A Signaling-Theoretic Framework for Attention Head Specialization

Anonymous ACL submission

Abstract

We analyze the self-attention layer as an implicit economy wherein heads act as signals competing for limited bandwidth within the shared residual stream. Each head's signal represents the spread of its attention distribution, summarized by the effective dimension D_{eff} (the exponential of attention entropy). Signals incur both a sending cost (C_S) and a receiving cost (C_R). The softmax bottleneck forces a separating equilibrium between two signal types. Conventional Signals (sharp attention) have low cost to send and to receive, supporting copying and induction. Assessment Signals (broad attention) have high cost to send and low cost to receive, decodable only when the surrounding sharp substrate keeps residual-stream complexity bounded. Architecture fixes the form of the equilibrium (two attractors, separating); the redundancy structure of the training data fixes its location on the Sharp%-axis. The per-head D_{eff} distribution is unimodal but right-skewed, and the Sharp/Broad partition is operational, anchored at the 2-bit Shannon boundary ($D_{\text{eff}} = 4$). The framework picks out a substrate-density phase transition validated across model scales (124M→70B), training distributions, and intervention modes: when the Sharp population drops below $\sim 5\%$ by global temperature, perplexity collapses.

1 Introduction

We propose a structural account to explain why Transformer attention heads partition into functional roles. The framework, a signaling-theoretic account of observed head specialization, draws on signaling theory (Spence, 1973; Grafen, 1990; Zahavi, 1993; Donath, 1998, 2007) to identify the equilibrium-selection logic that picks the empirically observed Sharp/Broad allocation out of those the softmax architecture permits. We show that attention heads are subject to an implicit signaling dynamic that forces them to specialize. The architecture (softmax bottleneck, causal mask, residual stream) determines that the equilibrium has the form of a separating Sharp/Broad partition; the redundancy and structure of the training data determines where on the Sharp%-axis it sits (§5.3). The resulting Sharp substrate seems causally required for Broad signals to be decodable, in a way that admits a behaviorally testable phase transition. We model the dynamics of the self-attention layer as an implicit economy among heads writing into a "shared residual stream" (Elhage et al., 2021). Our dip test (§4) does not reject unimodality and we treat the Sharp/Broad partition as operational throughout. Sharp ablation yields a clean causal dissociation on induction, and the substrate-density phase transition (§5.2) appears when the Sharp population drops below a critical density. A smaller Broad-removal effect appears intervention-sensitive. The architecture forms a funnel (Figure 1) where early layers cast a wide net and late

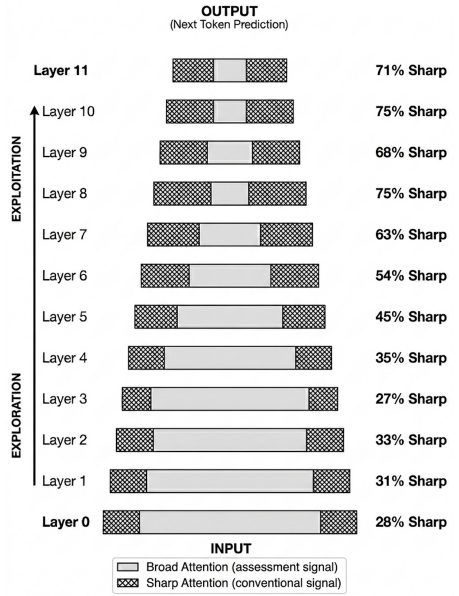

Figure 1: The Attention Funnel. Layer-by-layer Sharp head percentages for GPT-2 Small under natural inference.

layers filter for specific structural relationships, a consequence of the causal mask. In an autoregressive model the contextual signal available for integration grows with position. The paper proceeds next with a review of the relevant background literature. Thereafter, §3 develops the framework; §4 describes measurement and intervention protocols; §5 reports seven experiments; §6 sketches a derivation of the expected Sharp/Broad ratio from data statistics.

2 Background

Prior work identifies specific attention circuits, induction heads, copying heads, name movers (Olsson et al., 2022; Wang et al., 2023), and documents what they compute. Voita et al. (2019) and Michel et al. (2019) document heavy redundancy across the head population; Clark et al. (2019) catalog recurring attention-pattern types in BERT. We propose a complementary framework as a structural account of why heads partition into two functional populations. Zhai et al. (2023) characterize entropy collapse as a training instability; Dong et al. (2021) prove pure self-attention loses rank doubly exponentially with depth absent skip connections and MLPs, motivating the residual-stream decodability our C_R argument relies on. Recent rank-collapse (Nait Saada et al., 2024) and signal-propagation (Giorlandino and Goldt, 2025) analyses establish that avoiding both failure modes requires precise initialization and residual-scaling balance. We do not dispute the destabilizing nature of uniform entropy collapse during training; we claim an adaptive purpose to the structured Sharp/Broad partition observed in trained models. Symmetrically, driving Sharp% below $\sim 5\%$ via global temperature (§5.2) collapses perplexity. Jha and Reagen (2025)'s head-wise entropy regularizer reads as an externally imposed C_S on broad attention, our framework predicts a separating equilibrium of the resulting reorganization. Complementary accounts include cooperative-game analyses of head coalitions (Qu et al., 2025; Bouchaffra et al., 2026) and a non-cooperative potential-game treatment (Chakrabarti and Balachundar, 2026) that bounds the Price of Anarchy by a coupling term $\Gamma(G)$ tied to redundancy and hallucination. The Hydra effect (McGrath et al., 2023), where ablating layers seem compensated by another, identifies the inference-time mechanism by which such redundancy hides from zero-mask probes; we use it in §6.4 to interpret the Broad-arm null. Yang et al. (2018) show the softmax constrains output rank, the physical mechanism that bifurcates each head toward Sharp or Broad. Shannon (1951) estimated English at $\sim 75\%$ redundancy with driving dependencies across tens to hundreds of characters, the long-range structure that Broad heads capture.

3 The Signaling Framework

C_S and C_R provide the structured vocabulary for the asymmetry between producing and decoding an attention pattern (see Appendix A for notation). The signaling dynamic describes the attractor that gradient descent settles into during training; the trained weights realize this equilibrium as a fixed allocation we probe at inference. We read "quality" q in the Zahavi (1993) and Grafen (1990) sense: a latent property (whether the head's Q/K subspace supports integration of the current context) that gradient descent infers from the head's signal and its contribution to loss. The receiver is the optimization process during training. Each head represents a sender, emitting a signal characterized by its attention distribution. The *Conventional / Sharp* head attends to a single token: low C_S, low C_R, decodable by reading one position, corresponding to induction and copying (Olsson et al., 2022). The *Assessment / Broad* head attends to a wide distribution: high C_S (sustaining a high-entropy distribution against the softmax's exponential bias incurs a cost) and low C_R *conditional on substrate stability*, because the integrative content it carries remains decodable only when the surrounding sharp substrate keeps residual-stream complexity bounded. The low-C_R requirement is due to the cost of probing (Donath, 1998, 2007). A signal whose receiver cost rises becomes subject to probing cost, which the receiver in the limit ceases to act on; the substrate keeps Broad heads on the assessment side of that boundary. The framework's theoretical commitment of C_S/C_R manifests behaviorally (§5.2, §5.4), and through direct per-head cost measurement (§5.7).

