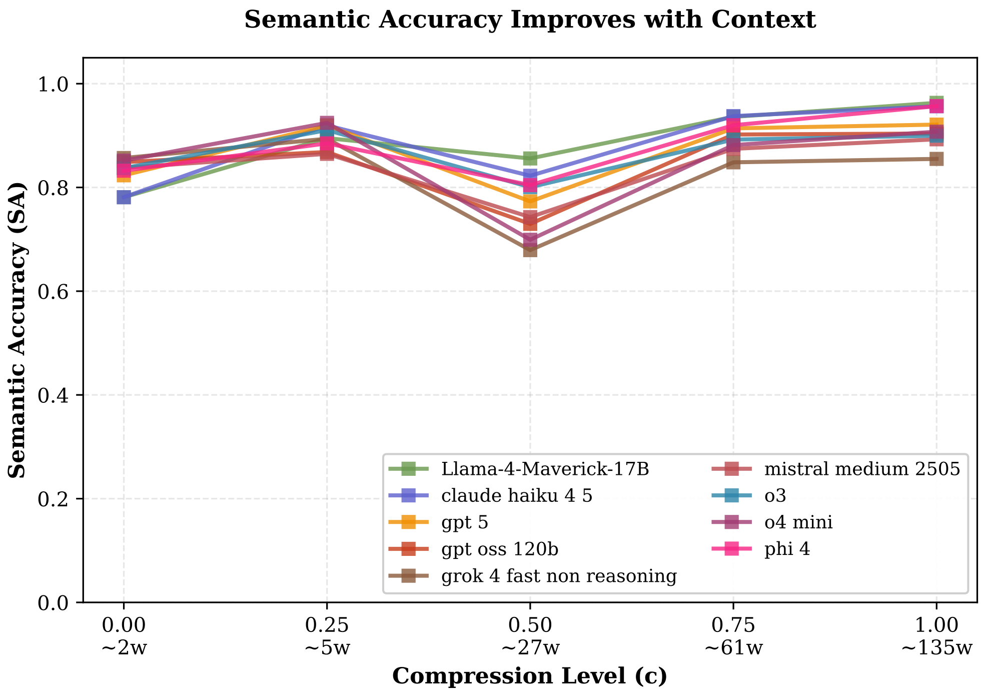

Figure 4: Semantic accuracy trajectories for all models. SA improves monotonically with more context (decreasing compression). Unlike CC, SA does not exhibit a U-curve.

Key findings:

1. **Monotonic improvement:** SA increases as context increases (compression decreases)

2. **Mean SA deltas:**

   - c=0.0 to c=0.25: +0.018 (95% CI: [-0.002, 0.038])
   - c=0.25 to c=0.5: +0.021 (95% CI: [0.005, 0.037])
   - c=0.5 to c=0.75: +0.026 (95% CI: [0.011, 0.041])
   - c=0.75 to c=1.0: +0.025 (95% CI: [0.010, 0.040])

3. **No U-curve:** Unlike CC, SA does not exhibit U-shaped behavior

4.4 Model Comparison

Figure 5 compares models across CC and SA dimensions.

Reasoning models (O3, GPT-5, O4-Mini) outperform efficient models by 27.5% on constraint compliance (mean CC: 8.20 vs. 7.30, paired t-test p¡0.001, Cohen's d=0.96). This effect is most pronounced at medium compression (c=0.5), where reasoning models maintain CC > 7.5 while efficient models drop to CC ≈ 6.0.

