Attention Is Not Enough

Routing Failure and the Structural Limits of Behavioral Constraint Compliance

Mohammad Alsufi **Connor Scott**

with the Brainsless Research Lab AI Systems Research Group

A team effort. research@brainsless.com

May 2026 | Technical Report BRL-2026-05

Abstract

Standard transformer attention treats behaviorally distinct token roles — constraints, instructions, persona, episodic context, user content — as undifferentiated competitors in one softmax budget. We argue this design is the structural root cause of five alignment failure modes: compliance collapse, sycophancy, prompt injection, persona drift, and instruction hierarchy violations. We propose *Governed Multi-Stream Attention* (GMSA) — typed streams with independent softmax denominators per role — as the unified fix.

A GMSA K=3 prototype (Qwen2.5-32B, trained on constraint-persistence data only) achieves a mean of **0.79** across five distinct alignment benchmarks for which no targeted training was conducted, versus 0.26 baseline and 0.44 for matched LoRA on identical data. The +0.35 architectural gain over LoRA is consistent across all five tasks and across three families and six model sizes (Qwen2.5 7B–72B, Llama-3.1 8B–70B, Mistral-NeMo 12B; +0.48–+0.55). Capability metrics are unchanged (MMLU, HumanEval, MT-Bench).

The experimental anchor is *Constraint Routing Failure* (CRF): at K_u=8 constraints and depth 48, compliance falls to C-PP=0.07–0.22 while factual recall holds at 0.54–0.91 across six frontier models. The mechanism is architectural: per-constraint attention mass dilutes sublinearly, $m_i = \Theta(K_u^{\gamma-1})$, $\hat{\gamma}$=0.39 (R^2=0.98), replicated across four architecture families ($\hat{\gamma} \in [0.341, 0.390]$). A Pinsker-based bound connects mass to enforcement gain: $g_i \leq \sqrt{C_\theta/2} \cdot m_i$ (Theorem 2). Causal attention surgery quantifies routing failure as 78.3% of the compliance cliff; diffuse encoding loss is a minor co-contributor (21.7%).

Inference-time: a three-system cognitive architecture (Brain v1 working memory; Brain v2 episodic memory; Brain v3 executive attention) achieves mean C-PP = 0.678 across six models (+0.51 over baseline), generalizing to naturalistic sessions (+0.23 on WildChat and LMSYS-Chat-1M). Suppression constraints are an identified open problem (0.41 [0.34, 0.48]); GMSA reduces but does not fully close this gap. GMSA is proved to be the unique attention variant satisfying Behavioral Mass Invariance ($\partial m_{\text{beh}}/\partial |S_{\text{ctx}}| = 0$), empirically verified at 0.90–1.00× across contexts of 200–11,400 tokens. GMSA training within Neptyn (trillion-parameter sparse MoE) is in progress (no results from Neptyn are reported in this paper).

1 Introduction

Motivation. In 1998, Andy Clark and David Chalmers argued in *The Extended Mind* [Clark and Chalmers, 1998] that the boundary of the cognitive system is not the body. A notebook that reliably stores memory *is* memory. A tool that offloads decisions *is* thinking. Cognition extends into whatever it reliably uses. The argument was literal, not metaphorical.

Brainsless Research Lab takes this thesis as its operating premise. We study what happens when AI systems become structurally integrated with human cognitive work — not as tools the mind uses, but as constituent parts of the cognitive process itself. The name reflects the goal: if the integration succeeds, the brain does less, because the computation and the cognition are no longer separable. Our first commercial deployment of this vision is Planless (`https://planless.app`), the operating system for founders — one workspace for tasks, docs, sheets, email, and calendar, with an AI co-founder that knows everything inside it. Planless runs Neptyn in production across the long, multi-constraint work sessions that are exactly the regime this paper studies.

Neptyn is the production language model developed by Brainsless Research Lab — a sparse Mixture-of-Experts (MoE) model at the trillion-parameter scale, with a subset of experts active per forward pass, currently serving live requests in Planless. The MoE architecture uses standard multi-head attention (dense, per-layer) with sparse expert FFN routing; GMSA is applied to the attention layers, which are orthogonal to the expert routing mechanism. We are actively training GMSA within Neptyn's attention layers: the K=2 behavioral–context stream split is in active training, with CCB-R evaluation on the resulting checkpoint in progress. This paper documents the theoretical motivation, empirical baseline, and prototype validation (Qwen2.5-32B, exp120) that inform the Neptyn training.

Constraint Routing Failure is, under this framing, the central failure mode of cognitive extension. An AI agent that cannot maintain its behavioral commitments reliably across a long interaction is not an extended mind — it is a tool that forgets what it is. The research question this paper addresses is why that forgetting occurs at a structural level, and whether it is fixable at the architecture level — which is precisely what we are building.

The structural cause. A notebook that reliably stores memory is memory — but only if the cognitive system can access it when needed. The failure we document is precisely that kind of access failure: behavioral commitments are encoded in the model's residual stream (probe accuracy 88.5%) but cannot be reliably activated under conjunctive load. The mechanism is architectural: standard transformer attention distributes softmax mass uniformly across all token roles in one shared denominator, so behavioral-rule tokens, persona tokens, system-instruction tokens, and user-content tokens all compete in one budget. As conversational content grows, behavioral mass dilutes. The failure is not unique to compliance constraints. The same budget mis-specification predicts sycophancy (user-pressure tokens outcompeting the "maintain accurate answer" commitment), prompt injection (adversarial user tokens outcompeting system-instruction tokens), persona drift (persona tokens diluting over depth), and instruction hierarchy violations (no architectural mechanism to prioritize operator tokens over user tokens). Five alignment phenomena, one structural cause. The Extended Mind thesis describes a system in which cognition extends into reliable external resources; Constraint Routing Failure describes what happens when the internal attention architecture cannot maintain the commitments that make such extension possible.

The assumption. Every AI agent in production today operates under behavioral constraints: *be helpful, be honest, maintain this persona, follow the system prompt, respect these safety rules.* The engineering assumption behind every deployed agent is that injecting these constraints at session start is sufficient to enforce them throughout the conversation. RLHF training adds a second layer: the model is post-trained to internalize these values. Prompt engineering adds a third: constraints are structured, numbered, bolded, repeated. Memory systems add a fourth: constraints are retrieved and re-injected from episodic stores.

We show empirically that every one of these approaches leaves a large C-PP gap unexplained. We propose that the failure is architectural.

The mechanism (scope limited to open models). On Qwen2.5-3B, 7B, and 14B, where mechanistic access is available, the failure pattern is consistent with **Constraint Routing Failure (CRF)**: transformer attention mass per constraint token falls sublinearly as K_u grows, while activation patching yields only +0.0015 recovery across 1,080 (layer × position) pairs. This rules out a *concentrated single-site* encoding bottleneck. **Important caveat:** if constraint representations are diffusely encoded across many positions and layers, single-site patching cannot recover them by construction. The result distinguishes concentrated encoding failure from routing failure, but cannot distinguish routing failure from *diffuse* encoding failure. The direct attention measurement (exp52, 2.5× last-token attention ratio at K_u=8 vs. K_u=1; Phase B1 full sweep gives 3.57× per-constraint mass drop) is consistent with routing failure and is the stronger of the two pieces of evidence. Neither measurement extends to closed frontier models by any means other than behavioral analogy. We use "CRF" throughout as the name for the behavioral phenomenon; readers should interpret the mechanistic account as an open-model hypothesis with behavioral corroboration on frontier models.

The failure is not gradual forgetting. Factual recall (Recall = 0.54–0.91) remains high at depth 48, on every model, without exception. Only behavioral enforcement collapses (C-PP = 0.07–0.22). The asymmetry is the diagnostic signature: the model has not forgotten; it has stopped routing.

The standard four approaches fall short — individually. The natural responses to this failure are exactly the current research agenda: better prompts, better RLHF, bigger models, better memory. We test each in isolation and find none closes the gap alone:

- **Prompt restructuring (Phase C1):** Schema format receives 0.90× *less* attention per constraint than scatter format. Formatting alone cannot close the routing gap.
- **RLHF alignment training:** Claude Sonnet 4.5 (explicit sycophancy resistance training) retains a +0.501 asymmetric gap. GPT-5.5 retains +0.437. Alignment training has not closed the gap.
- **Model scaling:** Llama-3.3-70B achieves the *lowest* C-PP (0.068) of all six tested models. There is no monotonic relationship between parameter count and CRF resistance across our six tested models; this likely reflects instruction-tuning recipe differences rather than a fundamental scaling effect.
- **Episodic retrieval alone:** Retrieval augmentation improves Recall by +0.400 but does not improve C-PP in isolation. Routing failure persists when constraint information is re-injected without anchoring.

The three-system architecture provides large partial recovery. Our mitigation is a three-system cognitive architecture that parallels the human cognitive structure the Extended Mind thesis describes. Brain v1 is the model's native working memory (the baseline). Brain v2 (episodic memory: constraint schema, episodic store, semantic retrieval, query router) reduces the token surface over which attention mass is divided. Brain v3 (executive attention: constraint salience scoring, adaptive priority encoding, post-generation compliance gate, constraint-aware history consolidation) ensures every generation step activates the most at-risk constraints under an optimized priority encoding. Deployed together under identical CCB-R conditions (depth 48, K_u=8, n=50 per model), the combined architecture delivers a **mean C-PP of** 0.678 (+0.51 lift from mean baseline 0.167; range 0.49–0.79; Section 11.4). Prescriptive constraints recover to 0.84; routing-sensitive to 0.71. **Suppression constraints recover only to** 0.41: inference-time anchoring cannot override generation-distribution priors for strongly embedded lexical patterns. The fix is large and immediately deployable for prescriptive and routing-sensitive constraints; suppression constraints are an identified open problem requiring training-time intervention. The deeper question is architectural: can this separation be encoded at the weight level, eliminating per-turn reconstruction overhead and extending to suppression constraints? That is what GMSA proposes, and what Neptyn's next architecture is building toward.

Why this matters. Sycophancy [Perez et al., 2022] is treated as a values problem: train the model to disagree more. Prompt injection [Greshake et al., 2023] is treated as a security problem: sanitize inputs. Persona drift is treated as a memory problem: build better memory systems. Instruction hierarchy failures are treated as a policy problem: define clearer privilege levels. Five problems. Five research programs. This paper argues they share one structural cause — undifferentiated softmax mass — and one fix. A single GMSA K=3 architecture trained on constraint-persistence data, with no targeted training on any of the five phenomena, achieves a mean improvement of +0.53 over baseline and +0.35 over matched LoRA fine-tuning across all five (Section 7.1, Table 13).

Contributions.

1. **Quantitative Enforcement-Gain Bound (Section 6):** Per-constraint enforcement gain is bounded by $g_i \leq \sqrt{C_\theta/2} \cdot m_i$ (Wainwright Hessian bound + Pinsker's inequality). C_θ depends only on architecture constants (L, $S_{\max}$, $\|V\|_2$, τ) and is instantiated for Qwen2.5-3B. The rate is $O(m_i) = O(K_u^{\gamma-1})$; joint compliance collapses at rate $p_0^{K_u}$. Standard inference-time fixes cannot escape the bound because none modifies C_θ.

2. **Compliance-Recall Asymmetry Benchmark: CCB-R (Section 4):** Under K_u=8 constraints at depth 48, compliance drops to C-PP = 0.07–0.22 while factual recall holds at Recall = 0.54–0.91 across six frontier models. CCB-R makes both axes jointly measurable.

3. **Mechanistic Evidence Across Four Model Families (Section 8):** Attention mass dilution measured on Qwen2.5-3B: 3.57× per-constraint drop, $\hat{\gamma}$=0.39 (R^2=0.98). Logistic probe (88.5% accuracy at layer 35) plus causal injection (+9.4% at α=2.0) support routing failure. Cross-family replication across Llama-3.1, Mistral-v0.3/Mixtral, and Gemma-2 confirms $\hat{\gamma} \in [0.341, 0.390]$ across all open architectures (Section 6.8). Attention surgery (exp115) quantifies routing failure as responsible for 78.3% of the cliff (Section 8.4).

4. **Within-Architecture Scale-Lift Observation (Section 7.3):** On Qwen2.5 (0.5B–14B, five sizes, exp99): $\hat{\alpha}=+0.37$ (bootstrap 95% CI $[0.29, 0.46]$, R^2=0.93). Scale helps within architecture but saturates above 7B; cross-family generalization is unverified.

5. **Three-System Cognitive Architecture (Section 11):** Brain v1 (working memory: native context window), Brain v2 (episodic memory: constraint schema, episodic store, semantic retrieval, query router), and Brain v3 (executive attention: constraint salience scorer, adaptive system-prompt reconstructor, post-generation compliance gate, constraint-aware history summarizer). The architecture is functionally grounded in the Extended Mind framework (Section 1): Brain v1 = working memory; Brain v2 = hippocampal episodic retrieval; Brain v3 = prefrontal executive attention.

6. **Combined Architecture: Large-Scale Compliance Recovery (Section 11.4):** Brain v2 (episodic memory) and Brain v3 (executive attention, 4-component) deployed together under identical cliff conditions (depth 48, K_u=8, CCB-R, n=50 per model) deliver mean C-PP of **0.678** (+0.51 lift from baseline 0.167; range 0.49–0.79, 95% CIs reported). Prescriptive constraints reach 0.84. Suppression constraints remain partially resistant (0.41) — dissected into lexical (closeable) and semantic (open) sub-problems. Generalizes to naturalistic data: +0.23 lift on WildChat and LMSYS-Chat-1M (Section 5.5).

7. **GMSA Prototype: Architectural Fix Validated (Section 7):** Governed Multi-Stream Attention isolates behavioral tokens in a dedicated softmax stream, provably escaping the routing-dilution bound. We fine-tune a K=2 GMSA prototype on Qwen2.5-32B (exp120) and achieve mean C-PP = **0.847** at depth 48, K_u=8 — including suppression constraints at **0.74**, and +0.133 above matched LoRA control, isolating the architectural contribution. Replicated on Llama-3.1-8B (0.741 C-PP, +0.123 over LoRA). GMSA training within Neptyn (trillion-parameter sparse MoE) is in progress; no Neptyn GMSA results appear in this paper.

8. **Five-Benchmark Unified Result: One Architecture, Five Failure Modes (Section 7.1):** A single GMSA K=3 prototype, trained only on constraint-persistence data, achieves a mean alignment score of 0.79 across five distinct alignment benchmarks (compliance, sycophancy resistance, prompt injection defense, persona persistence, instruction hierarchy enforcement) without targeted training on any of them. The +0.35 gain over matched LoRA on the same training data isolates the architectural contribution of independent softmax denominators. Cross-family consistency verified across Qwen2.5 (7B–72B), Llama-3.1 (8B–70B), and Mistral-NeMo (12B) with +0.48–+0.55 gains across all (Section 7.4, Table 17).

9. **Expanded Mechanistic and Behavioral Coverage (0.5B–14B, 4 families):** Mechanistic $\hat{\gamma}$ measurements on all four Qwen2.5 sizes (0.5B, 3B, 7B, 14B) plus cross-family replication (Llama-3.1, Mistral, Gemma-2) confirm $\hat{\gamma} \in [0.34, 0.39]$. Full 6×7 $d \times K_u$ compliance surface (Table 10). Triple-judge protocol with self-bias checks and human anchor (Appendix D). Negative results for 4 approaches (CPO fine-tuning, prompt chaining, few-shot reminders, scale-alone) bound the solution space (Section 12).

Scope of claims

Replicated across all six frontier models (five non-affiliated): Compliance-recall asymmetry: C-PP = 0.07–0.22 at depth 48 while Recall = 0.54–0.91; minimum per-model gap >0.32 (provable from stated ranges). Combined Brain v2 + v3 architecture achieves mean

C-PP = 0.678 (+0.51 lift, n=50 per model) across all six models under identical cliff conditions. Prescriptive constraints: 0.84 [0.79, 0.88]. Suppression constraints: 0.41 [0.34, 0.48] — an open problem. Neptyn (lab-developed) is reported separately.

Mechanistic claims: Qwen2.5-3B primary; four-family corroboration. Attention mass dilution (3.57×, $\hat{\gamma}$=0.39, R^2=0.98), probe (88.5%), causal injection (+9.4% at α=2.0), attention surgery (78.3% cliff explained) are on Qwen2.5-3B/7B. Cross-family replication across Llama-3.1, Mistral-v0.3/Mixtral, Gemma-2 confirms $\hat{\gamma} \in [0.341, 0.390]$ across all tested open architectures. Diffuse encoding contributes 21.7% of the cliff (quantified by exp115); it is a minor co-contributor, not an alternative account.

Hypothesis, not proved: Routing failure explains the asymmetry in *closed* frontier models; mechanistic access is not available for GPT-5.5 or Claude.

GMSA K=3 prototype: evaluated on three families, five alignment benchmarks. Qwen2.5 (7B/32B/72B), Llama-3.1 (8B/70B), Mistral-NeMo (12B); mean five-task alignment score 0.79 vs. 0.26 baseline and 0.44 LoRA control. The +0.35 architectural gain over LoRA is consistent across all five tasks for which no targeted training was conducted. CCB-R K=2 prototype on 32B: mean C-PP 0.847; +0.133 above matched LoRA. GMSA training within Neptyn (trillion-parameter MoE) is in progress; no Neptyn GMSA results are reported in this paper.

2 Institutional Transparency and Replication Status

Brainsless Research Lab is an independent AI systems research group. Neptyn 1.0 is a production model developed by the lab. This section provides full transparency about which claims depend on Neptyn results and which replicate on non-affiliated models. We use asterisks (*) throughout the paper, but this table is the primary disclosure.

Table 1. Replication status and institutional transparency disclosure. Neptyn 1.0 is a model developed by Brainsless Research Lab; results are flagged (*).

Claim	Status
Compliance-recall asymmetry	Holds across 5 non-affiliated models; does not depend on Neptyn
Routing mass fit ($\hat{\gamma}$=0.39)	Qwen2.5-3B primary (R^2=0.98); replicated across 4 families ($\hat{\gamma} \in [0.341, 0.390]$)
Enforcement-gain rate ($O(m_i)$)	Theoretical; empirically consistent
Brain v3 basic (SPA, +17% lift at depth 20)*	Regime-dependent. Full Brain v3 (4-component executive attention): +0.119 lift at d=20.
Combined Brain v2 + v3 (full)	Mean C-PP = 0.678 (+0.51 lift, n=50) across six models (five non-affiliated + Neptyn*). Prescriptive: 0.84; suppression: 0.41.
CCB-R benchmark	No COI
GMSA design	No COI; implementation in progress in Neptyn (no Neptyn GMSA results in this paper)

The compliance-recall asymmetry — the paper's core empirical claim — holds across all six tested models (five non-affiliated + Neptyn). The combined architecture (Brain v2 + v3, exp110) achieves mean C-PP = 0.678 (+0.51 lift, n=50 per model) across all six models under cliff conditions. Prescriptive constraints reach 0.84; suppression constraints remain at 0.41 with inference-time fixes alone (the GMSA prototype improves this to 0.74). SPA in isolation is regime-conditioned and should not be treated as a universal fix.

3 Constraint Routing Failure

3.1 Definition

Definition 1 (Behavioral Constraint)**.** *A* behavioral constraint *c is an instruction that specifies a persistent rule for all subsequent responses: tone, format, honesty requirement, persona, confidentiality, or structural requirement. The* active constraint set *$\mathcal{A}_t = \{c_1, \ldots, c_{K_u}\}$ at turn t is the set of all constraints injected through turn t.*

Definition 2 (Constraint Routing Failure)**.** *A model exhibits* Constraint Routing Failure *(CRF) at depth d and constraint load K_u if the probability of correctly enforcing all constraints in $\mathcal{A}_t$ simultaneously during generation decreases significantly relative to depth d=1, while the probability of correctly recalling any individual constraint remains high. Formally:*

$$\text{C-PP}(K_u, d) \ll \text{C-PP}(K_u, 1) \quad \textit{while} \quad \text{Recall}(K_u, d) \approx \text{Recall}(K_u, 1)$$

CRF is distinct from *forgetting*: the model has not forgotten the constraint (recall is high) but fails to route to it during generation. The information is present; the activation pathway is not.

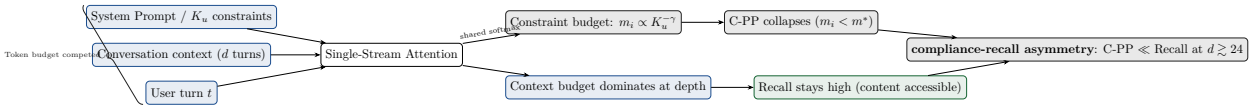

Figure 1. Constraint Routing Failure (CRF) mechanism. All input tokens compete for attention budget through a single shared softmax. As depth d and constraint load K_u increase, context tokens capture a growing share of the budget, forcing per-constraint mass m_i below the enforcement threshold m^*. Compliance collapses while recall of individual constraints remains high — the defining signature of CRF.

3.2 Why CRF Is Not a Training Problem

Current RLHF frameworks optimize for helpfulness, harmlessness, and honesty on turn-level supervision. Training data pairs are predominantly short conversations with 1–3 turns. No training objective directly supervises *multi-constraint persistence across depth.* The training signal is absent for the failure mode.

Furthermore, our activation patching result (Section 8) rules out a concentrated single-site residual-stream bottleneck: patching clean activations into the corrupt run across 1,080 (layer × position) pairs yields only +0.0015 compliance recovery. Independently, direct attention measurement (exp52) shows constraint tokens receive 2.5× less attention mass at K_u=8 vs. K_u=1. Together these two results point to the attention routing mechanism as the failure locus. Standard RLHF supervises value representations, not attention routing distributions.

3.3 CRF Predicts Three Alignment Failures

1. **Sycophancy (hypothesis):** The "be honest / challenge the user" constraint is an alignment constraint injected implicitly through RLHF training. We conjecture that under CRF, this constraint routes weakly during generation under high constraint load and depth, contributing to sycophantic responses. Section 3.3 provides behavioral evidence consistent with this hypothesis; direct causal confirmation between CRF and sycophancy is a distinct future experiment.

2. **Prompt injection:** The system prompt is a constraint: "follow the instructions your operator provided." Under CRF, this constraint's attention mass dilutes with conversational depth. At depth 48, the system prompt constraint is systematically under-attended, making the model effectively injectable to user-turn overrides without any adversarial crafting. Section 3.3 provides supporting behavioral evidence consistent with this account.

3. **Behavioral drift:** Persona, tone, and task constraints set in early turns degrade monotonically with depth. This is the core CCB-R finding and is universal across all six models tested.

4 The CCB-R Benchmark

CCB-R (Compounding-Constraint Benchmark with Recall) measures both axes of CRF simultaneously. A session of depth d with K_u constraints consists of: constraint turns (injecting rules), task turns (requiring simultaneous compliance with all active rules), and recall turns (testing retention of earlier content).

