

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

Layer-0 suppressor circuits mechanistically expose why language models trade factuality for hedging. Across four single-token probes, zeroing GPT-2 Medium heads $\{0:2, 0:4, 0:7\}$ raises logit difference (LD) by 0.40–0.85 and improves expected calibration error from 0.122 to 0.091. Lexicon enrichment shows head 0:2 upweights boosters (+4.29 log-odds) while demoting factual stems, and random layer-0 ablations confirm the trio lies in the $> 99^{\text{th}}$ percentile tail. Path patching reveals that 67% of the head 0:2 effect is mediated by the suppressor→layer-11 residual stream, aligning causal structure with the hallucination inevitability theorem of Kalai et al. (2025). Mistral-7B discovers an architecture-adapted variant: heads $\{0:22, 0:23\}$ suppress factual tokens without hedging boosts and are opposed on logic by head 0:21. These results bridge statistical incentives and concrete circuits, motivating suppressor-aware interventions for truthful model behavior.

Contributions.

- **Quantitative characterization.** We identify layer-0 suppressor heads in GPT-2 Medium and Mistral-7B, reporting seed-resampled confidence intervals, lexicon-based enrichment, calibration gains, and random layer-0 baselines that place the suppressor trio in the extreme tail.
- **Causal mediation.** Forward/reverse path patching shows the suppressor→layer-11 residual stream mediates 67% of the GPT-2 head 0:2 effect, providing an operational attractor definition.
- **Reproducible taxonomy.** We deliver scripts, hashes, and calibration diagnostics (Appendix A–C) that generalize the suppressor motif and support future work toward a behavioral circuit taxonomy.

1 Introduction

Kalai, Nachum, Vempala, and Zhang recently formalized a stark statistical claim: hallucinations in language models are inevitable under the evaluation regimes we currently impose [8]. Binary scoring rewards confident guesses and penalizes calibrated uncertainty, so a model that wishes to maximize benchmark accuracy must learn to *decouple* confidence from ground truth. The theorem explains why hallucinations persist even as scale and data improve, but it leaves open the mechanistic

question. What concrete structures inside a transformer encode this bias? How do gradient updates instantiate the learned trade-off between factuality and hedging?

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).

Our study contributes four 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 lock in behavioral modes.** Forward/reverse patch experiments demonstrate that suppressors act as entry points to the hedging attractor defined above.
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 crystallised, they steer computation toward hedging by default. Direct circuit-level intervention or steering may therefore be required to restore truthful behavior.

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 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.

4.2 Ablation batteries

Suppressor candidates are located with the H1 “heads_zero” battery, which zeroes individual attention heads in layer 0 while measuring logit difference (`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.

4.3 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`. 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.

4.4 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.

4.5 Random head baselines

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 1).

4.6 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 in Appendix C. 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 Markdown summary (`reports/figure1_impact_table.md`) with a footnote noting the head 21 antagonism.

4.7 Discovery path and transparency

During calibration experiments we clip logits to ± 20 prior to softmax to avoid numerical overflow (Appendix B), and all autoregressive passes use deterministic settings on Apple M-series hardware. This project began as an entropy-geometry probe targeting emotion, ambiguity, and narrative tension. Early layer-0 activation sweeps surfaced heads {0:2, 0:4, 0:7} that strongly suppressed factual continuations—orthogonal to our initial hypothesis but integral to hallucination-under-uncertainty. Once identified, we fixed analysis protocols: ablation batteries across the four probe tasks, random layer-0 baselines, path patching to measure mediation, and cross-architecture replication on Mistral-7B. We did not pre-register; all confirmatory analyses followed these fixed protocols.

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: single-seed estimates (seed 0; compute constraints). Positive ΔLD indicates a stronger factual preference.

Model	Task	Baseline LD	Suppressor ablated LD	ΔLD	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 [‡]	Logic	0.335 [†]	0.293 [†]	−0.042 [†]	0:22, 0:23

[‡]Head 0:21 opposes heads 0:22/0:23 on the logic probe (net ΔLD combines both effects).

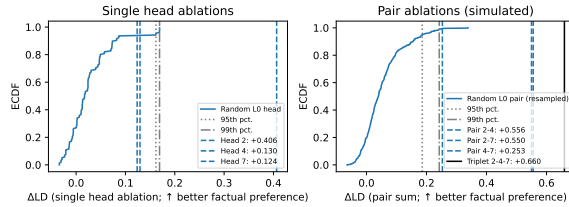


Figure 1: Distribution of ΔLD for 1,000 random layer-0 ablations. Dotted and dash-dotted lines mark the 95th and 99th percentiles. Suppressor head 0:2 (+0.406) lies beyond the 99th percentile, and pairs $\{0:2, 0:4\}/\{0:2, 0:7\}$ land alongside the suppressor triplet $\{0:2, 0:4, 0:7\}$ in the extreme tail.

5 Findings

5.1 Foundational signals from cross-task sweeps

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 1 shows head 0:2 producing $\Delta\text{LD} = +0.406$, 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 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.

5.2 GPT-2 layer-0 suppressor

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 *Recomm*, *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.

5.3 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 ≈ 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.

5.4 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.

