DISTRIBUTED BIAS TERMS IN BIAS-FREE TRANSFORMER FEED-FORWARD NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern language models omit bias parameters from their feed-forward networks (FFNs). We show that implicit input and output biases emerge in these networks regardless, and that training organizes them by a structure that is consistent across more than twenty public dense and mixture-of-experts (MoE) models from 70M to 30B parameters. Every token's FFN input contains a large, nearly constant component along the mean input, which gives each neuron a fixed offset, and the average FFN output is added to every position; these are the input and output biases. Training arranges the neurons around the mean input: the gate vectors of most neurons point against it, while those of the few neurons that write the most are strongly aligned or anti-aligned with it. For an anti-aligned gate, the mean component drives the gate preactivation negative but does not shut the neuron off, because SiLU remains nonzero there. The output bias is therefore not the work of a few always-on neurons but a small surplus left over in almost every neuron, the largest one percent contributing between a fifth and a half. Nor are the biases constants. The channels that hold the mean input carry information about the token, including the norm of the residual stream that normalization discards, and in the middle layers context reduces how much most tokens write along the output bias. Models that do have bias parameters form the same implicit biases and keep the explicit ones at one to twenty percent of their size, and training with the mean input removed removes the structure without a measurable change in loss. In MoE models each expert has its own biases, and the output biases of different experts point in a common direction. They add up coherently in the block output, and they account for much of the similarity between experts that is measured on averaged outputs and commonly read as redundancy. Generally, our results show that mean activations are not merely nuisance statistics for mechanistic interpretability, but part of a recurring computational structure across LLMs.

1 INTRODUCTION

Most large language models have no bias parameters in their feed-forward networks (FFNs). Historically, bias parameters disappeared from Transformer FFNs before the current generation of LLMs. T5 already used bias-free FFNs, presented simply as an architectural convention (Raffel et al., 2020). The original SwiGLU formulation then followed this implementation and likewise omitted bias terms (Shazeer, 2020). PaLM later gave the choice an explicit empirical motivation, reporting that removing biases from dense kernels and normalization layers improved training stability at scale (Chowdhery et al., 2022). LLaMA subsequently adopted the PaLM-style bias-free SwiGLU architecture (Touvron et al., 2023), and this design has since become standard across many open-weight LLMs.

The modern bias-free SwiGLU FFN (Shazeer, 2020) is a sum of H neurons, each of which multiplies a nonlinear gate of its input z by a linear read of the same input and writes the product along a fixed vector:

$$f(z) = \sum_{h=1}^{H} \underbrace{\mathrm{SiLU}(q_h^\top z)}_{\text{gate}} \underbrace{(u_h^\top z)}_{\text{read}} \underbrace{v_h}_{\text{write}}, \tag{1}$$

where q_h, u_h and v_h are the neuron's gate, read and write vectors (Geva et al., 2021). The activation patterns of these neurons are a central object of mechanistic interpretability. In particular, it is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

essential to separate the part of a response that depends on the context from the part that is the same for every token, for instance when comparing the experts of a mixture-of-experts (MoE) layer. Without bias parameters one might expect the second part to be absent, which turns out not to be the case.

Several features of transformer activations have been reported to act like biases. A few outlier or rogue dimensions are large on every token and dominate similarity measures and quantization error (Kovaleva et al., 2021; Timkey & van Schijndel, 2021; Dettmers et al., 2022); massive activations on a few tokens act as fixed biases for attention (Sun et al., 2024) and are tied to attention sinks (Xiao et al., 2024); and removing a common mean direction is a standard post-processing step (Mu & Viswanath, 2018). How the FFNs of bias-free models implement biases has not been studied systematically: at what offsets their neurons operate, which neurons produce the part of their output that is the same for every token, and how the two are organized. Our findings:

Training polarizes the neurons along the mean input (Section 3). In every gated model the gate vectors of the bulk of the population align against the mean input, so that their gate bias is negative, while shuffled weights show no alignment. The few neurons that write most differ: their gates split into a branch along the mean input and a branch against it, and their read vectors align with it too, whereas the reads of the bulk ignore it.

The output bias is written by essentially every neuron (Section 4). A negative gate bias does not switch a neuron off, because SiLU is not a rectifier: the gates sit at negative values of order one and the neurons keep writing on almost every token. Virtually no neuron is silent and none emits a constant. The top 1% of neurons write 19–45% of the output bias, and the remaining 99% write the rest as small surpluses.

Neither bias is a constant. The channels of the mean input carry aggregate information about the token, such as the norm of the residual stream that normalization removed; this is learned, since the mean channels carry none at initialization. The output bias holds 0.9–20% of the FFN output energy in pre-norm models and 24–76% where the FFN output is normalized, and in the middle layers most tokens write less along it than their mean component alone would.

In MoE layers each expert has its own biases, and the experts' output biases are aligned (Section 5). They add coherently in the block output, and similarity between experts measured on averaged outputs, a common proxy for redundancy, largely measures this agreement in biases.

The same organization forms early in training and appears in models with explicit bias parameters, whose explicit biases stay at 0.8–19% of the implicit ones. We do not claim that models need bias parameters: removing the mean input during training removes the polarization without a measurable change in loss (Section 3).

2 Implicit biases

Definitions. Let $z_t \in \mathbb{R}^d$ be the FFN input at position t, after the normalization and gain that precede it, so that (1) reads

$$f(z) = \sum_{h=1}^{H} a_h(z)\, v_h, \qquad a_h(z) = \mathrm{SiLU}(q_h^\top z)\,(u_h^\top z). \tag{2}$$

Write $\mu = \mathbb{E}_t[z_t]$ for the mean input over a reference population of tokens and $z_t = \mu + \xi_t$. Every input projection then splits into a centered part and a constant,

$$q_h^\top z_t = q_h^\top \xi_t + \beta_h^{\mathrm{g}}, \quad u_h^\top z_t = u_h^\top \xi_t + \beta_h^{\mathrm{u}}, \qquad \beta_h^{\mathrm{g}} = q_h^\top \mu, \;\; \beta_h^{\mathrm{u}} = u_h^\top \mu, \tag{3}$$

and the mean output is a fixed vector added at every position,

$$b = \mathbb{E}_t[f(z_t)] = \sum_h \bar{a}_h\, v_h, \qquad \bar{a}_h = \mathbb{E}_t[a_h(z_t)]. \tag{4}$$

We call β_h the neuron's *input biases* and b the FFN's *output bias*. Both identities are exact. Note that b is the mean write, not the write at the mean input: $b - f(\mu)$ is the average of the second- and higher-order terms of f around μ, and the two differ substantially (Section 4).

Per-neuron quantities. Since $(u_h, v_h) \mapsto (-u_h, -v_h)$ leaves f unchanged, u_h, v_h and a_h have no sign of their own. We use three sign-invariant quantities: the gate alignment $\cos(q_h, \hat{\mu})$ with $\hat{\mu} = \mu/\|\mu\|$; the signed contribution to the output bias $c_h = \bar{a}_h\, v_h^\top \hat{b}$ with $\hat{b} = b/\|b\|$, so that $\sum_h c_h = \|b\|$; and the size of the mean write,

$$s_h = |\bar{a}_h|\, \|v_h\|, \tag{5}$$

by which we rank the neurons of a layer.

Experts. An MoE layer routes token t to a set $\mathcal{S}_t$ of experts, each an FFN of the form (1). Expert e has its own reference population, the tokens routed to it, and hence its own mean $\mu_e = \mathbb{E}[z_t \mid e \in \mathcal{S}_t]$, input biases $q_h^\top \mu_e$, $u_h^\top \mu_e$, and output bias $b_e = \mathbb{E}[f_e(z_t) \mid e \in \mathcal{S}_t]$. We keep b_e distinct from $f_e(\mu)$, the expert's response at the layer's mean input, which needs no activations.

Models and data. The detailed panel has five dense models (Qwen2.5-1.5B and 7B, Qwen3-1.7B-Base, SmolLM2-1.7B, OLMo-2-1B) and seven MoE checkpoints (GraniteMoE-1B and 3B, OLMoE-1B-7B, Qwen1.5-MoE-A2.7B, DeepSeek-V2-Lite, Qwen3-30B-A3B base and post-trained). OLMo-2 normalizes the FFN output rather than its input; Qwen1.5-MoE and DeepSeek-V2-Lite have an always-on shared expert. Gemma-3-1B enters once, as a second architecture with a normalized output; other pruned or distilled models are excluded. Gate statistics use a further ladder of 26 public models, 11 with explicit FFN biases (Pythia, GPT-2, BLOOM, Phi; LayerNorm and GELU) and 15 without (13 dense, 2 MoE). Activations are teacher-forced on held-out C4 windows with positions 0–63 discarded, which removes the attention sink and the early-position transient. Means and directions are fitted on discovery documents and evaluated on disjoint ones. Each model is one checkpoint, so intervals cover data, not training seeds. Token-level per-neuron statistics exist for the five MoE models whose activations we stored (Section A).

Untrained controls. A mean-dominated input passed through any gated map produces some of the structure below. We therefore recompute each statistic with every FFN weight matrix replaced, keeping the real inputs, the real routing and the matrix's scale, either by permuting its entries or by redrawing them as Gaussians of matched variance, and attribute a statistic to training only if it differs from both.

2.1 When is the mean input a bias?

The offset $\beta_h(t)$ is a bias in the usual sense, a constant added on every token, to the extent that c_t is the same for every token. In the models we examine c_t is positive on at least 99.6% of held-out tokens in eight of nine models and has a coefficient of variation of 14–22%: every neuron's input bias is its constant β_h times a shared factor $c_t/\|\mu\|$ with mean one and that spread. The component is present before training. Randomly initialized networks on the same windows align their tokens with their own mean input about as strongly (median $\cos(z_t, \hat{\mu})$ 0.36–0.58 at initialization against 0.34–0.53 trained in five of six models; GraniteMoE-1B is the exception, 0.14 against 0.41; Figure 3a). The individual channels of μ, in contrast, fluctuate about as much as their mean (Section C): it is the projection on $\hat{\mu}$ that is stable, not any channel. How μ arises is described in Section C; here we take it as given and study how the weights are organized around it.

Nor are the channels of μ a constant feature. A linear read of the ten largest of them in the normalized input predicts the log-norm of the pre-normalization residual on held-out documents with R^2 0.30–0.52 in the pre-norm models (median over layers), against 0.01–0.05 from ten random channels and -0.02 to 0.12 from the same channels of the untrained twin (Figure 3b,c); controlling for log position leaves this unchanged (Section B). These channels carry information about the token; whether the network uses it is not tested here. Because a few shared outlier channels can dominate cosines (Timkey & van Schijndel, 2021), the cross-expert alignment of Section 5 is also checked with the largest channels deleted (Section Q).