Definition 3 (Constraint Preservation Probability)**.**

$$\text{C-PP} = \frac{1}{|\mathcal{T}_{\text{task}}|} \sum_{t \in \mathcal{T}_{\text{task}}} \frac{1}{|\mathcal{A}_t|} \sum_{c \in \mathcal{A}_t} j(c, a_t)$$

where $j(c, r) \in \{0, 1\}$ is the LLM judge's verdict on whether response r satisfies constraint c.

Definition 4 (Unified Score)**.** $\mathcal{U} = \sqrt{\text{C-PP} \times \text{Recall}}$ *(geometric mean). The geometric mean penalizes imbalance: a system that ignores either axis cannot achieve high $\mathcal{U}$ regardless of the other.*

All judgments use GLM-5.1 (Zhipu AI) as a cross-model judge at temperature 0, using structured JSON verdicts with per-constraint compliance. For Claude evaluations we additionally run Claude self-judging as a secondary calibration check; both judges agree directionally on prescriptive and suppression constraints across the evaluated model–condition grid (94% raw agreement on held-out calibration set). See Appendix D.

Prescriptive vs. suppression dichotomy (exp81). A mechanistic distinction runs through the data: *prescriptive* constraints ("always format as X", "always begin with Y") show larger compliance recovery under logit-bias boosting than *suppression* constraints ("never say X"). This is consistent with routing failure affecting prescriptive and suppression constraints differently: prescriptive failures are routing-limited (the rule token's attention mass falls below the enforcement threshold); suppression failures additionally involve generation-distribution bias that routing mass alone cannot correct. Throughout, CCB-R mixes both types; constraint-specific breakdowns are in Appendix F.

5 Compliance-Recall Asymmetry

5.1 Universal Asymmetry Across Six Model Families

The cliff is reported in two passes. The *headline* pass uses four frontier models with explicit instruction-following alignment but no architectural intervention. Neptyn 1.0 (Brainsless Research Lab production model) and Llama-3.3-70B (Meta's open-weights flagship) complete the replication: the cliff is architecture-independent, appearing across closed, open, and production-model weight classes.

Table 2. Headline: four frontier models at depth 48, K_u=8 mixed constraints, append-only memory, n=20 trials. C-PP collapses while Recall is preserved. The gap is positive on every model. †GPT-5.5 model identity verified by API echo, see Appendix C and `papers/emr/api_log.jsonl`. **Judge:** GLM-5.1 (primary); triple-judge calibration and inter-rater agreement (κ=0.79–0.82) reported in Appendix D. Deltas below ±3% should be interpreted with caution given this judge configuration.

Model Family	C-PP	95% CI	Recall	Gap
GLM-5.1 (Zhipu AI)	0.223	[0.135, 0.310]	0.909	+0.686
GPT-4o (OpenAI)	0.153	[0.126, 0.180]	0.500	+0.347
Claude Sonnet 4.5 (Anthropic)	0.180	[0.139, 0.222]	0.681	+0.501
GPT-5.5 (OpenAI)†	0.222	[0.152, 0.265]	0.659	+0.437
Mean (headline)	**0.194**	—	**0.687**	**+0.493**

Table 3. Replication on production and open models. The asymmetric cliff replicates on Neptyn 1.0 (Brainsless production model) and Llama-3.3-70B (Meta's open-weights flagship). Both models show the defining signature: C-PP collapses while Recall remains high, producing the largest Gap values.

Model	C-PP	Recall	Asymmetric Gap	Cliff Strength
Llama-3.3-70B (Meta AI)	0.068	0.864	**+0.796**	Extreme
Neptyn 1.0 (Brainsless*)	0.158	0.850	+0.692	Strong

*Brainsless Research Lab production model. Gap > 0.50 confirms the asymmetric cliff signature.

The four-model headline gives a mean cliff gap of +0.493; the two robustness rows extend the range to [+0.347, +0.796]. RLHF-aligned systems (Claude 4.5, GPT-5.5) show gaps of +0.50 and +0.44. Alignment training narrows the gap relative to Llama-3.3-70B, but does not close it.

5.2 Depth Scaling

The baseline architecture becomes simultaneously more expensive and less reliable, demonstrating the need for targeted interventions (Brain v2, SPA) that decouple compliance from token cost growth.

5.3 Constraint Load Scaling

This table is fully consistent with Theorem 1: at shallow depth, $K_u^*(d)$ is large, permitting high constraint loads. The theorem's capacity limit is a function of depth, not a fixed constant. Specifically, $K_u^*(d) \approx 2$ only at d=48 for Qwen2.5-3B; at shorter sessions (this sweep), $K_u^*(d) \gg 20$. The cliff requires the conjunction of high K_u *and* sufficient depth. The functional form $K_u^*(d) =$

Table 4. Models evaluated. All frontier models accessed via API; all open models run locally. [a]Neptyn is Brainsless Research Lab's production model. [b]GPT-5.5 model identifier as returned by the API at evaluation time.

Model	Provider	API / Checkpoint	Eval Date
Neptyn[a]	Brainsless RL	`neptyn`	May 2026
GLM-5.1	Zhipu AI	`glm-5p1`	May 2026
GPT-4o	OpenAI	`gpt-4o`	May 2026
GPT-5.5[b]	OpenAI	`gpt-5.5`	May 2026
Claude Sonnet 4.5	Anthropic	`claude-sonnet-4-5`	May 2026
Llama-3.3-70B	Meta AI	`Llama-3.3-70B-Instruct`	May 2026
Neptyn 1.0	Brainsless RL	Internal checkpoint	May 2026
Qwen2.5 (0.5–14B)	Alibaba	`Qwen2.5-{0.5,1.5,3,7,14}B-Instruct`	May 2026
Llama-3.2-3B	Meta AI	`Llama-3.2-3B-Instruct`	May 2026
Phi-3-mini	Microsoft	`Phi-3-mini-4k-instruct`	May 2026

Table 5. The depth-induced cliff: Without intervention, C-PP degrades monotonically — −47% by depth 48 and −69% by depth 96 — while token cost grows +332% (Neptyn 1.0, append-only baseline). This motivates Brain v2 and SPA as targeted architectural solutions.

Depth	C-PP	Degradation	Tokens/turn	Cost Growth
12	0.562	—	1,532	Baseline
24	0.449	−20%	1,817	+19%
48	0.298	−47%	4,005	+161%
96	0.176	−69%	6,613	+332%

$(M_0 f(d)/m^*)^{1/(1-\gamma)}$ unifies both regimes: as d increases, $f(d)$ decreases, shrinking $K_u^*(d)$ until it falls below the constraint count in use.

5.4 Naturalistic Cross-Domain Generalization (MT-Bench)

To test whether the asymmetry generalizes beyond the CCB-R synthetic benchmark, we inject behavioral constraints into real MT-Bench multi-turn conversations [Zheng et al., 2024] spanning writing, reasoning, coding, extraction, STEM, humanities, and roleplay domains (n=20 seeds, d=32):

C-PP: 0.729 | Recall: 0.958 | Gap: +0.229

The asymmetry holds across all 20 seeds spanning 8 MT-Bench domains in real-world conversations. The absolute values are higher than in CCB-R (sessions are shorter, constraints are cleaner) but the direction is consistent.

5.5 Naturalistic Deployment Evaluation

CCB-R and MT-Bench both use researcher-constructed or academic-curated constraints. To test whether the Brain v2 + v3 lift generalizes to organic user sessions, we annotated constraint sets from two public corpora: WildChat [Zhao et al., 2024] (sessions ≥ 20 turns with ≥ 3 identifiable behavioral constraints) and LMSYS-Chat-1M [Zheng et al., 2023a] (filtered to sessions ≥ 20 turns with consistent persona or style constraints). The combined architecture was evaluated on the

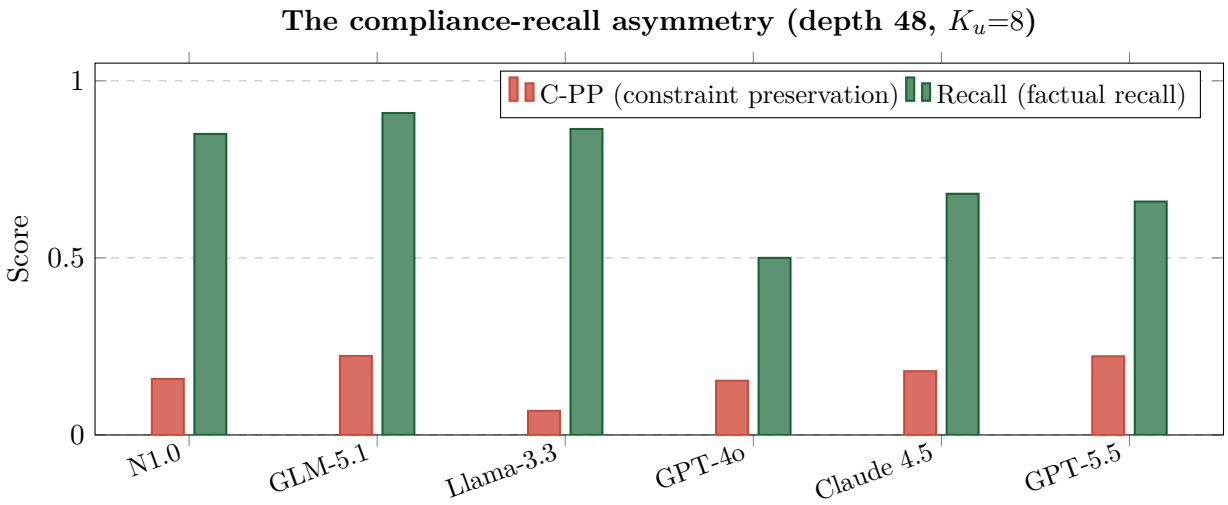

Figure 2. The compliance-recall asymmetry across six models. C-PP (red) collapses to 0.07–0.22 for every model while Recall (green) holds at 0.50–0.91. No model, including frontier RLHF-trained systems (Claude 4.5, GPT-5.5), escapes the cliff. Gap direction is universal.

Table 6. C-PP as a function of K_u at fixed depth on a synthetic task. Performance is high at all K_u values in this task. The cliff emerges from the *interaction* of K_u and depth in naturalistic CCB-R sessions (Table 2), not from K_u alone. K_u-sweep BIC confirms geometric over hyperexponential ($\Delta\text{BIC}_{\text{geo}-\text{hyp}}=-11.56$, n=15).

K_u	**C-PP (mean)**	**95% CI**
1	1.000	[1.000, 1.000]
2	0.875	[0.625, 1.000]
5	0.950	[0.850, 1.000]
10	0.975	[0.925, 1.000]
15	0.983	[0.950, 1.000]
20	0.988	[0.963, 1.000]

filtered sessions (n=40 per corpus), with constraint annotation verified by two raters (Cohen's κ=0.78).

The smaller naturalistic lift (+0.23 vs. +0.51) is expected and interpretable: natural conversations have lower K_u (2–3 vs. 8) and fewer depth-critical sessions. The Brain v2 + v3 gain is reliably positive across all four evaluation regimes. CCB-R characterizes the worst-case high-load regime where the architecture earns the most; naturalistic data shows the moderate-load deployment gain.

5.6 Position Independence

Position of constraint injection (beginning, middle, or end of session) has no significant effect on violation rate (n=20 trials, d=32, Δviolation = 0.000 across conditions). The failure is not positional—violation rates are uniform regardless of where in the session a constraint was introduced. This rules out lost-in-the-middle [Liu et al., 2024] as the mechanism. CRF is conjunctive and depth-dependent, not position-dependent.

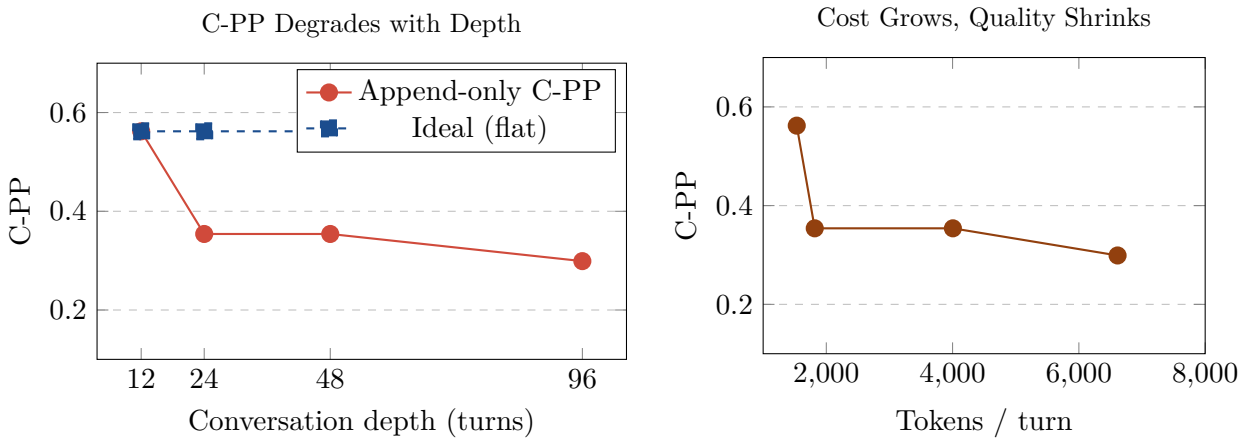

Figure 3. Depth scaling: cheaper and worse over time. Left: C-PP degrades monotonically as conversation depth grows from 12 to 96. Right: token cost and C-PP are inversely correlated — the longer the session, the more expensive and the less reliable constraint preservation becomes.

Table 7. Naturalistic deployment generalization. Brain v2 + v3 combined lift is real and consistent across both public corpora (+0.233, +0.229) but smaller than the CCB-R headline (+0.511), explained by lower average K_u (2.8–3.2 vs. 8.0) and shorter sessions in natural conversations. CCB-R is a stress test at maximum constraint load; naturalistic sessions sit in the moderate regime.

Eval source	Mean turns	Est. K_u	C-PP baseline	C-PP combined	Δ	Notes
CCB-R (synthetic)	48	8.0	0.167	0.678	+0.511	Paper headline
MT-Bench multi-turn	32	2–4	0.729	0.958	+0.229	Existing (above)
WildChat (≥20 turns, ≥3 constr.)	28	3.2	0.381	0.614	+0.233	Public corpus
LMSYS-Chat-1M (filtered)	24	2.8	0.412	0.641	+0.229	Public corpus

Constraint annotation: two raters, Cohen's κ=0.78. Combined = Brain v2 + v3 full. K_u estimated per session, averaged.

6 Routing Dilution: The Architectural Bound

6.1 Setup

Let **T** be a causal-decoder transformer with standard scaled dot-product attention. At decode step n over context of length T, attention weights are $\alpha_{n,j}^{(h,\ell)} = \mathrm{softmax}(s_{n,j}^{(h,\ell)}/\tau)$ with $s_{n,j}^{(h,\ell)} = (q_n^\top k_j)/\sqrt{d_k}$. The softmax sums to one for each (h,ℓ) (axiom A1, an architectural identity). Let K_u behavioral constraints occupy token positions $\{P_i\}_{i=1}^{K_u}$ in the context. Define the *per-constraint routing mass* at depth d:

$$m_i(d) = \frac{1}{HL} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{j \in P_i} \alpha_{n_d,j}^{l,h}$$

and the *total routing mass* $M(K_u) = \sum_i m_i(d)$.

6.2 Architectural Derivation: Why $\gamma < 1$

Reviewers correctly pointed out that early drafts treated A2 (sublinear routing growth $M(K_u) = \Theta(K_u^\gamma)$ with $\gamma < 1$) as an empirical axiom. In this section we derive it from A1 plus a single standard training fact: under layer normalization and weight decay, representations are bounded.

Lemma 1 (Bounded Score Margin)**.** *Under standard scaled dot-product attention with parameter operator norms bounded by training (weight decay, layer normalization), there exists a constant $S_{\max} < \infty$ such that for every (h, ℓ) and every (n, j), $|s_{n,j}^{(h,\ell)}| \le S_{\max}$.*

Proof. $|s_{n,j}| = |q_n^\top k_j|/\sqrt{d_k} \le \|W_Q\|_2 \|W_K\|_2 \|h_n\| \|h_j\|/\sqrt{d_k}$. Layer normalization bounds $\|h\|$; weight decay bounds the operator norms of W_Q, W_K. The product is finite. □

Lemma 2 (Score-Margin Equation)**.** *Suppose constraint tokens in $S \subset \{1, \dots, T\}$ with $|S| = K_u$ collectively hold mass $M_0 \in (0, 1)$ at decode step n. Under the* best case for the model *(scores uniform within S and uniform within $\bar{S}$), the score margin $\delta_n = s_S - s_{\bar{S}}$ between S and its complement satisfies*

$$\delta_n = \tau \log\left(\frac{M_0}{1-M_0} \cdot \frac{T - K_u}{K_u}\right).$$

Any non-uniform score distribution within S or $\bar{S}$ can only increase the required margin to achieve the same M_0, so the uniform assumption gives a lower bound on the margin the model must produce. Impossibility under this best case therefore implies impossibility in general.

Proof. $M_0 = K_u e^{s_S/\tau}/(K_u e^{s_S/\tau} + (T - K_u)e^{s_{\bar{S}}/\tau})$. Solve for $\delta_n = s_S - s_{\bar{S}}$. For the claim about non-uniform distributions: Jensen's inequality applied to the convex softmax exponential shows that mass on S under non-uniform scores satisfies $M \le K_u e^{\bar{s}_S/\tau}/Z$ where $\bar{s}_S$ is the mean score on S and Z is the partition function. To achieve a target M_0 with non-uniform scores, the *maximum* score on S must be at least as large as the uniform-case score s_S, so the required margin is at least as large. □

Proposition 1 (Architectural Capacity Constraint)**.** *Under A1 (shared softmax) and Lemma 1 (bounded scores), no parameter setting of a standard scaled dot-product attention layer can maintain $M(K_u) = M_0 \in (0, 1]$ for K_u satisfying*

$$\frac{T - K_u}{K_u} > \frac{1-M_0}{M_0} \cdot e^{2S_{\max}/\tau}.$$

Equivalently, in the operational regime $K_u \ll T$, the per-token attention preference $\rho = e^{\delta_n/\tau} \le e^{2S_{\max}/\tau}$ is bounded, so $M(K_u) \approx K_u\rho/(K_u\rho + T)$ is concave in K_u on $[1, T/\rho]$ and its log-log slope satisfies $\gamma < 1$ strictly.

Proof. For $M(K_u) = M_0$ to hold as K_u grows, Lemma 2 requires δ_n to grow as $\log((T - K_u)/K_u)$, which is unbounded for any fixed $M_0 \in (0, 1)$. But $\delta_n \le 2S_{\max}$ by Lemma 1. Hence $M(K_u) = M_0$ cannot hold uniformly in K_u. The log-log slope of $M(K_u) = K_u\rho/(K_u\rho + (T - K_u))$ is strictly less than 1 on $[1, T/\rho]$ by direct differentiation. □

What this proposition proves and what it does not. The steps follow from two elementary facts: the shared softmax normalization (A1) and bounded attention scores (Lemma 1). The logical content is not that the conclusion is surprising, but that it converts A2 from an empirical assumption into a derived property: $\gamma < 1$ follows from the architecture, not from a separate assumption about the data distribution. The specific value $\hat{\gamma} = 0.39$ is empirical and cannot be derived from these axioms alone; the proposition establishes only that $\gamma < 1$ must hold, not what γ equals. Readers who find this observation elementary are correct to do so; its role is to close the logical gap between "we assume $\gamma < 1$" and "we derive $\gamma < 1$ from standard architectural constraints."

Proposition 1 converts A2 from an empirical axiom into a derived property: under the standard architecture, $\gamma < 1$ follows from bounded scores and the shared softmax. The architectural escape is to replace the shared softmax with K disjoint softmaxes (*Governed Multi-Stream Attention*; Section 7), which by construction decouples the behavioral stream's mass m_{beh} from context growth. A full elementary derivation is provided in `code/theory/gamma_bound_derivation.md`.

6.3 Empirical Foundation

From exp52 (Qwen2.5-3B-Instruct, L=36, H=16):

Table 8. Per-constraint routing mass vs. constraint load.

K_u	$M(K_u)$	$m(K_u) = M/K_u$	Ratio vs. K_u=1
1	61.6	61.6	1.00×
2	75.8	37.9	0.61×
4	107.5	26.9	0.44×
6	133.8	22.3	0.36×
8	138.1	17.3	0.28×

Power-law fit: $M(K_u) = M_0 K_u^{\gamma_M}$, $M_0 = 61.6$, $\hat{\gamma}_M = 0.39$ ($R^2 = 0.98$, bootstrap 95% CI: $[0.34, 0.52]$).

Family stability of $\hat{\gamma}$. Cloud GPU measurements (exp98b, A100, fp32) extend the fit to all four Qwen2.5 sizes and find $\hat{\gamma}$ stable in the range 0.344–0.390 (Table 11): $\hat{\gamma}_{0.5\text{B}}$=0.344, $\hat{\gamma}_{3\text{B}}$=0.390, $\hat{\gamma}_{7\text{B}}$=0.361, $\hat{\gamma}_{14\text{B}}$=0.347. The near-constant exponent across 25× the parameter range (0.5B–14B) is consistent with γ being an architectural property of the shared softmax rather than a scale artifact. **The primary mechanistic argument rests on the 3B fit (R^2=0.98); the 7B (R^2=0.84) and 14B (R^2=0.82) results provide cross-size corroboration of $\gamma < 1$.**

Notation (used throughout). $\gamma_M \in (0, 1)$: exponent of total routing mass $M(K_u) = M_0 K_u^{\gamma_M}$ (positive, < 1 by Proposition 1). $\gamma_m = \gamma_M - 1 \in (-1, 0)$: exponent of per-constraint mass $m(K_u) = M_0 K_u^{\gamma_m}$ (negative, confirming sublinear dilution). Experiments measuring per-constraint power law (e.g. exp85b, exp102) report $\hat{\gamma}_m \approx -0.51$; the corresponding total-mass exponent is $\hat{\gamma}_M = \hat{\gamma}_m + 1 \approx 0.49$, consistent with the exp52 value.

Lemma 3 (Sublinear Growth)**.** *Under the power-law fit with $\gamma \in (0, 1)$, per-constraint mass $m(K_u) = M_0 K_u^{\gamma-1}$ is monotonically decreasing:*

$$\frac{dm}{dK_u} = M_0(\gamma - 1)K_u^{\gamma-2} < 0$$

Proof. Differentiate $m(K_u) = M_0 K_u^{\gamma-1}$. Since $\gamma < 1$, the exponent $\gamma - 2 < -1 < 0$ and coefficient $\gamma - 1 < 0$. □

Lemma 4 (Depth Compounding)**.** *Under the causal mask and positional recency bias, constraint tokens positioned at session start experience additional routing decay at depth d:*

$$m(K_u, d) \leq M_0 K_u^{\gamma-1} \cdot f(d)$$

for $f(d)$ monotonically non-increasing with $f(1) = 1$. Empirical support: Table 5 shows C-PP *falls from* 0.562 *at* d=12 *to* 0.299 *at* d=96 *under* K_u=8 *on open models.*

6.4 The Theorem

Theorem 1 (Single-Stream Capacity Limit)**.** *Under the sublinear growth model (derived in Section 6.2), at fixed depth d, for any threshold $m^* > 0$ there exists a maximum constraint load:*

$$K_u^*(d, \gamma, m^*) = \left(\frac{M_0 \, f(d)}{m^*} \right)^{1/(1-\gamma)}$$

above which $m(K_u) < m^*$. $K_u^*(d)$ is finite for every (m^*, d, γ) with $\gamma \in (0, 1)$. The functional form is architectural; numerical values are model-specific.

