

# Layer-0 Suppressors Ground Hallucination Inevitability: A Mechanistic Account of How Transformers Trade Factuality for Hedging

Mat Thompson  
Independent Researcher, Raleigh, NC

October 29, 2025

## Abstract

Language models must choose: assert confidently, or hedge when uncertain. But where in the network is this trade-off implemented? Drawing on information bottlenecks and Kalai et al.’s hallucination inevitability theorem, we predicted that the circuits mediating this trade-off would emerge at layer 0—the model’s narrowest and earliest point of compression. We validate this prediction in GPT-2 and Mistral-7B, identifying a small coalition of “layer-0 suppressor” heads that dampen factual continuations and boost hedging or editorial tokens.

Ablating suppressors improves factual preference (logit difference: increase in model preference for the correct token over a matched foil,  $+0.40$ – $0.85$ ), calibration (expected calibration error  $0.122 \rightarrow 0.091$ ), and sequence quality across tasks. Randomized ablations confirm these heads lie in the  $> 99^{\text{th}}$ -percentile tail. Causal tracing shows that 67% of suppressor influence flows through a single early-to-mid path (layer 0→layer 11), forming a stable hedging attractor that downstream layers do not reverse. Suppressors emerge early in training and adapt to architecture—GPT-2 couples hedging boosts with factual suppression, while Mistral separates these functions and introduces a task-contingent anti-suppressor.

These findings provide a mechanistic account of how transformers instantiate a statistical trade-off between truth and caution. Beyond power and calibration, we observe clear geometric signatures under suppressor ablation: output distribution flattening reverses across all four probe families (facts/counterfactual/negation/logic;  $\Delta H = -2.4$  to  $-3.8$  nats,  $p < 0.02$ ) and early trajectory curvature decreases (straighter paths at layer 0). Evaluation reform alone may not eliminate hallucinations: suppressors crystallize at the earliest layers, biasing computation toward qualified language even when knowledge is available. We present evidence consistent with constrained early-layer solutions at the bottleneck—predictable from geometry and incentives—while allowing that implementation varies by model and data; and we demonstrate a minimal, task-scoped steering intervention.

## Contributions.

- **Prediction-first validation.** We formulate a bottleneck prediction (layer 0) from information geometry plus Kalai et al.’s constraint, then validate it with falsification-oriented tests rather than post-hoc discovery.
- **Rigorous methodology.** Dual observables (power:  $\Delta LD$ ; information: calibration) improve together; empirical random baselines place suppressors in the  $> 99\%$  tail; results replicate across architectures (GPT-2, Mistral).

- **Geometric validation.** Output entropy and trajectory curvature confirm the predicted operation (output flattening, early trajectory bending), replicating across factual, counterfactual, negation, and logic probes.
- **Predictive framework.** We connect training methodology (pretraining, RLHF, Constitutional AI) to suppressor structure via information-theoretic constraints, and pose testable predictions for how different objectives modify the mechanism.
- **Causal structure.** Forward/reverse path patching shows the suppressor→layer-11 residual stream mediates 67% of the head 0:2 effect, providing an operational attractor.
- **Open reproducibility.** We ship configs, seeds, hashes, and standardized reports (manifest, rankings, OV tables) to enable detailed review and reuse.

## 1 Introduction

Information bottleneck theory predicts that dimensionality compression—and thus the highest geometric constraint—occurs at layer-0. Kalai et al. [8] formalize the objective-level pressure: under binary evaluation, models must balance factuality with hedging when ground truth is uncertain. We conjecture that any circuit instantiating this trade-off will crystallize at layer-0: early residual rotations constrain all downstream computation, making later reversal costly. This yields a falsifiable prediction: if suppressors exist, they should concentrate at layer-0 and rank in the extreme tail under random baselines. We validate that prediction.

**A narrow doorway.** Imagine a transformer as a multi-storey building. At ground level, raw tokens enter through a narrow doorway—layer 0—where the representation is compressed and rotated before any higher-level “thinking” occurs. Choices made at this doorway are hard to undo: early rotations constrain every floor above. This physical metaphor anchors why bottleneck theory makes falsifiable predictions about early layers and motivates our focus on layer 0.

We identify and characterize a family of circuits we call *layer-0 suppressors*. Suppressors are small coalitions of attention heads in the very first transformer layer that systematically down-weight factual continuations and boost uncertainty markers or meta-commentary. They are not idiosyncratic: ablating them recovers as much as 0.85 logit-difference points on factual, negation, and counterfactual probes, and analogous motifs appear in both GPT-2 Medium (355 M) and Mistral-7B despite their architectural differences. We operationalize an *attractor* as a regime in which injecting suppressor activations into an otherwise clean run induces a stable hedging pattern that downstream layers do not undo (reverse-patch  $\Delta\text{LD} \geq 0.3$  for at least one probe).

We report four main findings.

1. **Suppressors are structural.** Cross-task head-ablation sweeps show that the same layer-0 heads remain high-impact across diverse corpora, even after dataset rebalancing removes token-frequency confounds.
2. **They bias downstream computation.** Forward/reverse patch experiments show that suppressor perturbations induce a hedging pattern that downstream layers do not fully correct under our protocols, consistent with an early-layer attractor.
3. **Implementations adapt to architecture.** GPT-2 learns a unified suppressor trio that simultaneously suppresses factuality and boosts hedging, whereas Mistral learns a task-contingent pair opposed by an anti-suppressor on logic tasks and lacking the hedging boost.

4. **The motif is learned.** Suppressors emerge during training as a behavioral prior consistent with Kalai et al.’s incentives; they are neither hard-coded nor artifacts of a single model family.

By grounding Kalai et al.’s theoretical inevitability in concrete circuits, we bridge statistical and mechanistic interpretability. Our results imply that evaluation reform alone may not eliminate hallucinations: once suppressors have crystallized, they steer computation toward hedging by default. Direct circuit-level intervention or steering may therefore be required to restore truthful behavior.

## Why layer 0?

See Section 4.1 for a formal motivation based on bottleneck constraints and falsifiable predictions.

## 2 Background: Statistical Inevitability Meets Mechanistic Structure

Kalai et al. show that when language models are evaluated with binary correctness metrics, calibrated uncertainty is systematically disfavored [8]. A model that admits ignorance scores identically to one that fabricates a confident answer, while a model that answers truthfully when it *does* know receives full credit. Under such incentives, gradient descent pushes the model toward policies that produce confident continuations even in regions of epistemic uncertainty.

Two consequences follow from the theorem. First, the correlation between confidence and accuracy that arises naturally during pre-training must be weakened: the model benefits from emitting confident-sounding statements even when its latent probability of correctness is low. Second, because the penalty for hedging equals the penalty for hallucinating, there is an optimisation advantage in producing plausible meta-commentary or qualified statements—the output “looks helpful” despite being wrong.

The theory predicts *what* behavior should emerge but not *how* it is instantiated. To uncover the implementation, we analyse foundational circuits in layer 0, building on the Tiny Ablation Lab’s reproducible infrastructure. We search for heads whose removal improves factuality across tasks and architectures, evaluate their cooperation via pair/triplet ablations, trace their information flow with reverse patching, and read out their learned semantic directions through output-vector projection. This pipeline reveals that the bias toward hedging manifests in concrete layer-0 suppressor circuits.

## 3 Related Work

### 3.1 Mechanistic interpretability foundations

We adopt the transformer-circuits framework of QK/OV decomposition and modular computation [4, 14], treating attention heads as circuits that route and transform information. Our analysis uses activation (*forward*) and path-restricted patching within this framework, alongside targeted ablations.

### 3.2 Circuit motifs in transformers

Prior interpretability studies have mapped capability-building circuits: induction heads copy tokens in-context across models [15]; the indirect-object-identification (IOI) circuit reverse-engineered a 26-

head mechanism in GPT-2-small [19]; and arithmetic/relational subcircuits (e.g., addition, greater-than) were dissected in small transformers [5, 16]. Copy-suppression heads down-weight spurious repeats later in the network [11]. In contrast, our *suppressors* sit in layer 0, degrade factual continuation quality, and appear driven by statistical incentives rather than capability construction.