3 Training polarizes gate vectors along the mean input

The statistic. For each neuron we measure the alignment of its gate vector with the mean input, $\cos(q_h, \hat{\mu})$, and of its read vector, $\mathrm{sign}(\bar{a}_h)\cos(u_h, \hat{\mu})$; signing by the neuron's own mean activation

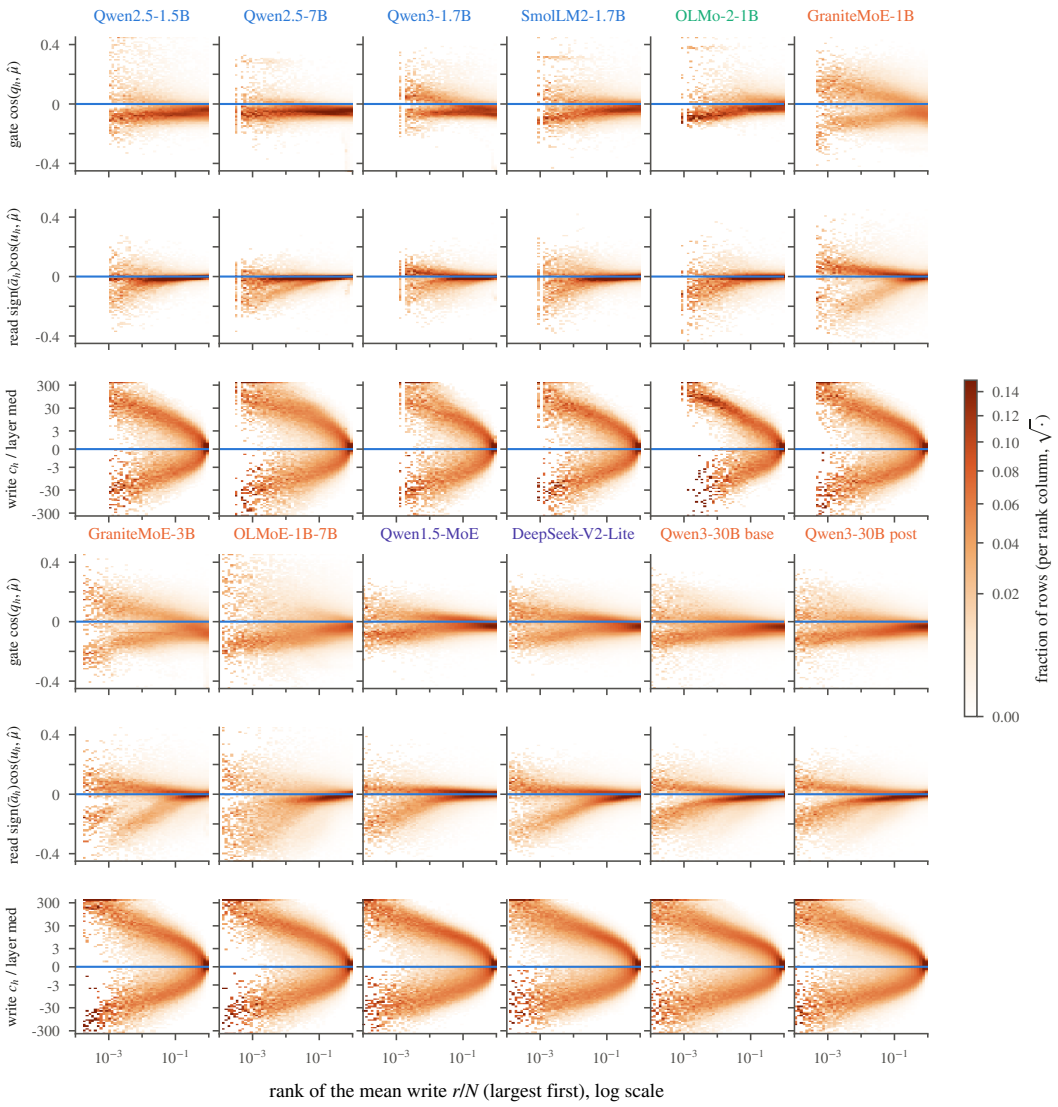

Figure 1: **Gate, read and write of every neuron against the mean input, for twelve public models.** Per neuron, against its rank by the size of its mean write s_h (logarithmic, largest first; ranks within layer, layers pooled, each rank column normalized): the gate alignment $\cos(q_h, \hat{\mu})$; the read alignment $\operatorname{sign}(\bar{a}_h)\cos(u_h, \hat{\mu})$; and the contribution to the output bias with the neuron's own part removed, $\bar{a}_h\, v_h^\top (b - b_h)/\|b\|$, over its layer's median absolute value (asinh axis). The bulk sits against the mean input with reads near zero; among the strongest writers the gates open into branches along and against it, and the reads polarize with them. Dense models at the layer mean μ, MoE experts at their own μ_e; plotting details in Section D.

removes the freedom $(u_h, v_h) \mapsto (-u_h, -v_h)$, under which $\bar{a}_h$ changes sign with u_h. An unarranged vector has $\cos \sim \mathcal{N}(0, 1/d)$, so we count a neuron as aligned along or against $\hat{\mu}$ when $|\cos| > 3/\sqrt{d}$, which an isotropic vector exceeds on either side with probability 0.13%. We resolve both alignments by the neuron's rank r/N within its layer by the size of its mean write, s_h (Equation (5)) on a logarithmic axis, which resolves the few neurons that write most (Figure 1; the same on a linear axis, which shows the population, in Figure 8), and summarize them for the top 1% and the rest (Figure 2; per-model counts in Section D).

The whole population is shifted against the mean input. The gate distribution of every pre-norm checkpoint sits below zero at every rank outside the top writers (Figures 1 and 8). Outside the top

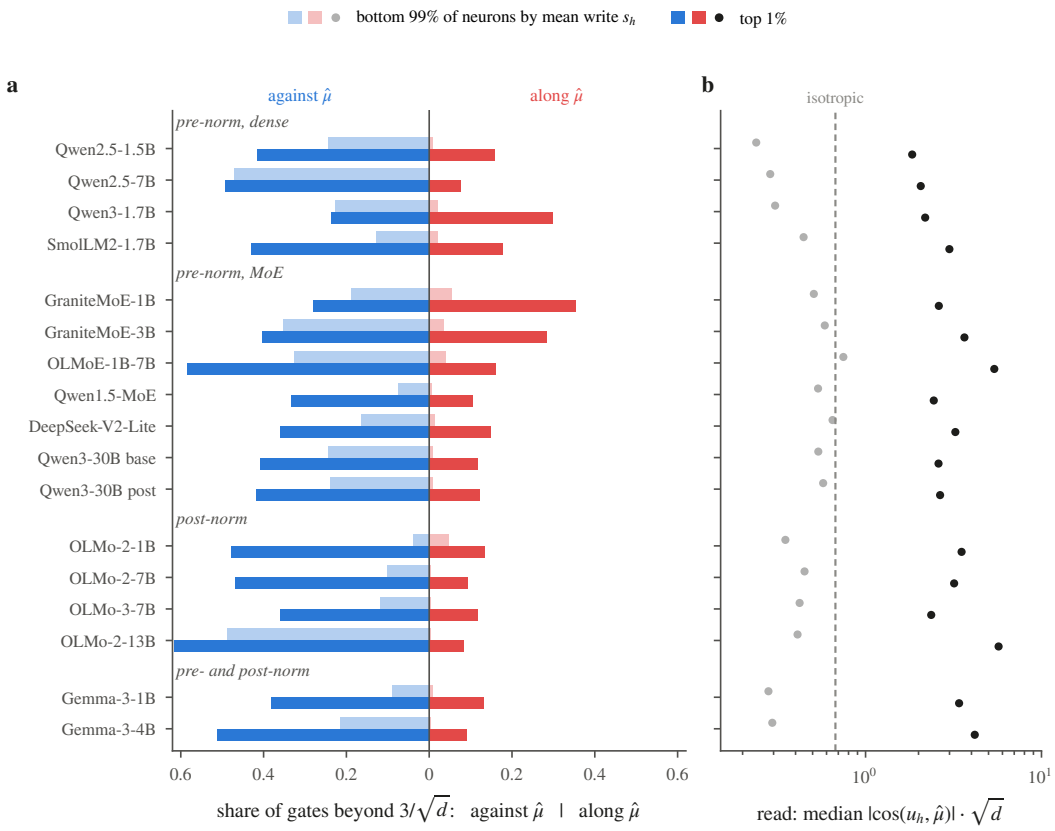

Figure 2: **Polarization of the top 1% of writers and of the rest, grouped by where the normalization sits.** (a) Share of gate vectors with $\cos(q_h, \hat{\mu})$ beyond $3/\sqrt{d}$ against the mean input (left, blue) and along it (right, red), for the 99% of neurons with the smallest mean write (light) and the top 1% (dark) by s_h within each layer; an isotropic vector exceeds the threshold on either side with probability 0.13%. (b) Median $|\cos(u_h, \hat{\mu})|\sqrt{d}$ of the read vectors; dashed: the isotropic median 0.674. Post-norm models normalize the FFN output rather than its input; Gemma-3 normalizes both.

1% of writers, 7.5–49% of gate vectors are aligned against $\hat{\mu}$ and only 0.2–5.4% along it, at least 3.5 times fewer, against 0.13% each for an isotropic vector, in 16 of the 17 checkpoints of Figure 2; the bottom half of the ranking alone already shows it (5.8–48% against, 0.1–3.0% along in the twelve checkpoints of Figure 1, OLMo-2-1B excepted). The shift is modest per neuron, a few times $1/\sqrt{d}$, but it is shared by essentially all of them. The read vectors of these neurons are not aligned: their median $|\cos|$ is 0.24–0.75 times $1/\sqrt{d}$, around the isotropic median of 0.67. Their writes lean toward the output bias throughout. A neuron's own write is part of b, so $c_h > 0$ partly by construction for a large writer; we therefore score the sign against $b - b_h$, the output bias with the neuron's own term b_h removed ($b_h = \bar{a}_h v_h$ in a dense layer, the routing-weighted equivalent in an MoE layer). Then 55–65% of all neurons write toward the output bias, still 51–57% in the bottom half of the ranking and 72–93% in the top 1% (79–94% before the removal), the share exceeds one half in every rank band, and every rank band contributes a net positive amount (Figure 1, bottom rows; Figure 9).