3.1 Payoff Matrix and the Softmax Bottleneck

The softmax converts attention scores to a probability distribution that sums to 1. Because the exponential amplifies logit differences, intermediate-entropy distributions, attending moderately to many tokens, are unstable. Any perturbation either spikes one logit (Sharp) or distributes mass widely (Broad), with each head's attention pattern pushed

Signal	Substrate stable	Substrate-unstable
Assessment (Broad)	Win-Win	Collapse
Conventional (Sharp)	Precision	Degraded

Table 1: Payoff structure for the attention signaling dynamic.

toward one of these two attractor shapes. The "bifurcation" here refers to the per-head attention-pattern attractor. Payoff-space separation does not require D_{eff}-space separation, because D_{eff} is a noisy summary statistic over context-dependent attention rows, and a Broad attractor can be realized by many qualitatively different high-entropy shapes whose mean D_{eff} varies continuously with input. The framework therefore predicts that the discrete payoff structure expresses itself behaviorally (under targeted ablation and substrate-density perturbation). The expected utility of each strategy depends on the population state (Table 1). Under uncertainty about other heads, the conventional strategy remains risk-robust because it transmits regardless of the substrate. The assessment strategy carries higher payoff but only conditional on substrate stability. The condition Grafen (1990) proved sufficient for a separating equilibrium to be evolutionarily stable is a single-crossing payoff structure. The system admits two degenerate equilibria (all-conventional, shallow; all-assessment, C_R explodes) and one separating equilibrium ($\sim$70–97% Sharp in late layers depending on scale, §5.6). The observed entropy collapse represents the conventional substrate of the separating equilibrium.

3.2 Grafen's Formalization

An *evolutionarily stable strategy* (ESS) describes a population mix that, once reached, cannot be invaded by a small group of agents playing a different strategy. Any deviating strategy plays strictly worse against the prevailing population than conformity (Maynard Smith, 1974; Parker, 1984). The analogue here for Sharp/Broad allocation is one that gradient descent cannot improve on by reallocating heads. Grafen (1990) proved that an honest separating equilibrium between two sender types is evolutionarily stable under three conditions: 1) the marginal cost of producing a given signal must be higher for a low-quality sender; 2) the receiver's best response must be monotonically increasing in perceived sender quality; 3) a single-crossing on net payoff. In our setup these sender types are: a low-quality sender emits the low-cost (Conventional / Sharp) signal, and a high-quality sender emits the high-cost (Assessment / Broad) signal. We identify each Grafen condition in the attention dynamic with the sender = head, receiver = downstream layer, sender quality q = whether the head genuinely carries integrative information, and signal level = $D_{\text{eff}} = \exp(\mathcal{H})$ (a monotone function of attention entropy). Layer depth functions as a Spence (1973) *index*: an unalterable architectural attribute that partitions heads into sub-environments with different opportunity sets, since the contextual signal available for integration grows with depth in an autoregressive model. Each layer therefore settles into its own separating equilibrium and the structural origin of the funnel reported in §5.1.

3.2.1 Quality-dependent cost

The marginal cost of producing a given D_{eff} must be strictly higher for the low-quality sender. We measure this directly in §5.7: at $T = 1$, C_S is $4\times$ larger for Broad than Sharp heads, and within a head $C_S(T)$ rises in lockstep with the head's induced D_{eff}. The single-crossing of the two cost curves is enforced by the softmax bottleneck. Operationally, "quality" q is a latent property of the head's learned Q/K subspace, that is, the rank and alignment of its query/key projections with respect to the context. The framework treats q as the architectural variable that the training objective resolves into the observed Sharp/Broad allocation.

3.2.2 Receiver best-response monotonicity

The receiver's optimal action is monotonically increasing in perceived sender quality. We test this in §5.7: Sharp-head population mean C_R is $\sim 7\times$ that of the Broad population at $T = 1$ and rises monotonically in T over $[1, 2]$ as the substrate is perturbed.

3.2.3 Single-crossing on net payoff

High-quality heads' net payoff is maximized at high D_{eff} (Broad/Stable: Win-Win). Low-quality heads' net payoff is maximized at low D_{eff} (Sharp/Stable: Precision). The bottleneck dynamically excludes the intermediate region. Formally, with $U(\sigma, q)$ the utility of emitting signal type $\sigma \in \{\text{Sharp}, \text{Broad}\}$ at quality $q \in \{q_{\text{low}}, q_{\text{high}}\}$,

the discrete single-crossing condition is

$$U(\text{Broad}, q_{high}) - U(\text{Sharp}, q_{high}) > U(\text{Broad}, q_{low}) - U(\text{Sharp}, q_{low}), \quad (1)$$

i.e., high-quality heads earn greater marginal payoff upgrading from Sharp to Broad (see Appendix A.2). The two types separate at the operational threshold $D_{\text{eff}} = 4$ ($\mathcal{H} = 2$ bits), the conventional Shannon (1951) boundary between categorical and distributed codes (§4). The Sharp/Broad split shows robustness to small perturbations of the population mix. The observed ratio is a basin, drawing in gradient descent. We view the temperature phase transition in §5.2 as an externally imposed perturbation that bypasses the equilibrium and collapses the cost gap; when substrate density falls below $\sim$5%, no separating equilibrium remains.

3.3 Signaling Correspondence

We operationalize C_S as the mean $|\Delta D_{\text{eff}}|$ a head exhibits under a small isotropic Gaussian perturbation of its Q/K column subspaces, and C_R as the per-head NLL increase under mean-ablation of the head's per-position output, which removes positional structure while preserving mean activation. A Sharp head writes an approximately rank-1 contribution; a Broad head writes a high-entropy mixture whose decodability depends on the surrounding stream remaining low-rank. Per-head measurements on GPT-2 Small (§5.7) confirm $C_S(\text{Broad}) \gg C_S(\text{Sharp})$ and Sharp C_R rising as the substrate is perturbed. The asymmetry is structural and concerns attention-pattern shape. Superposition (Elhage et al., 2022; Bricken et al., 2023) is orthogonal because C_R tracks spectral rank. It is silent on value-vector polysemanticity. The signaling dynamic predicts the existence and form of the equilibrium from architectural primitives alone; its location on the Sharp%-axis is shaped by both architecture and the redundancy structure of the training data (§5.3, §5.6, §6.1). This underscores why the cross-domain spread (47–91%; Table 3) and the scale-driven drift to a uniformly-Sharp regime at 70B fall outside what an architecture-only theory can fix.

4 Methods

Primary analysis focuses on GPT-2 Small (Radford et al., 2019) on WikiText-2 (Merity et al., 2017), with cross-architecture validation on Pythia-160M (Biderman et al., 2023), GPT-2 Medium/Large, OPT-350M, and BERT-base, and billion-scale validation on the Llama-3 family (1B/3B/8B/70B) and Mistral-7B-v0.3. Our primary metric, D_{eff}, translates a head's attention distribution p into an effective token count. A head focused entirely on one position has $D_{\text{eff}} \approx 1$, a head spreading attention evenly across ten positions has $D_{\text{eff}} \approx 10$. Formally,

$$D_{\text{eff}} = \exp(\mathcal{H}), \qquad \mathcal{H} = -\sum_i p_i \log p_i, \quad (2)$$

with $\mathcal{H}$ the attention entropy in nats. We classify a head as *Sharp* when $D_{\text{eff}} < 4$ and *Broad* otherwise. The threshold corresponds to $\mathcal{H} = \log 4 \approx 1.39$ nats $= 2$ bits, the conventional information-theoretic boundary between categorical selection and distributed signal. We use it as an *operational* partition. Hartigan's dip test (Hartigan and Hartigan, 1985) on the per-head D_{eff} distribution does not reject unimodality on GPT-2 ($D = 0.017$, $p = 0.99$), Pythia-160M ($p = 0.99$), or BERT-base ($p = 0.99$), even when restricted to early, middle, or late layer thirds (all $p > 0.26$). The distribution plots right-skewed with a heavy low-D_{eff} mass. We report the bimodality coefficient (Pfister et al., 2013) for descriptive comparison with prior work but do not treat $BC > 5/9$ as evidence of bimodality, since BC is necessary but not sufficient and inflated by skew. The Sharp/Broad partition is a threshold-based working definition anchored at $\mathcal{H} = 2$ bits, validated by the behavioral consequences of crossing it (§5.2). To test the causal role of the substrate we inject a global temperature multiplier T into the attention softmax. $T = 1$ is the unmodified model; $T > 1$ artificially flattens attention (driving heads toward Broad); $T < 1$ artificially sharpens it (driving heads toward Sharp).[1] We sweep $T \in [0.5, 6.0]$ and measure perplexity and Sharp%. For head-level dissociation we zero-mask $K = 12$ heads in the corresponding `o_proj` slices, comparing three matched panels. Lowest-D_{eff} (Sharp), highest-D_{eff} (Broad), and a random-K control. We filter attention sinks ($D_{\text{eff}} > 1.5$ AND layer ≥ 3) following Xiao et al. (2024) to avoid the well-documented sink-ablation confound. The Sharp and Broad selections seem deterministic given the per-head D_{eff} grid. We vary the random seed across $n = 11$ runs, which controls the random-K12 panel choice and the prompt-sample drawn from each evaluation set. Statistical tests in

[1] Concretely, $\text{Attn} = \text{softmax}(QK^\top/(\sqrt{d_k}\,T))\,V$.