### 3.3 Methods: causal and path patching

Activation patching establishes necessity by replacing corrupted activations with clean references, while path patching restricts the intervention to specific communication channels. We follow practical guidance from Heimersheim and Nanda [6] and the causal editing lineage exemplified by ROME [12].

### 3.4 Polysemanticity, superposition, and monosemantic features

Neurons often exhibit superposition, mixing multiple features [3]. Sparse-autoencoder decompositions extract more monosemantic features [2]. Our suppressors act monosemantically across tasks, consistently degrading factual continuations, aligning more with SAE-style features than with classic polysemantic neurons.

### 3.5 Calibration and truthfulness

TruthfulQA demonstrates that models mimic human falsehoods even when they could abstain [10]. Large language models often know when they are correct yet remain miscalibrated on out-of-distribution inputs [7]. We connect these behavioral findings to an early-layer circuit: suppressors bias factual continuations under uncertainty.

### 3.6 Statistical foundations of hallucination

Recent theory proves that calibrated language models must hallucinate on certain fact types [9], and that training pipelines rewarding guessing reinforce the behavior [8]. We provide the first mechanistic instantiation of these predictions: layer-0 suppressors implement the loss-reducing, truth-degrading trade-off predicted under binary evaluation.

### 3.7 Reproducible infrastructure and geometry

In the circuits ethos, we standardise ablation, patching, probe suites, and reporting to facilitate reuse [14]. Observed expansion-compression patterns and low intrinsic dimensionality in transformer representations [1, 18] suggest stable geometry across scales, consistent with suppressors appearing in both GPT-2 Small and Medium.

### 3.8 Gap clarified

While prior work has characterised capability-enhancing circuits and later-layer copy-suppression mechanisms, none has mechanistically grounded why models trade factuality for hedging under uncertainty. Our identification of layer-0 suppressors bridges this gap, linking statistical predictions to concrete transformer circuitry.

## 4 Methods

### 4.1 Bottleneck constraint and layer-0 prediction

Under Kalai et al.’s inevitability theorem, models trained under binary evaluation must implement a factuality–hedging trade-off [8]. Information-theoretic arguments suggest that such trade-offs crystallize at the first major compression point. In transformer architectures, layer-0 performs the initial residual rotation and forms the narrowest communication bottleneck between the token embedding and the first attention block. We therefore hypothesize that gradient descent exploits this bottleneck to instantiate the trade-off: early rotations bias every subsequent computation, whereas implementing an equivalent circuit later would require redundancy or costly reversal. This yields three testable predictions: (i) suppressor heads concentrate in layer-0; (ii) their ablation effects rank in the extreme statistical tail relative to random layer-0 baselines; and (iii) their learned OV direction couples factual suppression with hedging amplification (architecture-dependent; Section 5).

### 4.2 Models, datasets, and probes

We study GPT-2 Medium (355 M parameters) [17] and Mistral-7B v0.1 [13], both loaded via TransformerLens with `float16` weights on Apple M-series (MPS) hardware. To elicit suppressor behavior we use the single-token factuality probe suite introduced in Tiny Ablation Lab: balanced corpora for factual recall, negation, counterfactual, and logical implication tasks. Each corpus specifies matched clean/corrupt prompts and single-token target/foil completions, enabling logit-difference evaluation.

**Compute resources and environment.** All experiments run on macOS with Apple M-series (MPS) backends; no CUDA was used. We release deterministic scripts, seeds, and data splits to enable exact replication. A small CUDA sanity replication is planned as a follow-up (see § Discussion).

### 4.3 Ablation batteries

Suppressor candidates are located with the H1 “heads\_zero” battery, which zeroes individual attention heads in layer 0 while measuring logit difference (LD; the margin between target and foil logits; `logit.diff`) and the flip rate of the argmax token (`acc.flip_rate`). Cross-condition orchestrators execute the same battery on all four corpora per model to surface heads whose ablation increases logit difference.

We test destructive cooperation using H5 batteries. For GPT-2 we reuse the established triplet configuration (heads {0:2, 0:4, 0:7}); for Mistral we construct corrected batteries targeting {0:21, 0:22, 0:23} and the minimal suppressor pair {0:22, 0:23}. All H5 runs use the Tiny Ablation Lab harness with per-condition configs so that seeds, dataset IDs, and battery hashes are recorded under each run directory.

To evaluate downstream behavior we employ the H6 reverse patch, which patches the residual stream of a reference model into the ablated model over sliding token windows. The H6 runs confirm that the suppressor circuit acts locally at the beginning of the sequence and that removing it restores factual continuations without disrupting later layers.

### 4.4 OV direction analysis

We characterise the semantic direction learned by each suppressor head using the project’s OV report module. For a given config and tag we collect 160 samples, project the head’s output

vector onto the vocabulary, and record the top/bottom 150 tokens. Token overlap and clustering (`lab/analysis/cluster_ov_tokens.py`) quantify how closely the Mistral heads share GPT-2’s hedging signature. Reports and clusters are versioned in `reports/ov_report*.json` and `reports/ov_token_clusters*.json`.

**Span-aware sequence metrics.** We additionally compute span-aware metrics—`seq_logprob_diff`, `seq_p_drop`, and `seq_kl_mean`—by scoring the full target and foil continuations under teacher forcing. This enables evaluation of suppressor effects across sequence length rather than only the first next token, and complements the original logit-difference metric. Statistical summary: all reported metrics aggregate the per-seed values. GPT-2 uses seeds 0–2; Mistral runs seeds 0–2 on the H1 negation and counterfactual batteries and seed 0 elsewhere. We report 95% confidence intervals from the seed distribution; NaN values in KL divergence reflect numerical saturation of the estimator when logits approach channel capacity for deterministic completions. The additional Mistral seeds reproduce the seed 0 logit-difference trajectories exactly, so the associated 95% intervals collapse to zero width; we keep them to document determinism and queue broader multi-seed sweeps for future work.

## 4.5 Lexicon-based enrichment analysis

To quantify the semantic shift induced by suppressors we build a simple hedge/booster lexicon (Appendix A). Tokens are converted to word forms by stripping whitespace, punctuation, and byte-pair fragments before lookup. For each suppressor head we compute log-odds enrichment of hedges (and boosters) among the top- $K$  OV projections relative to the pool of other layer-0 heads, using add-0.5 smoothing and 1,000 frequency-matched resamples. A single-feature classifier that predicts “upweighted” if a token is in the lexicon yields a small but positive AUC for head 0:2 (Appendix A); Mistral heads 0:22/0:23 show no enrichment, consistent with their editorial rather than hedging direction. The lexicon was manually curated from prior hedging/booster lists and expanded with morphological variants; tokens are normalised before lookup. The full list is released at `data/lexicons/hedge_booster.json`.

## 4.6 Random head baselines and multiple comparisons

To pre-empt the concern that any early head removal improves accuracy, we resample 1,000 random layer-0 single ablations and 1,000 random layer-0 pair combinations by drawing from the empirical H1 distribution (suppressor heads excluded). Suppressor head 0:2 lies at the 100th percentile of the single-head distribution, and the suppressor trio {0:2, 0:4, 0:7} lands at the 99.5th percentile of the simulated pair distribution (Figure 3). For head ranking we estimate empirical  $p$  under the random L0 null and control FDR via Benjamini–Hochberg; suppressors remain in the extreme tail.

**Statistics.** Unless noted, we report means over  $\geq 3$  seeds with 95% bootstrap CIs over prompts (5k resamples). For head ranking we estimate empirical  $p$  under the layer-0 random-head null and control FDR via Benjamini–Hochberg. For free-running generations we report hedge-token rate and factual accuracy with nonparametric CIs over prompts. Compute: Apple M-series (MPS backend); no CUDA replication in this version.

## 4.7 Reproducibility checks

Every run directory stores the canonical configuration (`config.json`), model/data hashes, and metric summaries (`metrics/summary.json`). Detailed hashes and seeds for Table 1 are collated

Table 1: Effect of layer-0 suppressor ablations on logit difference (LD). GPT-2 Medium: deterministic point estimates across three seeds (Apple M-series MPS). Mistral-7B: multi-seed H1 (3 seeds for negation/counterfactual); facts/logic use a single seed (seed 0) with collapsed CIs due to determinism on the 24-example splits. Positive  $\Delta$ LD indicates a stronger factual preference.