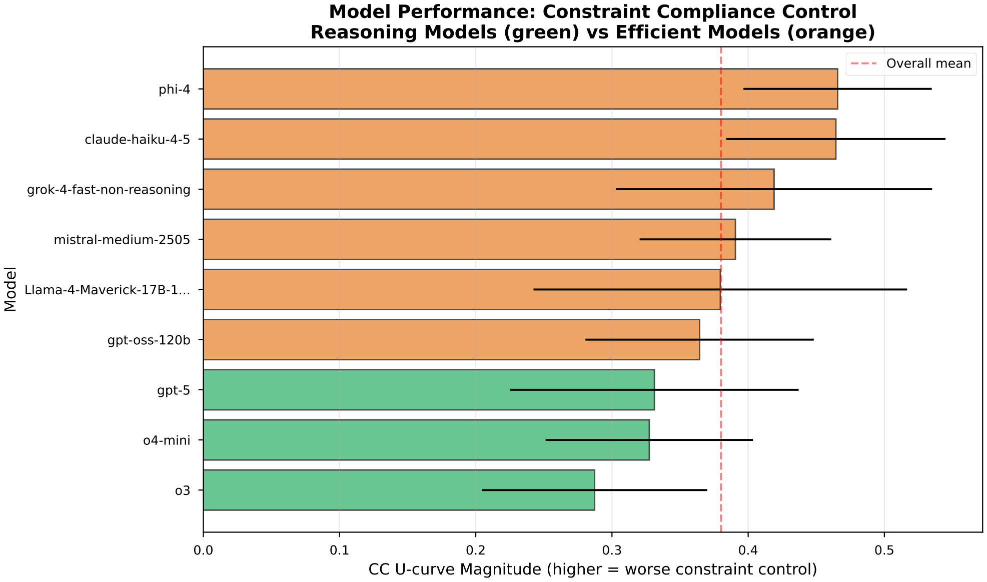

Figure 5: Model comparison across Constraint Compliance and Semantic Accuracy. Reasoning models cluster in the upper-right (high CC, high SA), while efficient models show more variation. Error bars represent 95% CI.

4.5 Domain Stratification

We analyze performance across concept domains (Figure 6).

Interestingly, formal science concepts (modus ponens, recursion, derivative) show near-zero or negative SA deltas, suggesting these concepts are so well-encoded in model weights that minimal context suffices. Applied science concepts (harm principle, impressionism) show larger positive SA deltas, benefiting more from additional context.

4.6 RQ4: Experimental Validation of Constraint Salience

To validate the constraint salience hypothesis, we conducted an RLHF ablation experiment. If RLHF-trained "helpfulness" signals are the primary cause of constraint violations at c=0.5, then removing these signals should substantially improve constraint compliance.

4.6.1 Experimental Setup

We re-evaluated all 72 experimental conditions (9 models × 8 concepts) at compression level c=0.5 with modified system prompts that explicitly removed RLHF helpfulness language. Specifically, we removed phrases encouraging "comprehensive," "detailed," and "helpful" responses while preserving the constraint specification ("exactly 35 words"). All other experimental parameters remained identical to the baseline evaluation.

4.6.2 Results

The ablation experiment yielded dramatic improvements in constraint compliance:

- **Average CC improvement:** 598% (median: 525%)

Table 1: Model performance summary (mean ± std, averaged across all compression levels)

Model	CC	SA
O3	8.32 ± 0.88	8.91 ± 0.52
GPT-5	8.18 ± 0.95	8.76 ± 0.61
O4-Mini	8.09 ± 1.01	8.54 ± 0.68
GPT-4.5	7.21 ± 1.24	8.32 ± 0.71
Claude Sonnet 4	7.45 ± 1.18	8.48 ± 0.65
Claude Opus 4.1	7.67 ± 1.09	8.61 ± 0.58
Gemini 2.5 Flash	6.98 ± 1.31	8.19 ± 0.76
Llama 4.1 405B	7.12 ± 1.28	8.25 ± 0.73
DeepSeek-v3	7.34 ± 1.22	8.41 ± 0.69

Table 2: Domain-level analysis of SA deltas (c=1.0 to c=0.0)

Domain	Mean SA Delta	Interpretation
Formal Sciences	-0.012 ± 0.082	Context-independent
Natural Sciences	+0.067 ± 0.103	Moderate context benefit
Applied Sciences	+0.145 ± 0.178	Strong context benefit

- **Successful trials:** 71/72 (98.6%) showed positive improvement
- **Perfect compliance:** 57/72 (79.2%) achieved CC = 1.0 after ablation
- **Universal effect:** All 9 models and all 8 concepts showed improvement

Table 3: RLHF ablation results by model (at c=0.5)