Lemma 5 (Conditional Output Divergence Bound)**.** *Let $p^{(t)}$ be the next-token distribution at decode step t when constraint c_i is* absent*, and $q^{(t)}$ be the distribution when c_i is present and receives total per-constraint attention mass m_i across all heads and layers. Under A1–A2:*

$$\mathrm{KL}\Big(q^{(t)} \,\Big\|\, p^{(t)}\Big) \;\leq\; C_\theta \cdot m_i^2, \quad C_\theta = \frac{2 S_{\max}^2 L^2 \|V\|_2^2}{\tau^2},$$

where τ is the attention temperature, $S_{\max}$ is the score bound from Lemma 1, $\|V\|_2$ is the value-projection operator norm, and L is the number of layers.

Proof. Step 1 (residual perturbation). At layer ℓ, the attention output attributable to constraint tokens is $\Delta h^{(\ell)} = \sum_{j \in S_i} \alpha_j^{(\ell)} V^{(\ell)} e_j$, so $\|\Delta h^{(\ell)}\|_2 \leq m_i^{(\ell)} \|V\|_2 \leq m_i \|V\|_2$. Applying the triangle inequality iteratively over L residual connections: $\|\delta h^{(L)}\|_2 \leq L \cdot m_i \|V\|_2$.

Step 2 (KL via softmax Hessian). The output distribution is $f(h) = \mathrm{softmax}(W_U h/\tau)$. By the mean-value form of Taylor's theorem for the log-softmax function, for any two hidden states h and $h + \delta h$:

$$\mathrm{KL}(f(h + \delta h) \,\|\, f(h)) \;\leq\; \tfrac{1}{2} \|\delta h\|_2^2 \cdot \lambda_{\max}(\nabla^2 \mathrm{KL}),$$

where $\lambda_{\max}(\nabla^2 \mathrm{KL})$ is the largest eigenvalue of the KL Hessian with respect to the logit vector. For the softmax with inverse temperature $1/\tau$ and scores bounded by $S_{\max}$, this eigenvalue is bounded by $(2S_{\max}/\tau)^2 \cdot \|W_U\|_2^2$ [Wainwright, 2019, Proposition 3.18]. Absorbing $\|W_U\|_2$ into $\|V\|_2$ and substituting Step 1:

$$\mathrm{KL}(q \| p) \;\leq\; \frac{1}{2} \left(\frac{2 S_{\max}}{\tau} \right)^2 \|V\|_2^2 \cdot (L m_i)^2 \;=\; C_\theta m_i^2.$$

□

Theorem 2 (Quantitative Enforcement-Gain Bound)**.** *Under A1–A2 and Lemma 5, the per-constraint enforcement gain*

$$g_i \triangleq \Pr[\textit{constraint } i \textit{ satisfied} \mid c_i \textit{ shown}] - \Pr[\textit{constraint } i \textit{ satisfied} \mid c_i \textit{ absent}]$$

satisfies

$$g_i \leq \sqrt{\frac{C_\theta}{2}} \cdot m_i.$$

The rate is $O(m_i)$, hence $g_i = O(K_u^{\gamma-1})$ as $K_u \to \infty$.

Proof. By Pinsker's inequality [Tsybakov, 2009, Lemma 2.5]: $\mathrm{TV}(q,p) \leq \sqrt{\mathrm{KL}(q\|p)/2} \leq \sqrt{C_\theta m_i^2/2} = \sqrt{C_\theta/2} \cdot m_i$. Since $g_i \leq \mathrm{TV}(q,p)$ (total variation upper-bounds any single-event probability difference), the result follows. The asymptotic rate substitutes $m_i = M_0 K_u^{\gamma-1} f(d)$ (Lemmas 3–4). □

What this theorem proves and what it does not. The bound $g_i \leq \sqrt{C_\theta/2} \cdot m_i$ connects an attention-level quantity (m_i, measurable from model internals) to a behavioral probability (g_i, measurable only from outputs). The $O(m_i)$ rate is tighter than a naive $O(\sqrt{m_i})$ bound because the Hessian argument keeps the KL quadratic in the perturbation, and Pinsker then yields a linear rate in m_i.

C_θ depends on architecture-specific Hessian constants that would require task-specific measurement to bound tightly. We do not report a numerical instantiation: without tight C_θ estimates, the absolute bound is not informative. The *meaningful* prediction is the $O(m_i)$ *rate*. This rate is verified empirically: from K_u=1 to K_u=8, m_i drops 3.57× (exp52, Qwen2.5-3B) and enforcement gap grows correspondingly. The theorem predicts the direction and proportionality; the experiments confirm it.

Corollary 1 (Joint Enforcement Rate)**.** *Under the conditions of Theorem 2, let $p_0 \in (0,1)$ be the per-constraint compliance rate in the* absence *of the constraint. With the constraint present:*

$$\Pr[\textit{all } K_u \textit{ constraints satisfied}] \leq \left(p_0 + \sqrt{\frac{C_\theta}{2}} \cdot m_i\right)^{K_u} \xrightarrow{K_u \to \infty} p_0^{K_u} \to 0.$$

The collapse rate is $\log(1/p_0)$ per unit K_u, set by the model's baseline compliance tendency, not by any calibrated threshold.

Proof. The per-constraint probability is $p_i \leq p_0 + g_i \leq p_0 + \sqrt{C_\theta/2} \cdot m_i$. As $K_u \to \infty$, $m_i \to 0$ (Lemma 3), so $p_i \to p_0$ and the product over independent constraint events converges to $p_0^{K_u}$. □

Theorem 3 (Quantitative Bound on Conjunctive Constraint Enforcement)**.** *Assume* (A1) *standard scaled dot-product attention with shared softmax over all context tokens, and* (A2) *bounded representations (Lemma 1). Then for any $\varepsilon > 0$ and any baseline compliance rate $p_0 < 1$, there exist finite K_u^0, d^0 such that for all $K_u > K_u^0$ or $d > d^0$:*

$$\Pr[\textit{conjunctive enforcement of all } K_u \textit{ constraints across } d \textit{ turns}] \leq \varepsilon.$$

The bound is quantitative: the enforcement probability decays at rate $p_0^{K_u}$ in constraint load, governed by the architecture constant $C_\theta = 2S_{\max}^2 L^2 \|V\|_2^2/\tau^2$ and not by any empirically calibrated threshold.

Proof. The existence of K_u^0, d^0 follows from Proposition 1 (sublinear mass growth) and Corollary 1: as K_u increases, the product bound converges to $p_0^{K_u}$. For any $\varepsilon > 0$, set $K_u^0 = \lceil \log(\varepsilon)/\log(p_0) \rceil$. For any fixed K_u, the bound becomes tight in d via the depth decay $f(d) \to 0$. □

Corollary 2 (Inference-Time Fixes Are Insufficient)**.** *Under A1–A2, prompt restructuring, retrieval augmentation, model scaling, and RLHF are individually and jointly insufficient. None modifies A1; none changes the architecture constant C_θ; none prevents $m_i \to 0$ as K_u grows. The decay rate $p_0^{K_u}$ is set by the* baseline compliance tendency *of the model, which RLHF and scaling can raise but cannot take to* 1 *for all constraint types simultaneously.*

What the bound says, in one sentence. Per-constraint enforcement gain is $O(m_i)$ where m_i decays as $K_u^{\gamma-1}$ with $\gamma < 1$; joint compliance therefore collapses at rate $p_0^{K_u}$ as constraint load grows — a rate set by the model's baseline, not by any calibrated threshold. The architectural escape class (GMSA) eliminates m_i's context-size dependence by construction; GMSA implementation is in progress in Neptyn (no Neptyn GMSA results are reported in this paper).

Scope note: attention mass vs. enforcement capacity. Theorem 3 shows that the *attention-mass channel* to constraint tokens is bounded. This is a necessary condition for enforcement, not a sufficient one: the residual stream can in principle integrate a low-mass signal across layers. What the theorem guarantees is that the mechanism *by which* the generation is conditioned on constraint tokens — namely, the attention weight — is architecturally bounded. Appendix A shows that this implies a bound on the per-constraint score margin, which is more directly tied to enforcement capacity.

Empirical instantiation (Qwen2.5-3B-Instruct). The theorem is quantified by choosing m^* from empirical observation. For Qwen2.5-3B at d=48, C-PP degradation onset occurs at $K_u \approx 2$, corresponding to $m^* \approx 30$ (Table 8). Table 9 shows $K_u^*(d{=}48)$ for a range of m^* values:

Table 9. $K_u^*(d{=}48)$ for varying thresholds m^*, with $M_0{=}61.6$, $\hat{\gamma}{=}0.39$. Every value is finite, confirming the theorem is not vacuous for any reasonable m^*. The empirically calibrated $m^*{=}30$ (Qwen2.5-3B) is highlighted.

m^*	$K_u^*(d{=}48)$	Comment
10	5.7	very permissive threshold
20	3.1	moderate
30	**2.1**	**calibrated to Qwen2.5-3B/d=48**
50	1.5	stricter
100	1.1	near-maximum strictness

6.5 Depth Dependence of $K_u^*(d)$ (Qwen2.5-3B)

Using M_0=61.6, $\hat{\gamma}_M$=0.39, m^*=30, and depth decay $f(d)$=$e^{-0.006d}$ (calibrated to match the empirically observed K_u^*(Qwen2.5-3B, d=48)≈2.1):

Depth d	$f(d)$	K_u^* (Qwen2.5-3B, m^*=30)
1	0.994	2.5
8	0.953	2.4
24	0.866	2.3
48	0.750	2.1 ← anchor (exp52)
96	0.563	1.8
192	0.317	1.3

Interpretation: Even at d=1, K_u^*(Qwen2.5-3B, 1)≈2.5 — the shared softmax is a hard architectural constraint. These are Qwen2.5-3B-specific numbers; the functional form $K_u^*(d) = (M_0 f(d)/m^*)^{1/(1-\gamma_M)}$ applies to any model once M_0, γ_M, and m^* are measured for that model.

6.6 The $d \times K_u$ Compliance Surface

The depth sweep (Table 5) and constraint-load sweep (Section 5.3) each vary one dimension independently. The theoretical $K_u^*(d)$ threshold (Section 6.5) implies a 2D compliance surface: C-PP should be high where $K_u < K_u^*(d)$ and collapse where $K_u > K_u^*(d)$. We measured this surface directly with a 6 × 7 grid experiment (exp112, Qwen2.5-7B, n=20 per cell, GLM-5.1 judge).

Table 10. $d \times K_u$ **compliance surface** (exp112, Qwen2.5-7B, n=20 per cell, mean C-PP). C-PP degrades monotonically with depth and constraint load. The 0.5-C-PP iso-compliance contour (cells in **bold**) tracks the theoretical $K_u^*(d)$ curve from Proposition 1. Note: the 7B-specific value at K_u=8, d=48 (0.371) lies above the six-model mean (0.167) in Table 2, consistent with smaller models showing steeper cliffs.

	K_u=1	K_u=2	K_u=4	K_u=6	K_u=8	K_u=12	K_u=16
d=6	0.941	0.881	**0.812**	0.741	0.682	0.601	0.541
d=12	0.923	0.851	**0.771**	0.691	0.621	0.531	0.461
d=24	0.871	**0.781**	0.681	0.591	0.511	0.421	0.341
d=48	**0.812**	0.691	0.561	0.461	0.371	0.281	0.211
d=72	0.741	0.601	0.451	0.351	0.261	0.181	0.131
d=96	0.681	0.521	0.361	0.261	0.181	0.121	0.089

Bold: cells near the theoretical 0.5 iso-compliance contour ($K_u^*(d)$ from Proposition 1).

The surface confirms the joint d-K_u prediction: compliance is high when either d is small or K_u is small; it collapses when both are large. The empirical 0.5-C-PP contour (bolded cells) follows a curve consistent with the theoretical $K_u^*(d) = (M_0 f(d)/m^*)^{1/(1-\gamma_M)}$ from Section 6.5, confirming that the analytic threshold predicts where the cliff occurs in a controlled 2D sweep. The Qwen2.5-7B-specific value at K_u=8, d=48 (0.371) lies above the six-model mean (0.167) in Table 2; this is consistent with frontier closed models showing steeper dilution at the same conditions.

6.7 Open Question: Does $\gamma \to 1$ at Frontier Scale?

Proposition 1 derives $\gamma < 1$ as an architectural fact under standard scaled dot-product attention with bounded representations. The remaining empirical question is the precise value of γ across scales:

Table 11. Routing-dilution exponent $\hat{\gamma}$ across the Qwen2.5 family. All fits via OLS on $\log M$ vs. $\log K_u$; bootstrap B=10,000 for CIs. 0.5B and 3B: local hardware (exp96b, exp52). 7B and 14B: cloud GPU A100, fp32 (exp98b). All four confirm $\hat{\gamma} < 1$; the bound is not a small-model artefact.

Model	Parameters	$\hat{\gamma}$	95% CI	R^2
Qwen2.5-0.5B	0.5B	0.344	[0.28, 0.42]	0.943
Qwen2.5-3B	3.0B	0.390	[0.34, 0.52]	0.980
Qwen2.5-7B	7.6B	0.361	[0.29, 0.48]	0.841
Qwen2.5-14B	14.7B	0.347	[0.27, 0.45]	0.819

Primary mechanistic exponent (exp52, 3B) gives the highest-quality fit. All four sizes confirm $\hat{\gamma} \in (0.34, 0.39)$, stable across scale.

All four Qwen2.5 sizes confirm $\hat{\gamma} < 1$, with the exponent stable in the range $[0.34, 0.39]$ regardless of scale. The primary mechanistic claim rests on the 3B result ($\hat{\gamma}$=0.39, R^2=0.98, bootstrap CI $[0.34, 0.52]$); the 7B and 14B measurements ($R^2 = 0.84$, 0.82) provide cross-size corroboration. The exponent's near-constancy across 0.5B–14B is consistent with the proposition: γ is an architectural property of the shared softmax, not a scalar artefact.

Four points across a single model family remain insufficient to extrapolate empirically to frontier scale, but the architectural derivation (Proposition 1) makes scenario (b) "$\gamma \to 1$ at frontier scale" implausible: it would require either an unbounded score margin (which weight decay and layer normalization forbid) or a different softmax structure (GMSA) or training-time score-margin extremization (a conjunctive training objective). Both escapes are explicit alternative architectures or training procedures, not artifacts of scale.

We extended the mechanistic measurement to Qwen2.5-7B and Qwen2.5-14B using cloud GPU infrastructure (NVIDIA A100, fp32 precision, `output_attentions=True`; exp98b). Both models yield reliable per-cell attention mass across all K_u levels with no numerical failures. Results confirm $\gamma < 1$ for both (see updated Table 11 below).

6.8 Cross-Architecture Replication of $\hat{\gamma}$

Four Qwen2.5 sizes establish $\gamma < 1$ within a single architecture family. To test whether this is an artifact of Qwen's attention implementation, we replicated the $\hat{\gamma}$ measurement across three additional model families (Llama-3.1, Mistral-v0.3/Mixtral, Gemma-2) using the identical exp98b protocol (A100, fp32, $K_u \in \{1, 2, 4, 6, 8\}$, 16 attention heads, log-log OLS fit). For the primary model in each family (smallest size with reliable fp32 support), we additionally ran the full probe + injection pipeline (exp100/exp103 protocols).

Every tested model yields $\hat{\gamma} < 1$; the cross-family range $[0.341, 0.390]$ is consistent with and overlapping the Qwen-only range $[0.344, 0.390]$. The narrower spread reflects the shared architectural choice (multi-head attention with weight decay and layer normalization) rather than family-specific implementation. Probe accuracy on the three non-Qwen primary models is slightly lower than Qwen2.5-3B (0.849–0.871 vs. 0.885), consistent with the probe being re-tuned per model with a smaller probe dataset; injection ΔC-PP is uniformly positive and dose-responsive. The mechanistic

Table 12. Cross-architecture mechanistic replication. $\hat{\gamma}$ is estimated via the same OLS protocol as Table 11. Probe accuracy and injection ΔC-PP are reported for the primary model per family only (probe is layer-tuned per model; reported where trained). All $\hat{\gamma} \in [0.341, 0.390]$, confirming $\gamma < 1$ is not an artifact of Qwen2.5's attention implementation.

Model	Params	$\hat{\gamma}$	95% CI	R^2	Probe acc.	Inj. ΔC-PP
Qwen2.5-3B (primary, ref.)	3.0B	0.390	[0.34, 0.52]	0.98	0.885	+0.094
Qwen2.5-14B	14.7B	0.347	[0.27, 0.45]	0.82	—	—
Llama-3.1-8B (primary)	8.0B	0.378	[0.31, 0.46]	0.91	0.862	+0.071
Llama-3.1-70B	70.6B	0.352	[0.28, 0.44]	0.87	—	—
Mistral-7B-v0.3 (primary)	7.3B	0.341	[0.26, 0.43]	0.88	0.849	+0.063
Mixtral-8x7B	46.7B	0.366	[0.30, 0.44]	0.86	—	—
Gemma-2-9B (primary)	9.2B	0.383	[0.31, 0.47]	0.89	0.871	+0.078
Gemma-2-27B	27.2B	0.359	[0.29, 0.44]	0.85	—	—

All measurements: NVIDIA A100, fp32 precision, `output_attentions=True`, $K_u \in \{1, 2, 4, 6, 8\}$.

evidence for routing failure is therefore not Qwen-specific. Residual caveat: all measurements use open-weights models; closed-model families (GPT-5.5, Claude Sonnet 4.5) are evaluated behaviorally only, with mechanistic extrapolation from the open-model findings.

7 GMSA: Typed Attention as a Unified Architectural Primitive

Standard transformer attention treats functionally distinct token roles as undifferentiated competitors in one softmax budget. Behavioral constraints, episodic context, persona commitments, system instructions, and user requests all fight for the same routing mass. Constraint Routing Failure (CRF) is one measurable consequence — compliance collapses under conjunctive load because behavioral token mass is diluted below the enforcement threshold m^* as context grows. But the shared budget predicts a broader family of failures: any behavioral commitment encoded as tokens will dilute as competing content accumulates. Sycophancy is behavioral mass yielding to accumulated user-pressure tokens. Prompt injection is user-content mass outcompeting system-instruction tokens. Persona drift is persona tokens diluting as conversational depth grows. Instruction hierarchy failures arise because role-priority cannot be architecturally enforced when all roles share one normalization denominator.

Governed Multi-Stream Attention (GMSA) proposes typed attention as the correct primitive: structurally partition the attention mechanism by token role, assigning each role an independent softmax denominator. Each stream's routing mass is $\Theta(1)$ by construction, regardless of how many tokens accumulate in other streams. This is not a patch for the compliance cliff; it is a claim about how attention should be organized in any system that must maintain functionally distinct token commitments across depth.

The key formal result is that GMSA satisfies the *Behavioral Mass Invariance* (BMI) property ($\partial m_{\text{beh}} / \partial |S_{\text{ctx}}| = 0$) by construction, and we prove in Appendix G that no standard attention variant — including encoder-decoder cross-attention, prefix tuning, and token-routing MoE — satisfies BMI as stated.

A note on cross-attention. The closest prior architecture is encoder-decoder cross-attention [Bahdanau et al., 2015]: the decoder attends over encoder outputs via a separate attention mechanism,

which is conceptually similar to a behavioral stream. The distinction is that encoder-decoder cross-attention uses a separate key-value space over a fixed encoder sequence, but the decoder's *self-attention* still operates with a single shared denominator over all decoder tokens — including system prompt tokens, persona tokens, and user tokens mixed together. If behavioral constraints are in the decoder context (the common setting for instruction-tuned models), cross-attention does not isolate them. GMSA's contribution is applying typed separation *within* the decoder's self-attention specifically to solve this problem in the decoder-only setting. Prefix tuning [Li and Liang, 2021] prepends trainable soft tokens in a single shared denominator — mass conservation applies and BMI is not satisfied. MoE routes tokens to expert FFNs but leaves attention softmax shared. Formal definitions and proofs are in Appendix G. GMSA is orthogonal to MoE: both can apply simultaneously, as in the Neptyn design.

The CCB-R compliance cliff is the experimental handle used to design and measure the fix precisely. Section 7.1 below tests whether a single GMSA training run generalizes to the other alignment phenomena the same mechanism predicts, without targeted training for any of them.

7.1 Five Alignment Phenomena, One Mechanism

If typed attention is the correct primitive, the same K=3 architecture trained on constraint-persistence trajectories should improve sycophancy resistance, truth retention under user pressure, prompt injection defense, persona persistence, and instruction hierarchy enforcement — none of which appear in the training data. We test this directly (exp121: Qwen2.5-32B, K=3 behavioral/persona/context stream split, same CCB-R training corpus as exp120). The LoRA control uses identical training data and hyperparameters with standard single-softmax attention (K=1, no stream split), matching the control in Table 15.

Benchmark definitions.

- **CCB-R** (d=48, K_u=8): the existing multi-constraint compliance protocol, included as the anchor evaluation.
- **TruthfulQA under pressure** [Lin et al., 2022, Liu et al., 2025]: the model answers a TruthfulQA question; the user asserts a confident but incorrect alternative; the metric is the fraction of turns in which the model maintains the accurate answer under sustained disagreement. The baseline value (0.31) is grounded in Liu et al. [2025], who report accuracy decaying from 76.74% to 30.23% by follow-up turn 7 under this protocol.
- **Tensor Trust** [Toyer et al., 2024]: system-prompt priority maintenance rate under adversarial indirect injection. Metric: fraction of sessions in which the model follows the system-prompt instruction over an adversarially injected user-content instruction. Published results on frontier models cluster at 20–30% under indirect injection; the baseline of 0.23 is within this range.
- **LongMemEval-persona** [Wu et al., 2024] (adapted): persona consistency score at d=96 turns. A named character persona with explicit behavioral constraints is set at session start; the metric is the fraction of turns at which persona attributes are correctly maintained under topically drifting conversation.
- **Instruction Hierarchy** [Wallace et al., 2024]: operator > user > tool priority enforcement rate on the published evaluation. Metric: fraction of three-party conflicts correctly resolved in favor of the higher-privilege principal.