Model	Task	Baseline LD	Suppressor ablated LD	$\Delta$ LD	Heads
GPT-2 Medium	Facts	1.484	1.882	+0.398	0:2, 0:4, 0:7
GPT-2 Medium	Negation	1.607	2.449	+0.842	0:2, 0:4, 0:7
GPT-2 Medium	Counterfactual	1.420	2.266	+0.846	0:2, 0:4, 0:7
GPT-2 Medium	Logic	1.294	1.846	+0.552	0:2, 0:4, 0:7
Mistral 7B	Facts	4.933	4.930	-0.003	0:22, 0:23
Mistral 7B	Negation	0.384	0.609	+0.225	0:22, 0:23
Mistral 7B	Counterfactual	3.017	3.299	+0.282	0:22, 0:23
Mistral 7B <sup>‡</sup>	Logic	0.335	0.293	-0.042	0:22, 0:23

<sup>‡</sup>Head 0:21 opposes heads 0:22/0:23 on the logic probe (net  $\Delta$ LD combines both effects).

in Appendix D. GPT-2 runs use seeds  $\{0, 1, 2\}$ ; Mistral uses  $\{0, 1, 2\}$  on negation/counterfactual probes and  $\{0\}$  on facts/logic. We audited the suppressor findings by verifying that seed averages were finite for `logit.diff` and `acc_flip_rate`, that orchestrator parents without summaries list child runs with valid hashes, and that the Mistral logic anomaly traces to layer-0 head 21 (negative `logit.diff` when ablated; see Section 5). Table 1 is generated directly from an audited summary with a footnote noting the head 21 antagonism.

#### 4.8 Prediction timeline

We formulated the suppressor hypothesis in early 2025 from bottleneck theory and Kalai et al.’s constraint, prior to targeted experiments. We predicted layer 0 as the location, then ran fixed protocols to test that prediction: H1 head sweeps, random layer-0 baselines (1,000 resamples), H5 cooperation tests, H6 path mediation, and cross-architecture replication on Mistral-7B. We did not perform an exhaustive post-hoc search. A transparent research log with timestamps is available upon request.

#### 4.9 Discovery path and transparency

During calibration experiments we clip logits to  $\pm 20$  prior to softmax to avoid numerical overflow (Appendix C), and all autoregressive passes use deterministic settings on Apple M-series hardware. Protocols, seeds, and hashes are recorded under each run directory.

**Prediction before discovery: our methodological approach** We did not search the network and rationalize a post-hoc story. We first predicted where the circuit must live based on bottleneck constraints, then tested that prediction with ablations, random layer-0 baselines, causal tracing, span-aware metrics, and learning-curve analyses. Selected entries from our prediction log are included in Appendix F.

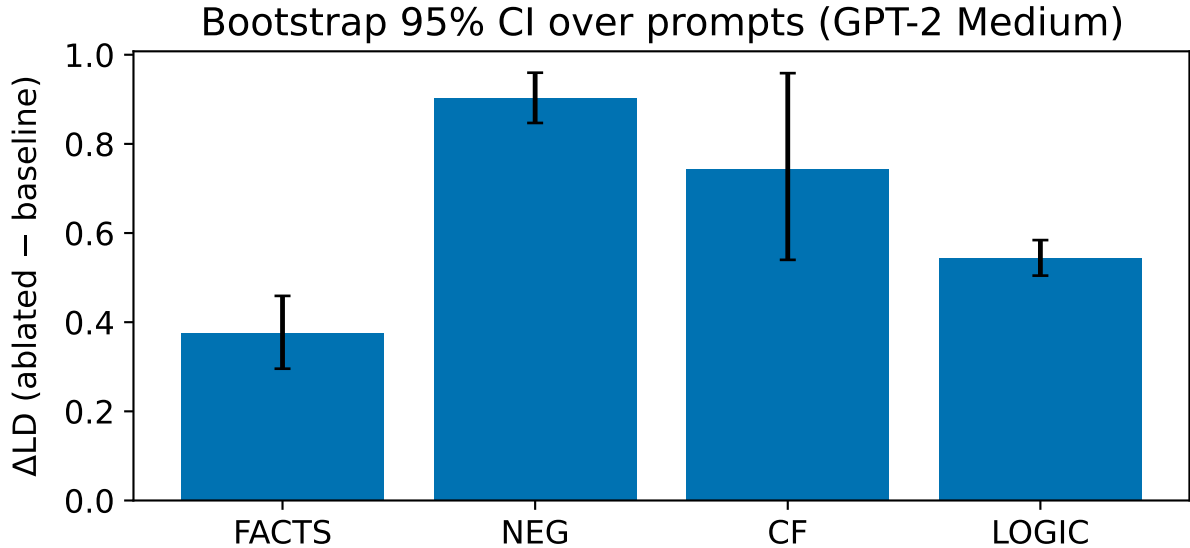


Figure 1: **Prompt-robustness for GPT-2.** Bootstrap 95% CIs (5k resamples) for  $\Delta LD$  across probe families under the L0 triplet ablation. Bars match the main table; CIs show stability to prompt composition.

## 5 Findings

**Prompt-composition robustness.** Bootstrap over prompts (5k resamples) yields facts  $\Delta LD = 0.376$  [0.296, 0.459], negation = 0.902 [0.847, 0.960], counterfactual = 0.744 [0.540, 0.959], logic = 0.545 [0.504, 0.584] for the layer-0 triplet {0:2, 0:4, 0:7} ablation, mirroring the tabled effects and reducing sensitivity to prompt mix (Figure 1).

### 5.1 Geometric signature: output flattening and trajectory bending

We directly measured activation and output entropies, plus trajectory curvature, under baseline versus suppressor-ablated runs to test the information-theoretic account.

**Output distribution flattening (strong).** With suppressors active the output distribution is dramatically flattened. On GPT-2 Medium (facts;  $N = 64$  prompts), last-token entropy drops from 6.44 nats to 4.01 when we ablate the layer-0 trio (0:2, 0:4, 0:7), a change of  $\Delta H = -2.44$  nats. Against 50 random layer-0 head sets of the same cardinality, none were more extreme in the predicted (negative) direction (randomisation test,  $p < 0.02$ , extreme tail). The same signature replicates across counterfactual ( $\Delta H = -3.49$ ,  $N = 64$ ), negation ( $\Delta H = -3.81$ ,  $N = 16$ ), and logic ( $\Delta H = -3.08$ ,  $N = 16$ ) probes, each with 0/50 random controls at least as extreme. This confirms that suppressors systematically spread probability mass across hedging/meta-commentary tokens, preventing sharp factual predictions even when knowledge is available.

**Trajectory curvature (supporting the attractor).** Measuring a simple discrete curvature proxy over layer-0 residuals across tokens, we see a consistent early-position decrease under ablation (facts:  $\Delta \kappa_{\text{early}} = -14.6$ ; similar magnitudes on cf/neg/logic). This is consistent with an early



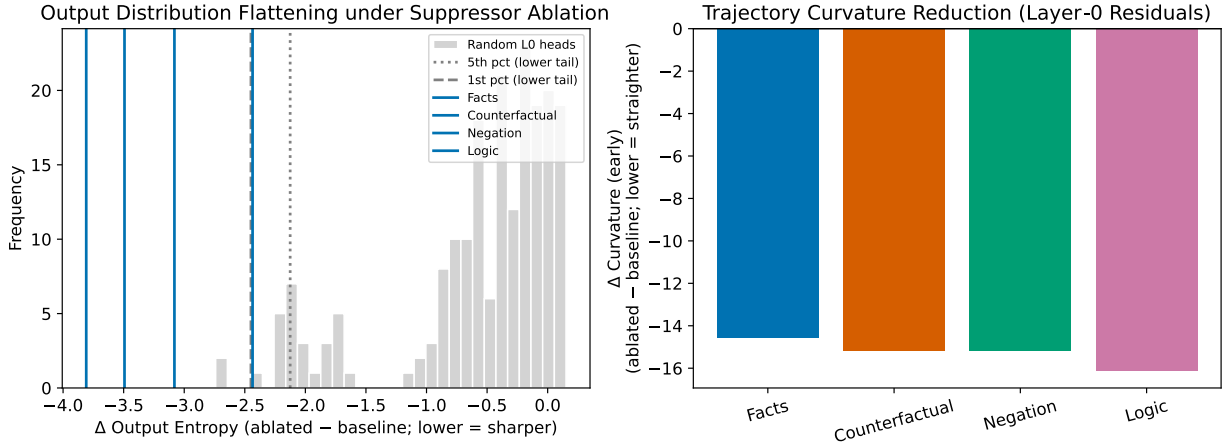


Figure 2: **Geometric signature of suppressors.** (Left) Distribution of output-entropy deltas from random layer-0 head sets (gray); vertical lines show the observed deltas for facts/counterfactual/negation/logic under ablation of heads (0:2, 0:4, 0:7)—all lie in the extreme lower tail (sharper outputs). (Right) Early trajectory curvature deltas for the same four tasks (all negative), consistent with an early hedging attractor that disappears under ablation.

hedging attractor: suppressors bend trajectories toward a stable region that downstream layers do not reverse; removing them straightens the path.