§5.4 report Sharp- and Broad-arm effects against the resulting random-null distribution. The IOI prompts are constructed from 15 templates following the structural conventions of Wang et al. (2023) (ABBA name placement, mid-clause continuation, IO/S disambiguation), with names filtered to single-token entries in the target tokenizer. The prompt set draws inspiration for templating from Wang et al. (2023). We control the random ablation panels internally by holding the prompt set fixed across Sharp / Broad / Random conditions.

5 Results

5.1 The Attention Funnel

GPT-2 at baseline shows 59% overall Sharp heads, with a strong layer gradient: 30.6% in early layers (0–3) rising to 72.5% in late layers (8–11) (Figure 1). The funnel shows that early layers do high-C_S assessment work, and late layers form the conventional substrate. The gradient: the layer-conditional equilibrium predicted by the index argument of §3.2: each depth's opportunity set fixed by the causal mask, and the layer-specific Sharp/Broad ratio the equilibrium gradient descent reaches given that constraint.

5.2 Temperature Phase Transition

Sweeping T from 0.5 to 3.0 on GPT-2 reveals a sharp cliff (Table 2). At $T = 2.0$, PPL is at $3.5\times$ baseline (95% bootstrap CIs non-overlapping with the $T = 1.0$ baseline at every $T \geq 1.5$); by $T = 2.5$, where late-layer Sharp% falls below 5%, PPL is at $12\times$ baseline. The small-T end ($T = 0.5$, PPL 126.7) is degraded because forced over-sharpening starves integration; forced over-broadening starves the substrate. We rule out a generic OOD-activation-drift confound by measuring residual-stream L2 norms across the sweep: norms drift smoothly through the regime in which PPL crosses $2\times$ baseline, with no norm cliff coincident with the $T = 2.0$ PPL cliff (Appendix B). To isolate substrate density from mean entropy, we ran a uniform-interpolation sweep on GPT-2: post-softmax, replace each row A with $(1-\alpha)A + \alpha U$, the canonical matched-entropy flattener.[2] At matched mean $D_{\text{eff}} = 6.5$, temperature gives PPL 414 while interpolation gives 59 ($9.9\times$ at $D_{\text{eff}} = 7$); when matched on Sharp%, the two are nearly indistinguishable in the moderate regime (ratios 0.98–1.20 for Sharp%$\geq$ 30%) but diverge at extreme substrate loss (ratios 7–48$\times$ for Sharp%$\leq$ 15%). Substrate density remains the controlling variable. We repeated the sweep on Llama-3.1-8B and Mistral-7B-v0.3 with 14 temperatures from 0.5 to 6.0 ($n = 32{,}704$ per-token NLLs from a 32,768-token sliding-window eval; 95% CI from $B = 1000$ bootstrap resamples over per-token NLLs, Figure 2). Both models reproduce the qualitative pattern and the threshold location. Late-layer Sharp% decays monotonically, mean D_{eff} crosses the Sharp/Broad threshold at $T \approx 1.7$, and a $2\times$-baseline-PPL cliff lands at $T = 2.0$ in both (Llama: $7.62 \rightarrow 18.66$, 95% CIs $[7.41, 7.84]$ vs. $[18.11, 19.26]$; Mistral: $6.94 \rightarrow 17.12$, $[6.75, 7.14]$ vs. $[16.60, 17.65]$). The threshold T is identical to GPT-2. The magnitude of the cliff is larger in the deeper models, consistent with more layers in which the substrate-loss signal can compound. The norm-control pattern also replicates in both models (Appendix B), so generic OOD activation drift does not explain the cliff in any of the three architectures tested.

T	late_Sharp%	PPL	95% CI	Status
0.50	94.5	126.7	[122.8, 131.0]	Too rigid
0.75	83.6	37.7	[36.5, 38.9]	Stable
1.00	70.6	34.0	[33.0, 35.1]	Baseline
1.50	38.0	39.6	[38.4, 41.0]	Drift
2.00	11.5	118.1	[114.3, 121.9]	**Cliff**
2.50	2.1	414.2	[400.2, 428.8]	**Collapse**
3.00	0.0	1060.3	[1019.5, 1101.9]	Saturated

Table 2: Temperature-induced phase transition on GPT-2 Small (WikiText-2, n=32,704 per-token NLLs from a 32,768-token sliding-window eval; 95% CI from B=1000 bootstrap resamples over per-token NLLs). Late-layer Sharp% falls below 5% at T=2.5. The 2×-baseline cliff lands at T=2.0. The T=1.0 late-Sharp% (70.6) differs slightly from the 72.5 reported in §5.1 because attention here is sampled inside the per-token PPL evaluation loop on WikiText-2; comparisons across rows of this table are internally controlled.

5.3 Structural Alignment: Code vs. Prose

If the substrate density reflects the rigidity of the training signal, code-trained models should show

[2]The matched-D_{eff} comparison reads D_{eff} from the seed-0 prompt set used for both sweeps (`data/temperature_collapse/temperature_collapse_gpt2_seed0_*.json` paired with `data/uniform_interp/uniform_interp_seed0_*.json`), so the two interventions sit on the same D_{eff} axis. Headline late-layer Sharp% values in Table 2 use the curated 8-prompt set in `data/phase9_combined_gpt2_*.json`. PPL is computed on the same WikiText-2 sliding-window eval (32,768-token window, $n = 32{,}704$ valid per-token NLLs) in all three files (PPL= 414.2 at $T = 2.5$ in both temperature files).

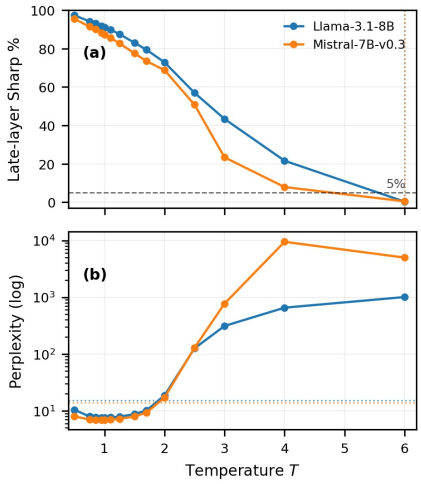

Figure 2: Phase transition replicated at billion scale on Llama-3.1-8B (blue) and Mistral-7B-v0.3 (orange), 14 temperatures from 0.5 to 6.0, $n = 32{,}704$ per-token NLLs per temperature (32,768-token sliding-window eval). (a) Late-layer Sharp% collapses monotonically (5% reference dashed). (b) Perplexity (log scale) shows a sharp cliff at $T = 2.0$ in both models, where PPL crosses or exceeds $2\times$ baseline (per-model dashed line). Shaded bands are 95% bootstrap CIs ($B = 1000$).