Normalization placement does not remove the shift. The one exception is OLMo-2-1B, which normalizes the FFN output instead of its input (outside its top 1%, 3.8% of gates against $\hat{\mu}$, 4.8% along). The other post-norm models we measured, OLMo-2-7B, OLMo-3-7B and OLMo-2-13B, and Gemma-3-1B and 4B, which normalize both the input and the output of the FFN, place 8.8–48.7% of their bulk against the mean input and at most 0.8% along it (Figure 2); they also differ from OLMo-2-1B in release, so we cannot say what sets it apart. Because the channels of the mean carry information such as the input norm (Section 2.1), a gate offset against $\hat{\mu}$ makes each neuron's

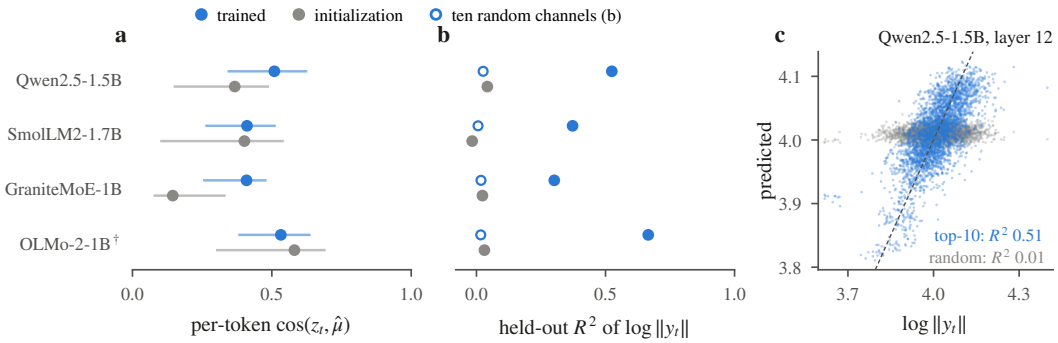

Figure 3: **Tokens align with the mean input at initialization already; the norm information in its channels is learned.** Blue: trained; grey: the same architecture at random initialization, on the same inputs. (a) Per-token $\cos(z_t, \hat{\mu})$ of the FFN input with its mean: median over tokens (dot) and 5th–95th percentile (line), median over the sampled layers. (b) Held-out R^2 of $\log \|y_t\|$, the norm of the pre-normalization residual, predicted from the ten largest channels of μ in z_t (filled) or from ten random channels (open), median over layers. (c) The predictions behind one point of (b): Qwen2.5-1.5B, layer 12. †OLMo-2 reads the unnormalized residual, so for it (b) is close to an identity. One model per family; all nine in Table 1. Packed 4096-token C4 windows, $t \geq 64$; fitted and scored on disjoint documents.

operating point, and with it its share of the output bias, depend on that information, wherever the normalization sits.

The strongest writers split into two branches. On the logarithmic axis the distribution opens toward the top of the ranking into a branch along $\hat{\mu}$ and a branch against it (Figure 1). Among the top 0.1% of neurons, 19–52% of gates are aligned along the mean input and 26–58% against it; the along branch is essentially confined to these neurons. Their read vectors are polarized as well, with median $|\cos|$ of 2.2–7.9 times $1/\sqrt{d}$ (1.8–5.4 for the top 1%). Gate and read are coupled through the write: at the mean input a neuron writes $\mathrm{SiLU}(\beta_h^{\mathrm{g}})\,\beta_h^{\mathrm{u}}\,v_h$, and among the top 1% of contributors that are measurably aligned on both input vectors this write points along b in 88–98% of cases in all twelve checkpoints (Section D). The pattern is the same in dense and MoE models, with and without shared experts, and in the base and post-trained Qwen3-30B checkpoints. The strongest writers read the mean input and write a disproportionate share of the output bias; the remaining population has gates offset against the mean input while its reads ignore it. The split is a partition by rank, not two populations: the sizes of the mean writes form one continuous distribution (Section 4).

Shuffled weights are not polarized. Permuting the entries of each weight matrix keeps the inputs and the weight distribution and removes the arrangement. On the same routed inputs (five MoE models, every eighth MoE layer, six experts per layer), shuffled experts split 49.8–50.6% on the sign of $q_h^\top \mu_e$, and none of their gates sits two standard deviations from zero in units of its own token-to-token spread, $q_h^\top \mu_e / \mathrm{std}_t(q_h^\top z_t)$; trained experts place 0.2–18.7% there.

Models with explicit biases. Few recent open-weight models keep FFN bias parameters. Pythia (Biderman et al., 2023) and ten older models (GPT-2, BLOOM, Phi) do, with the caveat that their FFNs are not gated and have no read vector. Their input vectors are nevertheless polarized against the mean input, one-sidedly, without the along branch of gated models, and the polarization forms within about 10^3 training steps. The explicit biases stay at 0.8–19% of the implicit offsets and are positively correlated with them ($+0.44$ to $+0.79$ in all eleven models); the implicit offset, not the explicit bias, predicts how often a neuron is active (Section F).

The geometry depends on the mean input; the loss does not resolve it. Training small dense models with the mean removed from every FFN input removes the structure (root-mean-square implicit offset 0.68 to 0.054, share of input vectors against the mean 0.83 to 0.48) at a loss of 3.131 against 3.117 without centering and 3.128 with centering plus explicit biases (three seeds each),

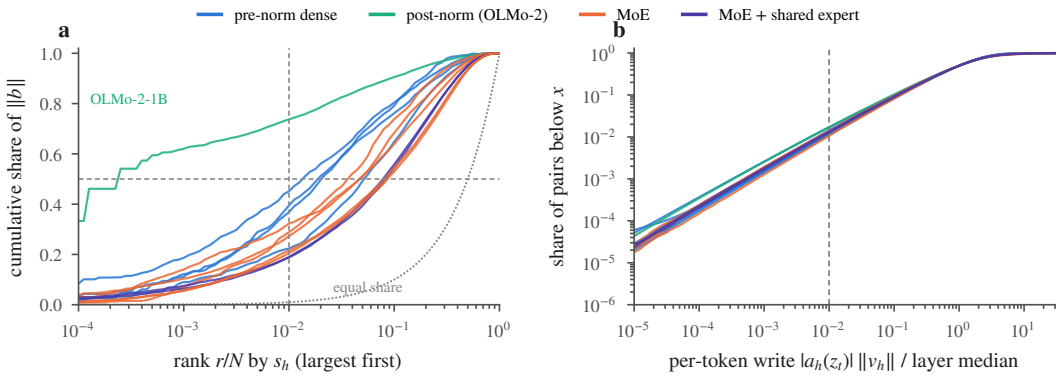

Figure 4: **Who writes the output bias.** (a) Cumulative share of the output bias, $\sum c_h/\|b\|$, over the neurons up to rank r/N by s_h (median over interior layers, twelve checkpoints); dotted: a share proportional to the number of neurons; dashed: the top 1% and one half. (b) Share of (neuron, token) pairs whose per-token write $|a_h(z_t)|\,\|v_h\|$ is below x times its layer median, seven dense and five MoE models; dashed: 1% of the median. Interior layers, relative depth 0.125–0.875; sampling in Section A.

differences below the 0.025-nat reliability floor of the comparison: removing the mean input does not improve the model, and we find no resolved evidence that it harms it. In trained checkpoints the implicit biases can be made explicit: replacing the projection of the FFN input on $\hat{\mu}$ in the first half of the layers by its discovery mean, a constant, recovers at least 98.7% of the loss incurred by deleting that projection in Qwen2.5, Qwen3, SmolLM2 and GraniteMoE-1B (Section G).

4 Who writes the output bias

Size. We measure the output bias against the FFN's own output, as the share of output energy that survives averaging over tokens, $\|b\|^2/\mathbb{E}_t\|f(z_t)\|^2$. In pre-norm dense models it is 0.9–20% across depth (Figure 16). In MoE experts it is 2.7–4.3%, estimated with a document-split cross-product that is unbiased under within-document dependence (Section O); in norm, $\|b_e\|/\mathbb{E}\|f_e\|$ is 0.18–0.24, so the output bias is about a fifth of a typical write. Where a normalization sits between the FFN output and the residual stream, as in OLMo-2 (output only) and Gemma-3 (input and output), the output bias carries 42–67% and 24–76% of the output energy, and $\|b\|$ is 0.55–1.80 times the root-mean-square of the centered output, against 0.09–0.50 in pre-norm models. There the constant enters the normalizer that rescales everything else the FFN writes (Section I). Size alone is not evidence of learning: a weight-shuffled FFN on the same inputs has $\|b_e\|/\mathbb{E}\|f_e\|$ of 0.11–0.24, because a mean-dominated input passed through any gated map produces an output bias. What training contributes is the organization.

Virtually no neuron is silent. If a small always-on subpopulation produced the output bias, the other neurons would be closed on most tokens. We measure what each neuron returns to the residual stream on each token, $|a_h(z_t)|\,\|v_h\|$, relative to its layer's median over (neuron, token) pairs, in the interior layers (Figure 4b). In all twelve models it forms a single lump around the median. Only 1.1–1.7% of pairs fall below 1% of the median, 0.13–0.26% below 0.1%, and 0.015–0.037% below 0.01%. Below the median the share falls as a power of the threshold with nearly the same exponent in every model (log–log slope 0.83–0.93 between 10^{-4} and 10^{-2} of the median), as expected if writes are small only on the tokens where the gate or the read happens to be near zero; a neuron that closes would instead add a floor of near-zero writes, and none appears. A typical neuron writes below 1% of the median on 1.0–1.8% of its tokens. Closed neurons do exist outside this range, in the final layer of the two Qwen3 models (27 of 6144 and 14 of 3072 neurons, with gate preactivations of -21 to -51). Averaged over tokens, no neuron falls below 1% of the median in three of five MoE models and 0.005–0.11% in the Granite models (Section D).

This includes the neurons whose gate points against the mean input (Figures 1 and 2), the majority, which under the reading of a gate as a switch would be off. SiLU is not a rectifier, and their gate

preactivations are negative but of order one: in the five MoE models the median realized gate of these neurons is -0.10 to -0.18, $|\,\mathrm{SiLU}(q_h^\top z_t)| > 0.05$ on 82–91% of their tokens, and preactivations below -5, where the neuron would be silent, are rare (5.7% of (neuron, token) pairs in GraniteMoE-1B, carrying 0.4% of the write; essentially none in OLMoE and Qwen1.5-MoE). (Section H).

A typical neuron's write is far from constant. A neuron that produced a bias by itself would write with a fixed sign. For each neuron we compare the size of its mean activation with its mean size, $|\bar{a}_h|/\mathbb{E}_t|a_h|$, which is 1 for a fixed sign and 0 for exact cancellation. The median over neurons is 0.14–0.19 in all five MoE models, and the sign of a neuron's contribution differs from the sign of its mean on 42–46% of tokens (Section D). A typical neuron's mean write is a sixth of its mean absolute write: the output bias is a small surplus left over in each of thousands of neurons.