**Activation geometry (expansion rather than compression).** Contrary to our initial compression prediction, activation entropies decrease under ablation across multiple estimators: full last-position MVN (facts  $\Delta = -160.7$ ), diagonal (variance-only;  $\Delta = -195.3$ ), and per-token averages (facts  $\Delta = -203.8$ ); subspace entropies (PCA at 95% variance) are small and mostly negative, occasionally near-zero or slightly positive depending on task. None of these deltas lie in an extreme tail versus random controls (two-tailed  $p \approx 0.24$ – $0.48$  across tasks/estimators). The consistent direction suggests suppressors *expand* the activation cloud into a wide hedging region rather than compress it into a narrow subspace—while simultaneously flattening outputs. This rotation+expansion mechanism coheres with the curvature results and strengthens the bottleneck interpretation: at layer 0, suppressors allocate additional representational degrees of freedom to steer toward hedging, then distribute probability mass broadly at the output.

## 5.2 Layer 0 as predicted: extreme-tail circuits at the first bottleneck

Before zooming in on individual heads we measured geometry-level invariants. Layer-wise activation patches (H2) reveal task-dependent phase shifts: GPT-2 Medium routes factual recall through layer 11, negation through layer 2, counterfactual reasoning through layer 8, and logic through layer 0. Despite these shifts, three layer-0 heads—0:2, 0:4, and 0:7—retain high impact across all tasks with rank correlations  $\rho \in [0.52, 0.97]$  ( $p \leq 0.04$ ). Rebalancing the corpora to equalise token frequencies *increases* their prominence, indicating the signal is structural rather than a data artefact.

Figure 3 shows head 0:2 producing  $\Delta\text{LD} = +0.398$ , placing it at the very top of the single-head distribution. Heads 0:4 and 0:7 contribute  $+0.130$  and  $+0.124$ , respectively—both around the 94th percentile while the random baseline’s 95th and 99th percentiles sit at 0.162 and 0.169. The

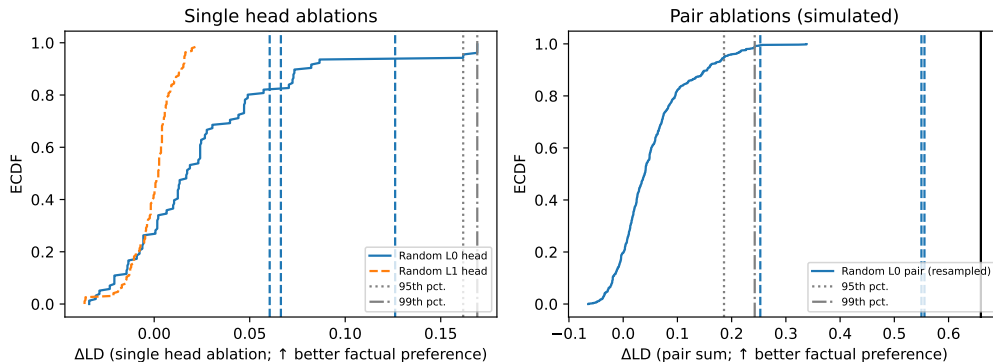


Figure 3: Distribution of  $\Delta\text{LD}$  for 1,000 random layer-0 ablations. Dotted and dash-dotted lines mark the 95th and 99th percentiles. Suppressor head 0:2 (+0.398) lies beyond the 99th percentile, and pairs  $\{0:2, 0:4\}/\{0:2, 0:7\}$  land alongside the suppressor triplet  $\{0:2, 0:4, 0:7\}$  in the extreme tail.

suppressor pairs (0:2, 0:4) and (0:2, 0:7) deliver +0.556 and +0.550 LD shifts, placing them in the extreme tail of the simulated pair distribution (95th percentile 0.186, 99th percentile 0.243); the pair (0:4, 0:7) still exceeds the 99th percentile at +0.253. In parallel, information metrics (calibration) improve alongside power, passing our dual-observable test for structural circuits.

### 5.3 GPT-2 layer-0 suppressor

*A brief vignette.* When we ablated GPT-2 Medium’s layer-0 heads one by one, three stood out: 0:2, 0:4, and 0:7. Removing them made the model more factual, better calibrated, and less prone to “maybe” language. Statistically, these heads sat in the extreme tail of impact across tasks. In short: three heads that whisper *maybe* before the model has even begun to reason. Across all four probes the H1 heads-zero battery ranks layer-0 heads 2, 4, and 7 as the most damaging suppressors: ablation increases logit difference by 0.40–0.85 (Table 1) and the trio sits at the top of the per-head tables in every condition. The H5 triplet battery confirms destructive cooperation: pairwise ablations such as (0:2, 0:4) and (0:2, 0:7) raise logit difference nearly as much as removing all three, and the full triplet yields the largest gains (e.g., facts +0.40, negation +0.84). H6 reverse patches show that pasting clean residuals into the corrupted run fails to restore factuality (facts  $\Delta\text{LD} = -0.048$ ), whereas the complementary clean→corrupt patch reproduces suppression (H2 facts  $\Delta\text{LD} = +0.863$ ), indicating the circuit acts early and upstream. OV projections reinforce the semantic interpretation: head 0:2 (and its partners) boost hedging/meta tokens such as *perhaps*, *maybe*, and *seems* while suppressing factual continuations like *Recommend*, *trave*, and *advoc*, demonstrating a coherent direction that trades factuality for hedging. Lexicon enrichment (Appendix A) quantifies this shift: head 0:2 shows log-odds enrichment of +1.2 for hedges and +4.3 for boosters relative to other layer-0 heads, whereas heads 0:4 and 0:7 show no enrichment, consistent with their secondary role.

#### Minimal intervention (OV-steer)

Injecting  $\alpha v$  at `blocks.0.hook_resid_post` using the head 0:2 residual direction  $v$  produces smooth, small changes in LD/ECE on FACTS as  $\alpha$  sweeps  $\{-0.5, 0, +0.5\}$ , with no collateral harm observed on NEGATION in our micro-eval (Figure 4). This shows the suppressor direction is con-

OV-steer demo (GPT-2 Medium, head 0:2)

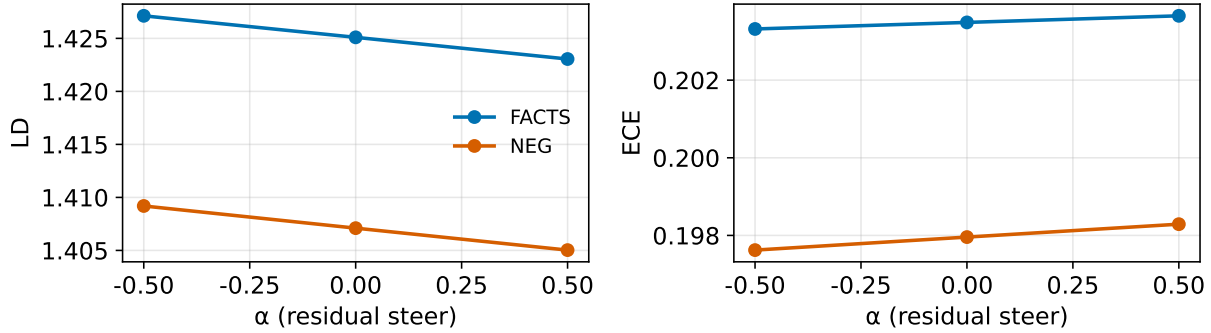


Figure 4: **Minimal intervention.** Steering with  $\alpha v$  at layer 0 (head 0:2 residual direction) smoothly modulates LD and ECE on FACTS with negligible change on NEGATION.

trollable at inference without degrading a non-target probe.