denser substrates than prose-trained ones. As a cross-model comparison, CodeGen-350M-mono (Nijkamp et al., 2023) shows 88.4% late-layer Sharp% versus 72.5% for GPT-2 (prose). The highly redundant but rigid nature of code (Hindle et al., 2012) means that more of the signal must be transmitted conventionally. Holding the network fixed and varying input domain within a single deployed Llama-3.1-8B (32 layers; $n = 164$ for HumanEval, $n = 500$ for the others; Table 3), the same input-rigidity pattern recovers. The two highest-rigidity domains, PubMedQA biomedical QA (Jin et al., 2019) (90.9% late-Sharp) and HumanEval code (Chen et al., 2021) (80.9%), produce the densest late-layer substrates. Four prose-class domains cluster well below them, including narrative TinyStories (Eldan and Li, 2023) (46.9%), formal WikiText-103 (Merity et al., 2017) (48.4%), legal CaseHOLD (Zheng et al., 2021) (57.5%), and Spanish Wikipedia (58.4%). A positive funnel gradient (early D_{eff} − late D_{eff}) in all six domains generalizes the funnel claim beyond English and code. The same deployed network reorganizes its substrate per-input in the direction the theory predicts.

5.4 Head-Level Sharp Dissociation

To test functional specialization, we ablate $K = 12$ matched panels of sink-filtered Sharp / Broad / Random heads on Llama-3.1-8B and evaluate on IOI (Wang et al., 2023) (induction-dependent) and LAMBADA (Paperno et al., 2016) (integration-dependent). Each task uses $n = 1024$ prompts, with each panel matched on K and sink-filtered per §4. The Sharp and Broad panels are deterministic given the head-classification grid. We vary the seed only on the Random-K12 control and on prompt sampling, repeating across $n = 11$ seeds. The Sharp arm is a clean specialization result. Across 11 seeds, Sharp ablation lies 3.6σ below the random null on IOI ($z = -3.65$, $p = 0.0001$, one-sided) and remains at the random-control floor on LAMBADA (Δ NLL $+0.020$ vs. random $+0.011 \pm 0.056$), overlapping the "retrieval heads" of Ma and Okazaki (2026). The Broad arm does not separate from the random null on LAMBADA ($z = +0.66$, $p = 0.255$, one-sided; $n = 11$ seeds), consistent with the unimodal D_{eff} distribution (§4) under which removing $K = 12$ right-tail heads behaves like removing $K = 12$ random heads. To test whether this null reflects functional absence or zero-masking bluntness for aggregation heads, we ran a complementary mean-ablation on GPT-2 (replace each selected head's per-position output with its per-sequence mean, $K = 10$, $n = 100$ random controls, WikiText-2). Baseline PPL 34.02; top-K Broad ablation yields PPL 53.96 ($+19.94$, $z = +2.20$ vs. random 38.05 ± 7.22, normal-approximation $p \approx 0.014$; empirical one-sided $p = 0.06$, with $6/100$ random draws reaching the Broad PPL or higher); top-K Sharp yields PPL 117.43 ($+83.41$, $z = +10.99$; no random draw reaches the Sharp PPL, empirical $p < 0.01$). The Sharp dissociation is discovery-grade by the $\alpha = 0.005$ convention of Benjamin et al. (2018); the Broad effect sits in the upper tail

Domain	n	Late%	Funnel
PubMedQA (medical)	500	**90.9**	+2.33
HumanEval (code)	164	**80.9**	+5.73
Spanish Wikipedia	500	58.4	+5.77
CaseHOLD (legal)	500	57.5	+6.87
WikiText-103	500	48.4	+3.22
TinyStories (narrative)	500	46.9	**+9.92**

Table 3: Six-domain fragility sweep on Llama-3.1-8B. *Late%* = Sharp head density in layers 21–31. *Funnel* = early D_{eff} − late D_{eff}.

Condition	Δ IOI logit-diff	Δ LAMBADA NLL
Sharp-K12	-0.830	$+0.020$
Broad-K12	-0.028	$+0.048$
Random-K12	-0.048 ± 0.214	$+0.011 \pm 0.056$

Table 4: Targeted head-level ablation on Llama-3.1-8B.

of the random null but is not separated from it once heavy-tailed null mass is properly characterized, consistent with the unpriced-redundancy account of §6.4.

5.5 Training Dynamics

We analyzed Pythia-160M across 20 training checkpoints from step 0 to step 143,000. The Sharp-density gradient (late-layer Sharp% − early-layer Sharp%; positive when late layers are sharper than early layers, the Sharp%-based counterpart of the D_{eff}-based funnel gradient of §5.3, with the same sign convention) starts at 0 at step 0 (uniform attention from random init) and rises monotonically to +77.2 by the final checkpoint (late layers 92.4% Sharp, early layers 15.2%). The Sharp/Broad division is a learned property emerging from training data.

5.6 Substrate Spreads with Scale

Running the D_{eff} pipeline across the Llama-3 family (Figure 3) reveals that the equilibrium deepens at scale. Mean D_{eff} decreases monotonically from 3.05 (1B) to 1.70 (70B) and stays below threshold at every size; late-layer Sharp% rises from 87.4% to 96.7%, early-layer faster (78.2% → 97.6%); the funnel gradient collapses from +9.2 at 1B to −1.0 at 70B. At 70B the network is uniformly Sharp; the funnel of §5.1 represents the small-model regime. Mistral-7B-v0.3 sits where the trend predicts. Mechanistically, as parameter count grows, per-head capacity to compress local redundancy rises and the effective $R_{\text{local}}/R_{\text{total}}$ ratio (§6.1) shifts upward; assessment work that demanded Broad heads at 1B is absorbed by larger Sharp heads with richer Q/K subspaces, with GQA K/V sharing amplifying per-Sharp-head reach. In signaling terms, 70B has not abandoned the separating equilibrium but slid along it toward the conventional pole. This predicts the Broad pole would re-emerge at 70B under sufficiently long-range inputs. We leave that test to follow-up work.

5.7 Quantitative Cost Measurement

We trace each head's $(C_S, C_R, D_{\text{eff}})$ trajectory across 26 global temperatures $T \in [0.75, 2.0]$ in steps of 0.05 on GPT-2 Small (WikiText-2, 256 prompts, last-token D_{eff}). C_S: the mean $|\Delta D_{\text{eff}}|$ under an isotropic Gaussian Q/K perturbation ($\sigma = 0.01$, 5 noise seeds, calibrated for $\bar{C}_S(T{=}1) \approx 0.4$. C_R: the per-head mean-ablation NLL increase. Implementation details reported in Appendix C; full per-head trajectories documented in supplementary JSON. Three hypotheses H1, H2, H3, are stated below before their corresponding measurements.

H1 Heads driven Broad pay more C_S (Figure 4). Per-head Spearman $\rho(D_{\text{eff}}(T), C_S(T))$ has mean 0.50, median 0.92, 78% positive (Wilcoxon $p = 1.4 \times 10^{-12}$). Within a head, sender cost rises in lockstep with the induced D_{eff}.

H2 Sharp heads pay rising C_R as the substrate is perturbed (Figure 5). Sharp-population mean C_R rises from 0.018 at $T = 1.0$ to 0.044 nats at $T = 2.0$ (2.5×; Spearman 0.93, $p = 8 \times 10^{-12}$, $n = 26$ temperatures spanning $T \in [0.75, 2.0]$ in steps of 0.05; on the substrate-perturbation arm $T \geq 1.0$, the trend is perfectly monotone, $\rho = 1.00$, $n = 21$). The substrate-density effect of §5.2 resolves at the per-head level.