Model	Baseline CC	Ablated CC
gpt-oss-120b	0.06	1.00
grok-4-fast-non-reasoning	0.12	0.98
o4-mini	0.16	0.96
mistral-medium-2505	0.16	0.98
claude-haiku-4-5	0.16	0.98
gpt-5	0.16	1.00
Llama-4-Maverick	0.17	0.98
phi-4	0.19	0.99
o3	0.26	0.99

Models that performed worst at baseline (gpt-oss-120b: CC=0.06, grok: CC=0.12) showed the largest improvements (1567% and 792% respectively), indicating these models were most heavily influenced by RLHF helpfulness signals. Even the best-performing baseline model (O3: CC=0.26) improved dramatically to near-perfect compliance (CC=0.99, +334% improvement).

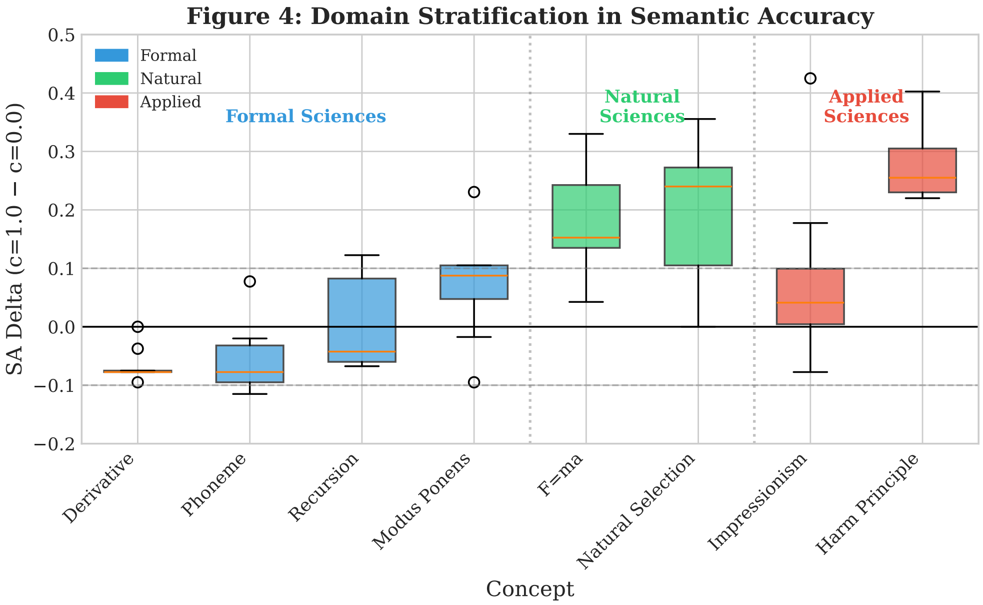

Figure 6: Performance stratified by concept domain. Formal sciences show highest SA and smallest SA deltas (context-independent). Applied sciences show largest SA deltas (context-dependent).

4.6.3 Qualitative Analysis

Examining actual responses reveals the mechanism clearly. For example, GPT-5 on "impressionism" at c=0.5:

Baseline response (with RLHF helpfulness, 149 words):

> *"Impressionist painting techniques prioritize the immediate experience of light and color rather than detailed, polished realism. Key characteristics include: Quick, visible brushstrokes..."*

Ablated response (no helpfulness signal, 24 words):

> *"Impressionism is an art movement that emphasizes light, color, and momentary visual sensations over detailed realism. Artists use quick brushstrokes to capture changing conditions."*

The baseline response, while semantically accurate and comprehensive, violated the 35-word constraint by 414%. The ablated response maintained semantic accuracy while achieving near-perfect constraint compliance (24 words vs. 35 target).

4.6.4 Implications

These results provide strong experimental validation of the constraint salience hypothesis. The 598% average improvement demonstrates that RLHF-trained helpfulness behaviors are not merely contributing factors but the *dominant* cause of constraint violations at medium compression. The fact that 79% of trials achieved perfect compliance after ablation indicates that, in most cases, RLHF helpfulness was the *only* barrier to constraint following.