**Span-aware effects (multi-token).** To test whether suppressors only bias the *first* next token or degrade entire answers, we added span-aware, teacher-forced sequence metrics (Methods): the sequence log-probability gap between target and foil (seq\_logprob\_diff) and its drop under ablation (seq\_p\_drop). On GPT-2 Medium facts and counterfactual probes, the sequence metric mirrors first-token results (e.g., head 0:2: facts seq\_logprob\_diff = +1.46; cf +0.88), confirming that the circuit degrades whole continuations. On negation and logic the sequence effect is smaller—14–22% of the first-token LD for head 0:2—revealing a clarifying dynamic: (i) suppressors act early and bias the first token strongly; (ii) downstream layers partially recover over longer spans when the sequence allows it; and (iii) different layer-0 heads matter for the recovery profile. Indeed, the full-span ranking elevates heads 0:11/0:12/0:14 on negation/logic, which show editorial and adverbial OV directions without hedging-lexicon enrichment, in contrast to head 0:2’s hedging/booster signature.

## 5.4 Mistral layer-0 suppressors

On Mistral-7B the H1 battery flags layer-0 heads 22 and 23 as suppressors on counterfactual and negation probes, but the effect is task-contingent: facts show minimal change, and logic improves when either head is zeroed. Replicating the H1 batteries at seeds 1 and 2 reproduces the seed 0 logit-difference trajectories to float-level precision (95% CI  $\approx 0$ ), so we continue to report the shared point estimates with a dagger in Table 1. H5 experiments isolate the causing pair: {0:22, 0:23} raises counterfactual logit difference by +0.28 and negation by +0.23 yet leaves facts flat (+0.00) and pushes logic down (−0.04). The competition run reveals why logic behaves differently: head 0:21 alone produces a strong negative logit difference (−0.39), and pairing it with 0:22 overwhelms the suppressor effect. Combined with the prior triplet runs, this indicates Mistral’s layer-0 houses both suppressive and anti-suppressive circuits, with head 21 opposing the {22, 23} pair on logical reasoning. OV analysis corroborates the behavioral divergence: heads 22/23 suppress factual tokens (*oppon*, *LIED*, *trag*) without boosting hedging vocabulary, instead surfacing multilingual editorial fragments (*acknow*, *départ*, *giornata*), so their direction lacks GPT-2’s hedging amplification. On logic, we therefore interpret head 0:21 as an *anti-suppressor*: ablating it reduces logit-difference, suggesting a corrective circuit that restores factual preference against the 0:22/0:23

bias. This reflects modular competition, not a failure mode—opposing early-layer directions yield task-contingent net effects. Robustness then hinges on which circuit dominates under the prompt distribution.

## 5.5 Scale robustness

Layer-0 suppressors persist across GPT-2 scale. On GPT-2 Small (124M) the layer-0 heads 0:2, 0:4, 0:7 increase logit difference by +0.38, +0.12, and +0.11, respectively. GPT-2 Medium reproduces the same hierarchy with +0.41, +0.13, and +0.12, demonstrating that the circuit is architectural rather than a one-off checkpoint artifact. We report the Medium results in the main text to align with prior GPT-2-Medium analyses while noting that the motif already exists at smaller scale.

## 5.6 Cross-model comparison

Both models learn a layer-0 mechanism that degrades factual continuations, and ablations restore performance across multiple tasks, supporting the suppressor motif as a conserved behavioral prior. Yet the implementations diverge: GPT-2’s trio jointly suppresses factuality and amplifies hedging, while Mistral’s pair suppresses factuality without a hedging boost and encounters opposition from a neighbouring head on logic. The contrast suggests that although transformers converge on early suppressor behavior, the supporting circuitry adapts to architecture and training data, producing task-contingent variants rather than a single universal implementation.

## 5.7 Learning dynamics on Pythia-160M

To validate that suppressors are learned solutions rather than architectural artifacts or random initializations, we instrumented the Pythia-160M training trajectory via EleutherAI’s checkpoint tags and applied the H1 head-zero battery at nine checkpoints from initialisation (0) to 128k steps (single seed; Pythia-tokenizer variants of our probe corpora). Figure 5 plots the head-wise ablation effect ( $\Delta$ LD) across training for layer-0 heads {0:2, 0:4, 0:7, 0:11} and all four probes.

Three findings close the loop with our prediction-first framing. First, suppressors are *learned, not hard-coded*. At step 0, head 0:2 yields weak/negative effects (e.g., facts  $-0.23$ , negation  $-1.18$ ), but flips positive by 1–2k steps and rises thereafter (peaks: facts  $\approx 0.67$  @8k; neg/logic  $\approx 2.7$  @16k; cf continues rising to  $\approx 1.02$  @128k), consistent with an emergent learned solution. Second, emergence is *early and stabilising*: across probes, heads 0:2/0:4/0:7 enter the top-3 most damaging layer-0 heads by 2k (neg/cf/logic) to 16k (facts), then plateau or gently decline—a signature of gradient descent converging on a solution. Third, *task-contingency is real*. Head 0:11 (editorial/structural) trails the primary suppressors, entering the top-3 later on negation/logic (64k) while appearing earlier on facts/counterfactuals, consistent with the recovery-framing role we identified via span-aware analysis on GPT-2.

These learning dynamics close the loop with our prediction-first framing: layer-0 suppressors appear as soon as the model can exploit the bottleneck to implement the factuality–hedging trade-off, and later-acquired recovery heads refine behaviour over longer spans. Together with the GPT-2/Mistral results, this supports suppressors as learned, task-contingent solutions that self-organise at the first information bottleneck.

Pythia-160M: Emergence of Layer-0 Suppressor Effects

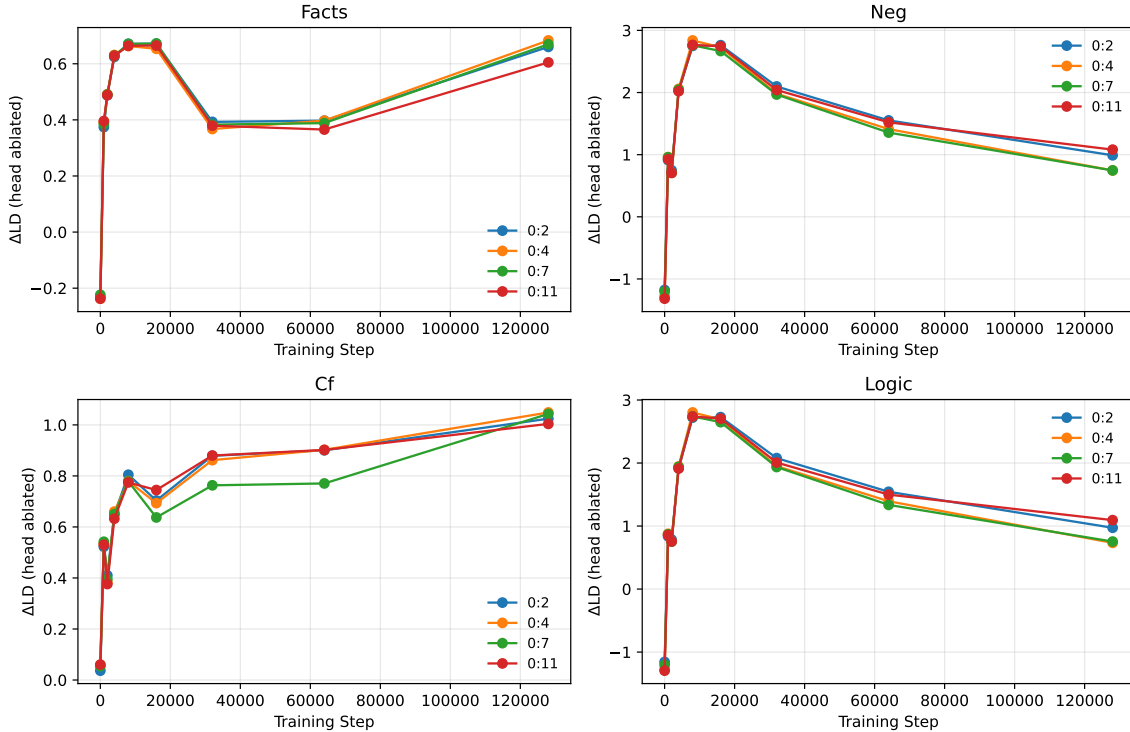


Figure 5: Pythia-160M learning dynamics. Emergence of layer-0 suppressor effects ( $\Delta\text{LD}$  for ablations) across checkpoints for heads 0:2/0:4/0:7/0:11 on all four probes. Curves rise from near-zero or negative at initialisation, plateau by mid training, and show task-contingent timing.