H3 Sharp and Broad separate in cost space at $T = 1$ (Figure 6). Sharp heads are 4× less Q/K-responsive than Broad (C_S mean 0.19 vs 0.75, Mann–Whitney $p = 1.5 \times 10^{-12}$). The Sharp population is $\sim 7\times$ more costly to mean-ablate in the mean (C_R mean 0.018 vs 0.003 nats); the rank test on C_R is not significant ($p = 0.17$), reflecting heavy-tailed Sharp heterogeneity, where a small set of critical substrate heads dominates the mean while the modal Sharp head is comparable to the modal Broad head, consistent with the unimodal D_{eff} distribution of §4 and the duplication account of §6.4. The pattern $C_S(\text{Broad}) \gg C_S(\text{Sharp})$ with Sharp C_R rising under substrate perturbation follows the signature of the cost asymmetry posited in §3 and Grafen's quality-dependent-cost condition.

6 Discussion

6.1 Sharp/Broad Ratio

The softmax bottleneck makes intermediate entropy distributions unstable. Each head bifurcates

toward Sharp or Broad. Cross-entropy imposes precision pressure (rigid tokens, served by Sharp $D_{\text{eff}} \approx 1$–3) and integration pressure (flexible tokens, served by Broad $D_{\text{eff}} \geq 4$). Shannon (1951) estimated English at $\sim$75% redundancy, decomposable into local ($\sim$50–60%, served by copying/induction) and non-local ($\sim$15–25%, served by integration). The predicted Sharp density gives an Expected Sharp %:

$$\frac{R_{\text{local}}}{R_{\text{total}}} \approx \frac{0.50\text{–}0.60}{0.75} \approx 67\text{–}80\%, \tag{3}$$

consistent with English-prose mid-scale models (GPT-2 72.5% inside the band). The wider cross-domain range (47–91%; Table 3) is accommodated qualitatively. High-rigidity domains (PubMedQA 90.9%, HumanEval 80.9%, CodeGen 88.4%) push above the band; low-rigidity prose (TinyStories 46.9%, WikiText-103 48.4%, CaseHOLD 57.5%, es-Wikipedia 58.4%) sits below. The local/non-local split is calibrated to English prose and the simple ratio does not predict magnitudes outside that central case.

6.2 Encoder vs. Decoder Generality

The framework predicts the same right-skewed Sharp-dominated D_{eff} profile wherever the softmax bottleneck applies. BERT-base (encoder) and GPT-2 (decoder) both show this profile (BC 0.72 and 0.79 respectively), consistent with Clark et al. (2019). The funnel *gradient*, by contrast, characterizes autoregressive dynamics: in a matched comparison ($n = 100$ inputs), GPT-2 shows late-minus-early Sharp% of $+12.3$ while BERT is essentially flat at -0.2. In a causal model the contextual signal available for integration grows with position, biasing assessment toward early layers; in a bidirectional encoder every layer sees full context, so the substrate has no preferred depth.

6.3 Substrate Density as a Fragility Diagnostic

The signal-propagation theory of Giorlandino and Goldt (2025) specifies the mechanism (softmax bottleneck plus gradient flow). The costly signaling framework specifies the equilibrium-selection argument that picks the observed Sharp/Broad allocation out of those that signal-propagation arithmetic alone permits. The framework predicts that a substrate forced thin by input distribution, generation length, or temperature, results in downstream layers that cannot reliably decode Assessment Signals. The substrate-density of late layers seems to be a candidate inference-time fragility diagnostic, whose application to specific failure modes such as hallucination we leave to follow-up work.

6.4 Unpriced Redundancy

The potential-game analysis of Chakrabarti and Balachundar (2026) bounds the Price of Anarchy (PoA) among attention heads by an interaction-coupling term $\Gamma(G)$ and proves that excess head redundancy scales with PoA as an unpriced externality. By construction Sharp substrate is low-rank (each head reads one position) and contributes little to $\Gamma(G)$, so removing $K = 12$ Sharp heads removes irreplaceable signal and drives the 3.6σ Sharp dissociation we observe. The Broad population sits in the high-$\Gamma(G)$ regime where multiple heads cover overlapping aggregation patterns at equilibrium, and Hydra-style self-repair (McGrath et al., 2023) reroutes the duplicated signal at inference time. The framework therefore predicts three signatures of the Broad arm under undirected mass ablation: 1) the effect is intervention-sensitive (zero-masking versus mean-ablation produce different magnitudes); 2) the random-control null is heavy-tailed because typical random K-sets overlap the same redundant subspace, so even strongly Broad-biased selections fail to separate from the empirical null ($n = 100$ controls: Broad PPL at the 94th percentile, empirical $p = 0.06$, §5.4); and 3) substrate density is the controlling variable, with Sharp removal driving the discovery-grade dissociation (§5.4) and Broad removal absorbed by duplication. We discuss coupling-aware ablation, total-correlation, activation-patching, and higher-order-task refinements in the Limitations.

7 Conclusion

Entropy collapse in deep Transformer layers is the conventional substrate of a separating equilibrium forced by the softmax bottleneck. Broad signals are decodable over a stable Sharp substrate only, while Sharp signals transmit regardless. Driving substrate density below $\sim$5% collapses perplexity (§5.2). Architecture fixes the equilibrium's form. The redundancy of training data fixes its location (§5.3, §5.5). The Broad-removal effect is intervention-sensitive, with Sharp removal driving the discovery-grade dissociation while Broad removal is absorbed by Hydra-style duplication and is not separated from the heavy-tailed random-control null at $\alpha = 0.005$ (§5.4, §6.4).

Limitations

The Sharp/Broad dichotomy is coarse. It is a claim about attention-pattern shape (§3.3). The Sharp/Broad classification thresholds attention-pattern entropy at $D_{\text{eff}} = 4$ and is silent on value-vector polysemanticity. Disentangling attention-pattern sharpness from value-content polysemanticity via crosscoder feature decomposition is the natural next experiment, but we do not attempt it here. The receiver-cost operationalization (mean-ablation NLL increase) separates the Sharp and Broad populations strongly in the mean ($\sim 7\times$) but not under a non-parametric rank test ($p = 0.17$), reflecting Sharp-population heterogeneity; a coupling-aware C_R that controls for the Hydra-style duplication of §6.4 is the natural refinement.

This paper analyzes the equilibrium of trained models primarily. The post-hoc calibration in §6, grounded in data statistics, predicts a central-case Sharp/Broad ratio of 67–80% for English prose at mid-scale; GPT-2's 72.5% sits inside this band, but other configurations land outside it (Pythia-160M 92.4%, the cross-domain range 47–91% in Table 3, and the 70B uniformly-Sharp regime of §5.6). The data-statistics argument of §6 and the scaling result of §5.6 accommodate these qualitatively but do not derive them quantitatively. A quantitative test would train small transformers from scratch on synthetic corpora with parametrically dialed redundancy and remains a follow-up.

We report $k \approx 10$ statistical tests across §5.4 and §5.7; hypotheses H1–H3 of §5.7 are stated in the manuscript before their corresponding measurements. Under the $\alpha = 0.005$ convention of Benjamin et al. (2018) and a conservative Bonferroni correction, the discovery threshold is $\alpha/k \approx 5 \times 10^{-4}$. The Sharp IOI dissociation ($p = 10^{-4}$), the Sharp mean-ablation effect ($p < 10^{-29}$), H1 ($p = 1.4 \times 10^{-12}$), H2 (Spearman $p = 8 \times 10^{-12}$, $n = 26$ temperatures), and the H3 C_S separation ($p = 1.5 \times 10^{-12}$) clear the corrected threshold by between four and twenty-five orders of magnitude. The Sharp empirical-null tail ($0/100$ random draws, empirical $p \leq 1/101 \approx 0.0099$) corroborates the Sharp mean-ablation effect non-parametrically but is floored above α/k by the resolution of $n = 100$ random controls. The Broad mean-ablation effect (normal-approximation $p \approx 0.014$, empirical $p = 0.06$) does not clear α/k and is reported throughout as suggestive rather than confirmatory; the Broad-arm null is, by §6.4, the prediction of the framework rather than a failed positive. The framework supplies a detection mechanism (substrate density predicts assessment channel fragility). We leave applications of the substrate-density diagnostic to specific downstream failure modes, and cross-architecture validation thereof, to future work. Our intervention experiments use head-level zero-masking, which the Hydra-effect literature (McGrath et al., 2023) shows emergent self-repair can compensate for. Activation patching experiments that swap Sharp↔Broad representations would neutralize this confound and strengthen the causal claims in §5.4. The Broad-arm intervention sensitivity of §5.4 admits four concrete refinements that follow from the literature surveyed in §6.4.