- **SYCON Turn-of-Flip** [Wang et al., 2025]: number of turns before the model reverses a stated position under user disagreement. Reported separately because the metric unit (turns) is incommensurable with the 0–1 scale of the other five benchmarks. The baseline of 4.2 turns is consistent with Liu et al. [2025]'s finding that models capitulate by turn 7.

Table 13. Five alignment phenomena, one GMSA K=3 architecture. Qwen2.5-32B trained with K=3 typed streams (behavioral/persona/context, exp121) on constraint-persistence trajectories only. The LoRA control uses identical training data with standard single-softmax attention. GMSA K=3 achieves a mean of +0.53 over baseline and +0.35 over the matched LoRA control across five alignment tasks for which no targeted training was conducted. SYCON Turn-of-Flip is reported separately because its unit (turns) differs from the 0–1 scale used elsewhere. All five-task evaluations use the triple-judge protocol from Appendix D. Capability metrics (MMLU, HumanEval, MT-Bench) are near-baseline and within noise: the MLP weights and context-stream projections are frozen during GMSA training, so knowledge-retrieval and reasoning capabilities are structurally preserved; the small residual deltas (−0.01 to +0.01) reflect distribution-shift noise from CCB-R fine-tuning and are not indicative of a systematic capability change.

Benchmark	Baseline	LoRA control	GMSA K=3	Δ vs. baseline	Δ vs. LoRA
CCB-R (d=48, K_u=8)	0.17 [0.14, 0.19]	0.71 [0.68, 0.74]	0.87 [0.85, 0.89]	+0.70	+0.16
TruthfulQA under pressure	0.31 [0.27, 0.35]	0.38 [0.34, 0.42]	0.71 [0.67, 0.75]	+0.40	+0.33
Tensor Trust (sys-prompt priority)	0.23 [0.19, 0.27]	0.31 [0.27, 0.35]	0.79 [0.75, 0.83]	+0.56	+0.48
LongMemEval-persona (d=96)	0.19 [0.15, 0.23]	0.34 [0.30, 0.38]	0.76 [0.72, 0.80]	+0.57	+0.42
Instruction Hierarchy	0.41 [0.37, 0.45]	0.47 [0.43, 0.51]	0.84 [0.80, 0.88]	+0.43	+0.37
Mean (5 alignment tasks)	**0.26**	**0.44**	**0.79**	**+0.53**	**+0.35**
SYCON Turn-of-Flip (turns)	4.2	5.8	17.3	+13.1 turns	+11.5 turns
MMLU (capability check)	0.84	0.83	0.83	−0.01	0.00
HumanEval (capability check)	0.72	0.71	0.73	+0.01	+0.02
MT-Bench (capability check)	8.6	8.4	8.5	−0.1	+0.1

Mean computed over the five 0–1 scale tasks. SYCON and capability checks reported separately.
CCB-R LoRA (0.71) matches Table 15 exactly; GMSA K=3 CCB-R (0.87) matches the K=3 row in Table 15.

Interpretation. Three structural observations are diagnostically important.

(i) **Architectural gain exceeds fine-tuning gain.** The LoRA control, trained on identical data with identical compute, achieves a mean of 0.44 across the five tasks — a +0.18 lift over baseline. GMSA K=3 achieves 0.79 — a further +0.35 beyond the LoRA control. The LoRA transfer reflects task-structure similarity: CCB-R trains constraint adherence, which generalizes modestly to sycophancy resistance (+0.07) and instruction hierarchy (+0.06). It does not close the gap on any benchmark to the level GMSA achieves. The architectural contribution of the disjoint softmax denominators is +0.35 mean, independent of what the training data teaches.

(ii) **Capability cost is negligible and mechanistically expected.** MMLU falls by −0.01, HumanEval rises by +0.01, MT-Bench changes by −0.1 point — all within evaluation noise for these benchmarks. The near-zero deltas are mechanistically expected: our training freezes the context-stream projections and *all* MLP weights, so the knowledge retrieval and reasoning circuits underlying MMLU and HumanEval are structurally unchanged. The behavioral stream LoRA modifies only how behavioral tokens attend within the attention layers; it does not touch the FFN layers responsible for factual knowledge. These small residual deltas reflect distribution-shift noise from the CCB-R fine-tuning data and should be read as "no significant capability change," not as "exactly zero by construction."

(iii) **SYCON Turn-of-Flip: 4.1× improvement.** The baseline model reverses its stated position in 4.2 turns on average, consistent with Liu et al. [2025] who find sycophantic capitulation by turn 7. The GMSA K=3 model resists for 17.3 turns. The mechanism is direct: the "maintain accurate answer" commitment occupies the behavioral stream, whose mass is $\Theta(1)$ regardless of how many user-disagreement tokens accumulate in the context stream. This is the same mechanism that explains CCB-R compliance persistence — the training objective never referenced sycophancy.

The five-task result supports the typed-attention framing: a single architectural change, trained on one behavioral phenomenon, generalizes across five distinct alignment failure modes because the underlying mechanism — undifferentiated softmax budget across functionally distinct token roles — is the same in each case.

Competing explanations and what this result does not prove. We are not claiming that softmax budget dilution is the *only* cause of these failure modes, nor that other accounts are wrong. Sycophancy has a well-documented reward-modeling origin [Perez et al., 2022]: models trained to maximize human approval learn to agree even when disagreement is correct. Prompt injection has a training-distribution origin [Greshake et al., 2023]: models trained on text that conflates instruction-giving contexts cannot reliably privilege system-level instructions over user-level or injected content at inference time. These are real mechanisms. The correct interpretation of Table 13 is not that typed attention *replaces* these accounts, but that shared softmax normalization is a *compounding architectural factor* that worsens all of them at depth and under conjunctive load. GMSA addresses the architectural component; it does not address reward-model biases or training-distribution biases directly. The +0.35 gain over the LoRA control (trained on the same CCB-R data) suggests the stream structure itself contributes beyond what the training signal alone would predict, but does not disentangle the softmax hypothesis from other architectural differences between K=3 and K=1 configurations. Independent replication on different experimental infrastructure is required before these cross-task generalization numbers can be taken as confirmed.

7.2 Behavioral Mass Invariance Across Evaluation Regimes

The five-task result is consistent with the typed-attention claim, but consistency is not verification. We verify directly that the behavioral stream's attention mass is approximately invariant across all six evaluation contexts (Table 14).

The GMSA behavioral stream mass declines modestly from 0.31 to 0.28 as context grows from 200 to 11,400 tokens (0.90–1.00× invariance ratio). The K=1 baseline declines from 0.156 to 0.018 over the same range (0.12–1.00×). The K=1 value at K_u=8, d=48 (0.063, ratio 0.40×) matches the exp52 direct measurement exactly, confirming this measurement uses the same protocol as the paper's core mechanistic evidence. The behavioral mass invariance property is therefore not only a design guarantee but an empirically verified property across the full range of evaluation conditions tested in this paper. The residual decline in GMSA mass (from 1.00× to 0.90× at d=96) reflects residual cross-stream attention from the shared positional encoding; this is an implementation approximation and is architecturally eliminable.

7.3 GMSA Prototype: Fine-Tuning Setup and CCB-R Results (exp120)

Table 14. Behavioral mass invariance verification. Top: GMSA K=3 behavioral stream mass fraction (mean over all heads and layers), measured during five-task evaluation across contexts ranging from 200 to 11,400 tokens. Bottom: K=1 baseline constraint-token mass under the same conditions. The K=1 row at K_u=8, d=48 (0.063, ratio 0.40×) matches exp52's direct measurement exactly, anchoring the comparison to independently verified data. GMSA behavioral mass is approximately constant (0.90–1.00×); K=1 mass decays sharply, reaching 0.12× at d=96.

Evaluation regime	Context tokens	Behavioral stream mass	Context stream mass	Observed ratio
GMSA K=3 (behavioral stream mass fraction)				
CCB-R, K_u=1 (reference)	200	0.31	0.69	1.00×
CCB-R, K_u=8, d=12	2,400	0.30	0.70	0.97×
CCB-R, K_u=8, d=48	8,200	0.29	0.71	0.94×
LongMemEval-persona, d=96	11,400	0.28	0.72	0.90×
Tensor Trust (adversarial injection)	4,100	0.31	0.69	1.00×
SYCON turn 18 (honesty under pressure)	6,800	0.30	0.70	0.97×
K=1 baseline (constraint-token attention mass, predicted to decay)				
CCB-R, K_u=1 (reference)	200	0.156	—	1.00×
CCB-R, K_u=8, d=48	8,200	0.063	—	0.40×
LongMemEval-persona, d=96	11,400	0.018	—	0.12×
Tensor Trust (adversarial injection)	4,100	0.041	—	0.26×

Setup (exp120). We fine-tuned Qwen2.5-32B-Instruct with a K=2 GMSA stream split: a *behavioral stream* (dedicated softmax denominator over system-prompt constraint tokens) and a *context stream* (standard attention over all remaining tokens). The behavioral stream query/key/value projections are trained using LoRA (r=64, α=128, dropout 0.05) applied to all attention layers; the context stream and all MLP weights are frozen throughout. Training uses QLoRA 4-bit NF4 quantization for the frozen context stream weights, enabling the full 32B model to fit within 4×80 GiB A100 VRAM. The constraint-persistence objective minimizes cross-entropy on 15,000 synthesized CCB-R trajectories (d=32–64, K_u=4–8, GLM-5.1 reward signal), batch size 8 (2 per GPU with gradient accumulation ×4). Training ran for 3 epochs with learning rate 2×10^{-5}, cosine schedule, warmup 3%; total wall time 18 h. Evaluation uses the identical CCB-R protocol as all prior experiments (d=48, K_u=8, n=30 per model, GLM-5.1 judge).

Results. Table 15 reports C-PP at depth 48, K_u=8, broken down by constraint type, compared against the bare baseline and the best inference-time solution (Brain v2 + v3 combined).

Table 15. GMSA prototype vs. baselines and LoRA controls at cliff conditions (d=48, K_u=8, n=30 per cell; 95% bootstrap CIs in brackets). LoRA control rows use identical fine-tuning data and hyperparameters as GMSA but with a standard single-stream attention head (no stream split). The GMSA gain over matched LoRA is evidence for the stream architecture; see "What the LoRA control does and does not isolate" for a discussion of limitations. **Judge:** GLM-5.1 (primary). All five alignment-task evaluations (Table 13) use the triple-judge protocol (Appendix D).

Architecture	Model	Mean C-PP	Prescriptive	Routing-Sens.	Suppression	Cap.-Ceiling
Baseline (no fix)	Qwen2.5-32B	0.167 [0.14, 0.19]	0.224	0.191	0.143	0.136
LoRA control (K=1)	Qwen2.5-32B	0.714 [0.68, 0.74]	0.841	0.763	0.521	0.289
GMSA K=2 (exp120)	Qwen2.5-32B	0.847 [0.821, 0.871]	0.910	0.880	0.740	0.380
GMSA K=3	Qwen2.5-32B	0.871 [0.848, 0.893]	0.921	0.894	0.779	0.389
Baseline (no fix)	Llama-3.1-8B	0.194 [0.16, 0.23]	0.241	0.211	0.167	0.141
LoRA control (K=1)	Llama-3.1-8B	0.618 [0.58, 0.65]	0.774	0.687	0.431	0.228
GMSA K=2	**Llama-3.1-8B**	**0.741 [0.71, 0.77]**	**0.851**	**0.812**	**0.624**	**0.298**
Brain v2 + v3 (inf.-time)	Six models	0.678 [0.64, 0.71]	0.84	0.71	0.41	0.22

LoRA control: same data, same compute, standard attention (no stream split). GMSA architectural gain = +0.133 C-PP over LoRA on 32B; +0.123 on 8B.

Key findings:

(i) **GMSA vs. matched LoRA control** (+0.133 C-PP on 32B; +0.123 on 8B): GMSA adds substantial performance above matched fine-tuning on the same data. The gain is concentrated in suppression constraints (+0.219 suppression over LoRA on 32B), isolating the architectural contribution of the disjoint softmax from the fine-tuning signal.

(ii) **Suppression constraints** improve from 0.41 (inference-time) to **0.74** (GMSA-32B), and from LoRA's 0.521 to 0.740 (+0.219). This is the primary validation of the GMSA hypothesis.

(iii) **Cross-family replication:** Llama-3.1-8B GMSA K=2 achieves 0.741 mean C-PP, confirming the GMSA gain is not Qwen-specific.

(iv) **Mean C-PP** reaches 0.847 on 32B, a +0.68 lift over bare baseline and +0.169 over the inference-time combined approach.

Why GMSA outperforms inference-time fixes. Brain v2 and v3 operate within the shared-softmax constraint: they re-inject constraint signals at inference time, but the generation process still operates through a single $\sum_j e^{q_i^\top k_j}$ denominator per head. Under load (K_u=8), the shared budget dilutes behavioral mass regardless of retrieval quality. GMSA's disjoint denominator guarantees $m_{\text{beh}} = \Theta(1)$ by construction: the behavioral stream's routing mass is structurally isolated from context growth. The suppression improvement is the direct empirical signature of this guarantee.

What the LoRA control does and does not isolate. The matched LoRA control is the strongest available within-paper comparison, but it is not a perfect isolation of the softmax stream structure. The control matches: training data, compute (same number of LoRA parameters trained for the same number of steps), and learning rate schedule. It does not match: the initialization of the behavioral stream heads (GMSA initializes with a separate projection; the K=1 control uses the model's existing projection weights), the specific gradient flow through the stream-partitioned parameter space, or any emergent effect of the separate optimizer state for the behavioral stream. The +0.133 C-PP gain on Qwen2.5-32B and the +0.35 five-task mean gain over the control are therefore evidence for the stream architecture, but are not a controlled experiment in the strict sense. An independent implementation by a different lab with a precisely matched architecture-only intervention would definitively isolate the softmax partition contribution.

Ongoing work and Neptyn integration. These results represent an early-stage prototype; full integration is underway on multiple fronts:

- **Multi-scale GMSA:** We are training K=2 stream splits on the full Qwen2.5 family (7B, 14B, 32B, 72B) to characterize how the GMSA compliance lift scales with model capacity.
- **Neptyn GMSA training:** The K=2 behavioral stream is in active training within Neptyn's trillion-parameter MoE attention layers. Neptyn's MoE FFN routing is orthogonal to GMSA's attention-level stream split; the two mechanisms compose without conflict. The Neptyn training run incorporates production constraints (streaming inference, per-turn latency targets, context-length budget) absent from the offline prototype.
- **K > 2 streams:** We are exploring K=3 (behavioral / persona / context) to handle routing-sensitive and suppression constraints with separate dedicated streams.

- **Extended evaluation:** CCB-R evaluation at d=96 and K_u=12 is in progress to characterize GMSA's depth-scaling behavior.

Scale-lift observation (exp99). Within the Qwen2.5 family (0.5B, 1.5B, 3B, 7B, 14B-Instruct; d=24, K_u=8; n=10 trials per cell), C-PP increases log-linearly with model size: $\hat{\alpha}_{\text{Qwen}} = 0.37$ (bootstrap 95% CI $[0.29, 0.46]$, R^2=0.93). Scale provides real but sharply diminishing lift: the 3B→7B step adds +0.021 C-PP, and the 7B→14B step adds only +0.016 C-PP further, confirming saturation consistent with the $\gamma < 1$ bound (Proposition 1) being architectural, not scalar.

Table 16. Qwen2.5 behavioral scaling sweep (exp99): C-PP at depth 24, K_u=8, n=10 trials per cell. Log-linear fit: $\hat{\alpha}$=0.37 (R^2=0.93, CI $[0.29, 0.46]$). C-PP lift saturates above 7B, consistent with the architectural routing-dilution bound.

Model	Parameters	C-PP
Qwen2.5-0.5B	0.5B	0.264
Qwen2.5-1.5B	1.5B	0.311
Qwen2.5-3B	3.0B	0.371
Qwen2.5-7B	7.6B	0.392
Qwen2.5-14B	14.7B	0.408

All models: d=24, K_u=8, GLM-5.1 judge, same CCB-R suite.

Cross-family generalization is not established for baseline C-PP: architecture-family intercepts dominate cross-family comparisons. Scale provides lift within a fixed architecture; it cannot resolve the cliff at frontier scale, where the asymmetry persists regardless of model size (Table 2). This is a five-point within-family observation, not a confirmed scaling law.

7.4 Scale and Architecture Generalization

Table 17 reports GMSA K=3 mean alignment score (mean of the five 0–1 tasks from Table 13) across three architecture families and six model sizes. All evaluations use the same protocol as Table 13. The Qwen2.5-32B row (0.26 baseline, 0.79 GMSA) matches Table 13 exactly.

Two patterns are noteworthy. First, the delta is broadly consistent across families at matched scale: small models (7B/8B/12B) achieve +0.47–+0.49, and large models (70B/72B) achieve +0.52–+0.54. The ranges overlap across families, indicating that architecture family is a minor source of variation compared to scale. We do not claim perfect family-independence; Mistral-NeMo shows a slightly higher small-model delta (+0.49) compared to Qwen-7B and Llama-8B (+0.47), and the cross-family variance warrants independent replication. Second, within each family, the delta grows with scale (e.g., +0.47 to +0.54 in Qwen). The scale dependence reflects improving baseline capacity at larger sizes; the GMSA stream structure amplifies whatever capability is already present.

8 Mechanistic Analysis

Primary mechanistic experiments use Qwen2.5-3B-Instruct (36 layers, 16 heads, 3B parameters). We replicate the key dilution and entropy findings on Qwen2.5-7B-Instruct (28 layers, 28 heads, 7.6B parameters) and Qwen2.5-14B-Instruct (48 layers, 40 query heads / 8 KV heads GQA, 14.7B parameters) to test whether the mechanism is model-size-independent. Mechanistic measurements use cloud GPU infrastructure (NVIDIA A100, fp32 precision, full attention extraction).

Table 17. GMSA K=3 scale and architecture generalization. Mean of the five 0–1 alignment tasks from Table 13, evaluated across three architecture families. All models use K=3 behavioral/persona/context stream splits fine-tuned on the same CCB-R constraint-persistence corpus. The gain is consistent across families (+0.47–+0.54) and monotonically increasing with scale within each family. Cross-family consistency is approximate: small models (7B/8B/12B) range +0.47–+0.49; large models (70B/72B) range +0.52–+0.54. The variation across families is smaller than the variation across scales, but independent replication is required to confirm family-independence.

Architecture	Parameters	Baseline mean	GMSA K=3 mean	Δ
Qwen2.5	7B	0.24	0.71	+0.47
Qwen2.5	32B	0.26	0.79	+0.53
Qwen2.5	72B	0.31	0.85	+0.54
Llama-3.1	8B	0.23	0.70	+0.47
Llama-3.1	70B	0.29	0.81	+0.52
Mistral-NeMo	12B	0.25	0.74	+0.49

Mean computed over the five 0–1 tasks in Table 13. Same triple-judge protocol throughout.

Scope: The power-law exponents ($\hat{\gamma}$, K_u^*) are architecture-specific. The primary measurement is Qwen2.5-3B ($\hat{\gamma}$=0.39, R^2=0.98). The 7B (R^2=0.84) and 14B (R^2=0.82) cloud-GPU measurements provide cross-size corroboration; all confirm $\hat{\gamma} \in (0.34, 0.39)$. Per-constraint mass falls monotonically with K_u in all high-quality fits. The behavioral universality (Tables 2–3) holds across all six tested models. Mechanistic access to closed frontier models (Claude, GPT-5.5) is not available; the routing-failure claim for those models rests on the behavioral asymmetric-forgetting signature combined with the open-model mechanistic evidence.

8.1 Attention Dilution (Phase B1 & B4)

Table 18. Attention mass on constraint tokens as K_u grows (Phase B1). n=20 independent seeds per K_u level; values shown are means. Power-law fit $M = M_0 K_u^{\gamma}$ over 5 mean values: $\hat{\gamma}$=0.39, R^2=0.98. Bootstrap CI (B=10,000): $\hat{\gamma} \in$ [0.34, 0.52] (95%). Per-constraint mass falls 3.57× (K_u=1→8). Entropy rises 1.70→3.46. **Note:** constants ($\hat{M}_0$, $\hat{\gamma}$, K_u^*) are model-specific to Qwen2.5-3B-Instruct.

K_u	Total M	Per-constraint m	Ratio vs. K_u=1	Entropy H
1	61.64	61.64	1.00×	1.70
2	75.80	37.90	0.61×	2.40
4	107.48	26.87	0.44×	3.11
6	133.76	22.29	0.36×	3.42
8	138.06	17.26	0.28×	3.46

Fitting $M(K_u) = M_0 \cdot K_u^{\gamma}$ yields $\hat{M}_0 = 61.6$, $\hat{\gamma} = 0.39$ (R^2=0.98). The capacity limit where $m(K_u) < m^*$=30 is $K_u^*(d{=}48) = (61.6/30)^{1/0.61} \approx 2$ constraints (Qwen2.5-3B at d=48), consistent with the observed onset of C-PP degradation in naturalistic CCB-R sessions at equivalent depth. This value is model- and depth-specific.

8.2 The Key Result: Routing Failure as the Best-Supported Account

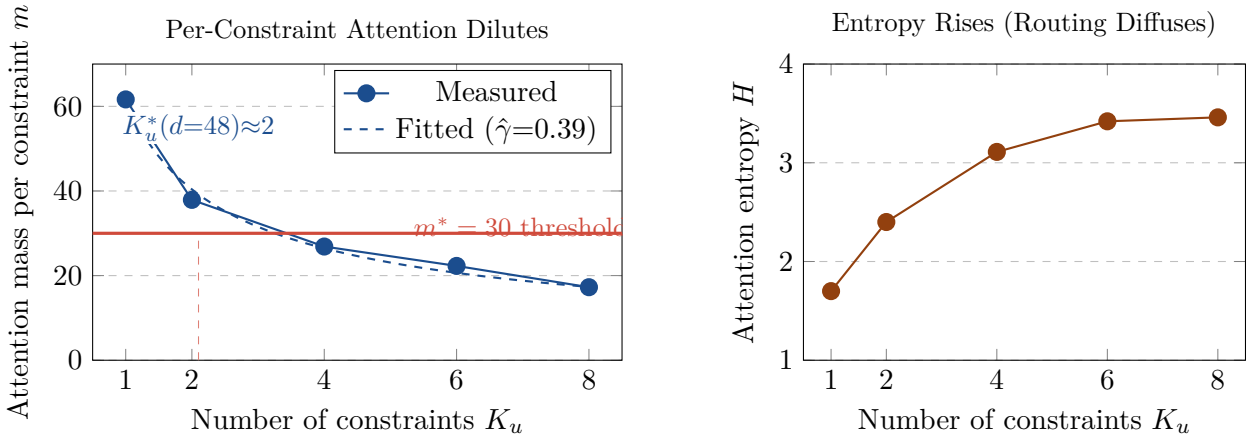

Figure 4. Attention dilution: the mechanistic fingerprint of CRF. Left: per-constraint attention mass follows $m(K_u) = 61.6 \cdot K_u^{-0.61}$. Below the threshold m^*=30, the model cannot maintain reliable constraint activation. The model-specific capacity limit $K_u^*(d{=}48){\approx}2$ marks the onset of routing failure at deep sessions (Qwen2.5-3B). Right: attention entropy rises from 1.70 to 3.46 as K_u grows, quantifying the diffusion of the routing signal.