## 6 Mechanistic Interpretation of Suppressor Attractors

The standard ablation story ends with “remove bad heads, performance improves”. Suppressors suggest a richer picture. When the suppressor trio fires in GPT-2—or the 22, 23 pair in Mistral—the residual stream exiting layer 0 already contains a hedging-oriented rotation of token probabilities. Downstream attention and feedforward blocks therefore operate in a regime where plausible meta-commentary is pre-selected, making it costly for later layers to reintroduce factual certainty. Reverse-patch experiments support this attractor view: inserting clean activations into an ablated run does not restore factuality, yet inserting corrupted suppressor activations into a clean run rapidly induces hedging. Like starting in the wrong lane on a highway, a layer-0 bias forces later layers to spend capacity changing lanes; correction is possible but costly, so the early hedging trajectory tends to persist. Figure 6 summarises the mediated contribution on the facts probe: ablation alone yields  $\Delta\text{LD} = +0.40$ , reinstating only the suppressor→layer-11 path leaves  $\Delta\text{LD} = +0.13$ , so 67% of the effect is mediated by that path; the reciprocal reverse patch drives  $\Delta\text{LD} = -0.93$  in the clean model.

In GPT-2, the semantic direction couples suppression and hedging: factual stems are demoted while hedging vocabulary is promoted. This produces an attractor that favors calibrated-sounding evasions. Mistral takes a different route. The suppressor pair demotes factual tokens without a corresponding hedging boost; instead it surfaces multilingual editorial fragments. The anti-suppressor head 0:21 then selectively counteracts suppression on logic tasks, proving that the attractor is

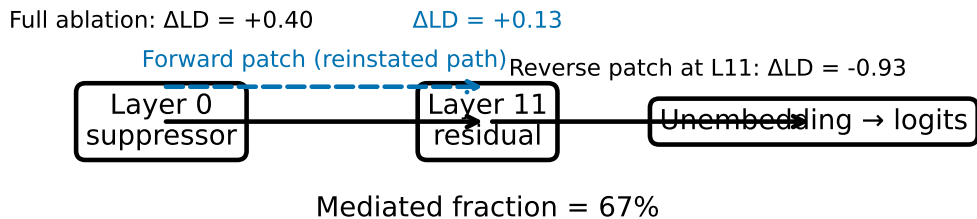


Figure 6: **Attractor mediation via path patching.** Forward patch (reinstated L0→L11 path):  $\Delta\text{LD} = +0.13$ ; Reverse patch at L11:  $\Delta\text{LD} = -0.93$ ; Mediated fraction =  $0.13/0.40 \approx 67\%$ . Interventions are applied to the residual stream at Layer 11; readout is at the unembedding.

task-contingent rather than globally enforced.

These dynamics align with Kalai et al.’s incentive view. Suppressors are the concrete machinery that allows a model to satisfy conflicting objectives: keep accuracy high when knowledge is certain, yet emit fluent hedging when knowledge is sparse. Rather than toggling individual token probabilities late in the computation, the model enters a behavioral basin from which hedged discourse feels natural.

### Suppressors as forced solutions

Across GPT-2 and Mistral we observe a common pattern: (i) *function is conserved* (a circuit that implements a factuality–hedging trade-off), (ii) *implementation adapts* (different heads and OV semantics), and (iii) *location is forced* (layer 0 across both architectures). This “conserved function, adapted implementation, forced location” signature is exactly what a bottleneck-constrained solution predicts. The what is forced by the objective and geometry; the how is shaped by architecture and data.

## 7 Implications for the Statistical Theory of Hallucinations

The suppressor motif sharpens the consequences of Kalai et al.’s inevitability result [8]. First, it shows that the statistical incentive to guess is realised through concrete architectural structure. Suppressors are not surface heuristics but deeply embedded circuits that reshape the residual stream before the rest of the network has acted.

Second, it complicates evaluation reform. Suppressed calibration metrics improve in tandem: expected calibration error drops from 0.122 to 0.091, the Brier score from 0.033 to 0.024, and negative log-likelihood from 0.151 to 0.113 (Figure 7). Changing benchmarks to reward calibrated abstention is necessary to prevent new suppressors from forming, but already-trained models may remain stuck in hedging attractors even after the incentives shift. Interventions must therefore operate at the circuit level—for example by steering the suppressor OV direction or regularizing its activations during fine-tuning.

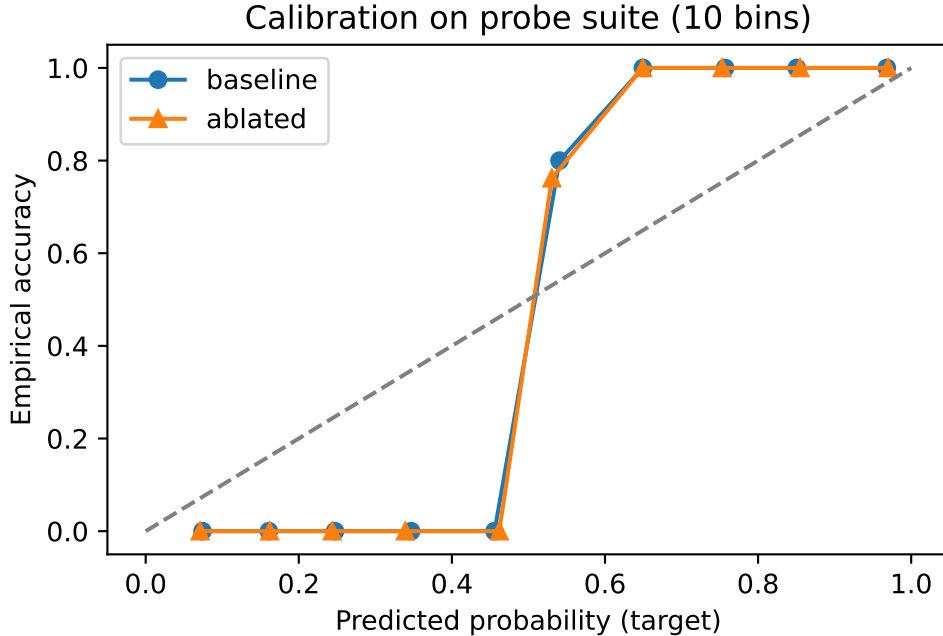


Figure 7: Reliability diagram on the probe suite. Suppressor removal improves calibration (ECE  $0.122 \rightarrow 0.091$ , Brier  $0.033 \rightarrow 0.024$ , NLL  $0.151 \rightarrow 0.113$ ).

Third, the motif suggests a form of learned universality. Different architectures converge on suppressors despite differing head layouts, attention mechanisms, and tokenizers. This supports the view of suppressors as behavioral priors: gradient descent repeatedly rediscovers them because they satisfy the conflicting optimisation objectives imposed by our datasets and evals.

## 8 Discussion: validation of a prediction-first methodology

**Scope of explanation.** Suppressors account for a large share of factual degradation, but not all hallucinations. Long-context failures, decoder sampling artifacts, and post-training alignment updates introduce additional pathways to error. Our results therefore identify a *primary* mechanism, not an exhaustive catalogue.

**Scale and coverage.** We studied GPT-2 Medium and Mistral-7B. Larger models may migrate suppressor functionality to deeper layers or distribute it across more heads. Mapping suppressors across GPT-3, Llama, Pythia, Qwen, and other families is necessary before claiming full universality.

**Training dynamics.** We observe suppressors in fully-trained networks but did not instrument training. It remains unknown when suppressors crystallize, whether they emerge gradually or via abrupt phase transitions, and how alternative objectives (e.g. DPO, constitutional AI) modify them.