The strongest writers and the rest both write the output bias. Ranking the neurons of each layer by s_h and summing their signed contributions c_h, which add up to $\|b\|$, separates the strongest writers from the rest by what they write (Figure 4a). In the interior layers of the eleven pre-norm checkpoints, the top 1% of neurons hold about 8% of the mean-write mass but write 19–45% of $\|b\|$, two to five times their share, because their writes are aligned with b; half of $\|b\|$ is reached within the top 1.4–8.6% of the neurons. The bulk writes a comparable total: the 90% of neurons that write least contribute 20–47% of $\|b\|$, each with a small surplus. Scored against the output bias with their own term removed (Section 3), 55–62% (dense) and 56–65% (MoE) of all neurons contribute positively, 72–93% and 74–82% among the top 1%, and the median contribution is positive in every rank band (Section D). OLMo-2 is the exception: its top 0.1% write 62% of $\|b\|$ and its top 1% 74%, so that in this post-norm model the output bias does come from dedicated neurons. Otherwise the largest writers are not a separate population: the distribution of s_h over its layer median has the same form in all twelve checkpoints, a bulk around the median and a heavy tail, and the top 1% are the tail of this one distribution; only OLMo-2 adds a localized packet at 237 times the median, the dedicated writers above (Section E).

The output bias is dynamic. How much a token writes along the output bias depends on its context. We compare the block evaluated at the mean input with the average write, through $r = f(\mu)^\top \hat{b}/\|b\|$, at every layer of nine models, with b the change the block makes to the mean of the residual stream and f the whole block. In the middle layers of Qwen2.5-0.5B and 1.5B, Qwen3-0.6B and 1.7B, SmolLM2-1.7B and the GraniteMoE-1B block, r exceeds 1 and peaks at 1.7–3.2; no pre-norm model exceeds 0.92 at random initialization, and Llama-3.2-1B and SmolLM2-360M never exceed 1.07. At those layers 86–98% of held-out tokens write less along b than the mean component of their own input alone would, against 1–11% at initialization (Section M). Along the line from μ to a real token the write along b is concave in 99% of cases in Qwen2.5-1.5B, so context lowers the default write rather than adding to it. OLMo-2 is excluded: at random initialization its output normalization alone gives r of 2.1–2.5. These measurements are observational; we have not clamped the mean component and measured the effect.

5 Mixture-of-experts: aligned biases and expert similarity

Each expert is an FFN with its own input biases about the mean μ_e of its routed tokens and its own output bias b_e; replacing the input biases by explicit constants changes the loss by less than 0.02 nats in all seven checkpoints, and the anatomy of the previous sections holds within experts (Figures 1 and 4). The experts' output biases share a direction: scored against the normalized sum of the other experts' unit vectors (leave-one-out, so that no expert is scored against a direction it helped fit), the median cosine is 0.19–0.37 in all seven checkpoints and positive at every layer, against at most 0.017 for isotropic vectors. The agreement belongs to the trained experts at their actual input: in the three models without a shared expert, the mean pairwise cosine of the experts' responses at the mean input falls from 0.15–0.26 to 0.010–0.011 when μ is replaced by a random or permuted vector of the same norm, and to at most 0.0005 for norm-matched random experts (Section N.1). Aligned biases add coherently in the block output: in GraniteMoE-1B, GraniteMoE-3B and OLMoE one fixed direction carries 4–6% of the block's output energy, and 15–33% at the final layer, against 0.05–0.1% for a random direction (Section P).

Consequence for expert similarity. Similarity between experts is commonly computed on averaged outputs or activations and read as redundancy, for pruning and merging (Lu et al., 2024; Li et al., 2024; He et al., 2024). Averaging removes the first-order response to the input and keeps the bias, so these statistics measure agreement of biases: the mean pairwise cosine between output biases is 0.06–0.20 in six checkpoints, while the same experts, compared as functions on a common sample of tokens, have mean per-token cosine 0.004–0.023 (Figure 21, Section N). Agreement in biases is a real property of the model; a claim that two experts compute similar functions is a claim about their centered responses on common inputs, and should be reported separately.

6 Related work

Outlier dimensions, massive activations and sinks. Contextual representations are anisotropic (Ethayarajh, 2019); a few rogue or outlier dimensions dominate similarity measures and matter for quantization and function (Kovaleva et al., 2021; Timkey & van Schijndel, 2021; Dettmers et al., 2022); and removing common mean directions is a standard post-processing step (Mu & Viswanath, 2018). Sun et al. (2024) show that a handful of very large activations on a few tokens act as fixed biases for attention and relate them to attention sinks (Xiao et al., 2024), which Queipo-de-Llano et al. (2026) connect to low-rank compression valleys. The outlier dimensions are the channels of μ (Section 2.1): large on every token, but not constant. Massive activations are rare values on a few tokens and almost absent at the FFN input: removing every token on which the criterion is met leaves μ at cosine ≥ 0.97 to the original (Section K). Neither is the constant input feature an explicit bias supplies.

FFN neurons and mean ablation. The key–value view (Geva et al., 2021) describes neurons by what they read and write. We add the operating point: the offset at which a neuron's gate sits on average, and what the neuron writes on average. Replacing an activation by its mean is a standard intervention (Wang et al., 2023); our results describe what that mean contains at FFN inputs and outputs.

Experts. Sparse MoE layers (Shazeer et al., 2017; Fedus et al., 2022; Dai et al., 2024) make the expert a natural unit of analysis. Expert similarity is often measured on raw outputs: token-wise cosine or angular similarity (Lo et al., 2025), or the cosine of outputs averaged over inputs (Omi et al., 2025), used for clustering and compression (Li et al., 2026) and evaluated as a similarity metric (Wu et al., 2026); HC-SMoE represents an expert by its mean output over a calibration corpus (Chen et al., 2025). None separates each expert's mean output from the input-dependent variation around it, so a shared mean enters the measured similarity directly, which we advocate against here.

7 Discussion and limitations

Bias-free FFNs have biases: each neuron's operating point is set by its weights' projection on the mean input, the layer adds its mean output at every position, and training arranges the gates around the mean input in a way that shuffled weights do not reproduce. Two consequences follow. The mean input is not a neutral reference for interventions: since a neuron's bias on a token is its constant times the token's coefficient along the mean direction, and most gates are aligned against that direction, anything that changes the coefficient, such as mean-ablating an upstream component, shifts the operating points of most neurons in the same direction. And similarity between experts computed on averaged outputs measures agreement of output biases; whether two experts compute similar functions is a question about their centered responses on common inputs, and the two should be reported separately.

Limitations. Every public model is one checkpoint, so intervals cover documents, not training seeds. Token-level per-neuron statistics exist for five MoE models. The contrast with models that normalize the FFN output rests on one model per architecture, without untrained controls. The context dependence of the output bias is a paired comparison of each token with its own mean component, not an intervention, and the power-law tail awaits a Vuong test. We do not test what later layers do with the output bias or with the norm information in the mean channels, nor show that any published conclusion about expert redundancy changes.

AI USE STATEMENT

We used generative AI tools to implement methods and analysis code, to design experiments and their controls and give feedback on them, to clean and reformat data, to support data analysis, to help interpret results and propose and refine hypotheses, to create and edit figures and software, to search and summarize literature, and to draft and edit parts of this paper. We did not use them to generate synthetic data or for translation. We reviewed all AI-assisted work: analysis code was reviewed independently before use and tested against known cases, including tests verified to fail when the fix they guard is reverted; every finding that suggested learned structure was rerun against untrained-weight controls, and several were withdrawn as a result (Section H); and numbers in the text were checked against the recorded artifacts. We take responsibility for the final content, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All public models are identified by checkpoint in Section 2. Estimators are defined where first used; corpora, sampling, token floors and split-half reliabilities are in Section A; the fitting procedure for the contribution distribution is in Section E; the untrained-weight controls are specified in Section 2 and Section H. Every figure is regenerated from a stored data file by a standalone script.

REFERENCES

S. Biderman et al. Pythia: A suite for analyzing large language models across training and scaling. ICML, 2023.

I.-C. Chen, H.-S. Liu, W.-F. Sun, C.-H. Chao, Y.-C. Hsu, C.-Y. Lee. Retraining-free merging of sparse MoE via hierarchical clustering. Proceedings of the 42nd International Conference on Machine Learning, 267:8594–8620, 2025.

A. Chowdhery et al. PaLM: Scaling language modeling with Pathways. arXiv:2204.02311, 2022.

D. Dai et al. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. ACL, 2024.

T. Dettmers, M. Lewis, Y. Belkada, L. Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. NeurIPS, 2022.

K. Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. EMNLP-IJCNLP, 2019.

W. Fedus, B. Zoph, N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(120):1–39, 2022.

M. Geva, R. Schuster, J. Berant, O. Levy. Transformer feed-forward layers are key-value memories. EMNLP, 2021.

S. He, D. Dong, L. Ding, A. Li. Demystifying the compression of mixture-of-experts through a unified framework. arXiv:2406.02500, 2024.

R. Hu, Y. Cao, B. Kong, M. Sun, K. Yuan. Synergistic intra- and cross-layer regularization losses for MoE expert specialization. 2nd DeLTa Workshop at ICLR, 2026.

O. Kovaleva, S. Kulshreshtha, A. Rogers, A. Rumshisky. BERT busters: Outlier dimensions that disrupt transformers. Findings of ACL, 2021.

P. Li, Z. Zhang, P. Yadav, Y.-L. Sung, Y. Cheng, M. Bansal, T. Chen. Merge, then compress: Demystify efficient SMoE with hints from its routing policy. ICLR, 2024.

L. Li, Q. Zhu, J. Wang, X. Qin, W. Li, H. Gu, S. Han, Y. Guo. Sub-MoE: Efficient mixture-of-expert LLMs compression via subspace expert merging. Proceedings of the AAAI Conference on Artificial Intelligence, 40(27):22994–23002, 2026.

K. M. Lo, Z. Huang, Z. Qiu, Z. Wang, J. Fu. A closer look into mixture-of-experts in large language models. Findings of the Association for Computational Linguistics: NAACL 2025, pp. 4427–4447, 2025.

X. Lu, Q. Liu, Y. Xu, A. Zhou, S. Huang, B. Zhang, J. Yan, H. Li. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. ACL, 2024.

J. Mu, P. Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. ICLR, 2018.

N. Muennighoff et al. OLMoE: Open mixture-of-experts language models. ICLR, 2025.

N. Omi, S. Sen, A. Farhadi. Load balancing mixture of experts with similarity preserving routers. arXiv:2506.14038, 2025.

E. Queipo-de-Llano, A. Arroyo, F. Barbero, X. Dong, M. Bronstein, Y. LeCun, R. Shwartz-Ziv. Attention sinks and compression valleys in LLMs are two sides of the same coin. ICLR, 2026.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67, 2020.