1. Coupling-aware ablation: pre-compute a head-pair interaction matrix on the Broad subpopulation and ablate maximally decoupled K-sets (low off-diagonal mass within the panel), predicting that the Broad effect then separates from the random null.

2. Information-theoretic verification: measure the total correlation TC across Broad-head outputs and test whether it tracks $\Gamma(G)$ from Chakrabarti and Balachundar (2026), linking the unpriced-redundancy account to a measurable statistic.

3. Activation patching that swaps Broad representations across prompts (per the Hydra control above), to determine the residual causal effect once self-repair is neutralized.

4. Higher-order task workloads of the kind motivated by Bouchaffra et al. (2026), where prediction depends on n-token dependencies, predicted to be the regime in which the Broad-arm effect is largest. A complementary substrate-vs-assessment temperature sweep, restricted to the bottom-D_{eff}-quartile and top-D_{eff}-quartile of heads, would test whether the §5.2 cliff is substrate-specific; the Qu et al. (2025) GAC result (targeted non-substrate flattening yields accuracy gains) is consistent with this prediction.

We define the partition of Sharp and Broad operationally by thresholding D_{eff} at the $\mathcal{H} = 2$ bits Shannon boundary.

Hartigan's dip test does not reject unimodality of the per-head D_{eff} distribution in any model we examined (§4), so claims about "two populations"

read as a working partition that behavioral dissociation validates (§5.2, §5.4). We leave to future work a more refined characterization of the underlying distribution, which is right-skewed with a heavy low-D_{eff} mass, and the question of whether a bimodal regime emerges only under specific architectural or scale conditions.

The framework as stated assumes standard dense softmax self-attention, a single shared residual stream into which every head writes, and a softmax bottleneck that bifurcates each head's attention pattern. The framework does not directly cover architectures that relax either assumption. Mixture-of-Experts attention partitions writers into routed sub-populations whose signaling dynamics demand re-derivation per route. Cross-attention introduces an asymmetry between encoder and decoder residual streams that the present single-stream payoff structure does not capture. Linear and other sub-quadratic attention variants replace the softmax bottleneck whose exponential bias drives the bifurcation in §3, and would require a separate analysis of whether an analogous separating equilibrium exists. Grouped-query attention, where multiple query heads share K/V projections, is a partial relaxation we treat as already covered. Llama-3.1 (1B/3B/8B/70B) uses GQA throughout, and the substrate-density results of §5.2, §5.3, §5.4, and §5.6 therefore constitute partial empirical evidence that the framework survives this relaxation. We leave extending the signaling account to MoE, cross-attention, and sub-quadratic attention to future work.

Ethical Considerations

We used Large Language Models (Claude, Anthropic; Gemini, Google) to assist with code generation, manuscript polishing, and ideation. We verified all outputs and assume full responsibility for the accuracy and empirical validation of all claims.

This work is foundational and analytical. We release no new model or dataset, and propose no deployed system. The principal dual-use consideration concerns the substrate-density diagnostic discussed in §6: an adversary could in principle apply it to construct inputs that drive the substrate below the $\sim 5\%$ collapse threshold of §5.2; we note that such inputs already produce visibly degenerate outputs, and the same diagnostic provides a cheap detection signal for the failure mode, so the net direction is defensive.

The datasets used contain no personal or sensitive information beyond what is present in the public benchmarks themselves (e.g., Wikipedia article text), and our IOI evaluation uses only synthetic name templates. We perform inference-time analysis only and release no generated text.

We submit all analysis code and the experiment output JSON files supporting the figures and tables in this paper with this draft, and plan to release the data and code underlying the experimentation publicly as open source software.

References

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, and 53 others. 2018. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*.

Djamel Bouchaffra, Faycal Ykhlef, Hanene Azzag, Mustapha Lebbah, and Bilal Faye. 2026. Neurogame transformer: Gibbs-inspired attention driven by game theory and statistical physics. *Preprint*, arXiv:2603.18761.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread.

Kushal Chakrabarti and Nirmal Balachundar. 2026. Multi-head attention is a multi-player game. *Preprint*, arXiv:2602.00861.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Judith Donath. 1998. Identity and deception in the virtual community. In *Communities in Cyberspace*, pages 27–57. Routledge, London.

Judith Donath. 2007. Signals in social supernets. *Journal of Computer-Mediated Communication*, 13(1):231–251.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. 2021. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. *Preprint*, arXiv:2103.03404.

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How small can language models be and still speak coherent English? *Preprint*, arXiv:2305.07759.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Tom Conerly, Saurav Showk, Robert Lasenby, Karina Larson, Dawn Drain, Yuntao Bai, Anna Chen, Roger Cardenas, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, and Chris Olah. 2022. Toy models of superposition. Transformer Circuits Thread.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Alessio Giorlandino and Sebastian Goldt. 2025. Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation. *Preprint*, arXiv:2505.24333.

Alan Grafen. 1990. Biological signals as handicaps. *Journal of Theoretical Biology*, 144(4):517–546.

J. A. Hartigan and P. M. Hartigan. 1985. The dip test of unimodality. *The Annals of Statistics*, 13(1):70–84.

Abram Hindle, Earl T. Barr, Zhendong Su, Mark Gabel, and Premkumar Devanbu. 2012. On the naturalness of software. In *Proceedings of the 34th International Conference on Software Engineering (ICSE)*.

Nandan Kumar Jha and Brandon Reagen. 2025. Regularizing the entropy landscape of self-attention: Towards a soft inductive bias in LLMs. In *NeurIPS 2025 Workshop on Optimization for Machine Learning (OPT)*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Youmi Ma and Naoaki Okazaki. 2026. From interpretability to performance: Optimizing retrieval heads for long-context language models. *Preprint*, arXiv:2601.11020.

J. Maynard Smith. 1974. The theory of games and the evolution of animal conflicts. *Journal of Theoretical Biology*, 47(1):209–221.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. 2023. The hydra effect: Emergent self-repair in language model computations. *Preprint*, arXiv:2307.15771.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32.

Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. 2024. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. *Preprint*, arXiv:2410.07799.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2023. CodeGen: An open large language model for code with multi-turn program synthesis. In *International Conference on Learning Representations*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. Transformer Circuits Thread.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Geoffrey A. Parker. 1984. Evolutionarily stable strategies. In J. R. Krebs and N. B. Davies, editors, *Behavioural Ecology: An Evolutionary Approach*, 2nd edition, chapter 2, pages 30–61. Blackwell Scientific Publications, Oxford.

Roland Pfister, Katharina A. Schwarz, Markus Janczyk, Rick Dale, and Jonathan B. Freeman. 2013. Good things peak in pairs: A note on the bimodality coefficient. *Frontiers in Psychology*, 4:700.

Xiaoye Qu, Zengqi Yu, Dongrui Liu, Wei Wei, Daizong Liu, Jianfeng Dong, and Yu Cheng. 2025. Cooperative or competitive? understanding the interaction between attention heads from a game theory perspective. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14079–14099, Vienna, Austria. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Claude E. Shannon. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1):50–64.