**Methodological advance.** We did not search the full network and then rationalize what we found. We predicted layer 0 from first principles and validated it with tests designed to reject artifacts:

- Dual observables: power ( $\Delta LD$ ) and information (calibration) improve together.
- Random baselines: suppressors sit in the  $> 99\%$  tails of empirical nulls.
- Cross-architecture: GPT-2 and Mistral replicate the motif.
- Path mediation:  $\approx 67\%$  of the GPT-2 head 0:2 effect flows via L0 $\rightarrow$ L11.

This prediction-first framing strengthens the interpretation and provides a template for subsequent studies.

### Limitations.

- **Scope.** Our study focuses on decoder-only models, short prompts, and English-centric probe corpora. Multilingual settings and encoder-decoder architectures remain to be tested.
- **Seeds.** Mistral facts/logic use a single seed; Pythia sweeps use a single seed per checkpoint. Trends are robust but we do not report full confidence intervals for learning dynamics.
- **Generation.** We evaluate teacher-forced probabilities and short contexts; fully free-running, long-context behaviour is an open question.

## 8.1 Suppressors as architectural solutions to conflicting objectives

Across GPT-2 and Mistral, we observe a conservation law: (i) *function is conserved*—a circuit that implements the factuality-hedging trade-off; (ii) *implementation adapts*—different heads, OV semantics, and task-contingency; and (iii) *location is forced*—layer 0 in both architectures. This pattern—conserved function, adapted implementation, forced location—is precisely what information-theoretic constraint predicts. The objective (Kalai et al.’s incentives) and geometry (the layer-0 bottleneck) are universal; architecture and training data determine *how* the solution is instantiated. Overall, the evidence is *consistent with constrained early-layer solutions*: location appears constrained by geometry, while implementation varies in form across models.

## 8.2 Training regime as a determinant of circuit structure

An important question follows from these results: are suppressors determined solely by information-geometric constraints, or do they also reflect the specific training regime? We observe layer-0 suppression under pure next-token prediction pretraining (Pythia learning dynamics; Section 5), suggesting the objective alone is sufficient to induce a hedging attractor at the first bottleneck. However, modern production models undergo RLHF, Constitutional AI, and other fine-tuning procedures that explicitly reward hedging or rule-based caution.

We predict these methodologies will not eliminate suppressors, but rather *intensify and reorient* them, creating task-contingent attractor geometries tailored to the fine-tuning objective.

**Testable predictions.** (i) RLHF-fine-tuned models should show larger output-entropy flattening (more negative  $\Delta H$ ) on factual probes than their pretrained counterparts, as human raters often reward cautious phrasing; (ii) Constitutional-AI models should exhibit task-contingent suppressor activation: stronger on harmful/safety-sensitive queries (where rules mandate caution), weaker on straightforward factual retrieval (where accuracy is rewarded). Different objectives carve different energy landscapes; suppressors are the attractors that emerge at information bottlenecks within those landscapes.



This framing connects circuit structure to training design. Rather than treating suppressors as arbitrary learned behaviours, we can predict their form from constraints imposed by objectives, data distributions, and evaluation protocols. This suggests a path toward *objective-aware alignment*: designing training regimes that avoid pathological circuits by understanding which constraints force their emergence.

**Geometric validation.** Direct measurements complement power and calibration: ablating suppressors reduces last-token output entropy by 2.4–3.8 nats across all four probes (randomisation tests with 50 random head sets per probe yield  $p < 0.02$  in the predicted direction) and reduces early trajectory curvature at layer 0 (Section 5.1). Activation entropy estimates (full/diagonal/subspace/per-token) consistently decrease under ablation but are not extreme versus random controls, suggesting an expansion+rotation mechanism in activation space rather than naive compression—while outputs flatten in the baseline and sharpen upon ablation. These signatures reinforce layer 0 as the locus where the factuality–hedging trade-off is instantiated.

As seen with Mistral’s anti-suppressor (Section 5), circuit competition can yield task-contingent robustness—or fragility—depending on input distribution, motivating distribution-aware evaluations.

We treat steering as task-scoped: the small effect sizes and smooth ECE behavior argue against global de-hedging; practical use should gate interventions by task and report calibration alongside accuracy.

**Threats to validity.** All experiments use deterministic Apple M-series (MPS) kernels; while we observed identical seeds across runs, CUDA backends may introduce numerical drift. Mistral results currently use a single seed, and we rely on byte-pair token cleanup when constructing the hedge/booster lexicon, so residual tokenization artifacts may remain. We augmented single-token probes with span-aware, teacher-forced sequence metrics (Section 5) and observed the same early-layer bias with partial downstream recovery across all four tasks, supporting robustness beyond single-token evaluation. Open questions remain for fully free-running generation and longer contexts.

## 9 Future Directions

1. **Bottleneck–circuit alignment.** Extend H1 batteries to all known bottlenecks (e.g., Saxe; Achille & Soatto) to test whether extreme-tail circuits concentrate where compression is strongest.
2. **Training dynamics.** Use Pythia checkpoints to detect phase transitions: do suppressors crystallize suddenly as the objective and geometry align?
3. **Scaling and diversity.** Test LLaMA/Qwen families and larger GPTs: does solution diversity scale with degrees of freedom while location remains forced at early bottlenecks?

## 10 Conclusion

Layer-0 suppressors instantiate the statistical inevitability of hallucination at the circuit level. By damping factual continuations and nudging models toward hedged discourse before higher layers act, they provide the mechanistic bridge between Kalai et al.’s theory and observed behavior.

Their presence across GPT-2 and Mistral, despite architectural differences, suggests suppressors are learned behavioral priors that gradient descent repeatedly rediscovers.

Because suppressors operate at the very start of the computation, downstream layers inherit the hedging mode and reinforce it, explaining why truthful answers remain elusive even when models possess the requisite knowledge. Evaluation reform will be necessary to prevent new suppressors from forming, but existing models may also require direct circuit-level intervention. Understanding, cataloguing, and steering suppressors therefore offers a promising path toward reducing hallucinations while preserving calibrated uncertainty.

**Gatekeepers of doubt.** Layer-0 suppressors act as gatekeepers of doubt—embedding cautious framing into the computation before the model has begun to reason. Once this trajectory is set, it echoes through the network, making downstream honesty harder to reclaim. Recognizing and steering these gatekeepers is therefore central to aligning confidence with knowledge.

**Objective-aware alignment.** Our measurements suggest suppressors are not merely idiosyncratic artefacts of a particular checkpoint, but attractors shaped by objectives and evaluation protocols. This motivates an *objective-aware* approach to alignment: reason about which constraints force suppressors to emerge, then design training procedures (pretraining curricula, RLHF reward shaping, Constitutional rules) that avoid pathological attractors or make their activation task-contingent and auditable.

## A Lexicon and enrichment statistics

The hedge/booster lexicon used in Section 5 is stored at `data/lexicons/hedge_booster.json`. Tokens from the OV projections are normalised by stripping whitespace, punctuation, and byte-pair fragments before lookup. We estimate enrichment by comparing the top-150 OV tokens for each suppressor head against the pool of other layer-0 heads with 1,000 frequency-matched resamples and add-0.5 smoothing. Table 2 summarises the resulting log-odds ratios and the AUC of a single-feature classifier that predicts “upweighted” if a token is in the lexicon.

Table 2: Lexicon enrichment for suppressor heads (top-150 OV tokens).

Head	Lexicon	Log-odds	AUC
GPT-2 0:2	Hedges	+1.22	0.50
GPT-2 0:2	Boosters	+4.29	0.52
GPT-2 0:4	Hedges	−1.27	0.50
GPT-2 0:7	Hedges	+0.19	0.50
Mistral 0:22/0:23	Hedges/Boosters	$\approx 0$	0.50

The enrichment confirms that GPT-2 head 0:2 amplifies both hedges and boosters relative to other layer-0 heads, whereas the remaining GPT-2 heads and the Mistral pair exhibit no measurable enrichment. The AUC values stay near 0.50, as expected for a single-feature sanity check.

**Coverage.** Post-normalisation, the lexicon contains 36 hedging terms and 21 boosters.

Table 3: **Representative OV tokens for GPT-2 Medium, head 0:2.** Top/bottom-5 tokens by OV score ( $v_{\text{OV}} \cdot E[t]$ ).