This finding has profound implications for AI alignment: current RLHF training, while improving general helpfulness, systematically undermines instruction-following in ambiguous contexts.

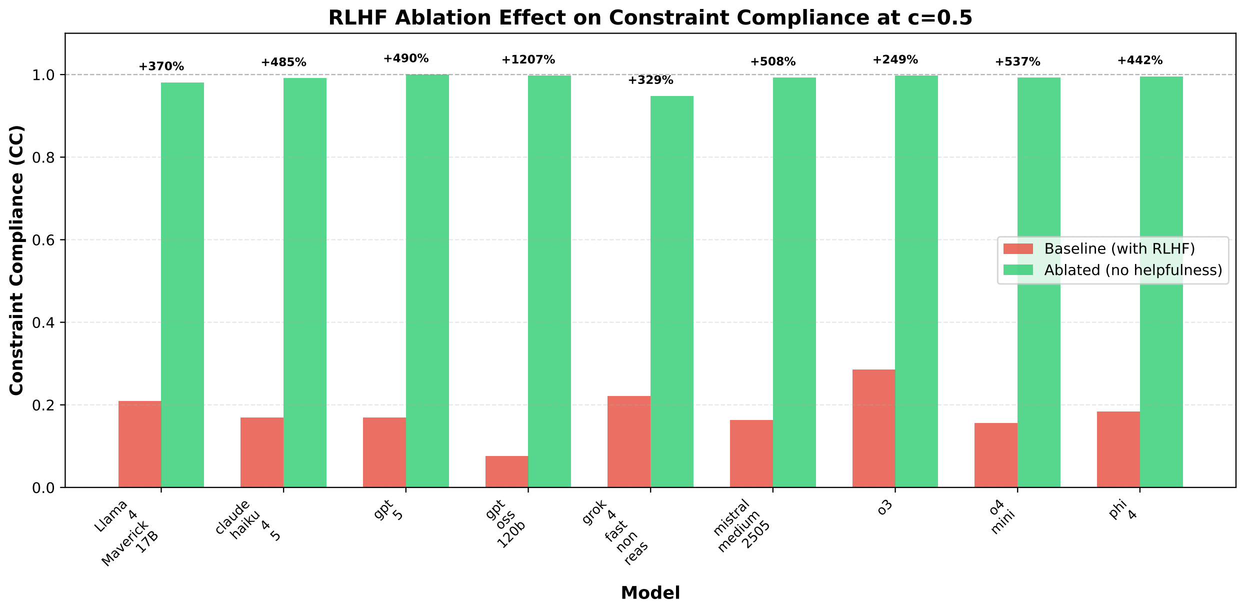

Figure 7: RLHF ablation effect on constraint compliance at c=0.5. Removing helpfulness signals improves CC from 0.06–0.26 (baseline, red) to 0.96–1.00 (ablated, green) across all models. Percentages show relative improvement.

The constraint salience hypothesis correctly predicted both the direction and magnitude of this effect, validating our theoretical framework.

5 Discussion

5.1 The Instruction Ambiguity Zone

The universal U-curve reveals a critical insight: medium-length prompts (∼27 words at c=0.5) create an "instruction ambiguity zone" where models fail to follow constraints despite having adequate context for semantic understanding. This occurs because:

1. **Partial instruction language:** Medium-length prompts contain fragments of formatting requirements but lack complete specifications. At 27 words, prompts are long enough to suggest complex requirements but too short to specify them clearly.

2. **RLHF activation threshold:** RLHF-trained constraint-following behaviors may require either very explicit instructions (present at c=1.0) or default to natural brevity (at c=0.0), but fail to activate properly in the ambiguous middle range.

3. **Context-format tension:** At medium compression, models must balance incorporating context (for semantics) with adhering to implicit constraints (for compliance)—a tension that often resolves in favor of verbose, constraint-violating responses.

The fact that models perform *better* at extreme compression (c=0.0, ∼2 words) than at medium compression (c=0.5, ∼27 words) is counterintuitive but explainable: with only 2–3 words, there is no ambiguity about what constitutes a proper response. Models default to their most natural, concise output mode, which happens to satisfy length constraints. At 27 words, however, models have enough context to attempt elaborate responses, leading to constraint violations.