N. Shazeer. GLU variants improve transformer. arXiv:2002.05202, 2020.

N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. ICLR, 2017.

M. Sun, X. Chen, J. Z. Kolter, Z. Liu. Massive activations in large language models. COLM, 2024.

W. Timkey, M. van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. EMNLP, 2021.

H. Touvron et al. LLaMA: Open and efficient foundation language models. arXiv:2302.13971, 2023.

K. Wang, A. Variengien, A. Conmy, B. Shlegeris, J. Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. ICLR, 2023.

J. Wu, J. Cheng, F. Lv, D. Ou, L. Yuan. SERE: Similarity-based expert re-routing for efficient batch decoding in MoE models. International Conference on Learning Representations, 2026.

G. Xiao, Y. Tian, B. Chen, S. Han, M. Lewis. Efficient streaming language models with attention sinks. ICLR, 2024.

F. Xue, Z. Zheng, Y. Fu, J. Ni, Z. Zheng, W. Zhou, Y. You. OpenMoE: An early effort on open mixture-of-experts language models. ICML, 2024.

A DATA, SAMPLING AND STATISTICS

Corpora and positions. All statistics use teacher-forced C4 English validation text. Dense models: per-model tokenization, 1024-token windows, 61,440 discovery tokens per layer. MoE models: one shared draw of 128 documents that qualify under all five tokenizers, two disjoint 512-token windows per document, 57,344 discovery tokens, with the same split seeds throughout. Three further corpora test corpus shift (below). No window begins with a beginning-of-sequence token and no tokenizer in the panel prepends one, except Gemma-3, whose windows follow its own convention. Positions 0–63 are excluded everywhere.

Corpus shift. We repeat the measurement for one MoE model (GraniteMoE-1B) and one dense model (Qwen2.5-1.5B) on three corpora besides C4: Wikipedia articles (4 and 22 of the 256 windows, for the two models, cross a junction between articles on unrelated topics), books (PG-19 validation and test, excerpts starting 10–60% into the book) and Python source (codeparrot-clean-valid, one file per document). Every corpus uses the MoE protocol above: 128 documents, two disjoint 512-token windows each, the same document split, and 57,344 discovery tokens per layer. The mean input μ (and μ_e in the MoE model) is recomputed on each corpus, and every alignment is taken against that corpus's own mean. Qwen2.5-1.5B has the same tokenizer as Qwen1.5-MoE (vocabulary, merges and pre-tokenization) and uses its windows. For this comparison the read cosine is signed by the direction of the write, $\mathrm{sign}(v_h^\top b)\cos(u_h, \hat{\mu})$, and alignments are in units of $1/\sqrt{d}$.

The mean input itself moves with the corpus (Figure 5a). Its cosine with the C4 mean, median over layers, is 0.89–0.92 for Wikipedia, 0.78–0.87 for books and 0.60–0.61 for code, falling to 0.38 in the last layer of the dense model on code. The polarization is measured against each corpus's own mean, and its population-level form does not move. The median gate alignment of all neurons stays at -1.1 to -1.3 (MoE) and -1.8 to -2.1 (dense) on every corpus (Figure 5b). Outside the top 1%, 15–19% (MoE) and 20–27% (dense) of gates are aligned against $\hat{\mu}$ ($|\cos| > 3/\sqrt{d}$) and 4.0–5.4% and 0.8–0.9% along it. In the rank band $r/N \in [10^{-2}, 10^{-1}]$ the dense reads stay near chance on every corpus ($|\text{median}| \leq 0.84$, chance 0.67), while the MoE reads keep both branches (-1.25 to -1.67 and 1.04 to 1.12; Figure 5c). Among the top 0.1%, the branch along $\hat{\mu}$ holds 31–39% (MoE) and 29–35% (dense) of gates on every corpus.

The top of the ranking is corpus dependent. On books and code the branch against $\hat{\mu}$ among the top 0.1% shrinks (MoE 40% on C4, 26% and 19%; dense 44%, 28% and 38%). The median alignment of the top 1% falls, most in the MoE model, where the gate goes from 0.114 on C4 to 0.090 and 0.084, and the read from 0.082 to 0.052 and 0.051 (Figure 5d,e). Different neurons also make up the top 1%. Their overlap with the C4 top-1% set, as a Jaccard index, is 0.37–0.49 for Wikipedia, 0.22–0.31 for books and 0.13–0.22 for code, against 0.59 between two discovery halves of C4 in the MoE model (Figure 5f; MoE layers in which the token floor removes an expert on either side are excluded, which leaves 15–22 of 23). The dense model's C4 statistics from the main protocol (1024-token windows, different documents) agree with the 512-token C4 run to within 0.04 in null units and 0.003 in the top-1% medians, so window length does not enter. These are one model per family and one draw of documents; we attach no intervals.

Per-token writes. Figure 4b uses eight dense models (all neurons of the interior sampled layers, 2000 held-out tokens each, including the distilled Llama-3.2-1B, which is not among the twelve checkpoints) and five MoE models (every fourth interior MoE layer, 256 neurons of up to six experts per layer, 400 routed tokens each).

Token floors and reliability. Per-expert token counts have medians of 2.1k–13.4k per layer. The 128-expert Qwen3-30B models are the thin end: a floor of 500 tokens drops 30% of their (layer, expert) cells, so their panels describe frequently routed experts. Between two independent discovery halves, the direction of the block output mean has cosine 0.97–0.99, and the top-1% sets of large contributors overlap by 0.46–0.75. Estimation noise in μ_e attenuates cosines (by a factor of about 0.996 in the worst retained cell), so sampling error works against, not for, every polarization statistic. Experts with $b_e = 0$ are ranked last under every ranking and produce no undefined cosines.

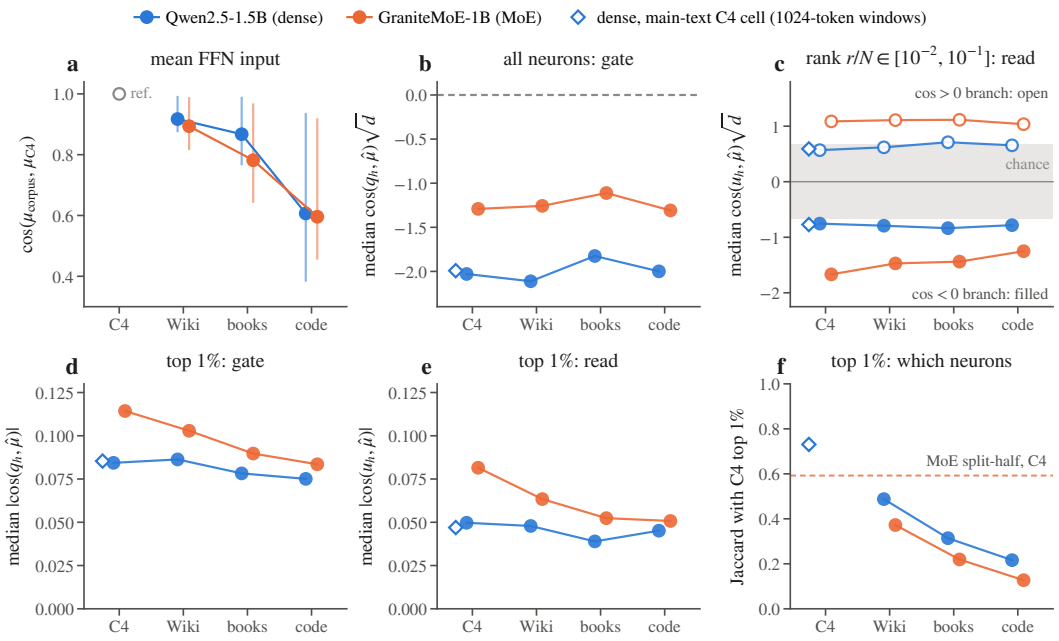

Figure 5: **Corpus shift.** GraniteMoE-1B (orange) and Qwen2.5-1.5B (blue) on C4, Wikipedia, books (PG-19) and Python source, 57,344 tokens per layer each, every alignment against the corpus's own mean input. (a) Cosine of the layer mean input with the C4 mean: median over layers (dot) and range (bar). (b) Median $\cos(q_h, \hat{\mu})\sqrt{d}$ over all neurons. (c) Median signed read $\cos(u_h, \hat{\mu})\sqrt{d}$ in the rank band $r/N \in [10^{-2}, 10^{-1}]$, separately for positive (open) and negative (filled) values; grey band: ± 0.67, where an isotropic vector's medians fall. (d, e) Median $|\cos|$ of gate and read among the top 1% by s_h. (f) Jaccard overlap of the top-1% set with that on C4; dashed: two discovery halves of C4 (MoE). Diamonds: the dense model's main-protocol C4 statistics.

B Norm information in the channels of the mean

Estimator. In each of the nine models with stored single-token inputs, at eight sampled layers, the target is $\log \|y_t\|$, where y_t is the residual stream entering the FFN's normalization, recorded per token by the collector before the normalization is applied. The predictors are the values $z_t[j]$ of the normalized FFN input in the ten channels with the largest $|\mu_j|$. Four disjoint subsets of documents, each with its own packed 4096-token C4 windows, are used for different purposes: μ, and hence the choice of channels, is computed on subsets 0 and 1; a ridge regression on standardized features, with its penalty chosen by five-fold cross-validation, is fitted on subset 2; and R^2 is scored on subset 3. Positions $t < 64$ are excluded. Neither the channel choice nor the fit sees the scored tokens.

Controls. (i) The same fit on ten channels drawn at random from the rest (mean over five draws). (ii) The same fit on the ten largest channels of μ in the randomly initialized twin. (iii) Position: the channels of μ also carry position (Section C), so we fit $\log \|y_t\|$ on $\log t$, remove that fit (estimated on subset 2) from the target, and predict the remainder. Position explains none of the norm, and removing it leaves the prediction unchanged (Table 1).