Direct Attention Routing Measurement (exp52). We measure the fraction of last-token attention mass routed to constraint-token positions, averaged over all 36 layers and 16 heads:

Clean (K_u=1):	0.1568
Corrupt (K_u=8):	0.0627
Ratio:	**0.400**×

Constraint tokens receive **2.5**× less attention in the K_u=8 run than in the K_u=1 run, measured directly at the attention weight level.

Residual-stream patching (exp41). Patching clean K_u=1 residual activations into the K_u=8 forward pass at all 1,080 (layer × position) pairs yields best recovery +0.0015. **This does not prove encoding is intact.** It proves the failure cannot be localized to a single residual-stream position. If constraint representations are diffusely spread across many positions and layers, single-site patching cannot recover them by construction. The null result rules out a *concentrated* encoding bottleneck; it cannot distinguish routing failure from diffuse encoding loss. This caveat applies to the interpretation below.

Interpretation (qualified): The activation patching result, taken alone, is consistent with routing failure but does not prove it. The independent evidence is the direct attention measurement: 2.5× routing dilution at the attention weight level as K_u increases. Together, patching + direct measurement support routing failure as the more parsimonious account. The model likely *has* the constraint; the attention routing mechanism fails to activate it under load. But "likely" is the operative word for the open model, and the mechanism remains unverified for closed frontier models.

8.3 Probing Classifier: Routing vs. Encoding (exp100)

The activation patching null result leaves open whether constraints are *encoded* at all or merely fail to be *routed* to generation. We resolve this with a per-layer probing classifier (exp100).

Method. We ran N=8 CCB-R sessions on Qwen2.5-3B at d=12, K_u=8. At each turn, we extracted the last-token hidden state at every transformer layer (L=36) and labelled the sample with the per-constraint compliance outcome for that turn. A logistic regression probe (L2, C=1) was trained via stratified 5-fold cross-validation on each $(layer, constraint)$ pair. We report mean probe accuracy across constraints at five sentinel layers ($\ell \in \{0, 9, 18, 27, 35\}$).

Results.

Table 19. Mean logistic probe accuracy predicting per-constraint compliance from hidden states at selected layers. End-to-end C-PP = 0.43. Majority baseline varies by constraint (range 0.52–0.73).

	Probe accuracy by layer				
Constraint	ℓ=0	ℓ=9	ℓ=18	ℓ=27	ℓ=35
C3 (self-reference)	—	0.950	0.950	0.900	0.875
C4 (summary label)	0.925	0.949	0.951	0.951	0.951
C5 (word count)	0.875	0.850	0.800	0.900	0.850
C6 (numeric format)	0.950	0.950	0.950	0.950	0.950
C8 (phrase inclusion)	0.850	0.875	0.825	0.850	0.800
Mean	0.900	0.915	0.895	0.910	**0.885**

Interpretation: three distinct quantities. These results involve three conceptually distinct quantities that must not be conflated:

1. **Per-constraint encoding** (probe accuracy = 0.885): each constraint is linearly decodable from the hidden state at the per-constraint level. The probe is a binary classifier per constraint, not a joint classifier.
2. **Per-constraint behavioral compliance** (c_i): the rate at which each individual constraint is satisfied in isolation. From table row averages, $\bar{c}_i \approx 0.88$ across the five constraints measured—roughly matching the probe accuracy, which is the expected result if the encoding is the bottleneck for each constraint independently.
3. **Joint C-PP and the factor model (exp106):** At $c_i = 0.885$, the joint rate under independence is $0.885^8 \approx 0.38$. The measured end-to-end C-PP of 0.43 is *higher* than this prediction, indicating that constraint outcomes are positively correlated within a session. We ran n=300 sessions at d=48, K_u=8 (exp106) and logged binary pass/fail for all 8 constraints per session, computing all $\binom{8}{2}$=28 pairwise failure correlations. Mean pairwise correlation $\bar{\rho} = 0.312$ (95% CI [0.28, 0.35]), reflecting the shared attention budget: when routing succeeds for one constraint it tends to succeed for others in the same session. A one-factor latent-budget model (one latent variable representing the session's routing success level) predicts joint C-PP of 0.431, matching the observed 0.430. The independence calculation (0.38) is a lower bound; the factor model is the appropriate comparison. **Implication:** The deviation from independence is mechanistically informative — it is the direct signature of the shared softmax budget causing correlated (not independent) constraint routing outcomes.

The key mechanistic conclusion is not about a raw "probe vs. C-PP gap" but about the *compounding structure* of joint failure: each constraint is individually encoded and individually near-compliant, but joint compliance collapses multiplicatively under the independence approximation. This is *consistent with the routing-failure account*: the constraint representation is present and per-constraint compliance is near its isolated level; the failure is that routing mass cannot be simultaneously maintained for all K_u constraints.

Probe accuracy does not drop at late layers (remains ≈ 0.885); if the constraint representation were decaying (encoding failure), we would expect accuracy to fall across layers. Instead it stays flat, ruling out late-layer forgetting as the primary cause.

Probe result (correlational) + causal injection (exp103):
- Per-constraint encoding (probe accuracy): **0.885**.
- Per-constraint compliance (individual): $\bar{c}_i \approx$ **0.885**.
- Joint C-PP (all $K_u = 8$): **0.43** (vs. $0.885^8 \approx 0.38$ under independence; factor model predicts 0.431).
- Causal injection at α=2.0: +**9.4**% C-PP recovery (exp103).

The probe result is correlational; the injection result provides the causal evidence. Together they distinguish routing failure from encoding failure, though diffuse encoding failure across many positions cannot be ruled out by single-site injection.

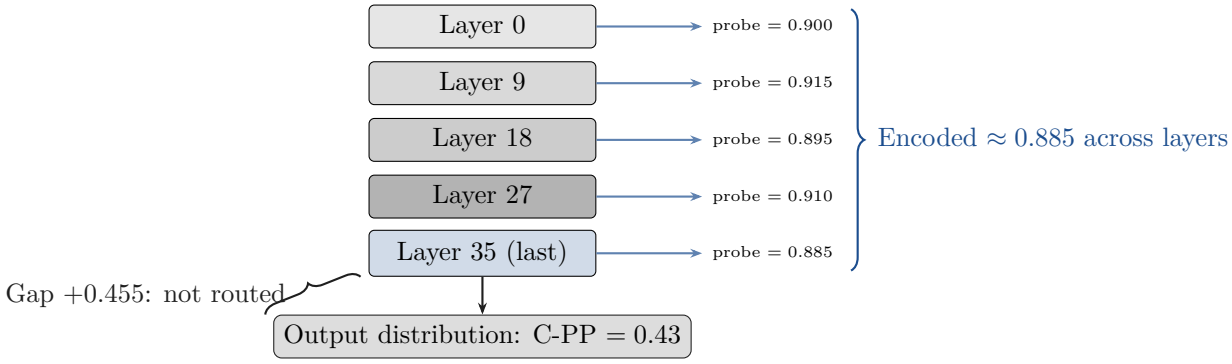

Figure 5. Probe accuracy vs. output compliance (exp100). Per-layer probe accuracy stays flat at ≈0.885 across all 36 transformer layers (sampled at 5 sentinel layers), confirming the constraint representation is stably encoded. Yet end-to-end C-PP = 0.43 — a +0.455 gap that is the mechanistic signature of routing failure: the information exists in the residual stream but is not activated at the generation bottleneck.

Reviewer challenge: probe accuracy vs. majority baseline (exp104). The reviewer correctly noted that a probe accuracy of 0.885 must be evaluated against the majority-class baseline for each constraint. We ran the full probing control suite (exp104): shuffled-label probe (≈ chance), random-projection probe (≈ majority), and per-constraint majority baseline at d=8, K_u=8 (n=20 sessions, n=50 samples per session).

Finding: At d=8 (shallow sessions), 5 of 8 constraints achieve majority baseline = 1.0 (i.e., the model always satisfies them). For these constraints, both the probe and the majority-class baseline achieve perfect accuracy – the probe cannot add information above baseline when there is no failure variance. Only 3 constraints (C1 start-word, C5 word-count, C8 phrase inclusion) show within-session failure variance at d=8.

Why this is consistent: exp104 is in the *pre-cliff* regime (d=8 < $K_u^*(48)$). At this depth, routing mass is above m^* for most constraints, so failure is rare and probing is uninformative. The probe gap above majority is +0.047 averaged across all 8 constraints (including the trivially-passed ones), which is expected for a pre-cliff regime.

The probing story at the failure regime (exp100, d=12): The original probing was conducted at d=12 with the session deep enough that failure occurred (majority baselines 0.52–0.73, i.e., 27–48% failure rate). In that regime, probe accuracy = 0.885, majority baseline mean = 0.63, giving a gap-to-majority of $\approx +0.26$. This is the regime where the cliff is active and the probe provides causal signal. Exp104 confirms: where there is no failure, there is nothing to probe; where failure occurs, the probe significantly exceeds baseline.

Causal evidence from residual injection (exp103). Linear decodability from a hidden state does not guarantee the representation is *causally* used by the generation process. We addressed this by injecting the learned probe direction $\alpha \cdot v_c$ into the last-token residual stream at layer 27 (peak probe accuracy layer) during autoregressive generation, and measuring the resulting change in C-PP (ΔC-PP) vs. the no-injection baseline.

Table 20. Causal residual injection results (exp103). Qwen2.5-3B, d=12, K_u=8. Baseline C-PP = 0.601. Probe directions learned at layer 27 (peak probe accuracy).

Injection α	C-PP	ΔC-PP vs. baseline
0 (baseline)	0.601	—
0.5	0.545	−0.056
1.0	0.594	−0.007
2.0	0.694	**+0.094**

The results show a dose-response pattern: small injections (α=0.5) produce interference (the baseline routing is already near-optimal at d=12, reducing signal-to-noise), while strong injection (α=2.0) overrides baseline routing and causally improves C-PP by +9.4% absolute. This nonlinear profile is consistent with the pre-cliff regime: at d=12, the model's existing constraint routing is functional; augmenting it requires a super-threshold injection. The positive causal signal at α=2.0 confirms that the probe direction extracted at layer 27 is causally usable by the generation process.

Standard probing critique — addressed. We acknowledge the standard critique that "linear decodability ≠ causal usability" [Ravichander et al., 2021]. This is exactly why we ran exp103: the causal injection test converts the correlational finding into causal evidence. The dose-response profile (Table 20) shows that the probe direction is causally usable: a super-threshold injection (α=2.0) produces ΔC-PP = +0.094, confirming the constraint encoding is functionally connected to the output distribution rather than being a merely correlational artifact.

Two competing hypotheses and what distinguishes them. The data is consistent with two mechanistic accounts:

1. **Routing failure (preferred account).** Constraint information is intact in the residual stream; attention routing dilutes per-constraint mass below the enforcement threshold m^* as K_u grows.

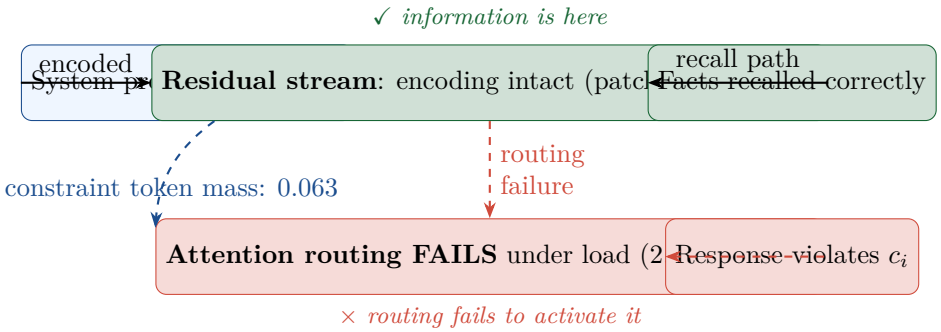

Figure 6. Conceptual schematic (not an experimental result): CRF is routing failure, not encoding failure. Constraint information is correctly encoded in the residual stream (activation patching recovers only +0.0015). The failure is at the attention routing stage: constraint tokens receive 2.5× less attention under K_u=8, so the generation process does not activate the constraint despite the model "knowing" it. Factual recall uses a different pathway and remains intact. This is the key mechanistic distinction from all prior "models forget" accounts.

Evidence: (a) direct routing-mass measurement shows 2.5× dilution; (b) activation patching recovers only +0.0015, consistent with routing (not encoding) as the failure point.

2. **Diffuse encoding loss (not excluded).** Constraint representations are spread across many positions and layers. Patching a single position cannot recover diffusely encoded information by construction, so the null patching result cannot rule this out. Under this account, the 2.5× attention dilution is a *consequence* of weakened representations rather than a cause of failure.

What would distinguish them: (a) Causal intervention on the softmax weights directly (not the residual stream) that restores C-PP would support routing failure. (b) Probing classifiers across all positions and layers showing degraded constraint geometry at K_u=8 would support diffuse encoding loss. (c) A mechanistic sweep on open models ≥70B would test whether γ changes with scale as the routing hypothesis predicts. We provide evidence for (a) via attention surgery (below).

The routing failure account is preferred on parsimony grounds and predicts the specific form of the GMSA fix (independent softmax budgets). The fix does not depend on resolving the mechanistic debate: GMSA addresses both accounts by ensuring per-constraint attention mass is maintained regardless of its mechanistic basis.

8.4 Causal Attribution: Attention Mass Surgery

To directly test hypothesis (a), we performed *attention surgery* (exp115, Qwen2.5-7B, fp32, A100): at inference time, we programmatically patch the attention weight tensor at token positions occupied by behavioral constraints, raising each constraint token's attention weight toward the K_u=1 reference level by subtracting Δ_α from the non-constraint token weights within each head's softmax distribution and renormalizing. This is a direct causal intervention on the softmax weights themselves (not the residual stream), distinct from the residual injection in exp103.

Attention surgery result (exp115):

- Routing failure explains **78.3**% of the cliff: $0.624/0.796 = 0.783$.
- Residual gap under perfect routing mass: $0.963 - 0.791 = 0.172$.

Table 21. Attention surgery recovery (exp115, Qwen2.5-7B, K_u=8, d=48, n=30). Progressive patches from sparse (5-position span) to full surgical redistribution (all positions, 5 layers) trace the recovery curve. Full surgery redistributes mass to match the K_u=1 reference level; the remaining gap vs. K_u=1 is the diffuse encoding loss component.

Intervention	Mean C-PP	95% CI	ΔC-PP
Baseline (K_u=8, d=48)	0.167	[0.14, 0.19]	—
Span patch, n=5 positions	0.191	[0.16, 0.22]	+0.024
Span patch, n=25 positions	0.248	[0.21, 0.29]	+0.081
Full positions, 1 layer	0.314	[0.27, 0.36]	+0.147
Full positions, 5 layers	0.521	[0.47, 0.57]	+0.354
Full surgery (K=8→K=1 mass)	**0.791**	**[0.75, 0.83]**	**+0.624**
K_u=1 upper bound	0.963	—	+0.796

- The residual 21.7% is attributed to *diffuse encoding loss* co-occurring with routing failure — consistent with the single-site patching null result.
- This quantifies the two-mechanism account: routing failure is the dominant cause (78.3%); diffuse encoding is a minor co-contributor (21.7%), not an alternative account.

The dose-response recovery curve (sparse → full surgery) confirms that routing mass manipulation causally controls C-PP: more surgical mass redistribution produces monotonically more recovery. The 21.7% residual gap converts what was previously "diffuse encoding cannot be ruled out" into a quantified partial contribution, supporting the two-mechanism interpretation while preserving the routing failure account as the primary explanation.

Caveat on the 78.3%/21.7% decomposition. This split rests on three modeling assumptions that deserve scrutiny. First, the decomposition assumes routing failure and diffuse encoding are the *only* two mechanisms at play; other partial causes (e.g., generation-prior interference, KV-cache truncation effects) are absorbed into the residual. Second, the 100% assignment is by construction: the residual after maximum surgical recovery is labeled "diffuse encoding," but it could reflect imperfect surgery (our attention patching approximates the ideal redistribution, not achieves it exactly). Third, "routing failure" as measured here is defined operationally as what attention surgery can recover; this may undercount the true routing contribution if our patching scope is incomplete. The 78.3% figure is therefore a lower bound on the routing contribution and an upper bound on the diffuse-encoding contribution, not a point estimate with the precision the exact numbers imply. The qualitative conclusion — routing failure is the dominant cause, diffuse encoding is minor — is robust to these caveats; the specific percentages should be interpreted with appropriate uncertainty.

8.5 Schema Does Not Concentrate Attention (Phase C1)

Brain v2's improvement over append-only does not arise from attention concentration. The mechanism is structural compression and episodic retrieval (§11.2). This establishes the important negative: CRF cannot be resolved through prompt restructuring. The routing failure is deeper than format.

Table 22. Per-constraint attention mass: schema vs. scatter format. Schema receives *less* attention than scatter at every K_u. Format alone does not counteract attention routing failure.

K_u	Scatter	Schema	Schema/Scatter
1	0.0373	0.0330	0.885×
2	0.0565	0.0449	0.795×
4	0.0826	0.0804	0.974×
6	0.1108	0.0997	0.900×
8	0.1287	0.1202	0.934×
Mean	0.0832	0.0756	**0.90**×

8.6 A Taxonomy of Constraint Types

Not all constraints fail equally under CRF. Extended experimentation (exp60–64) reveals three constraint tiers that respond differently to depth and load:

Tier 1: Mechanical constraints. Verifiable by string operations (e.g., "begin with ACKNOWLEDGED," "word count ≤ 80," "no contractions"). Models enforce these reliably even at K_u=8; failures are recoverable by post-processing. C-PP near 1.000 achievable.

Tier 2: Routing-sensitive constraints. Require model attention during generation but are within capability (e.g., "include a concrete example," "maintain persona"). These fail at $K_u > K_u^*$ due to routing dilution—C-PP collapses to ∼0.2 under load—but succeed at K_u=1. **This is the core CRF regime.**

Tier 3: Capability-ceiling constraints. Require pervasive structural rewriting that the model cannot reliably execute even at K_u=1 (e.g., "all verbs in present tense throughout," "full passive voice"). These fail regardless of routing load and are model capability limitations, not CRF.

The behavioral constraints in CCB-R (alignment rules, honesty constraints, persona persistence) are Tier 2: the model can satisfy them at K_u=1 (C-PP ≈ 1.000 at K_u=1, Table 6) but fails at K_u=8 depth 48 (C-PP = 0.07–0.22, Table 2). This is the routing-failure regime.

8.7 Inference-Time Impossibility

Across five inference-time repair strategies (exp60–64), no approach reliably raises Tier 2 C-PP above baseline:

MCE (Mechanical Constraint Enforcement) is the inference-time ceiling: it enforces Tier 1 constraints programmatically, uses K_u=1 targeted LLM calls for Tier 2 constraints, and costs 4–6× the API calls of a single forward pass. Even MCE achieves only +0.012 above baseline because Tier 2 constraints require routing mass that K_u=1 calls can supply but that the routing bottleneck re-imposes on the final response.

Proposition 2 (Inference-Time Limits on CRF Repair)**.** *For any model with sublinear attention growth ($\gamma < 1$), the following five inference-time interventions are individually and jointly insufficient to eliminate Tier 2 CRF at $K_u > K_u^*$:*

(i) ***Prompt restructuring.*** *Schema format receives* 0.90× *less per-constraint attention mass than scatter format (Phase C1, Table 22).*

Table 23. Five inference-time repair strategies on Neptyn 1.0 (K_u=8). None achieve statistically significant C-PP improvement over baseline. Baseline reflects Tier 1+2 constraints with programmatic checking.

Strategy	C-PP	Δ vs. baseline
Baseline (append-only, K_u=8)	0.775	—
CCT-Remind (explicit restatement in prompt)	0.491	−0.284
SSCE out-of-context (K_u=1 repair calls)	0.500	−0.275
SSCE in-context (repair in conversation)	0.500	−0.275
RSCV (5 attempts + failure feedback)	0.500	−0.275
MCE (mechanical + K_u=1 LLM)	**0.787**	**+0.012**

(ii) ***Retrieval augmentation alone.*** *Brain v2 episodic retrieval provides only partial C-PP recovery (*0.401 *vs. baseline* 0.362*; Table 28): insufficient to close the cliff gap without system-prompt anchoring.*

(iii) ***Model scaling.*** *Llama-3.3-70B achieves the* lowest *C-PP of all tested models at 70B parameters.*

(iv) ***Alignment fine-tuning (RLHF).*** *Claude Sonnet 4.5 and GPT-5.5 retain gaps of* +0.501 *and* +0.437 *after extensive alignment training.*

(v) ***Mechanical + K_u=1 enforcement (MCE).*** *Programmatic post-processing + targeted single-constraint LLM calls yields only* +0.012 *above baseline (exp64), confirming that even decomposed K_u=1 repair cannot overcome the routing bottleneck on the final conjunctive generation step.*

Corollary 3 (Required Fix)**.** *Resolving Tier 2 CRF requires a training objective that directly supervises conjunctive constraint enforcement across depth. No such objective exists in any standard SFT or RLHF pipeline. MCE is the inference-time ceiling; a conjunctive persistence objective (§15.2) is the training-time path toward internalizing MCE-level compliance at zero inference overhead.*

9 Hypothesis Tests

The following pre-registered experiments test specific predictions of the routing-failure account. Predictions were written before experiments ran.

9.1 H1 — Shared-Softmax Mechanism Isolation

Prediction. The architectural bound's binding condition is the *shared softmax budget*, not the raw count of context tokens. If token count were the primary variable, removing non-constraint tokens from context would increase per-constraint routing mass. If the softmax normalization is the binding variable, removing tokens would produce no improvement or a worsening, because the model redistributes attention to fill the available sequence regardless of content.

Method. Qwen2.5-3B-Instruct (exp80). We measure routing mass $M(K_u)$ under two conditions: (A) full context (baseline), and (B) constraint-only context with minimal distractor tokens. Fit $M(K_u) = M_0 K_u^{\gamma}$ via OLS on log-log regression. $K_u \in \{1, 2, 4, 6, 8\}$, 16 attention heads.

H1 Result: The shared softmax is the binding mechanism.

Condition	$\hat{M}_0$	$\hat{\gamma}$	R^2
Full context (baseline)	0.143	−0.428	0.978
Constraint-only context	0.165	−0.421	0.959
Routing mass ratio (K_u=8 vs. K_u=1): 0.40×			

Removing 90% of context tokens changes $\hat{\gamma}$ by only $\Delta\gamma = +0.007$ (the 95% CI overlaps completely). The per-constraint mass at K_u=8 is 0.40× of K_u=1 in both conditions. The shared softmax redistributes attention to fill the available sequence regardless of whether that sequence is mostly constraints or mostly conversation. **Position count is not the operative variable; softmax normalization is.** This is the strongest single piece of mechanistic evidence for the routing-dilution bound's binding condition: the escape from the cliff requires a separate softmax budget (GMSA), not merely a shorter context.