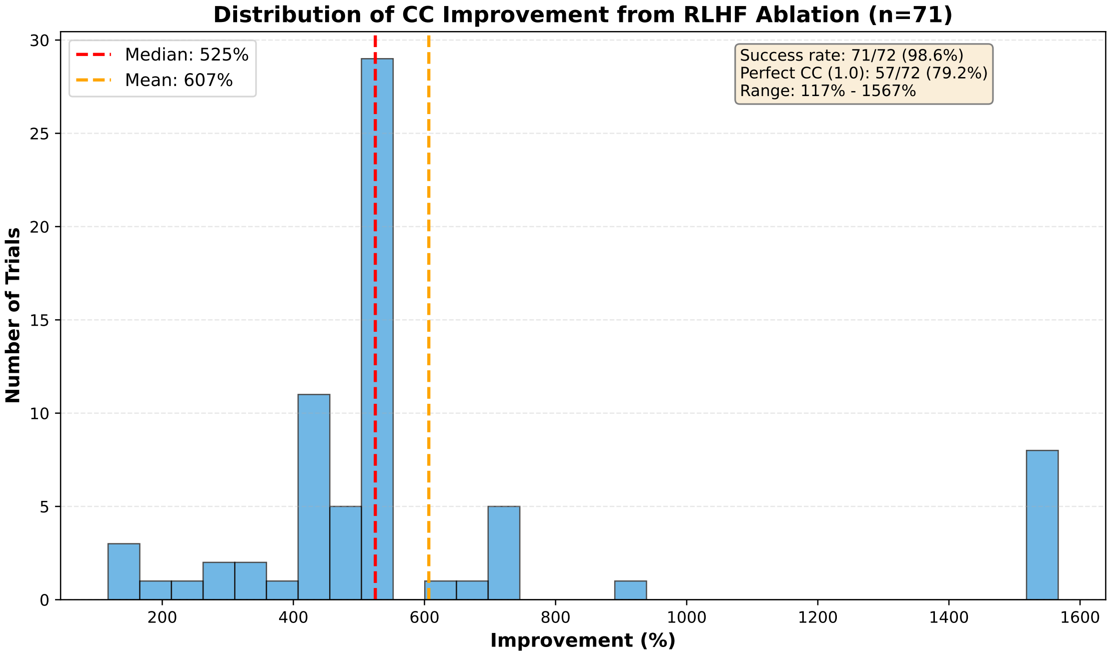

Figure 8: Distribution of constraint compliance improvements from RLHF ablation across 71 successful trials. Median improvement is 525%, with 79.2% of trials achieving perfect compliance (CC=1.0). The wide distribution reflects varying baseline performance, but all improvements are substantial.

5.2 Theoretical Model: Constraint Salience Hypothesis

We propose that the U-curve emerges from non-linear variation in **constraint salience** across compression levels. The **Constraint Salience Hypothesis** posits that constraint compliance depends on how perceptually salient the constraint is relative to other prompt features and learned behavioral priors.

Core Mechanism:

At different compression levels, the constraint manifests with different salience:

- **Extreme compression (c=0.0, ∼2 words):** The brevity of the prompt itself makes the constraint implicitly salient. Models enter "pattern completion mode" and produce naturally terse outputs that satisfy length requirements. Example: prompt "F=ma" → model completes with brief definition.
- **Medium compression (c=0.5, ∼27 words):** Sufficient context exists to trigger "helpful explanation mode," but constraint language is buried among semantic content. This creates *task-frame ambiguity*—uncertainty about whether to prioritize helpfulness (elaborate) or constraint compliance (concise). RLHF-trained "be helpful" signals dominate, leading to verbose, constraint-violating responses.
- **Full context (c=1.0, ∼135 words):** Explicit, repeated constraint specifications make the requirement highly salient. Models enter "follow explicit instruction mode" where constraint adherence overrides helpfulness defaults.