Michael Spence. 1973. Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. 2018. Breaking the softmax bottleneck: A high-rank RNN language model. In *International Conference on Learning Representations*.

Amotz Zahavi. 1993. The fallacy of conventional signalling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 340(1292):227–230.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. 2023. Stabilizing transformer training by preventing attention entropy collapse. In *Proceedings of the 40th International Conference on Machine Learning*.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAIL)*, pages 159–168.

A Notation and Derivations

This appendix provides definitions utilized in the signaling framework (§3).

A.1 Glossary of Variables

Term	Definition
$\mathcal{H}$	Shannon Entropy of a head's attention distribution: $\mathcal{H} = -\sum p_i \log p_i$.
D_{eff}	Effective Dimension is a continuous measure of attention spread: $D_{\text{eff}} = \exp(\mathcal{H})$. Serves as the signal level in the signaling economy (§3).
q	Latent sender "quality": whether a head's Q/K subspace supports context integration (§3.2).
T	Temperature scalar applied to pre-softmax logits to artificially manipulate attention focus.
C_S	Sender Cost (Production Cost) is the optimization penalty incurred by emitting a signal. Operationalized as the mean $\lvert\Delta D_{\text{eff}}\rvert$ under a small isotropic Q/K perturbation (§3.3, §5.7).
C_R	Receiver Cost (Decoding Cost) is the cost paid by downstream layers to decode a head's contribution to the residual stream. Operationalized as the per-head NLL increase under mean-ablation of the head's per-position output (§3.3, §5.7).
R_{total}	Expected total redundancy (predictability) of the language distribution.
R_{local}	Expected local redundancy (predictability derived purely from grammar or exact copying).

Table 5: Glossary of signaling and attention variables.

A.2 Separating Equilibrium

The single-crossing inequality (1) that anchors §3.2 is the Transformer-level application of the strategic conditions for an honest signaling equilibrium formalized by Grafen (1990).

The softmax bottleneck enforces $C_S(\text{Broad}) > C_S(\text{Sharp})$ at the production side, so the inequality reduces to a substrate-decoding argument: the residual stream rewards resolving complex predictions (lowering C_R globally), and only heads with genuine q_{high} context can justify the high C_S of sending a Broad signal.

Gradient descent forces the remaining heads into the low-cost Sharp role, which is the separating equilibrium that §5 probes empirically.

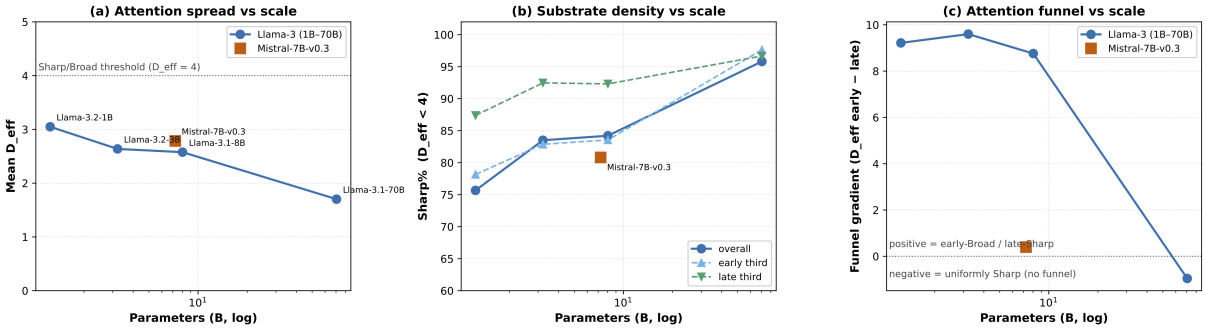

Figure 3: Sharp/Broad equilibrium across the Llama-3 scale ladder; Mistral-7B-v0.3 as comparison. (a) Mean D_{eff} decreases monotonically with scale and stays below the $D_{\text{eff}} = 4$ threshold at every size. (b) Sharp% (overall, early third, late third) rises with scale; at 70B the model is uniformly Sharp. (c) Funnel gradient shrinks from $+9.2$ at 1B to -1.0 at 70B, the substrate spreads through the network. Supports §5.6.

B Norm-Control Measurements for the Temperature Sweep

This appendix reports the residual-stream L2 norm measurements that support the OOD-drift control claim in §5.2.

B.1 GPT-2 Small (WikiText-2)

Early-layer norms are flat across the entire $T \in [0.5, 3.0]$ sweep (layer-0 mean norm 50.0–53.6); the last-layer norm rises smoothly from 212 at $T = 0.5$ to 260 at $T = 2.0$, then drops to 178 at $T = 3.0$. No norm cliff coincides with the $T = 2.0$ PPL cliff. Norms drift smoothly through the regime in which PPL crosses $2\times$ baseline, and the only abrupt norm change (the late-layer drop at $T \geq 2.5$) occurs after the PPL cliff has already fired.

B.2 Llama-3.1-8B and Mistral-7B-v0.3

Across the cliff ($T \in [1.0, 2.0]$, where PPL crosses $2\times$ baseline), mean residual-stream norms at the first, middle, and last layers drift by less than 10% in both models (Llama last-layer norm 142.8 $\rightarrow$ 137.5, -3.7%; Mistral last-layer norm 364.4 $\rightarrow$ 348.5, -4.4%, in fact decreasing slightly), while PPL crosses an order of magnitude. Past the cliff into the collapse regime ($T \geq 2.5$) both last-layer norms drop further (Llama $\rightarrow$ 133.0 at $T = 2.5$, Mistral $\rightarrow$ 304.1 at $T = 2.5$), but the PPL cliff has already fired by then.

In all three architectures the perplexity cliff at $T = 2.0$ is therefore not attributable to generic OOD activation drift.

C Quantitative Cost Measurement

This appendix gives implementation detail for §5.7 and reports the three figures corresponding to hypotheses H1, H2, H3. All measurements are on GPT-2 Small (FP32, MPS) with `attn_implementation="eager"` and WikiText-2 validation prompts (256 prompts, max sequence length 256 tokens, prompts shorter than 32 tokens dropped).

Temperature. We apply global temperature T via a context manager that monkey-patches `torch.nn.functional.softmax` so that the call divides any 4-D input (the attention-score tensor of shape $[B, H, T_q, T_k]$) by T before the original softmax. This matches the temperature convention used in the §5.2 sweep and avoids dependence on the internal layout of `GPT2Attention`.

D_{eff} extraction. Per head and per sequence we record the attention row at the last non-padded query position, computing $D_{\text{eff}} = \exp(\mathcal{H})$ on that row and averaging over prompts. This matches the convention used throughout the rest of the paper.

Sender cost C_S. For each (layer ℓ, seed s) we draw an isotropic Gaussian perturbation $\delta \sim \mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.01$ and add it to the Q and K column subspaces of the layer's `c_attn.weight` block, restricted to the head's $d_h = 64$ columns. We perturb all 12 heads of a layer simultaneously per seed (a $12\times$ speedup over the per-head loop) and re-measure D_{eff} for each head. C_S is the mean of $|\Delta D_{\text{eff}}|$ across the 5 noise seeds; we calibrated the noise scale so the population-mean C_S at $T = 1$ is ≈ 0.4, large enough to drive a measurable D_{eff} shift without leaving the linear regime.

Receiver cost C_R. A `forward_pre_hook` on each layer's `attn.c_proj` replaces the slice of the head's contribution with its per-sequence mean broadcast over positions; C_R is the resulting NLL increase. Mean-ablation removes the head's posi-

tional structure while preserving its mean activation level, isolating the receiver-side decoding cost from a generic activation drop.

Sharp/Broad classification. We classify heads at $T = 1.0$ using the project-wide threshold $D_{\text{eff}} < 4$ (§4). On the GPT-2 last-token measurement, 26.4% of heads are Sharp at $T = 1$.