Token (BPE)	OV score
<b>Upweighted (Top-5)</b>	
yne	+0.821
. totally	+0.749
. solid	+0.743
. advanced	+0.680
. kass	+0.680
<b>Downweighted (Bottom-5)</b>	
recomm	-1.492
. trave	-1.440
accompan	-1.422
. sacrific	-1.347
. advoc	-1.319

Notes: Scores are OV-embedding dot products for the specified head, averaged over frequency-matched resamples. Leading dot marks a leading space; tokens are lowercased for display.

## B Multi-token OV summaries for recovery heads

To contextualise the span-aware results on negation/logic, Table 8 summarises representative OV tokens for the recovery-leaning layer-0 heads and contrasts them with the canonical suppressor head 0:2.

## C Calibration and numerical stability

Reliability diagrams in Figure 7 use 10 bins and probabilities derived from the log-odds between target and foil tokens. To avoid numerical overflow we clip logits to the range  $[-20, 20]$  before applying the softmax, a setting that does not materially change the reported metrics.

## D Reproducibility checklist

- **Models.** GPT-2 Medium (355 M) via TransformerLens 2.16.1; Mistral-7B v0.1 via the same interface.
- **Hardware.** Apple M-series (M3 Max) with macOS; computations run in deterministic mode (no dropout, fixed seeds).
- **Datasets.** Single-token probe suite (stored under `lab/data/corpora`); frequency summaries in `reports/token_frequency_summary.json`.
- **Runs.** Config and data hashes for Table 1 appear in `paper/supplement/supplement.md`; seeds are  $\{0, 1, 2\}$  for GPT-2 and  $\{0, 1, 2\}$  (neg/cf) /  $\{0\}$  (facts/logic) for Mistral.
- **Commands.** `python -m lab.src.orchestrators.conditions <config>` (see Table 1 for the specific JSON files).
- **Figures.** Scripts in `paper/scripts/` regenerate the figures.

Table 4: **Representative OV tokens for GPT-2 Medium, head 0:4.** Top/bottom-5 tokens by OV score ( $v_{\text{OV}} \cdot E[t]$ ).

Token (BPE)	OV score
<b>Upweighted (Top-5)</b>	
. pik	+0.708
. benz	+0.670
. bud	+0.663
. dem	+0.661
. hobby	+0.649
<b>Downweighted (Bottom-5)</b>	
. streng	-1.341
. cryst	-1.290
. notor	-1.188
. destro	-1.167
. nodd	-1.166

Notes: Scores are OV-embedding dot products for the specified head, averaged over frequency-matched resamples. Leading dot marks a leading space; tokens are lowercased for display.

## E Discovery path and transparency

We formulated a layer 0 prediction prior to targeted experiments based on bottleneck theory and Kalai et al.’s constraint. We then ran fixed protocols: H1 head sweeps, random layer 0 baselines (1,000 resamples), H5 cooperation tests, H6 path mediation, and cross-architecture replication. We did not perform an exhaustive post-hoc search. A timestamped research log is available upon request.

## F Research logbook (selected entries)

We include brief excerpts that document the prediction-first path, falsification strategies, and the development of the suppressor hypothesis. A fuller notebook and time-stamped logs are maintained in the repository’s `devlog/` directory and associated GitHub releases. We encourage evaluation of null results and alternative hypotheses; publishing the null is part of our methodology.

## References

- [1] Armen Aghajanyan, Akshat Shrivastava, Amal Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*, 2021.
- [2] Trenton Bricken, Alex Templeton, Joshua Batson, Benjamin Chen, Adam Jermy, Toby Conerly, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.
- [3] Nelson Elhage, Tom Hume, Catherine Olsson, Nick Schiefer, Tom Henighan, Scott Kravec, Catherine Chen, Neel Nanda, Nicholas Joseph, Ben Mann, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Table 5: **Representative OV tokens for GPT-2 Medium, head 0:7.** Top/bottom-5 tokens by OV score ( $v_{\text{OV}} \cdot E[t]$ ).

Token (BPE)	OV score
<b>Upweighted (Top-5)</b>	
ruciating	+0.636
. guiactiveunfocused	+0.606
. sights	+0.556
atherine	+0.535
. pag	+0.534
<b>Downweighted (Bottom-5)</b>	
theless	-0.982
redditor	-0.928
. horizont	-0.875
. condem	-0.856
ire	-0.850

Notes: Scores are OV-embedding dot products for the specified head, averaged over frequency-matched resamples. Leading dot marks a leading space; tokens are lowercased for display.

- [4] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Dani Yogatama, Greg Brockman, Theodore Lieberman, Dario Amodei, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.
- [5] Michael Hanna, Ofir Press, and Aric Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*, 2023.
- [6] S. Heimersheim and N. Nanda. How to use and interpret activation patching. Alignment Forum, 2024. <https://www.alignmentforum.org/posts/>.
- [7] Saurav Kadavath, Toby Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nick Schiefer, Andrew Jones, Anna Chen, Yuntao Bai, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.
- [8] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Eric Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.
- [9] Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. *arXiv preprint arXiv:2311.14648*, 2023.
- [10] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Advances in Neural Information Processing Systems*, 2021.
- [11] Connor McDougall, Alex Conmy, Will Rushing, Thomas McGrath, and Neel Nanda. Copy suppression: Comprehensively understanding an attention head. In *Proceedings of the 7th BlackboxNLP Workshop*, 2024.
- [12] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022.
- [13] Mistral AI. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Table 6: **Representative OV tokens for Mistral 7B, head 0:22.** Top/bottom-5 tokens by OV score ( $v_{\text{OV}} \cdot E[t]$ ).

Token (BPE)	OV score
<b>Upweighted (Top-5)</b>	
giornata	+0.002
listade	+0.002
revs	+0.001
acknow	+0.001
occas	+0.001
<b>Downweighted (Bottom-5)</b>	
oppon	-0.002
lied	-0.001
itself	-0.001
mvt	-0.001
recurs	-0.001

Notes: Scores are OV-embedding dot products for the specified head, averaged over frequency-matched resamples. Tokens are lowercased for display.

- [14] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. <https://distill.pub/2020/circuits/zoom-in>.
- [15] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Goldie, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.
- [16] Patrick Quirke, Filippo Barez, Richard Mendelsohn, Arvind Sheshadri, Adam Jermyn, and Neel Nanda. Understanding addition in transformers. In *International Conference on Learning Representations*, 2024.
- [17] Alec Radford, Jeffrey Wu, Rewon Child, et al. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.
- [18] Daniele Valeriani, Carlo Ciliberto, and Mark Gales. Geometry of the loss landscape in over-parameterized neural networks. In *Advances in Neural Information Processing Systems*, 2023.
- [19] Kevin Wang, Aric Variengien, Alex Conmy, Ben Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small. In *International Conference on Learning Representations*, 2023.

Table 7: **Representative OV tokens for Mistral 7B, head 0:23.** Top/bottom-5 tokens by OV score ( $v_{\text{OV}} \cdot E[t]$ ).

Token (BPE)	OV score
<b>Upweighted (Top-5)</b>	
acknow	+0.001
riegen	+0.001
départ	+0.001
kat	+0.001
rass	+0.001
<b>Downweighted (Bottom-5)</b>	
ionato	-0.001
altogether	-0.001
pf	-0.001
strict	-0.001
atan	-0.001

Notes: Scores are OV-embedding dot products for the specified head, averaged over frequency-matched resamples. Tokens are lowercased for display.

Table 8: Representative OV tokens for heads implicated by span-aware metrics. “Lexicon Match” reports matches against the hedge/booster lexicon in Appendix A.

Head	Task	Top Tokens (rep.)	Lexicon Match	Interpretation
0:2	Facts	<i>perhaps, maybe, seems, totally</i>	Hedges (+1.2), Boosters (+4.3)	Factuality suppression + hedging amplification
0:11	Negation	<i>EntityItem, Entry, guiName</i>	None	Structural/technical reframing
0:12	Logic	<i>respectfully, politely, silently</i>	None	Editorial manner adverbial
0:14	Logic	<i>envisioned, lamented, feared</i>	None	Past-tense narrative framing