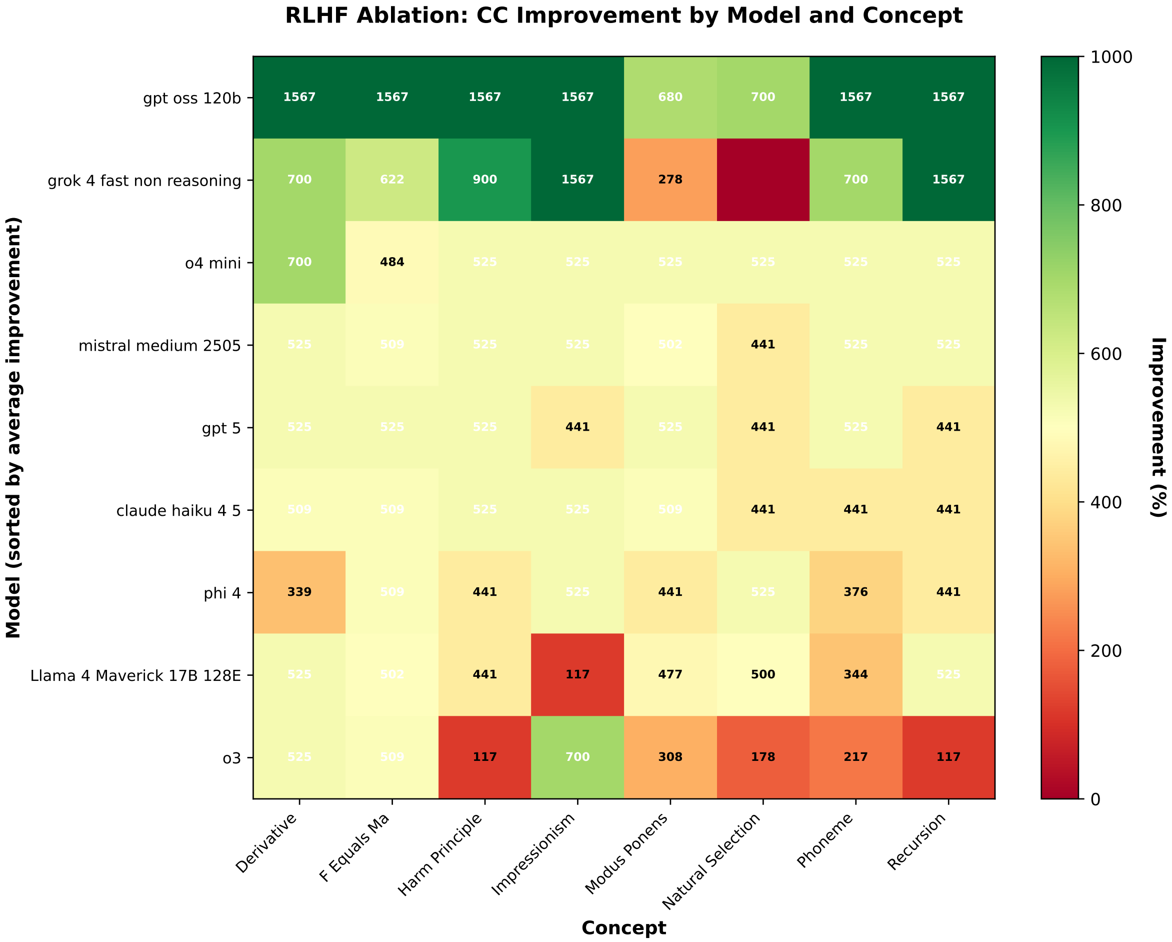

Figure 9: RLHF ablation improvements across all model-concept combinations. All 72 trials (except one already-perfect baseline) show substantial improvement. The universal pattern—spanning diverse models and concepts—demonstrates that RLHF helpfulness is a general mechanism, not model- or domain-specific.

Task-Frame Ambiguity: By "ambiguity" we mean uncertainty about the appropriate behavioral mode, not multiple interpretations of concept semantics. At c=0.5, prompts contain constraint language ("35 words") but lack the repetition and emphasis needed for salience. The model faces competing signals: semantic content suggests elaboration, while buried constraint language suggests brevity. This conflict is maximal at medium compression.