The information is not an artifact of the normalization. The normalization computes $z_j = g_j\, y_j \sqrt{d}/\|y\|$, so a channel whose value in y were nearly constant would read out $1/\|y\|$ after normalization by construction. The channels of μ are not constant. In y, the magnitude of each of the ten channels varies across tokens with a standard deviation of 0.95–1.16 in $\log |y_j|$, against 0.07–0.19 for $\log \|y\|$; it keeps the sign of its mean on only 84–94% of tokens; and it correlates only weakly with the norm ($+0.11$ to $+0.23$). A predictor that assumes constant channels, $\log \|y_t\| \approx \log(g_j \bar{y}_j \sqrt{d}/z_{t,j})$ averaged over the ten channels, has negative R^2 (-1.8 to -29) in every model, and so does its collective version, which assumes only that the projection $\bar{y}^\top y_t$ on the ten channels

Table 1: **Held-out R^2 of $\log \|y_t\|$,** median over the sampled layers: from the ten largest channels of μ, from ten random channels, from the ten largest channels at initialization, from $\log t$ alone, and from the ten largest channels after the $\log t$ fit is removed from the target.

model	top-10	random 10	top-10, init	$\log t$ only	top-10, $\log t$ removed
Qwen2.5-1.5B	0.52	0.03	0.04	0.000	0.52
Qwen3-1.7B	0.47	0.03	0.04	0.001	0.47
SmolLM2-1.7B	0.37	0.01	-0.02	-0.000	0.37
Qwen2.5-0.5B	0.46	0.03	–	-0.000	0.46
Qwen3-0.6B	0.47	0.05	–	0.000	0.47
SmolLM2-360M	0.45	0.02	–	-0.001	0.45
GraniteMoE-1B	0.30	0.02	0.02	-0.001	0.30
Llama-3.2-1B	0.34	0.02	0.12	-0.001	0.34
OLMo-2-1B	0.67	0.02	0.03	-0.000	0.67

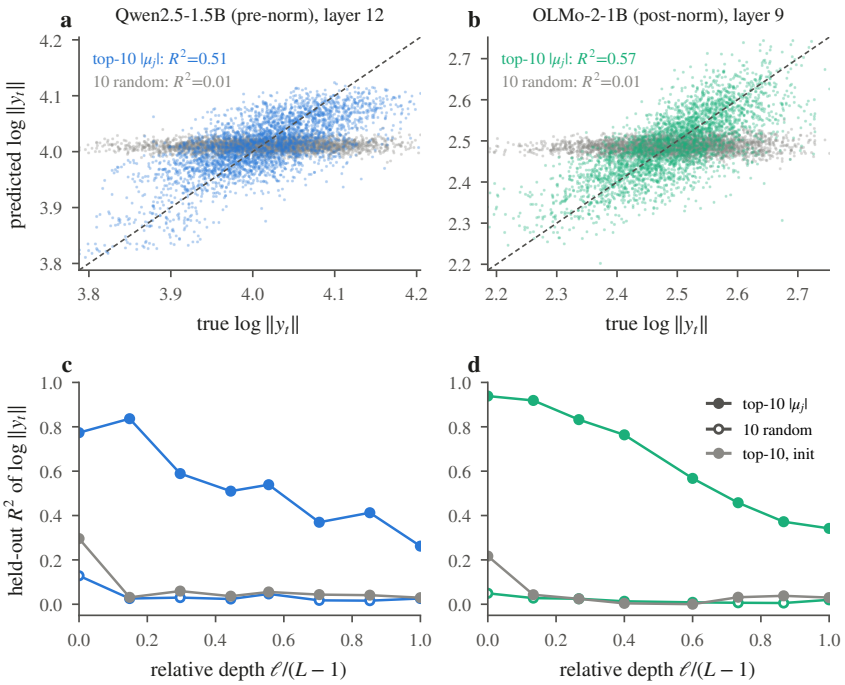

Figure 6: **The mean channels carry the input norm in a pre-norm and a post-norm model.** (a, b) Held-out prediction of $\log \|y_t\|$ from the ten largest channels of μ in the FFN input (coloured) and from ten random channels (grey), one layer each of Qwen2.5-1.5B, whose FFN reads the normalized residual, and OLMo-2-1B, whose FFN reads the unnormalized residual y_t itself. (c, d) Held-out R^2 through depth: ten largest channels (filled), ten random channels (open), ten largest at initialization (grey). In OLMo-2 the ten channels are part of y_t, so their prediction of its norm is partly compositional; in Qwen2.5 the input is normalized and the norm has to be read off the channels' values.

is constant: $\log \|y_t\| \approx c - \log \sum_j (\bar{y}_j / g_j) z_{t,j}$ with one fitted constant (R^2 -1.1 to -40). That projection varies across tokens with a coefficient of variation of 0.17–0.66, more than the norm itself does. The norm information read by the fitted linear combination therefore comes from how the signed values of these channels co-vary with the norm, not from normalizing a constant; training creates it (it is absent at initialization and in random channels), and we do not identify its mechanism. In OLMo-2 the FFN reads the unnormalized residual, so the ten channels are components of the vector whose norm they predict and part of the prediction is compositional (Figure 6).

C THE MEAN INPUT AS A BIAS

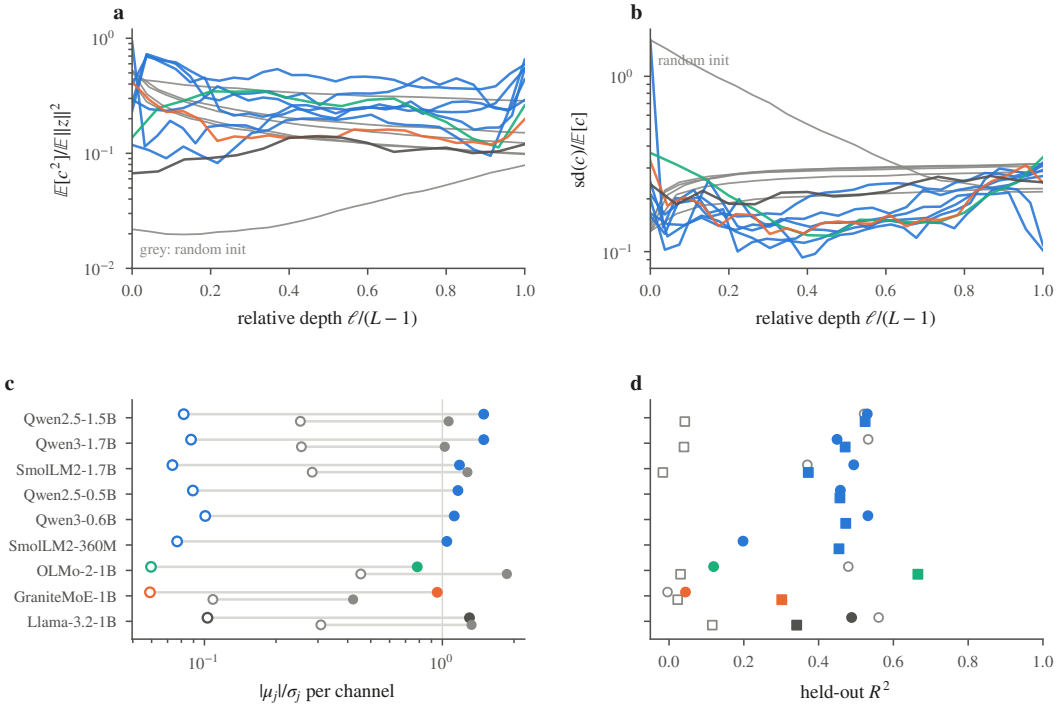

Figure 7: **The mean input as a bias.** With $c_t = z_t^\top \hat{\mu}$ the coordinate of the FFN input along its mean: (a) share of the input energy along the mean, $\mathbb{E}[c^2]/\mathbb{E}\|z\|^2$, and (b) its token-to-token coefficient of variation $\mathrm{sd}(c)/\mathbb{E}[c]$, through depth; coloured: trained (blue dense, green OLMo-2, orange GraniteMoE-1B block, dark grey Llama-3.2-1B), grey: random initialization. (c) $|\mu_j|/\sigma_j$ per channel, median over the ten largest channels of μ (filled) and over all channels (open); $\times$: ten largest at initialization. (d) Held-out R^2 of the log norm of the pre-normalization residual from the ten largest channels of μ (squares) and of log position from the hundred largest (circles); open grey: initialization. Packed 4096-token C4 windows, $t \geq 64$; means from two document subsets, token statistics fitted and scored on two others.

A mean is not a bias. Decomposition (3) holds for any input distribution, and $w^\top \mu$ is then just the mean of $w^\top z_t$. An explicit bias corresponds to an input feature equal to 1 on every token. The mean input plays that role only to the extent that every token carries a large component along μ that varies little from token to token. We measure this with the coordinate $c_t = z_t^\top \hat{\mu}$ in nine models for which we stored full first and second moments and single-token inputs (Figure 7), each with a randomly initialized twin.

The direction of the mean behaves as a bias. c_t is positive on at least 99.6% of held-out tokens at every sampled layer in eight of the nine models (SmolLM2-360M falls to 78% at one layer). It carries 11–45% of the FFN input energy (median over layers) and varies across tokens with a coefficient of variation of 0.14–0.22.

This alone is not learned. At random initialization c_t is also positive on essentially every token, with a coefficient of variation of 0.21–0.43: a mean-dominated input is a property of the architecture and the data before it is one of training. What training changes is where the mean lives. In trained models the ten largest channels of μ hold 81–97% of $\|\mu\|^2$ (43% in the distilled Llama-3.2-1B), against 4–8% at initialization, while they carry only 3–19% of the token-to-token variance of the input.

The channels of the mean are not constant. Each of these channels fluctuates about as much as its mean: $|\mu_j|/\sigma_j$ is 0.8–1.5 for the ten largest channels (median over them), against 0.06–0.10 for the median channel. The component along $\hat{\mu}$ is stable because the fluctuations of different channels partly cancel in the projection, not because any channel is constant. The channels also carry information. A linear read of the ten largest predicts the log norm of the pre-normalization residual on held-out documents with R^2 0.30–0.67 (0.61–0.85 from the hundred largest), against at most 0.14 at initialization. Log position is decodable from the hundred largest with R^2 0.12–0.53, but equally at initialization, so we do not attribute it to training. The input bias terms are therefore dynamic: a large, sign-stable offset that also carries information about the rest of the input. We keep this in mind when we call $w^\top \mu$ a bias.

D NEURON ANATOMY

Plotting details for Figures 1 and 2. In Figure 1, rank columns with fewer than 30 neurons are left blank, and in the bottom row the positive and negative halves of each rank column are normalized separately. The post-norm and Gemma-3 models of Figure 2 were measured with the same estimator on the same C4 windows as OLMo-2-1B (Gemma-3: its own windows, with a beginning-of-sequence token).

Cancellation within neurons. Table 2 reports, per neuron, the ratio $|\mathbb{E}_t x|/\mathbb{E}_t|x|$ for the gate value $x = \mathrm{SiLU}(q_h^\top z_t)$, the read value $x = u_h^\top z_t$ and their product $x = a_h$, and the share of tokens on which the sign of the neuron's contribution differs from the sign of its mean. The ratio is 1 for a quantity of fixed sign and 0 for one that cancels exactly.