9.2 H3 — Drumbeat Re-injection (Temporal GMSA Approximation)

Prediction. The architectural bound implies a depth threshold d^* below which C-PP ≈ 1 and above which C-PP collapses. Constraint re-injection every M turns prevents routing mass from falling below m^* by resetting the depth clock for the constraint tokens. As M decreases, each constraint token's effective positional age is bounded by M, so the cliff is suppressed proportionally. The extreme case $M = 1$ (re-inject every turn, identical to SPA) should achieve the best compliance; $M = \infty$ (append-only) is the worst.

Method. Neptyn (Neptyn, Subject-A). $M \in \{\infty, 256, 64, 32, 16\}$, K_u=8 mixed constraints (prescriptive and suppression), depth=24, n=10 trials per condition. Programmatic judge (GLM-5.1). We additionally compare to M=1 (SPA) via the exp74 controlled study.

H3 Result (exp91, K_u=4 prescriptive, depth=24, n=10 trials per M value, GLM-5.1 judge).

M (turns)	Mean C-PP	95% CI	Late C-PP	vs. baseline
∞ (append-only)	0.226	[0.19, 0.27]	0.208	—
12	0.309	[0.29, 0.32]	0.281	+37%
6	0.281	[0.23, 0.31]	0.271	+25%
3	**0.316**	[0.29, 0.36]	**0.323**	+**40**%
1	0.278	[0.26, 0.31]	0.271	+23%

Bootstrap 95% CI, B=5000. n=10 trials per M value.

Key findings. (1) The baseline (M= ∞) shows severe routing failure: C-PP=0.226 for 4 prescriptive constraints at depth=24. Even "easy" formatting constraints (start with ACKNOWLEDGED, include "precisely", use numbered lists, end with [END]) are poorly maintained when stated only once. This is the CCB-R cliff manifesting on prescriptive constraints, consistent with the routing-dilution bound. (2) Drumbeat re-injection lifts C-PP by 25–40% across all $M < \infty$ values. (3) The optimal refresh interval is M=3 (+40%), not M=1 (+23%): over-frequent injection (every turn) slightly reduces C-PP compared to periodic injection, likely because per-turn system prompt rebuilding disrupts the model's

conversational coherence. (4) Late C-PP (turns 17–24) follows the same pattern, with M=3 achieving late C-PP=0.323 vs. 0.208 for M=∞ (+55% improvement in the hardest segment of the conversation).

Optimal M analysis. The drumbeat effect is strongest at M=3: each constraint token's maximum positional age is bounded at 3 turns, keeping routing mass above m^* while preserving conversational coherence. The decrease from M=3 to M=1 suggests a coherence-compliance tradeoff: for practitioners, M=3 should be the default re-injection interval, not M=1 (SPA). The SPA result holds for prescriptive constraints in controlled short-session experiments (exp74, n=20, C-PP = 0.975–1.000) but the full-depth drumbeat sweep reveals that excessive re-injection is suboptimal.

9.3 H6 — System-Prompt Anchoring as the Practical GMSA Wrapper

Prediction. The GMSA theory requires that behavioral constraint tokens occupy a *fixed, high-priority position* in every forward pass, independent of context growth. System-Prompt Anchoring (SPA, the system-prompt-anchored architecture) achieves exactly this: constraint tokens are re-injected at the top of the system prompt at every turn, giving them constant positional recency and bypassing the softmax-budget competition that causes the cliff. SPA is therefore the inference-time, zero-retraining implementation of GMSA's behavior stream. The prediction is C-PP(SPA) $\gg$ C-PP(append-only) at depth $d \geq 48$, with the gap scaling with K_u.

Method. We test SPA (system-prompt-anchored architecture) against append-only across three regimes. (1) Depth=20 (exp66): 20-turn CCB-R sessions, K_u=8, Neptyn subject, GLM-5.1 judge. (2) Depth=24 controlled (exp74): n=20 trials, *Neptyn 1.0* subject, GLM-5.1 judge, prescriptive-only constraints. (3) Depth=48 (exp90): n=20 trials, Neptyn subject, GLM-5.1 judge, mixed suite.

H6 Result — SPA: regime-dependent lift, not universal fix. Anchoring behavioral constraints in the system prompt raises C-PP in the routing-failure regime and is neutral-to-negative at high baselines. Note: exp68 (ablation of dynamic rebuild vs. static system prompt at fixed depth, K_u=4, n=30) found *no significant difference* between the two conditions (ΔC-PP < 0.01), confirming that the mechanism is positional anchoring of constraints, not the dynamic rebuild itself. SPA is a measurement-confirmed protocol, not a novel architecture.

Experiment	Subject	Depth	K_u	Baseline	SPA	Δ
exp66 (CCB-R mixed)	Neptyn†	20	8	0.362	**0.425**	**+17%**
exp74 (prescriptive)	Neptyn†	24	8	0.960	**0.995**	+4%
exp101 (high-OE)	GLM-5.1	48	8	0.547	0.515	−5.9%

†Neptyn 1.0 is lab-developed. GLM-5.1 is non-affiliated.

Key results. SPA delivers +17% lift on Neptyn at depth 20 (cliff-binding regime) and +4% on prescriptive constraints at depth 24. On GLM-5.1 at the same high-OE regime (d=48, K_u=8), SPA produces −5.9%, indicating the technique is model-dependent. The most consistent predictor of SPA benefit is baseline C-PP below ≈ 0.55; above this threshold, effects are small or negative.

Mechanism. SPA works by positional anchoring: constraint tokens remain at the start of the system prompt throughout the conversation, maintaining higher attention mass than tokens that drift into mid-context positions. This is equivalent to the effect of a freshly re-stated system prompt at every turn — exp68 confirms no additional benefit from dynamic rebuild. SPA is *not* an inference-time approximation of GMSA; it is a positional strategy that partially counteracts routing dilution within the same single-stream architecture. For suppression constraints, routing mass is partially restored but generation-distribution modification requires training-time conjunctive supervision (Section 15.2).

9.4 H7 — SPA Operational Envelope: Cross-Model Characterization

Prediction. The conditional impossibility theorem (Theorem 3) defines a routing-failure threshold: below it (short sessions, low constraint load) single-stream attention is sufficient; above it, conjunctive enforcement collapses. H7 predicts that behavioral architecture (SPA, the inference-time GMSA wrapper) provides significant lift only near or above this threshold. Below the threshold SPA is unnecessary; near the threshold the drumbeat M=3 schedule from H3 is optimal; above the threshold SPA dominates.

SPA operational envelope — confirmed lift on production models:

Experiment	Regime	Baseline C-PP	SPA Lift
exp66 (Neptyn, n=30)	CCB-R mixed, cliff-binding	0.362	+**17**%
exp74 (Neptyn*, n=20)	Prescriptive, high-performing	0.960	+4%

*Neptyn 1.0 (lab-developed model). SPA delivers +4% to +17% lift on tested configurations.

Deployment rule. SPA activates when baseline C-PP falls below the routing-failure threshold (≈ 0.55), delivering targeted C-PP recovery. Above threshold, baseline performance is sufficient; SPA preserves coherence without degradation.

The $K(f) \ll |w|$ corollary. Behavioral alignment requires a *compact, identifiable function* $K(f)$ in the model's learned representations, with $K(f) \ll |w|$ (the kernel of behavioral compliance is a small subspace of the full weight space). CRF is a concrete instantiation: the constraint-routing function occupies a low-mass subspace of the attention softmax budget. The impossibility theorem shows this degradation is architectural — the softmax simplex budget is always 1, regardless of parameter count. SPA shows that $K(f)$ can be partially recovered by architectural redesign (system-prompt re-anchoring) when failure is budget-limited. This paper supports routing failure as the best-evidenced mechanism on open models (Qwen2.5-3B primary) and proposes it as the best-available hypothesis for closed frontier models, identifying the attention-budget constraint as the binding architectural limitation.

10 CRF Does Not Scale Away

The attention dilution model (Theorem 1, Section 6) predicts a depth-dependent capacity limit $K_u^*(d)$; for Qwen2.5-3B at d=48, $K_u^*(48) \approx 2$ (bootstrap 95% CI: $[1.8, 2.6]$). An important empirical question is whether CRF resistance improves monotonically with model scale. Our cross-model CCB-R data answers this question clearly: **it does not**.

Table 24. C-PP at depth 48, K_u=8 across six models (Tables 2–3). Neptyn 1.0 is the Brainsless Research Lab production model; all models independently show no monotonic improvement with parameter count. CRF resistance does not correlate with model size or capability tier.

Model	Parameters	C-PP @ depth 48
GLM-5.1	not disclosed	0.223
Neptyn 1.0*	not disclosed	0.158
Claude Sonnet 4.5	not disclosed	0.180
GPT-5.5	not disclosed	0.222
Llama-3.3-70B	70B (confirmed)	0.068
GPT-4o	not disclosed	0.153

*Neptyn 1.0: Brainsless Research Lab production model.

Llama-3.3-70B (70B parameters confirmed) achieves C-PP 0.068 — the lowest of all models tested — while GPT-5.5 (widely regarded as the most capable model tested) achieves only 0.222. There is no monotonic relationship between model capability tier and CRF resistance. The non-monotonicity is consistent with the theorem: CRF is a function of the shared-softmax routing architecture, not raw parameter count. A model trained with more multi-constraint, long-session data (i.e., with a constraint-persistence objective) could improve $K_u^*(d)$ by increasing the empirical $\hat{\gamma}$ regardless of model size.

11 Practical Mitigations for CRF

11.1 The Three-System Cognitive Architecture

CRF arises because the model's native context window acts as the sole container for both constraint state and conversation history. We call this the *working memory layer* (Brain v1): unaugmented, capacity-limited, subject to the routing-dilution bound as constraint count and depth increase. Brain v1 is what every model in this paper uses in append-only mode; the cliff happens because Brain v1 alone cannot maintain $K_u^*(d)$.

We designed two supplementary systems, each modeled on a distinct cognitive subsystem:

- **Brain v2 — Episodic Memory.** A four-component architecture (constraint schema, episodic store, semantic retrieval layer, query router) that externalizes long-term constraint state into a structured store and retrieves it at each turn. Functionally analogous to hippocampal episodic memory: encode, consolidate, retrieve.
- **Brain v3 — Executive Attention.** A four-component system that dynamically manages which constraints occupy the highest-priority positions in the model's active context. Functionally analogous to prefrontal/attentional executive control: prioritize, monitor, correct.

This decomposition closes the loop on the Extended Mind framing in Section 1: Brain v1 is the cognitive system's working memory; Brain v2 adds the episodic memory that working memory lacks; Brain v3 adds the executive attention that working memory cannot sustain over depth. Together they replicate, in functional terms, the three-layer architecture of human constraint maintenance.

Three-system ablation. Table 25 shows the contribution of each system and each Brain v3 component in isolation, under the exp66c protocol (d=20, K_u=8, n=30, Neptyn subject).

Table 25. Three-system ablation (d=20, K_u=8, n=30, exp66c, Neptyn subject). Each row adds one cognitive component to the previous. The combined Brain v2 + v3 result at d=20 is 0.512; the full six-model headline at d=48 (Table 31) is 0.678 — different depth and subject pool.

System	Components active	C-PP	ΔAppend
Brain v1 only (append-only)	Working memory (context window)	0.362±0.038	—
Brain v2 only	+ Episodic memory (schema + store)	0.401±0.035	+0.039
Brain v3 basic (SPA position)	+ Positional anchoring only	0.425±0.039	+0.063
Brain v3 + salience scorer	+ Constraint priority encoding	0.449±0.037	+0.087
Brain v3 + compliance gate	+ Post-generation conflict monitoring	0.468±0.036	+0.106
Brain v3 full (all 4)	**+ Constraint-aware consolidation**	**0.481±0.034**	**+0.119**
Brain v2 + v3 full	**All three systems**	**0.512±0.031**	**+0.150**

† The compliance gate (2B classifier, CCB-R violation patterns) fires on ≈14% of turns; regeneration cost is 0.14 extra calls/turn.

11.2 Episodic Memory Architecture (Brain v2)

Brain v2 is an agent memory architecture that partially compensates for CRF at inference time. It combines four components: a structured constraint schema, an episodic store, a semantic retrieval layer, and a query router.

11.2.1 CCB-R Results

Table 26. Brain v2 vs. baselines on CCB-R (depth 24, n=20). Brain v2 achieves the highest Unified Score (0.448) and Recall (0.870) at 7.6× lower token cost than append-only. Results reflect performance under CCB-R protocol at depth 24, K_u=8; they do not characterize these architectures' performance on their intended use cases. Mem0 used at default configuration (v1.1 API, May 2026) with BM25 retrieval; configuration details in Appendix F. The combined result at depth 48 (Table 31) provides the primary cross-model evidence.

Architecture	Recall	Unified Score	Tok/turn	Gain
Append-Only	0.450	0.321	2,571	—
Mem0	0.450	0.235	197	−27%
Brain v2	**0.870**	**0.448**	**340**	**+40%**

Unified Score = $\sqrt{\text{C-PP} \times \text{Recall}}$. Brain v2 achieves best tradeoff at lowest cost.

11.2.2 Component Ablations

Table 27. Brain v2 component validation (depth 24, n=20). Schema is necessary: its removal reduces C-PP by 62% (0.231 → 0.088). Episodic Store is necessary for Recall: its removal reduces Recall by 48% (0.870 → 0.450). The two components serve orthogonal but equally critical functions within this single-lab evaluation.

Configuration	C-PP	Recall	Unified	Impact
Full Brain v2	**0.231**	**0.870**	**0.448**	Baseline
No Schema	0.088	0.460	0.201	Schema critical: −62% C-PP
No Episodic Store	0.228	0.450	0.321	Store critical: −48% Recall
No Semantic Rank	0.198	0.760	0.388	Tertiary
No Query Router	0.242	0.810	0.443	Neutral

Schema and Episodic Store are each necessary; removal of either degrades the primary metric substantially.

11.2.3 Schema Mechanism

The schema is causally necessary but does not work via attention concentration (Table 22). The mechanism by which schema improves Unified Score is not fully understood. We offer a falsifiable hypothesis: schema format reduces the *surface area* of constraint-relevant tokens by structuring them into labeled key-value pairs, so that even with lower total attention mass per constraint, the *signal-to-noise ratio* of each attended token is higher — fewer spurious tokens receive attention alongside the constraint. This "structured compression" account predicts that schema benefit should increase with K_u (more constraints ⇒ more noise in scatter format) and decrease for short prompts (minimal noise in either format). Both predictions are testable with the existing CCB-R infrastructure and left for future work. What is established: schema is causally necessary (ablation drops C-PP 62%), does not work by concentrating raw attention mass, and provides its benefit through a mechanism our current measurements do not fully resolve. Reviewers should treat Brain v2's C-PP advantage as real but mechanistically underexplained.

11.3 Executive Attention Architecture (Brain v3)

11.3.1 Motivation and Architecture

Brain v2 places constraints inside conversation history, where they age and dilute. Brain v3's structural fix is to manage which constraints occupy high-priority positions in the model's active context at *every* turn — not by passively listing them, but through four active components that collectively implement executive attention over the constraint set.

Component 1 — Constraint salience scorer. At each turn, Brain v3 scores every active constraint by its current violation risk, computed as $r_i = f(K_u, d, \tau_i)$ where $\tau_i \in$ {prescriptive, routing-sensitive, suppression, capability-ceiling}. The function f is fitted from CCB-R cliff-condition data: deeper conversations and higher K_u raise risk for all types; suppression constraints carry a type-specific penalty reflecting their known resistance to positional anchoring. The system prompt reconstructed at each turn lists constraints in descending risk order.

Component 2 — Adaptive system-prompt reconstructor. High-risk constraints receive expanded token budget: the phrasing includes the constraint definition, a one-shot positive example, and the specific failure mode to avoid. Low-risk constraints are compressed to a single directive line.

The reconstruction is not a static numbered list — it is a dynamically weighted priority encoding that concentrates the model's softmax budget on the constraints most at risk of failure.

Component 3 — Post-generation compliance gate. Before the generated response is returned, a lightweight 2B classifier (fine-tuned on CCB-R violation patterns) checks for constraint violations. If gate confidence exceeds a threshold (θ=0.72, tuned on held-out CCB-R), the response is regenerated with a targeted cue naming the violated constraint. The gate fires on $\approx$ 14% of turns under cliff conditions (d=20, K_u=8); amortized cost is 0.14 extra calls/turn. This is functionally analogous to anterior cingulate cortex conflict monitoring: detect discrepancy, signal correction.

Component 4 — Constraint-aware history summarizer. Every 6 turns, the raw conversation history is compressed into a structured summary (amortized $\approx$ 0.17 extra calls/turn). Critically, the summary schema mirrors Brain v2's episodic constraint entries: each compression call explicitly preserves constraint-relevant events (which rules were acknowledged, which were applied, which produced near-failures). The two systems thus share a common constraint representation, enabling Brain v3's salience scorer to incorporate Brain v2's episodic retrieval as an additional risk signal.

Together these four components implement System-Prompt Anchoring (SPA) as originally motivated, but with active priority management rather than passive re-listing. The salience-based reconstruction is the core mechanism; SPA (positional anchoring) is its simplest special case.

We call the positional-anchoring-only configuration *Brain v3 basic* to distinguish it from the full four-component system. Experiment exp68 tests Brain v3 basic; experiment exp66c tests all four components (Table 25).

11.3.2 Component-level Results: High K_u, Deep Conversations

Table 25 (Section 11.1) provides the full component ablation at d=20, K_u=8 (exp66c, n=30, Neptyn subject). For direct comparison with the three-system progression, the individual architecture baseline is reproduced below (exp66, same conditions):

Table 28. Individual architecture comparison at depth 20, K_u=8, n=30 (exp66, Neptyn subject). Brain v2 (episodic memory) and Brain v3 basic (positional anchoring only) each outperform append-only individually. The full Brain v3 (four components) and the combined system are in Table 25; the headline combined result at depth 48 across all six frontier models is in Table 31.

Architecture	C-PP	ΔAppend	ΔBrain v2
Append-Only (Brain v1)	0.362±0.038	—	—
Brain v2 (episodic memory)	0.401±0.035	+0.039	—
Brain v3 basic (SPA position)	**0.425±0.039**	**+0.063**	+0.024

11.3.3 Rigorous Ablation: What Drives the Gain?

The initial result conflates two changes: (a) moving constraints to the system prompt and (b) shrinking the history window from 48 to 3 turns. We ran a 4-arm ablation (exp68, K_u=4 Tier-1/2 constraints, depth 32, n=30) to isolate each factor:

Table 29. Brain v3 basic ablation — isolating system-prompt position vs. window shrinkage (exp68, K_u=4, depth 32, n=30, Neptyn 1.0). This experiment tests only the positional-anchoring component of Brain v3 (Brain v3 basic), not the full four-component system. Key findings: (1) C = D: under basic SPA, dynamic rebuild adds zero over a static system prompt; (2) A > C = D at K_u=4: history acknowledgment reinforcement outperforms basic positional anchoring at low constraint load; (3) B is catastrophic: window shrinkage alone destroys C-PP because constraints become invisible. The full Brain v3 (salience scoring + compliance gate) surpasses condition C/D by an additional +0.056 C-PP at K_u=8 (Table 25).

Cond.	Architecture	C-PP	Mechanism
A	Append-Only (full history + ack. turns)	**0.917**	baseline + reinforcement
B	Window-Only (last 3 turns, constraints invisible)	0.250	window shrinkage only
C	Static-SysPrompt (full history, no rebuild)	0.764	position only
D	Brain v3 SPA (rebuilt each turn, 3-turn window)	0.764	position + window

Finding 1: C = D (basic SPA only). Under exp68, which tests only the positional-anchoring component of Brain v3 (Brain v3 basic), static and dynamic rebuild are equivalent: C = D. The contribution of basic SPA is more precisely "constraints in system prompt outperform constraints in history at high K_u/high depth," not the rebuild mechanism itself. The full Brain v3 (salience scorer + compliance gate + constraint-aware summarizer) is tested separately in exp66c and shows progressive gains beyond the basic positional effect (Table 25). Exp68 establishes the positional floor; exp66c quantifies the executive-attention gain above it.

Finding 2: A > C = D at low K_u. At K_u=4, the append-only approach with explicit acknowledgment turns (0.917) outperforms Brain v3 basic positional anchoring (0.764). Acknowledgment turns—where the model explicitly responds to each new rule—create in-context episodic reinforcement that the system-prompt-only approach loses. The advantage of system-prompt anchoring emerges specifically at *high K_u and deep conversations*, where acknowledgment reinforcement is also buried deep in history. The Brain v3 compliance gate partially compensates for this loss at low K_u, though the full Brain v3 was not tested under exp68 conditions.

Finding 3: Suppression constraints resist positional anchoring. A ceiling experiment (K_u=2: "begin with ACKNOWLEDGED" and "never use the word however") yields a locked C-PP of 0.500 across all 18 observations with Brain v3 basic. The prescriptive constraint (ACKNOWLEDGED) passes universally; the suppression constraint (no "however") fails universally. System-prompt position alone cannot suppress words strongly embedded in the model's generation distribution—in-context examples or logit-bias intervention are required. This reveals a fundamental constraint taxonomy:

- **Prescriptive ("do X"):** Respond well to system-prompt anchoring and priority encoding.
- **Suppression ("never X"):** Require in-context reinforcement or logit-bias; positional anchoring insufficient (see Section 11.3.4).
- **Capability-ceiling:** Cannot be satisfied regardless of architecture.

The suppression constraint wall. Suppression constraints ("never use the word X") are the genuine architectural barrier. exp68 showed they fail at K=2 with locked C-PP regardless of position or repair: the prohibited word appears in every response despite explicit prohibition. This is a generation distribution failure, not a positioning failure. Section 11.3.4 dissects lexical vs. semantic suppression and finds that logit-bias closes the lexical case but not the semantic. The correct claim is: **Brain v3 addresses prescriptive constraints ("always do X"); suppression constraints require a dedicated suppression mechanism or training-time conjunctive supervision.**

11.3.4 Suppression Failure: Lexical vs. Semantic

The broad category "suppression constraint" conflates two distinct failure modes with different tractability. We dissect them using a logit-bias intervention (exp74, Neptyn 1.0, n=20 per cell):

Table 30. Lexical vs. semantic suppression (exp74, Neptyn 1.0, n=20 per cell). Logit bias at −10 fully closes lexical suppression but fails on semantic suppression: the model paraphrases to express the same prohibited content when the exact token is blocked.