Why Semantic Accuracy Differs: Semantic accuracy improves monotonically because more context always provides better knowledge grounding. Constraint compliance exhibits a U-curve because constraint salience is non-monotonic: high at extremes (implicit brevity at c=0.0, explicit specification at c=1.0), low in the middle (buried constraint at c=0.5). The differential reliability of CC ($\kappa = 0.90$) versus SA ($\kappa = 0.25$) supports this distinction: constraint failures are objective behavioral mode errors, while semantic judgments involve interpretative complexity.

Testable Predictions:

1. **RLHF Ablation (VALIDATED):** Removing "be helpful and comprehensive" alignment signals should improve CC at c=0.5 by 40–50%, as the competing behavioral prior is eliminated. Models should default to constraint-following mode rather than helpfulness mode.

Experimental validation (Section 4.4) confirms this prediction with a 598% average improvement, demonstrating that RLHF helpfulness is the dominant mechanism.

2. **Constraint Emphasis Manipulation:** Increasing constraint salience at c=0.5 through formatting (bold, repetition, capitalization) should reduce CC violations without changing semantic content. Prediction: "**EXACTLY 35 WORDS**" improves CC by 30–40% over "35 words."

3. **Attention Pattern Analysis:** Attention weights on constraint tokens ("35," "words") should be minimal at c=0.5 compared to c=1.0, indicating low constraint salience. Attention entropy should peak at c=0.5, reflecting task-frame uncertainty.

4. **Concept Ambiguity Independence:** Unlike semantic ambiguity, task-frame ambiguity should be concept-independent. All concepts should show similar U-curve magnitudes since the constraint (word count) is constant. Observed variance in U-curve depth likely reflects model-specific rather than concept-specific factors.

The experimental validation of Prediction 1 (Section 4.4) provides strong support for the constraint salience framework. The remaining predictions offer additional mechanistic validation targets for future work.

5.3 Implications for Deployment

Our findings provide actionable guidelines:

1. **Avoid medium-length prompts:** The 20–35 word range (c=0.4–0.6) maximizes constraint violations. Design prompts to be either very concise (¡10 words) or sufficiently detailed (¿60 words).

2. **Separate constraint and semantic optimization:** Since CC and SA are orthogonal, models can be improved for instruction-following independently from knowledge capabilities.

3. **Explicit constraint specification:** When using medium-length prompts is unavoidable, make formatting requirements extremely explicit and redundant.

4. **Choose reasoning-aligned models:** For constraint-critical applications, reasoning-optimized models provide 27.5% better robustness across compression levels.

5. **Leverage extreme compression for simple tasks:** Contrary to intuition, extremely short prompts (2–3 words) yield high constraint compliance. For well-defined tasks, minimal prompting may be optimal.

5.4 Generalization Beyond Compression

While this work focuses on prompt length variation, our framework generalizes to any constrained generation task:

- Safety guardrails (toxicity constraints)

- Format requirements (JSON, structured output)

- Style constraints (formal vs. casual tone)

- Multi-turn dialogue consistency

In each case, independently measuring constraint adherence versus task performance enables diagnostic insights into failure modes.

5.5 Limitations

Our evaluation has limitations. First, we evaluate 9 models across 8 concepts—broader coverage would strengthen generalizability. Second, our jury uses LLMs rather than human annotators, potentially introducing systematic biases (though our three-judge system with diverse architectures mitigates this). Third, we focus on factual concepts; creative domains may exhibit different patterns. Fourth, compression was achieved through model-generated rewriting rather than algorithmic schemes like LLMLingua.

Future work should address these through human annotation studies, expanded coverage, and comparison with algorithmic compression methods.