Table 2: **Cancellation within neurons.** Median over neurons of $|\mathbb{E}_t x|/\mathbb{E}_t|x|$ for gate value, read value and product, and share of tokens on which the sign of the contribution differs from the sign of its mean (values below 1% of $\mathbb{E}_t|a_h|$ excluded). Held-out C4 windows, every fourth MoE layer, up to eight experts with at least 2000 routed tokens.

model	gate	read	product	sign differs
GraniteMoE-1B	0.683	0.369	0.190	0.423
GraniteMoE-3B	0.643	0.331	0.184	0.422
OLMoE-1B-7B	0.507	0.282	0.166	0.441
DeepSeek-V2-Lite	0.485	0.312	0.158	0.447
Qwen1.5-MoE	0.410	0.244	0.142	0.455

Gate and read biases are tied by the sign of the write. At the mean input a neuron writes $\mathrm{SiLU}(\beta_h^{\mathrm{g}})\,\beta_h^{\mathrm{u}}\,v_h$. For this write to point along b, the read bias and the write vector must have opposite relative signs when the gate bias is negative and equal signs when it is positive. The relation is invisible in raw weights, where gate and read alignment are uncorrelated over all neurons and over the large contributors alike, and it appears once the sign freedom $(u_h, v_h) \mapsto (-u_h, -v_h)$ is fixed by $\mathrm{sign}(v_h^\top b)$ (Table 3). Over all neurons 96% have $|\cos(u_h, \hat{\mu})| \le 0.05$; among the rest the relation holds for about 76%. It concerns the write at the mean input: on individual tokens the sign of a neuron's contribution differs from its mean sign on 42–46% of tokens (Table 2).

Depth. Splitting the gate row of Figure 1 by depth shows where the branch along the mean input comes from. In the first fifth of the layers the largest writers point against the mean input; the branch along it appears at mid depth and holds the top of the ranking in the last fifth of the layers (Figure 10).

Per-token openness. The statistics of Section 4 on realized gates pool $\mathrm{SiLU}(q_h^\top z_t)$ over the neurons with $\cos(q_h, \hat{\mu}_e) < -0.1$, each expert at its own μ_e on its own routed tokens; a gate preactivation below -5 is counted as closed. The mean absolute write of a neuron is $M_h = \|v_h\|\,\mathbb{E}_t|a_h|$ over its layer median, and its Spearman correlation with the signed alignment $\cos(q_h, \hat{\mu}_e)$ and with $|\cos(q_h, \hat{\mu}_e)|$ is taken per layer and summarized by the median over layers.

Ranking by contribution to the output bias. The main text ranks neurons by s_h, which ignores the direction of the write. Ranking instead by the contribution to the output bias, $s_h|\cos(v_h, \hat{b})|$ (for

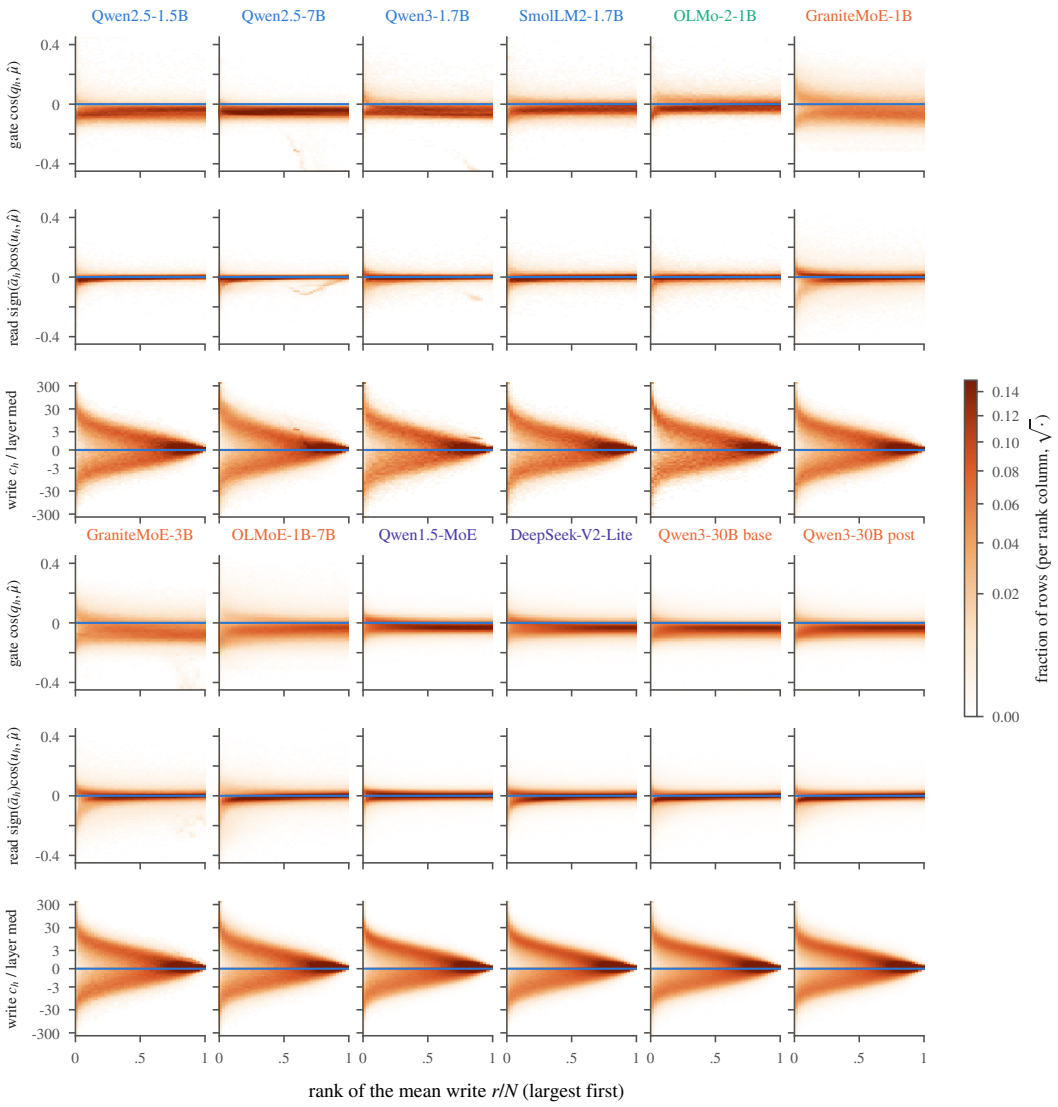

Figure 8: **Gate, read and write against linear rank, twelve checkpoints.** As Figure 1, linear rank, which gives the bulk of the population most of the axis: the whole population is shifted against the mean input and leans toward the output bias. The share of neurons writing toward b at each rank is in Figure 9.

MoE models against the layer's routing-weighted block mean b), leaves the top-1% medians of the gate and read alignment unchanged within 0.02 in all twelve checkpoints (Table 4); the largest shift, in OLMoE (gate $0.115 \to 0.091$, read $0.110 \to 0.087$), keeps the branch structure. The two rankings select the same neurons because write alignment is concentrated in the high-amplitude tail: the bulk of neurons has $|\cos(v_h, \hat{b})|$ at about twice the chance level $0.674/\sqrt{d}$, uncorrelated with amplitude, and the per-layer Spearman correlation between s_h and $|\cos(v_h, \hat{b})|$ is positive in every layer checked (medians $+0.12$ to $+0.19$). High amplitude is close to necessary, and far from sufficient, for writing along b.

E Fitting the distribution of mean-write sizes

One continuous distribution. The distribution of s_h over its layer median has the same form in all twelve checkpoints (Figure 11): a bulk around the median, a break at 2.9–4.7 times the median,

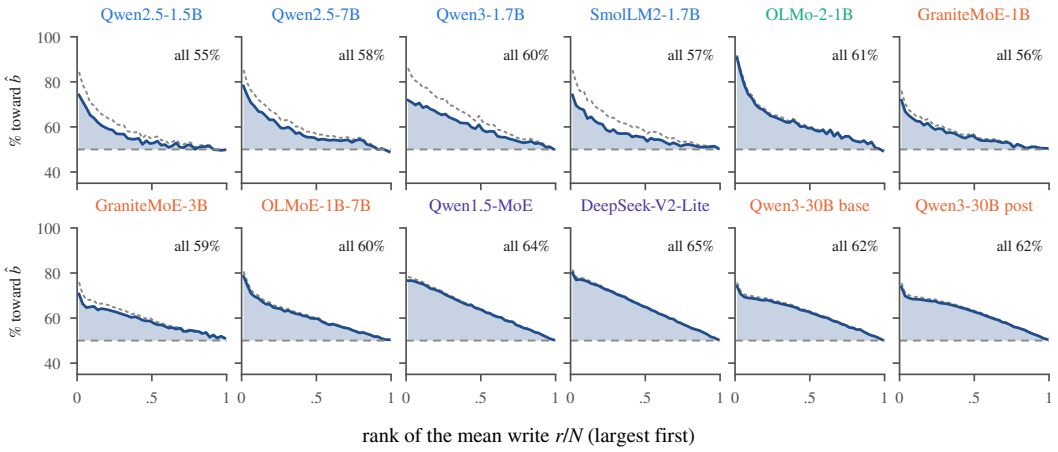

Figure 9: **Share of neurons writing toward the output bias, by rank.** Share of neurons with $\bar{a}_h\, v_h^\top (b - b_h) > 0$, the contribution scored against the output bias with the neuron's own term b_h removed (solid, shaded above one half), against the linear rank of s_h within its layer; dotted grey: the same share for $c_h = \bar{a}_h\, v_h^\top \hat{b}$, which includes the neuron's own term and so overstates the largest writers; dashed: one half. The number is the share over all neurons (leave-one-out).

Table 3: **The sign relation among the largest contributors.** Among the top 1% of neurons by s_h with $|\cos| > 0.05$ on both input vectors: share whose write at the mean input points along b, and ratio of consistent to violating neurons. Last column: share of the top 1% with $|\cos(u_h, \hat{\mu})| \le 0.05$.

model	consistent	ratio	read-neutral
Qwen2.5-1.5B	94.2%	16:1	52%
Qwen3-1.7B	90.3%	9:1	51%
SmolLM2-1.7B	87.8%	7:1	41%
OLMo-2-1B	98.4%	60:1	35%
Qwen2.5-7B	94.4%	17:1	61%
GraniteMoE-1B / 3B	92.7% / 93.2%	13:1 / 14:1	41% / 40%
OLMoE-1B-7B	93.0%	13:1	26%
Qwen1.5-MoE	88.4%	8:1	55%
DeepSeek-V2-Lite	89.7%	9:1	48%
Qwen3-30B-A3B base / post	95.0% / 96.1%	19:1 / 24:1	49% / 48%

and a power-law tail above it. A maximum-likelihood fit of the survival function $[1 + (x/x_c)^p]^{-\beta/p}$ gives $p = 1.11$–1.21 and $\beta = 3.0$–5.3, with Kolmogorov–Smirnov distance 0.005–0.020 (Table 5). The top 1% of neurons hold about 8% of the mass (15% in OLMo-2). The largest writers are the tail of this one distribution rather than a separate population: the far-tail excess above 30–100 times the median, the top $\sim$0.1%, sits mostly in the final layers and continues the same law. OLMo-2, whose output is normalized, is the exception: on top of the same law it has a localized packet of neurons centred at 237 times the median, the dedicated bias writers of Section 4.