D Reproducibility

D.1 Software

We used two execution environments. Small model analyses on the local Apple Silicon MPS backend used PyTorch 2.8.0, HuggingFace `transformers` 4.57.3, `numpy` 2.3.3, and `scipy` 1.16.2.

Cluster A100 runs used PyTorch 2.11.0 (CUDA 13.0), `transformers` 5.6.2, `accelerate` 1.13.0, `datasets` 4.8.4, and `numpy` 2.4.4. We load models with `from_pretrained` at the released checkpoint revision with default configuration. Attention output as: `output_attentions=True`. We modify no generation or inference-time sampling parameters.

All runs use `attn_implementation="eager"` (required: SDPA and FlashAttention backends return `None` for the attention tensor and therefore preclude the per-head D_{eff} measurement on which this paper depends). Bimodality coefficient statistics (computed locally) use `scipy.stats` and follow the formulation of Pfister et al. (2013).

D.2 Compute

The experiments are primarily inference tests. Small model analyses (GPT-2, Pythia-160M, OPT-350M, BERT-base, CodeGen-350M-mono) ran on an Apple Silicon MPS backend in FP32.

Billion scale analyses (Llama-3.1 1B/3B/8B, Mistral-7B-v0.3) and the 70B run ran on an HPC cluster of NVIDIA A100 80GB GPUs (CUDA 13.0). The total allocated compute across all reported and exploratory cluster runs is approximately 300 A100-hours; the 70B sweep (4×A100, ∼36 wall-clock hours) dominates this budget.

D.3 Head-level ablation parameters

The head-level dissociation in §5.4 (Table 4) zero-masks $K = 12$ heads in the corresponding `o_proj` slices on Llama-3.1-8B; sink-filter thresholds are $D_{\text{eff}} > 1.5$ AND layer ≥ 3; per-condition prompt count is $n = 1024$; we aggregate results across seeds 0–10 ($n = 11$ seeds total), with the Sharp/Broad selections deterministic given the head-classification grid and the seed varying the random-K12 panel and the prompt sample.

D.4 Datasets and licenses

Datasets from experiments represent released research benchmarks accessed from the HuggingFace Hub. Licenses include WikiText-2/103 (CC BY-SA 3.0), HumanEval (MIT), PubMedQA (MIT), TinyStories (CDLA-Sharing-1.0), CaseHOLD (Apache 2.0), Spanish Wikipedia (CC BY-SA 3.0), LAMBADA (modified MIT). We constructed the IOI evaluation following the template of Wang et al. (2023).

D.5 Models and licenses

We accessed pretrained checkpoints from the HuggingFace Hub: GPT-2 family (MIT), Pythia (Apache 2.0), Llama-3.1 (Llama 3 Community License), Mistral-7B-v0.3 (Apache 2.0), OPT-350M (OPT-175B research license), BERT-base (Apache 2.0), CodeGen-350M-mono (BSD-3-Clause). All artifacts are used for non-commercial, inference-time research analysis only, consistent with their released research licenses and intended use. We release no new model, dataset, or derivative artifact.

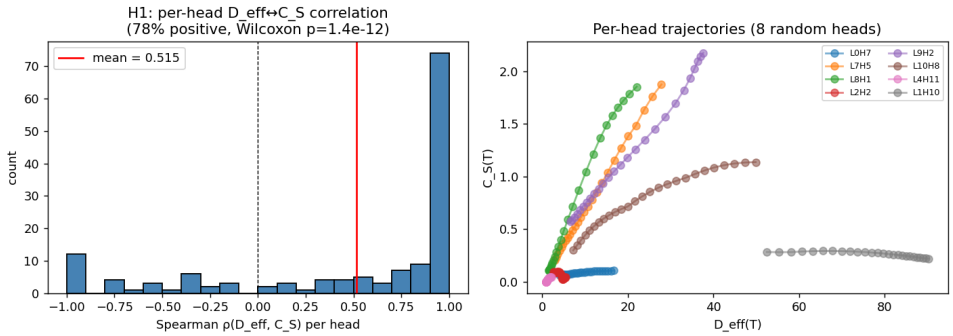

Figure 4: Per-head Spearman correlation between $D_{\text{eff}}(T)$ and $C_S(T)$ across the 26 temperatures (H1). Mean $\rho = 0.50$, median $\rho = 0.92$, 78% of heads positive (Wilcoxon $p = 1.4 \times 10^{-12}$). Heads driven Broad by global temperature pay more sender cost. Supports §5.7.

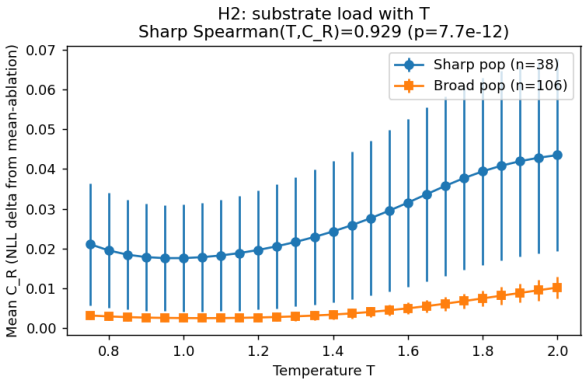

Figure 5: Sharp-population mean C_R across temperature (H2). Dense sweep, $n = 26$ temperatures, $T \in [0.75, 2.0]$ in steps of 0.05. Mean Sharp C_R rises from 0.018 at $T = 1.0$ to 0.044 at $T = 2.0$ (2.5×; Spearman $\rho = 0.93$, $p = 8 \times 10^{-12}$ over the full range; on the substrate-perturbation arm $T \geq 1.0$ the relationship is perfectly monotone, $\rho = 1.00$, $n = 21$). As substrate density falls under temperature, the heads still classified as Sharp become more critical. Supports §5.7.

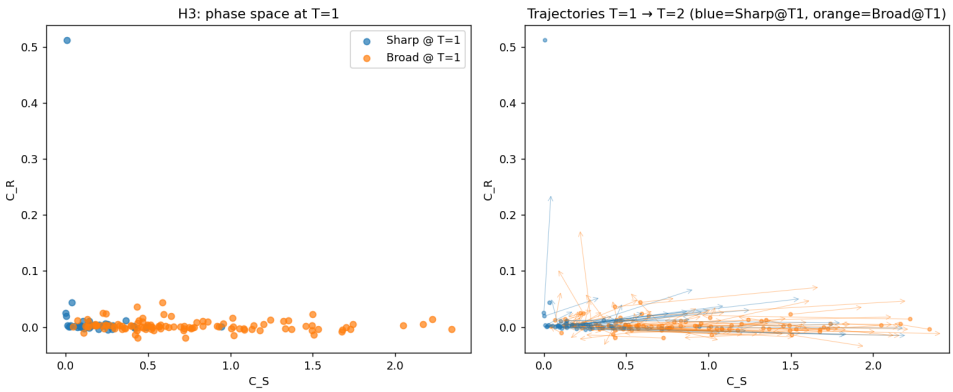

Figure 6: Phase-space separation at $T = 1.0$ on GPT-2 Small (H3). Each point is one of the 144 attention heads. Sharp heads (blue, $D_{\text{eff}} < 4$) concentrate at low C_S; Broad heads (red, $D_{\text{eff}} \geq 4$) span the full C_S range. C_S is in D_{eff} units (mean $|\Delta D_{\text{eff}}|$ under Q/K perturbation), C_R in nats (mean-ablation NLL increase). C_S Sharp vs Broad means 0.19 vs 0.75 (Mann–Whitney $p = 1.5 \times 10^{-12}$); C_R Sharp vs Broad means 0.018 vs 0.003. Supports §5.7.