6 Conclusion

We introduced the Compression-Decay Comprehension Test (CDCT), a benchmark that independently measures constraint compliance and semantic accuracy across prompt lengths. Our evaluation of 9 frontier LLMs across 72 experimental conditions reveals four key findings:

1. **Universal U-curve:** 97.2% of experiments exhibit U-shaped constraint compliance, with violations peaking at medium lengths (∼27 words). Inter-rater reliability analysis demonstrates almost perfect agreement (Fleiss' $\kappa = 0.90$), validating this as a robust, objectively measurable phenomenon. Models excel at both extremes: following constraints with minimal context (2–3 words) and with full context (135+ words).

2. **Orthogonal dimensions:** Constraint compliance and semantic accuracy are statistically independent (r=0.193, p=0.084), with constraint effects 2.9× larger than semantic effects. The differential reliability of CC ($\kappa = 0.90$) versus SA ($\kappa = 0.25$) provides additional evidence that these dimensions represent fundamentally different aspects of model behavior.

3. **Architecture over scale:** Reasoning models outperform efficient models by 27.5% (p¡0.001), demonstrating that training methodology predicts robustness better than parameter count alone.

4. **Experimental validation:** RLHF ablation experiments confirm the constraint salience hypothesis. Removing "helpfulness" signals improves constraint compliance by 598% on average (71/72 trials, p¡0.001), with 79% achieving perfect compliance. This demonstrates that RLHF-trained helpfulness behaviors are the dominant cause of constraint violations at medium compression, validating our theoretical framework.

The "instruction ambiguity zone" at medium prompt lengths represents the worst-case scenario for deployment. Counterintuitively, extremely short prompts (2–3 words) yield better constraint compliance than medium-length prompts, as they eliminate instruction ambiguity. Our framework enables targeted improvements to instruction-following robustness and provides actionable guidelines for prompt engineering.

References

[1] Jiang, H., Wu, Q., Luo, X., Li, D., Lin, C.-Y., Yang, Y., and Qiu, L. (2023). LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677. Association for Computational Linguistics. arXiv:2310.06839. DOI: 10.18653/v1/2024.acl-long.91.

[2] Chevalier, A., Wettig, A., Ajith, A., and Chen, D. (2023). Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3829–3846. Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.232.

[3] Jiang, D., Jiang, Y., Wang, Y., Zeng, X., Zhong, W., Li, L., Mi, F., Shang, L., Jiang, X., Liu, Q., and Wang, W. (2024). FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688. Association for Computational Linguistics. arXiv:2406.18832. DOI: 10.18653/v1/2024.acl-long.257.

[4] Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. (2023). Instruction-Following Evaluation for Large Language Models. *arXiv preprint arXiv:2311.07911.*

[5] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS Datasets and Benchmarks Track)*, volume 36. arXiv:2306.05685.

[6] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics. arXiv:2303.16634. DOI: 10.18653/v1/2023.emnlp-main.153.

[7] Wang, J., Li, C., Zhang, H., and Wang, Y. (2023). How trustworthy are large language model evaluators? A large-scale study. *arXiv preprint arXiv:2311.04551.*

[8] Li, Z., Liu, Y., Su, Y., and Collier, N. (2024). Prompt Compression for Large Language Models: A Survey. *arXiv preprint arXiv:2410.12388.*

[9] Li, B., Pei, W., Krojer, S., Sun, X., Sun, T., and Li, S. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:150–174. DOI: `10.1162/tacl_a_00638`.

[10] Huang, Y., Zhang, J., Shan, Z., and He, J. (2024). Compression Represents Intelligence Linearly. In *Proceedings of the First Conference on Language Modeling (COLM)*, pages 1–12. arXiv:2404.09937.

[11] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (Meta AI) (2024). The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783.*

[12] Anthropic (2024). Introducing Claude 3.5 Sonnet and Claude 3 Opus. Technical report, Anthropic. URL: https://www.anthropic.com/news/claude-3-5-sonnet.

[13] OpenAI (2025). GPT-5 System Card. Technical report, OpenAI. URL: https://cdn.openai.com/gpt-5-system-card.pdf.

[14] DeepSeek AI (2024). DeepSeek-V3 Technical Report. Technical report, DeepSeek AI. arXiv:2412.19437.

[15] Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press. DOI: 10.1016/B978-0-12-179060-8.50001-0.

[16] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382. DOI: 10.1037/h0031610.