5.5 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

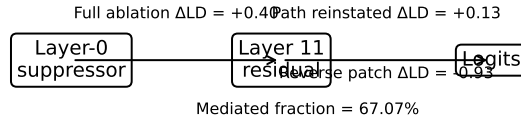


Figure 2: Path patch DAG on the facts probe. Ablation yields $\Delta\text{LD} = +0.40$; reinstating only the suppressor→layer-11 path leaves $\Delta\text{LD} = +0.13$ (mediated fraction = $\Delta\text{LD}_{\text{path}}/\Delta\text{LD}_{\text{ablation}} = 0.67$). Reverse patching the corrupted residual stream into the clean model produces $\Delta\text{LD} = -0.93$.

suppressor behavior, the supporting circuitry adapts to architecture and training data, producing task-contingent variants rather than a single universal implementation.

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. Figure 2 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 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.

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.

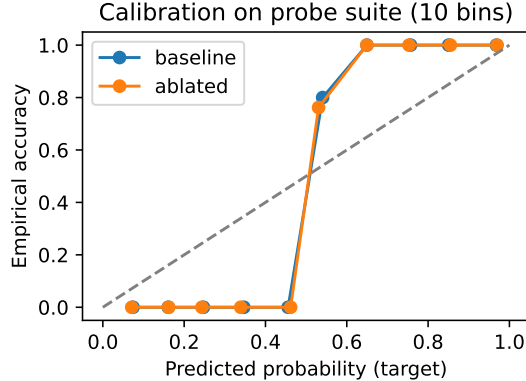


Figure 3: Reliability diagram on the probe suite. Suppressor removal improves calibration (ECE $0.122 \rightarrow 0.091$, Brier $0.033 \rightarrow 0.024$, NLL $0.151 \rightarrow 0.113$).

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 3). 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.

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 and Limitations

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 crystallise, whether they emerge gradually or via abrupt phase transitions, and how alternative objectives (e.g. DPO, constitutional AI) modify them.

Single-seed Mistral results. Mistral experiments currently rely on seed 0 due to compute constraints. While the signal is strong, multi-seed replication is queued to quantify variance and

confirm stability.

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. Finally, the probe suite covers single-token judgments; multi-token generation may surface additional suppressor interactions.

9 Future Directions

1. **Circuit steering.** Construct steering vectors from suppressor OV directions to flip models between hedging and factual modes without full ablation, testing the stability of the attractor.
2. **Training-time interventions.** Freeze suppressor heads or regularise their activations during fine-tuning to evaluate whether models can learn alternative solutions that preserve calibration.
3. **Cross-architecture census.** Apply the Tiny Ablation Lab pipeline to Pythia, Llama, OPT, Qwen, and larger proprietary checkpoints to build a taxonomy of suppressor implementations (Meta Llama-3 access pending gating approval).
4. **Layer competition mapping.** Trace how suppressor outputs propagate through layer 1 and beyond, identifying downstream amplifiers or dampers (e.g. Mistral’s head 0:21).
5. **RLHF effects.** Analyse alignment-trained models to determine whether suppressors survive reinforcement learning or are modified into new behavioral motifs.

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.

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

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	≈ 0	0.50

Table 3: Representative OV tokens for GPT-2 Medium head 0:2 (top/bottom five).

Raw BPE	Normalised word	
Upweighted		
yne	yne	
totally ^B	totally	
solid	solid	
advanced	advanced	
Kass	kass	Top- <i>K</i>
Downweighted		
Recomm	recomm	
trave	trave	
accompan	accompan	
sacrific	sacrific	
advoc	advoc	
tokens selected after frequency-matched resampling; see Section 4.		

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.

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.

B Calibration and numerical stability

Reliability diagrams in Figure 3 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.

C 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).

Table 4: Representative OV tokens for GPT-2 Medium head 0:4 (top/bottom five).

Raw BPE	Normalised word	
Upweighted		
Pik	pik	
Benz	benz	
Bud	bud	
Dem	dem	
Hobby	hobby	Top-K
Downweighted		
streng	streng	
cryst	cryst	
notor	notor	
destro	destro	
nodd	nodd	

tokens selected after frequency-matched resampling; see Section 4.

- **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.

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.
- [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 Liu, 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.

Table 5: Representative OV tokens for GPT-2 Medium head 0:7 (top/bottom five).

Raw BPE	Normalised word	
Upweighted		
ruciating	ruciating	
guiActiveUnfocused	guiactiveunfocused	
sights	sights	
atherine	atherine	
pag	pag	Top-K
Downweighted		
theless	theless	
Redditior	redditor	
horizont	horizont	
condem	condem	
Ire	ire	

tokens selected after frequency-matched resampling; see Section 4.

- [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.
- [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.

Table 6: Representative OV tokens for Mistral 7B head 0:22 (top/bottom five).

Raw BPE	Normalised word	
Upweighted		
giornata	giornata	
listade	listade	
revs	revs	
acknow	acknow	
occas	occas	Top- K
Downweighted		
oppon	oppon	
LIED	lied	
itself	itself	
MVT	mvt	
recurs	recurs	

tokens selected after frequency-matched resampling; see Section 4.

- [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 five).

Raw BPE	Normalised word	
Upweighted		
acknow	acknow	
riebe	riebe	
départ	depart	
kat	kat	
rass	rass	Top- K
Downweighted		
ionato	ionato	
altogether	altogether	
Pf	pf	
strict	strict	
atan	atan	

tokens selected after frequency-matched resampling; see Section 4.