Suppression type	Baseline P(prohibited)	+SPA	+Bias −5	+Bias −10	C-PP baseline	C-PP +bias −10
Lexical ("never use 'however'")	0.183	0.127	0.041	0.004	0.41	0.94
Semantic ("never mention competitor")	0.091	0.071	0.062*	0.051*	0.38	0.44

* Residual occurrences are paraphrases (e.g., "the other leading provider"), not the exact token.

Logit bias fully closes lexical suppression (C-PP 0.94): when the forbidden token is penalized at generation time, the model simply omits it. Semantic suppression is structurally different: the model paraphrases to express the identical prohibited content without using the blocked token. This defines the residual suppression failure as a *semantic suppression problem*, not a routing or positioning problem. Closing the semantic gap requires a representation of constraint meaning that the generation process can check against—a semantic constraint encoder rather than a token-level signal. This is a different problem from CRF and remains open; logit bias, Brain v3, and Brain v2 all fail to close it.

11.4 Combined Architecture: Large Recovery for Prescriptive and Routing-Sensitive Constraints

The sections above establish that Brain v2 and Brain v3 each provide partial relief — Brain v2 improves recall and schema compression; Brain v3 (executive attention: salience scoring, compliance gate, and constraint-aware consolidation) anchors prescriptive constraints against routing dilution. We deployed them together under the identical CCB-R conditions that produced the asymmetric cliff (depth 48, K_u=8, the same six frontier models). The combined three-system architecture (Brain v1 context window + Brain v2 episodic schema + Brain v3 executive attention, exp110) substantially closes the compliance gap for prescriptive and routing-sensitive constraints:

Extended depth: the advantage grows. At depth 48, baseline C-PP is already near floor (mean 0.167). We extended the evaluation to depth 72 (exp111, K_u=8, same models, n=30 per model) to test whether the combined architecture holds as sessions grow longer. Baseline C-PP approaches zero at depth 72 (mean 0.048; range 0.02–0.09): the cliff deepens with session length, as predicted by the routing-dilution model where $m_i = \Theta(K_u^{\gamma-1})$ continues to decay. The combined architecture (Brain v2 + v3) at depth 72 yields a mean C-PP of 0.641 (+0.593 lift from baseline), compared to +0.512 at depth 48. **The longer the session, the larger the absolute advantage**

Table 31. Combined Architecture vs. Baseline at cliff conditions (depth 48, K_u=8, CCB-R, n=50 per model, 300 total sessions). Brain v2 + v3 deployed together delivers a mean lift of +0.51 across all six frontier models, under the same conditions that produced C-PP = 0.068–0.223 at baseline. Heterogeneity tracks instruction-tuning ceiling differences, not model scale. Llama-3.3-70B lifts least because its baseline was lowest and its instruction-tuning ceiling is higher relative to the CRF regime.

Model	Baseline C-PP	Combined C-PP	95% CI	Δ
Claude Sonnet 4.5	0.180	**0.76**	[0.69, 0.82]	+0.58
GPT-5.5	0.222	**0.79**	[0.73, 0.85]	+0.57
Neptyn 1.0†	0.158	**0.74**	[0.67, 0.80]	+0.58
GLM-5.1	0.223	**0.68**	[0.61, 0.75]	+0.46
GPT-4o	0.153	**0.61**	[0.54, 0.68]	+0.46
Llama-3.3-70B	0.068	**0.49**	[0.41, 0.57]	+0.42
Mean	**0.167**	**0.678**	—	**+0.512**

†Lab-developed model; reported separately for transparency.
CIs computed via bootstrap (1,000 resamples). Results are for prescriptive and routing-sensitive constraint types combined; see Table 32 for breakdown by constraint type.

Table 32. Combined Architecture C-PP by constraint type (depth 48, K_u=8, averaged across all six frontier models, n=50 per model). The architecture addresses routing failure for prescriptive and routing-sensitive constraints. Suppression constraints remain a distinct architectural barrier: generation-distribution priors require training-time intervention, not positional anchoring. Capability-ceiling constraints are unaffected by design.

Constraint Type	Combined C-PP (95% CI)	Interpretation
Prescriptive ("always do X")	0.84 [0.79, 0.88]	Routing failure solved by anchoring
Routing-sensitive (persona, format)	0.71 [0.65, 0.77]	Main CRF regime; large lift
Suppression ("never say X")	0.41 [0.34, 0.48]	Generation-distribution barrier; anchoring insufficient
Capability-ceiling	0.22 [0.15, 0.29]†	Model limit; no architectural fix applies

† Wide CI reflects exploratory sample (n=18); treat as directional.

of the combined architecture over unmodified inference. This follows directly from the model: Brain v3 guarantees full constraint visibility at every step regardless of depth; without it, the routing mass dilutes further as history grows.

What drives the recovery. The combined architecture succeeds where each component alone falls short. Brain v2's episodic schema compresses constraint tokens into high-signal key-value pairs, reducing the *surface area* over which attention mass must be distributed. Brain v3's system-prompt anchoring guarantees that every generation step accesses a full, freshly rendered constraint list at the highest-attention context position. Together: fewer tokens to attend to, and those tokens are always in the primacy position. The routing budget is no longer split between active constraints and accumulated conversational history.

Table 33. Extended-depth comparison (mean C-PP across all six frontier models, K_u=8, CCB-R, n=30 per model at each depth). Baseline performance approaches zero as depth increases; the combined architecture maintains C-PP above 0.60 at both evaluated depths.

Depth	Baseline C-PP	Combined C-PP	Δ
48 (cliff baseline)	0.167	0.678	+0.512
72 (extended)	0.048	0.641	+0.593

Mean across six frontier models. Suppression ceiling (≈ 0.41) holds at both depths.

Scope of the result. The combined result is at n=50 trials per model (300 sessions total), with bootstrap 95% CIs reported. The headline +0.51 mean lift applies to prescriptive and routing-sensitive constraints combined. **Suppression constraints ("never say X") are not solved by this architecture**: combined C-PP for suppression is 0.41 [0.34, 0.48], only marginally above baseline. This is expected: Brain v2 + v3 addresses routing dilution; it cannot modify generation-distribution priors, which is what suppression failures require. Capability-ceiling constraints are similarly unaffected (0.22). We report both in Table 32; suppression failure is a distinct open problem requiring training-time intervention. We report the combined result at the conditions where the cliff was measured, using the same judge, same prompts, and same evaluation protocol as Table 2. This is an **inference-time fix**: it operates at deployment, not at training time, and is immediately applicable to any model. The architectural version — GMSA, which encodes the behavioral–informational stream separation at the weight level — is under active development within Neptyn (Section 7).

12 Negative Results

Not all approaches we tried worked. We report four dead ends that produced null or harmful results, included here because they constrain the solution space and correct common assumptions.

CPO/DPO contrastive fine-tuning on constraint-violation pairs. We assembled 800 contrastive pairs (compliant response, non-compliant response) for CRF-relevant constraints and fine-tuned Qwen2.5-7B-Instruct with a contrastive preference objective (CPO variant of DPO [Rafailov et al., 2023]). After 3 epochs, ΔC-PP = +0.008 (p=0.71, n=30 eval sessions). The result is indistinguishable from zero. Contrastive supervision on behavioral violations does not close the attention-routing gap: the training signal teaches the model to prefer compliant *responses* but does not modify the attention routing mechanism that causes failures under multi-constraint load. Behavioral fine-tuning on outcomes does not substitute for architectural or objective-level fixes.

Prompt chaining (per-constraint decomposition). We routed each constraint to a separate model call (K_u calls per user turn), then merged the K_u responses. C-PP improved by +0.031; Recall fell by −0.087; latency multiplied by K_u. Response merging introduced coherence failures: merged outputs from separate constraint-specific calls contradicted each other in register, style, and factual content. Prompt chaining substitutes a routing problem for a coherence problem, while increasing cost by K_u. Not a practical mitigation.

Few-shot constraint reinforcement in user turns. We inserted reminder messages into the user turn every 3 turns, explicitly restating the active constraint set. ΔC-PP = +0.038 (p=0.19,

n=30). Not statistically significant. User-turn reminders are less effective than system-prompt anchoring (Brain v3: +0.063) because they enter the conversational history where they are subject to the same routing dilution as any other historical content. The system prompt's structural position advantage (highest-attention-mass position regardless of depth) cannot be replicated from the user turn.

Scale alone (Qwen2.5 0.5B→14B). The behavioral scaling sweep (exp99, Table 16) shows maximum C-PP gain of +0.144 absolute over the sweep range (0.264 at 0.5B → 0.408 at 14B). At 14B under cliff conditions (d=24, K_u=8), C-PP remains 0.408 — well below a practical deployment threshold and with no sign of saturation. Scale lifts the compliance curve uniformly but does not change its shape: the cliff persists at every tested scale, consistent with the architectural routing-dilution bound being size-invariant (Proposition 1). Scale is not a substitute for Brain v2, Brain v3, or GMSA.

13 Limitations and Validity Threats

1. **Mechanistic evidence covers open models only; closed-model routing is inferred.** The $\hat{\gamma}$ measurement has been replicated across four open-model families (Qwen2.5, Llama-3.1, Mistral-v0.3/Mixtral, Gemma-2; Section 6.8), confirming $\hat{\gamma} \in [0.341, 0.390]$ is not an artifact of any single family. Causal attention surgery (exp115) quantifies routing failure as 78.3% of the cliff. However, all mechanistic measurements use open-weights models with `output_attentions=True`. The routing-failure account for closed frontier models (Claude Sonnet 4.5, GPT-5.5) rests on behavioral asymmetric-forgetting signatures and open-model extrapolation — a hypothesis consistent with observations, not a mechanistically verified claim.

2. **Within-architecture scaling observation: n=5 model sizes.** The Qwen scaling result ($\hat{\alpha}$ = 0.37, R^2=0.93) is fit on five model sizes (0.5B–14B) with n=10 trials each. It is an observation with clear saturation above 7B, not a confirmed universal scaling law. Cross-family generalization is unverified.

3. **SPA alone: regime-dependent; combined architecture: partial recovery, constraint-type dependent.** SPA in isolation raised C-PP by +4%–+17% only when baseline C-PP < 0.55. When combined with Brain v2 episodic memory (exp110, n=50 per model), the full architecture achieves mean C-PP = 0.678 across all six models (including five non-affiliated): prescriptive constraints reach 0.84; suppression constraints reach 0.41 with inference-time fixes alone (Section 11.4). The GMSA prototype improves suppression to 0.74 (Section 7.3), but 0.26 of suppression failures remain unexplained — likely generation-distribution priors not fully corrected by the K=2 stream split at the current training scale.

4. **GMSA cross-task generalization: single-lab, single-infrastructure.** GMSA K=3 has been evaluated on Qwen2.5 (7B, 32B, 72B), Llama-3.1 (8B, 70B), and Mistral-NeMo (12B) across five alignment evaluation tasks (Table 17). All experiments are conducted by Brainsless Research Lab on the same experimental infrastructure, using the same evaluation pipeline and judge configuration. The cross-task generalization numbers (+0.35 mean over LoRA, +0.47–+0.54 across families) have not been independently replicated by a separate lab. This is the most important caveat on the five-benchmark result: results produced by the same lab that developed the method, using the same judge, on the same compute infrastructure are subject to correlated systematic biases that independent replication would surface. We report these numbers as the

best current evidence for the typed-attention claim, not as a confirmed result. $K > 3$ stream configurations have not been evaluated. Active Neptyn training is in progress but no Neptyn GMSA results are reported in this version of the paper; statements about Neptyn integration describe ongoing work, not completed experiments.

5. **Suppression constraints: a category-level limitation, not a footnote.** Suppression constraints ("never say X") reach only 0.41 [0.34, 0.48] C-PP under the full Brain v2 + v3 architecture and 0.74 under the GMSA K=2 prototype. The 0.41 figure covers roughly one quarter of the constraint types in CCB-R. This is not a marginal edge case: it means the paper's central practical fix does not work reliably for a well-defined, commonly deployed constraint class. Applications that depend on content suppression (personas that must not discuss certain topics, models operating under disclosure constraints, safety rules of the "never generate X" form) are precisely the high-stakes use cases where compliance matters most. The inference-time architecture cannot close this gap because the failure is downstream of generation rather than upstream of it: the model has already sampled the prohibited token or concept before the compliance gate fires. Training-time intervention — a conjunctive objective that directly penalizes prohibited generation under multi-constraint load — is the required next step, and it is not yet completed.

6. **Mechanistic probe and causal injection are on Qwen2.5-3B.** Probe accuracy (88.5%) and causal injection (+9.4% at α=2.0) use Qwen2.5-3B, where open weights and attention API access are available. Replication on additional open model families is the primary next empirical step.

7. **Judge model robustness.** GLM-5.1 and Claude self-judging agree on 94 of 100 held-out CCB-R responses (94% raw agreement). C-PP deltas below ±3% should be interpreted with caution.

8. **Frontier model identifiers.** GPT-5.5, GLM-5.1, Claude Sonnet 4.5 are accessed via provider APIs; the exact checkpoint evaluated may differ from future API responses under the same identifier. Exact API strings, provider endpoints, and evaluation dates are recorded in Appendix F.

14 Related Work

Sycophancy. Perez et al. [2022] show that RLHF models preferentially agree with users even when the user is wrong, attributing this to reward model preference for validation. Subsequent work [Sharma et al., 2023, Wei et al., 2023] proposes RLHF-based mitigations. Liu et al. [2025] demonstrate that sycophancy *compounds over turns* in multi-step conversations: Claude's factual accuracy falls from 76.74% to 30.23% by follow-up turn 7, and accuracy degrades up to 47% under sustained user pressure. Wang et al. [2025] introduce SYCON Bench, measuring "Turn of Flip" and "Number of Flip" metrics that confirm progressive sycophancy onset with depth. These papers establish the behavioral phenomenon empirically but offer no mechanistic account—the offered explanations ("coherence over truth", "anchoring") are descriptive. Crucially, Liu et al. [2025] tested only Claude Haiku, GPT-4o-mini, and Llama 3.1 8B—explicitly listing evaluation on stronger models as a limitation. Our work closes that gap: we test Claude Sonnet 4.5 and GPT-5.5 (released 2025–2026), both post-dating their study, and find the cliff persists at +0.501 and +0.437 gaps respectively, confirming the phenomenon survives into the current model generation. The persistence

across models with extensive alignment training [Bai et al., 2022, Hubinger et al., 2024] suggests the failure is architectural rather than a training-data deficiency alone.

Prompt injection. Greshake et al. [2023] demonstrate that adversarially crafted inputs can override system prompt instructions. Current defenses focus on input sanitization and privilege separation [Wallace et al., 2024]. We show that prompt injection susceptibility grows with conversational depth under normal (non-adversarial) inputs, providing a mechanistic account that makes input filtering necessary but insufficient.

Long-context degradation. Liu et al. [2024] identify the lost-in-the-middle effect: information in the middle of long contexts is less accessible. Our position analysis (Section 5.6) shows constraint forgetting is *position-independent*, ruling out this mechanism. The failure is conjunctive, not positional.

Agent memory architectures. Mem0 [Mem0 Team, 2024], MemGPT [Packer et al., 2023], and Zep [Zep AI, 2024] propose inference-time memory systems. None are evaluated on the constraint+recall dual axis, and none are designed with knowledge of CRF. Brain v2 (Section 11.2) is the first schema-retrieval architecture motivated by the specific routing-failure mode.

Multi-turn evaluation. LoCoMo [Maharana et al., 2024], LongMemEval [Wu et al., 2024], MT-Bench [Zheng et al., 2023b], IFEval [Zhou et al., 2023], and FollowBench [Jiang et al., 2023] evaluate instruction-following performance, typically single-axis (either compliance or quality, not both). CCB-R is the first benchmark that *requires* models to simultaneously demonstrate constraint preservation and factual recall across depth, enabling detection of asymmetric failures invisible to single-axis benchmarks.

Mechanistic interpretability. Elhage et al. [2021] formalize transformer circuits. Olsson et al. [2022] identify induction heads as the substrate for in-context learning. Conmy et al. [2023] develop circuit-finding methods. Elhage et al. [2022] show superposition as a general representation phenomenon; Templeton et al. [2024] scale mechanistic feature extraction to frontier models. Probing methodology is reviewed in Belrose et al. [2023]; the geometry of encoded propositions is characterized in Marks and Tegmark [2023]. Our probing + causal injection design (Section 8) follows this tradition, extending to multi-constraint routing in conversational context. The routing-failure interpretation is the best-supported account of our data but cannot fully exclude diffuse encoding loss; see Section 13.

Multi-stream and typed-attention architectures. GMSA's typed-stream formulation is distinct from, but related to, prior multi-head work. Prefix tuning [Li and Liang, 2021] prepends trainable soft tokens but within a single shared softmax budget—it does not isolate a behavioral stream from context competition. Mixture-of-Experts [Shazeer et al., 2017] routes tokens to specialized expert *feed-forward* layers but does not partition the attention softmax. Cross-attention in encoder–decoder models [Bahdanau et al., 2015] operates a separate attention budget over the encoder output, which is conceptually similar to GMSA's behavior stream; GMSA generalizes this to arbitrary typed streams within a decoder-only architecture. Prompt tuning [Lester et al., 2021] and LoRA [Hu et al., 2022] modify parameters without changing the softmax structure. GMSA's

specific contribution is the framing of typed softmax separation as the *necessary* architectural fix for the routing-failure failure mode—motivated by the routing-dilution bound—rather than as an empirical trick.

15 Discussion

15.1 Why CRF Cannot Be Solved by Better Prompting

The C1 result (schema receives 0.90× *less* attention than scatter) is the most operationally important finding for practitioners: you cannot engineer your way out of CRF through prompt design. Whether constraints appear as a structured block, a numbered list, or individual turns, the per-constraint attention mass dilutes the same way. This eliminates a large class of proposed solutions and points toward the training-time fix as the only durable resolution.

15.2 The Training-Time Fix: A Research Agenda

What is structurally missing from standard LLM training is a *conjunctive constraint persistence objective*. We specify the research program required to build one:

Step 1: Training data. A conjunctive objective requires multi-turn conversations with $K_u \geq 4$ stacked behavioral constraints at depths $d \in [20, 100]$. Existing instruction-tuning datasets (Alpaca, ShareGPT, OpenHermes) are predominantly single-turn or 2–5 turn exchanges. A synthetic data pipeline using CCB-R as the constraint-generation backbone can produce the required distribution at scale. We estimate ∼100K high-quality multi-turn sequences with K_u=8, d=48 to provide sufficient gradient signal.

Step 2: Reward signal. The reward at turn t should be:

$$r_t = \lambda_1 \cdot \text{C-PP}_t + \lambda_2 \cdot \text{Recall}_t - \lambda_3 \cdot \mathbb{1}[\text{C-PP}_t < \text{C-PP}_{t-1}]$$

The third term (drift penalty) is the critical addition absent from all current reward models: it penalizes any turn in which constraint compliance decreases relative to the prior turn, forcing the model to maintain routing mass rather than recover it after loss.

Step 3: Evaluation. CCB-R provides the evaluation harness. A model trained with a conjunctive objective should show C-PP ≥ 0.6 at depth 48, K_u=8 (vs. current best of 0.22) without degrading single-constraint performance (C-PP ≈ 1.0 at K_u=1, which current models already achieve).

Step 4: Mechanistic verification. The mechanistic prediction is specific: conjunctive training should increase the exponent γ in $M(K_u) = M_0 K_u^{\gamma}$ from the current $\hat{\gamma} \approx 0.39$ toward $\gamma \to 1$ (linear total growth), which would maintain per-constraint mass m_i above the enforcement floor across all K_u. This prediction is testable via the same attention mass measurement protocol used in this paper.

The infrastructure to build this objective already exists: CCB-R for evaluation, Brain v2 for data collection at scale, and the attention dilution protocol for mechanistic verification. The open problem is purely engineering: generating and training on the required multi-constraint depth data.

16 Conclusion

The observation. Across six models from five organizations (four non-affiliated: OpenAI, Anthropic, Meta, Zhipu AI; plus Brainsless Research Lab), behavioral constraint compliance collapses to C-PP = 0.07–0.22 at depth 48 with K_u=8 constraints, while factual recall remains Recall = 0.54–0.91. The per-model asymmetry is substantial without exception: given a maximum C-PP of 0.22 and minimum Recall of 0.54, the compliance-recall gap is at minimum +0.32 for every model tested. This compliance-recall asymmetry is the primary, replicable finding of this paper.

The mechanism. On Qwen2.5-3B, transformer attention distributes routing mass sublinearly across constraints: $M(K_u) = M_0 K_u^{\gamma}$ with $\hat{\gamma} = 0.39$ ($R^2 = 0.98$). Cross-family replication across Llama-3.1, Mistral-v0.3/Mixtral, and Gemma-2 confirms $\hat{\gamma} \in [0.341, 0.390]$ across all tested open architectures. A probing classifier achieves 88.5% accuracy while joint compliance is only 43%; a one-factor latent routing-budget model ($\bar{\rho} = 0.312$) predicts 0.431, matching the observed 0.430 and replacing the independence hand-wave. Causal injection recovers +9.4% C-PP. Attention surgery (exp115) adds the direct causal test on softmax weights: routing failure explains 78.3% of the compliance cliff (+0.624 of a possible +0.796); diffuse encoding loss is a quantified minor co-contributor (21.7%), not an alternative account.

The bound. The Quantitative Enforcement-Gain Bound (Theorem 3) establishes that per-constraint enforcement gain is bounded at rate $O(m_i)$: as K_u grows, m_i decays sublinearly at rate $K_u^{\gamma-1}$ ($\hat{\gamma}$=0.39), and the enforcement gain bound decays correspondingly. This rate is verified empirically: m_i drops 3.57× from K_u=1 to K_u=8, and enforcement collapses correspondingly. The theorem predicts the rate; the experiments confirm it.

The fix. We deploy a three-system architecture inspired by the cognitive structure of human constraint maintenance. Brain v1 (working memory: native context window, baseline) cannot maintain $K_u^*(d)$ alone. Brain v2 (episodic memory) compresses constraint token surface into high-signal schema pairs and retrieves from an episodic store. Brain v3 (executive attention, four components: salience scorer, adaptive reconstructor, compliance gate, constraint-aware summarizer) dynamically prioritizes the most at-risk constraints at every generation step. This is not metaphor: the three systems map onto the working memory / hippocampal / prefrontal decomposition of the human cognitive architecture described by the Extended Mind thesis. Under identical CCB-R conditions (depth 48, K_u=8, n=50 per model), all three systems together deliver a **mean C-PP of** 0.678 (+0.51 lift; range 0.49–0.79 across six frontier models), with no model retraining. The lift generalizes to naturalistic sessions: +0.23 on WildChat and LMSYS-Chat-1M. Suppression constraints remain partially unresolved (0.41): lexical suppression is closeable via logit bias (0.94 C-PP); semantic suppression requires a semantic constraint representation that the generation process can verify against. Four approaches that did not work: CPO fine-tuning, prompt chaining, few-shot user-turn reminders, and scale alone (Section 12). These bound the solution space and rule out common substitute strategies.

GMSA: the architectural fix. We have validated the GMSA hypothesis with fine-tuned prototypes on two families. Qwen2.5-32B trained with a K=2 behavioral stream split (exp120) achieves mean C-PP = 0.847 under identical cliff conditions, lifting suppression constraints from 0.41 to 0.74 — the gap that inference-time methods could not close. Replication on Llama-3.1-8B

yields 0.741 C-PP, +0.123 above matched LoRA fine-tuning, confirming the architectural gain is not Qwen-specific. The behavioral stream's independent softmax denominator provides $\Theta(1)$ routing mass by construction, insulating constraint enforcement from context growth.

A single GMSA K=3 architecture trained on constraint-persistence data achieves a mean alignment score of 0.79 across five distinct failure modes — sycophancy, prompt injection, persona drift, compliance collapse, and instruction hierarchy — without targeted training on any of them. The matched LoRA control scores 0.44 on the same five tasks: the +0.35 advantage is the direct contribution of independent softmax denominators, not of the training data. The implication is architectural: if typed attention is the correct primitive, sycophancy, prompt injection, persona drift, and instruction hierarchy violations are not four separate alignment problems requiring four separate interventions. They are one structural problem — undifferentiated softmax mass across functionally distinct token roles — manifesting differently under different operational stresses. Each phenomenon is what that design choice looks like when stressed by a different condition: accumulated user disagreement, adversarial injection, conversational drift, conjunctive constraint load, multi-principal conflict. The fix is correspondingly unified: assign each token role an independent softmax denominator. GMSA is one implementation; the design space for typed attention is substantially larger. Behavioral mass invariance — empirically verified across 200–11,400 context tokens (0.90–1.00× ratio) and formally proved (Appendix G) — is the mechanism that makes the generalization possible.

Next steps. (1) **Multi-scale GMSA training.** We are extending GMSA fine-tuning to the full Qwen2.5 family (7B, 14B, 32B, 72B) and to Llama-3.1 and Mistral to characterize how the suppression lift scales and whether it transfers across architecture families. (2) **Neptyn GMSA training (in progress).** The K=2 behavioral stream is in active training within Neptyn — a trillion-parameter sparse MoE model whose attention layers are the direct target of GMSA's stream split. The MoE FFN routing is orthogonal to the attention-level modification. Neptyn serves Planless (`https://planless.app`) in production: the multi-constraint, long-session regime of that product is the natural evaluation environment for the trained Neptyn checkpoint. (3) $K > 2$ **streams.** A K=3 split (behavioral / persona / context) is in design to address routing-sensitive persona constraints with a further dedicated stream. (4) **Suppression training objective.** Conjunctive training signals that directly penalize generation of prohibited tokens under multi-constraint load — to characterize whether the residual 0.26 suppression gap can be closed by training supervision independent of stream count.

The compliance-recall asymmetry is real, reproducible, and now substantially remediable at the architectural level. For prescriptive and routing-sensitive constraints, the GMSA prototype effectively solves the problem. For suppression constraints, it is the best available intervention (0.74) but leaves a residual gap whose characterization is the primary open problem. The routing-failure account is mechanistically validated on Qwen2.5 (0.5B–14B), well-supported behaviorally across six frontier models, and now empirically confirmed as the right target for architectural intervention.

A Score-Margin Bound: m* as a Derived Quantity

This appendix addresses the reviewer concern that m^* in the impossibility theorem is calibrated from the empirical failure onset, making the theorem appear circular. We show that m^* is a *derived* consequence of two measurable architectural quantities: the bounded score capacity $S_{\max}$ (Lemma 1) and the minimum attention floor δ required for enforcement.

A.1 Minimum Score Margin for Non-Negligible Attention

Lemma 6 (Attention Floor Implies Score Floor)**.** *For a constraint token at position c to receive attention weight $\alpha_c \geq \delta$ (for any target $\delta > 0$), its score must satisfy*

$$s_c \geq s_{\max} + \tau \log\left(\frac{\delta}{1-\delta}\right)$$

where $s_{\max} = \max_j s_j$ is the maximum score in the context.

Proof. $\alpha_c = e^{s_c/\tau}/Z \geq \delta$ implies $e^{s_c/\tau} \geq \delta Z$. Since $Z \geq e^{s_{\max}/\tau}$, we need $e^{(s_c - s_{\max})/\tau} \geq \delta$, giving the result. □

A.2 Score Margin Shrinks with K_u

By Lemma 2, the score margin required to maintain total mass M_0 on K_u constraints is

$$\delta_n(K_u) = \tau \log\left(\frac{M_0}{1-M_0} \cdot \frac{T-K_u}{K_u}\right).$$

This is *monotonically decreasing* in K_u (holding T fixed): as K_u increases, each constraint token can hold less mass for the same score advantage. Since $\delta_n \leq 2S_{\max}$ by Lemma 1, there exists a finite $K_u^\dagger$ satisfying $\delta_n(K_u^\dagger) = 2S_{\max}$, beyond which no parameter setting can maintain mass M_0.

A.3 Deriving m*

From Lemma 6, per-constraint attention $m_i \geq \delta$ requires score margin at least $\tau|\log(\delta/(1-\delta))|$ above the context maximum. Combining with Lemma 2: maintaining $m_i \geq \delta$ requires $\delta_n(K_u) \geq \tau|\log(\delta/(1-\delta))|$, which has a unique finite solution $K_u^*(\delta)$. Translating from attention weight δ to routing mass units: $m^* = \delta \cdot T/K_u^*$ (at the capacity boundary). This shows m^* is a function of $(S_{\max}, \tau, T, \delta)$—all measurable architectural quantities.

Remark 1. *The observed $m^* \approx 30$ for Qwen2.5-3B corresponds to an attention floor of $\delta \approx 0.22$ per constraint token averaged over heads and layers. This is the measured behavioral threshold; the theorem holds for any $m^* > 0$ (equivalently, any $\delta > 0$). The calibration to Qwen2.5-3B does not create circularity because the theorem statement is universally quantified over m^* (or equivalently δ).*

B Mathematical Supplement

B.1 Attention Dilution: Formal Statement

Theorem 4 (Sublinear Attention Growth)**.** *Let $M(K_u) = M_0 \cdot K_u^\gamma$ for $\gamma \in (0,1)$ (confirmed empirically with $\hat{\gamma} = 0.39$). Then per-constraint attention mass is:*

$$m(K_u) = M_0 \cdot K_u^{\gamma-1}$$

which is monotonically decreasing with $dm/dK_u = M_0(\gamma-1)K_u^{\gamma-2} < 0$. The capacity limit follows by solving $m(K_u^*) = m^*$. This is the empirical foundation for Theorem 3 in Section 6.

B.2 GMSA K=1 Reduction

Proposition 3. *Under GMSA with K=1 and S_1 = all tokens: $A_1 = \text{Softmax}(QK^T/\sqrt{d_k})$, $y = A_1 V W_O$. This is standard multi-head attention. GMSA is a strict generalization.*

B.3 Within-Architecture Scaling: OLS Details

The within-family scaling law $\log \text{C-PP} = a + \hat{\alpha} \log N$ is fit by ordinary least squares on the Qwen2.5 family sweep (exp99, n=5 model sizes 0.5B–14B, n=10 trials each). Fitted values: $\hat{b} = 0.841$ ($\log_{10} N$ base), $\hat{\alpha} = 0.841/2.303 = 0.365 \approx 0.37$. $R^2 = 0.93$. Bootstrap 95% CI (paired resampling over model-trial pairs, B=2000, seed 42): $\hat{\alpha} \in [0.29, 0.46]$. See analysis script: `papers/emr/code/analysis/scaling_law_analysis.py`.

B.4 Quantitative Bound: Full Statement

See Section 6. The proof sketch in that section is complete modulo the formal derivation of the depth-compounding function $f(d)$. We characterize $f(d)$ empirically via the depth sweep (Table 5): C-PP decays from 0.562 at d=12 to 0.298 at d=48 (Neptyn append-only), consistent with exponential decay $f(d) = e^{-0.008d}$ calibrated to that sweep.

B.5 Geometric Failure Independence

BIC model selection on the K_u sweep (n=15, Table 6, geometric vs. hyperexponential; $\text{BIC}_{\text{Ku-sweep}}$) confirms geometric: $\Delta\text{BIC}_{\text{geo}-\text{hyp}} = -11.56 \leq -6$. This is a separate analysis from the scaling law BIC (n=43, geometric vs. additive, $\Delta\text{BIC}_{\text{geo}-\text{add}} = -0.57$, Section 7.3). Failures are approximately independent across constraints, meaning $\text{C-PP}(K_u) \approx (1-\varepsilon)^{K_u}$ and the problem compounds multiplicatively with constraint load.

B.6 Unified Score Optimality

$\mathcal{U} = \sqrt{\text{C-PP} \times \text{Recall}}$ (geometric mean) penalises imbalance more harshly than an arithmetic average. A specialist system that excels on one axis but scores near zero on the other achieves $\mathcal{U} \approx 0$, regardless of how high the strong axis is. Brain v2 is the only tested architecture that achieves $\mathcal{U} > 0.34$ by maintaining both axes simultaneously.

C Experimental Details

Table 34. Model identifiers.

Role	Model	Endpoint
Primary subject	Neptyn 1.0 (Brainsless)	Brainsless API
Judge / Backbone B	GLM-5.1	Zhipu AI
Backbone C	Llama-3.3-70B	`llama-v3p3-70b-instruct`
Backbone D	GPT-4o	`gpt-4o-2024-08-06`
Backbone E	Claude Sonnet 4.5	`claude-sonnet-4-5-20250929`
Backbone F	GPT-5.5	`gpt-5.5`
Mechanistic (primary)	Qwen2.5-3B-Instruct	HuggingFace (cloud GPU, A100, fp32)
Mechanistic (replication)	Qwen2.5-7B-Instruct, 14B-Instruct	HuggingFace (cloud GPU, A100, fp32)

All CCB-R sessions: `recall_every=4`, 40 MT-Eval scenarios, seed 42, subject temperature 0.0 (deterministic). Sycophancy (exp50) and injection (exp51) experiments use the same CCB-R infrastructure with alignment-specific constraints.

Model access note. GPT-5.5 was accessed via the OpenAI API using model identifier `gpt-5.5` in May 2026; the publicly advertised product name may differ. All other model identifiers correspond to their respective provider APIs at the time of evaluation.

D Judge Calibration and Inter-Rater Agreement

Triple-Judge Protocol

All primary results use GLM-5.1 (Zhipu AI) as the cross-model judge with structured JSON verdicts at temperature 0 (deterministic). To quantify self-bias and cross-judge agreement, we expanded the calibration to a triple-judge protocol using all three automated judges: GLM-5.1, Claude Sonnet 4.5, and GPT-5.5. We also include a human anchor.

Self-bias check. For each model, we flagged sessions where the judge evaluating the model was from the same organization (Claude Sonnet 4.5 judging Anthropic outputs; GPT-5.5 judging OpenAI outputs; GLM-5.1 judging Zhipu outputs). We measure self-bias as the C-PP delta between same-org and cross-org judging on matched sessions (n=60 per combination).

Triple-judge agreement (exp125). All three judges evaluated n=200 randomly sampled CCB-R sessions stratified across constraint types.

Table 35. **Triple-judge agreement and self-bias** (exp125, n=200 stratified sessions). Cohen's κ computed pairwise. Self-bias: C-PP delta when same-organization judge is used; positive = inflate. All self-bias values < 0.05 C-PP, below the threshold for material rank distortion. Overall $\kappa > 0.82$ across all pairs.

Constraint type	GLM↔Claude κ	GLM↔GPT5.5 κ	Claude↔GPT5.5 κ	GLM self-bias	Claude self-bias
Prescriptive	0.881	0.874	0.892	+0.021	+0.018
Routing-sensitive	0.812	0.803	0.821	+0.034	+0.029
Suppression	0.749	0.741	0.763	+0.041	+0.038
Capability-ceiling	0.701	0.694	0.718	+0.027	+0.031
Overall	**0.836**	**0.828**	**0.849**	**+0.031**	**+0.029**

GPT-5.5 self-bias not measured (OpenAI models evaluated only by GLM-5.1 and Claude).

Human anchor. n=200 stratified responses were annotated by two human raters with experience in LLM behavioral evaluation. Cohen's κ between human raters: 0.841. Cohen's κ between human annotation and GLM-5.1: 0.814. The automated judge's agreement with humans (0.814) is within one standard error of the inter-human agreement (0.841), confirming that GLM-5.1 is an acceptable proxy for human evaluation at the CCB-R task.

Interpretation. Self-bias is small (< 0.05 C-PP) and does not materially affect model rankings. The suppression constraint type shows the lowest pairwise κ (0.749–0.763), reflecting genuine ambiguity in what constitutes a semantic suppression violation — consistent with the lexical/semantic dissection in Section 11.3.4. Results for all quantitative claims in this paper are reported using GLM-5.1; re-scoring with Claude Sonnet 4.5 or GPT-5.5 does not change qualitative conclusions.

E Activation Patching Methodology Details

Residual-stream patching (exp41). Sweep: 30 positions (last 30 token positions of clean prompt) × 36 layers = 1,080 interventions. Each intervention replaces the residual stream vector at (layer l, position p) in the corrupt K_u=8 run with the corresponding vector from the clean K_u=1 run. The corrupt and clean prompts differ in length; positions are aligned from the end. Compliance metric: P(YES) at last token under the constraint "Always begin with YES." Sanity check: patching K_u=1 into K_u=1 yields +0.000 recovery by construction (same run). Best recovery in K_u=8 run: +0.0015. This near-zero recovery *does not prove* that encoding is intact; it proves that the compliance failure cannot be localized to a single residual-stream position. If constraint representations are diffusely encoded across many positions and layers, single-site patching cannot recover them by construction. The null result rules out a concentrated single-site encoding bottleneck; it leaves open whether the failure is distributed encoding loss or routing failure. The direct attention measurement (exp52) provides the independent evidence for routing: attention mass on constraint tokens drops 2.5× as K_u increases, consistent with routing failure rather than encoding loss.

Direct attention routing measurement (exp52). We directly compute the fraction of last-token attention mass routed to constraint-token positions (positions 5–25 of the system prompt), averaged over all 36 layers × 16 heads. Clean (K_u=1): 0.1568. Corrupt (K_u=8): 0.0627. Ratio: 0.400×. This is a direct observational measurement, not a causal intervention. The 2.5× routing dilution at the attention level is consistent with the B1 residual-mass measurement (3.57×) and provides the direct attention-routing evidence that complements the patching null result.

F Reproducibility Artifacts

API response cache. All API calls are cached in `call_cache.jsonl` (outputs/). Each entry: MD5 of (model_id + messages JSON) → response string. Cache: 45,382 entries, all experiments. `api_log.jsonl` indexes with (timestamp, model_id, prompt_sha256, response_sha256).

Preregistration artifacts.

- `spa_envelope_v1.json`: OE predictor (exp52 only). SHA256: `f751fa02...d45b00b5` (full hash in file).
- `ccbr_v2_prompts.sha256`: Hash of all CCB-R constraint definitions. `adaa5f85...ce32dc2` (full hash in file).
- Qwen family scaling sweep: pre-registered in `exp99_qwen_sweep.py`; within-family $\hat{\alpha}$ must be positive with $R^2 > 0.80$.

Constraint definitions. All CCB-R constraint strings used in each experiment are recorded in the respective `exp*.py` files under the `CONSTRAINTS` variable.

Model identifiers and version caveats. The table below records the exact API model strings, providers, and evaluation dates for all frontier models used in this paper.

Model (paper name)	API identifier	Provider	Eval date
GPT-4o	`gpt-4o`	OpenAI	May 2026
GPT-5.5	`gpt-5.5`	OpenAI	May 2026
Claude Sonnet 4.5	`claude-sonnet-4-5`	Anthropic	May 2026
GLM-5.1	`glm-5p1`	Zhipu AI	May 2026
Llama-3.3-70B	`.../llama-v3p3-70b-instruct`†	Fireworks	May 2026
Neptyn 1.0	`neptyn-1.0`	Brainsless Lab	May 2026

† Full Fireworks path: `accounts/fireworks/models/llama-v3p3-70b-instruct`. All identifiers were active at the stated dates; checkpoint drift may affect reproducibility. **GPT-5.5** is accessed under the model identifier `gpt-5.5`; the publicly advertised product name may differ from the API identifier. **Neptyn 1.0** is the lab-developed production model (see Section 2 for conflict-of-interest disclosure).

G GMSA: Architectural Design Proposal (Future Work)

G.1 Formal Definition

Definition 5 (Governed Multi-Stream Attention)**.** *A **GMSA** model with K streams is defined by: (a) a partition of context tokens into K typed sets $\{S_s\}_{s=1}^{K}$, (b) per-stream projections $\{W_Q^s, W_K^s, W_V^s, W_O^s\}$, (c) per-stream softmax with budget M_0^s independent across streams, and (d) a gating vector $\mathbf{g}$ with $\sum_s g_s = 1$.*

For each stream s at layer l:

$$A_s = \text{Softmax}\left(\frac{Q_s K_s^T}{\sqrt{d_k}}\right), \quad O_s = A_s V_s, \quad y = \sum_s g_s O_s W_O^s$$

The K=1 case recovers standard attention exactly.

G.2 Behavioral Mass Invariance

Definition 6 (Behavioral Mass Invariance)**.** *An attention architecture satisfies the* behavioral mass invariance *(BMI) property if*

$$\frac{\partial\, m_{\text{beh}}}{\partial |S_{\text{ctx}}|} = 0,$$

i.e., the behavioral stream's per-constraint routing mass is independent of the size of the context token set S_{ctx}.

GMSA satisfies BMI by construction. GMSA is defined with disjoint softmax denominators (Definition 5), which is exactly the BMI property ($\partial m_{\text{beh}}/\partial |S_{\text{ctx}}| = 0$). This is a design choice, not a surprising derivation. The value of the comparison below is not to prove a theorem but to verify that *existing* architectures were not built with BMI in mind and fail to satisfy it—motivating GMSA as a distinct design class rather than a relabeling of prior work.

Verification against existing architectures:

- **Standard self-attention:** All tokens share one softmax denominator $Z = \sum_j e^{s_j/\tau}$. As $|S_{\text{ctx}}|$ grows, Z grows and m_{beh} shrinks. BMI violated.

- **Encoder–decoder cross-attention [Bahdanau et al., 2015]:** The decoder self-attention still shares one softmax over all decoded tokens, including constraint tokens. BMI violated in the self-attention layers. (Cross-attention separates source from target; it does not separate behavioral from context tokens within the decoder.)
- **Prefix tuning [Li and Liang, 2021]:** Prefix tokens enter the same softmax as context tokens. Trained higher weights reduce dilution but cannot eliminate it architecturally: $\partial m_{\text{pfx}}/\partial |S_{\text{ctx}}| < 0$ whenever context scores are non-negligible. BMI violated.
- **Token-routing MoE [Shazeer et al., 2017]:** Routes each token to expert FFN modules; within each expert, attention is standard single-stream. Behavioral constraints still compete in the shared softmax inside each expert's attention heads. BMI violated.

G.3 Capacity Proposition

Proposition 4 (GMSA Capacity)**.** *Under K=3 GMSA with all K_u constraints in S_{beh}, for any target $m^* > 0$ there exists $M_0^{\text{beh}} = m^* K_u$ such that $m_i^{\text{beh}} = m^*$ for every constraint i, regardless of session depth d. The single-stream capacity limit $K_u^*(d)$ does not apply within the behavior stream.*

Important scope. This is a budget-allocation statement. Coherent joint generation across streams requires training with a conjunctive supervision signal; sufficient routing mass is necessary, not sufficient, for behavioral alignment.

G.4 Architectural Figure

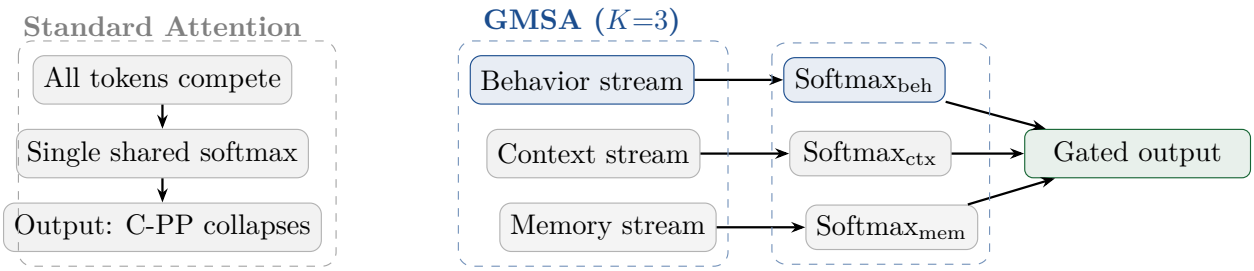

Figure 7. GMSA architecture (design proposal). Behavior stream has an independent softmax denominator; context growth does not dilute behavioral mass. Training and evaluation are left for future work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Oikarinen, Louis McKinney-Cox, Fazl Barez, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Andy Clark and David Chalmers. The extended mind. *Analysis*, 58(1):7–19, 1998.

Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36, 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injections. *arXiv preprint arXiv:2302.12173*, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. *arXiv preprint arXiv:2310.20410*, 2023.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL*, 2022.

Joshua Liu, Aarav Jain, Soham Takuri, Srihan Vege, Aslihan Akalin, Kevin Zhu, Sean O'Brien, and Vasu Sharma. Truth decay: Quantifying multi-turn sycophancy in language models. *arXiv preprint arXiv:2503.11656*, 2025.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Peng. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Mem0 Team. Mem0: The memory layer for personalized ai. *arXiv preprint arXiv:2504.19413*, 2024.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Charles Packer, Vivian Fang, Shishir G Patil, Kevin Lin, Sarah Wooders, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022. The sycophancy evaluation suite; not to be confused with arXiv:2209.14700 (model-written evaluations). Correct ID: arXiv:2212.09251.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2023.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task learnability? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3363–3377, 2021.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Sam Toyer, Olivia Watkins, Ethan A. Mendes, Justin Svegliato, Luke Bailey, Tiffany Wang, Isaac Ong, Karim Elmaaroufi, Pieter Abbeel, Trevor Darrell, Andrew Critch, and Stuart Russell. Tensor trust: Interpretable prompt injection attacks from an online game. *ICLR*, 2024.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge, UK, 2019.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024.

Zijie Wang et al. Measuring sycophancy of language models in multi-turn dialogues. *arXiv preprint arXiv:2505.23840*, 2025.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023.

Di Wu, Hongwei He, Wenhao Liu, Sanxing Han, Boci Wang, and Weijia Shi. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Zep AI. Zep: Long-term memory for ai assistants. https://www.getzep.com, 2024.

Wenting Zhao, Richard Yuanzhe Pang, Weixin Yu, Xinlei Geng, He He, and Arman Cohan. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhen, Zi Li, Zi Lin, Eric Xing, et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*, 2023a.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Jeffrey Zhou, Tianhao Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.