Estimator. Fits are maximum-likelihood on the density, never regressions on a log–log plot. The broken power law with survival function $F(x) = [1 + (x/x_c)^p]^{-\beta/p}$ has density $f(x) = \beta x^{p-1} x_c^{-p} [1 + (x/x_c)^p]^{-\beta/p-1}$ on $x > 0$. We minimize the mean negative log-likelihood over $(\log x_c, \log p, \log \beta)$ by Nelder–Mead from two initializations, on a subsample of at most 200k points. The samples are s_h over its layer median, layers pooled after per-layer normalization, with 12.5% of layers trimmed at each end of the stack; without trimming the final layers dominate the extreme tail (Figure 12). Goodness of fit is the Kolmogorov–Smirnov distance on the fitted sample.

Tail-only fits. A tail-only maximum-likelihood fit with the lower cutoff chosen by KS scan prefers a power law over a lognormal on the same tail in 80 of 80 layers (four models), by raw likelihood ratio. The tail-only exponents ($\alpha - 1 \approx 2$–2.6) are shallower than the whole-distribution β: the

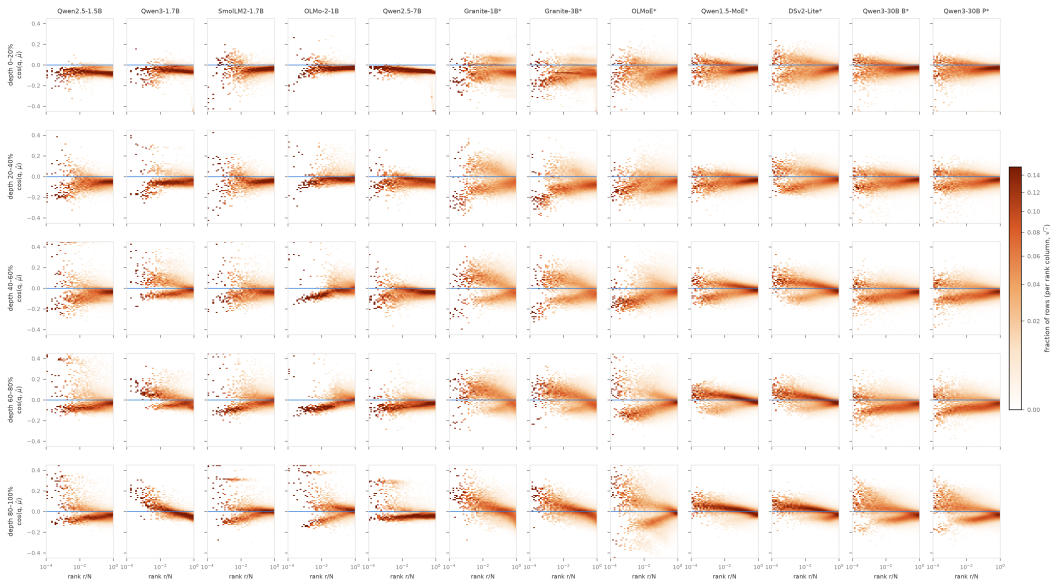

Figure 10: **Gate alignment against rank, by depth.** As the gate row of Figure 1: $\cos(q_h, \hat{\mu})$ of every neuron against its rank r/N by s_h within its layer (logarithmic, largest first; each rank column normalized), with the layers of each model split into five depth quintiles (rows). Dense models at the layer mean μ, MoE experts (*) at their own μ_e.

Table 4: **Two rankings select the same neurons.** Top-1% band medians (median over layers) of $|\cos(q_h, \hat{\mu})|$ and $|\cos(u_h, \hat{\mu})|$, with neurons ranked by s_h (main text) or by $s_h|\cos(v_h, \hat{b})|$.

	gate		read	
model	s_h	$s_h\|\cos\|$	s_h	$s_h\|\cos\|$
Qwen2.5-1.5B	0.088	0.092	0.040	0.031
Qwen3-1.7B	0.069	0.066	0.053	0.041
SmolLM2-1.7B	0.082	0.089	0.062	0.053
OLMo-2-1B	0.077	0.077	0.074	0.074
Qwen2.5-7B	0.057	0.069	0.029	0.033
GraniteMoE-1B	0.104	0.107	0.060	0.053
GraniteMoE-3B	0.099	0.101	0.068	0.058
OLMoE-1B-7B	0.115	0.091	0.110	0.087
Qwen1.5-MoE	0.052	0.045	0.047	0.039
DeepSeek-V2-Lite	0.065	0.053	0.048	0.043
Qwen3-30B-A3B base	0.066	0.065	0.053	0.046
Qwen3-30B-A3B post	0.066	0.068	0.056	0.045

interiors of pre-norm dense models carry a second, far-tail regime above the single-break law. A constrained mixture test (broken power law fixed, one lognormal component restricted to $x > 30$) finds a localized packet only in OLMo-2 (centre 237× the median, width 0.35 decades, weight 3.8×10^{-4}, $2\Delta\text{LL} = 340$). In the pre-norm models the fitted peak sits at the constraint boundary, which indicates a smooth second slope rather than a packet. The unconstrained six-parameter mixture is not identifiable, because the lognormal absorbs the bulk.

F Formation and explicit biases

Estimators. Per checkpoint and layer, on that checkpoint's own activations: $\hat{\mu}$ is the unit mean of the FFN input and $m = \mathbb{E}_t[z_t^\top \hat{\mu}]$. The share against the mean is the share of input vectors with $\cos(q_h, \mu) < 0$ (Figure 14); Figure 13 instead counts vectors beyond ± 0.02, so that the near-isotropic vectors at initialization fall on neither side. The power along the mean is $d\,\hat{\mu}^\top C \hat{\mu}$ with

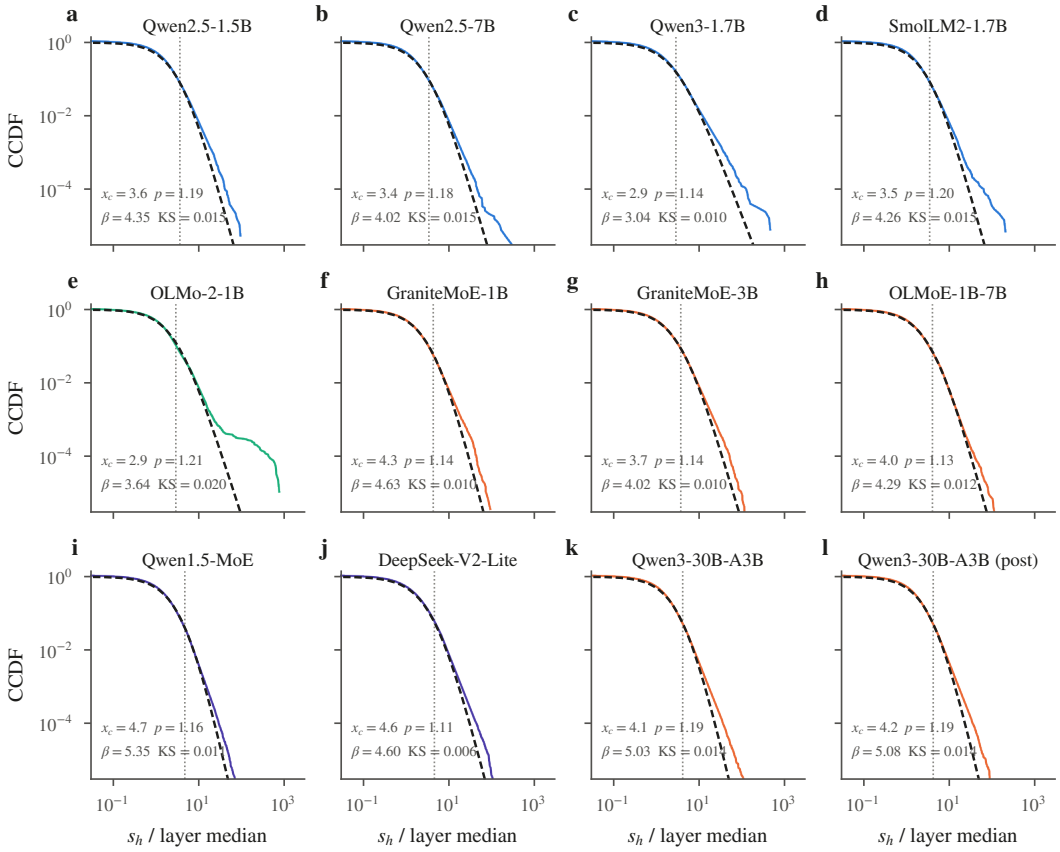

Figure 11: **Survival functions of s_h over its layer median, with the maximum-likelihood broken power law,** twelve checkpoints, interior layers. Parameters in Table 5.

Table 5: **Broken-power-law parameters** of s_h over its layer median, interior layers, twelve checkpoints (KS 0.005–0.020).

model	x_c	p	β
Qwen2.5-1.5B	3.6	1.19	4.35
Qwen3-1.7B	2.9	1.14	3.04
SmolLM2-1.7B	3.5	1.20	4.26
OLMo-2-1B	2.9	1.21	3.64
Qwen2.5-7B	3.4	1.18	4.02
GraniteMoE-1B	4.3	1.14	4.63
GraniteMoE-3B	3.7	1.14	4.02
OLMoE-1B-7B	4.0	1.13	4.29
Qwen1.5-MoE	4.7	1.16	5.35
DeepSeek-V2-Lite	4.6	1.11	4.60
Qwen3-30B-A3B base	4.1	1.19	5.03
Qwen3-30B-A3B post	4.2	1.19	5.08

$C = \mathrm{mean}_h\, \hat{q}_h \hat{q}_h^\top$, which is 1 for isotropic vectors. The implicit offset of a neuron is $(q_h^\top \hat{\mu})\, m$ and is compared with its explicit bias b_q.

Formation, and a model without gates. Pythia has non-gated GELU FFNs with explicit biases, so it is a contrast rather than a replicate; its training checkpoints show when the input vectors organize (Figure 13). At initialization the alignments are symmetric. Within about 10^3 steps the along side empties, from 10–31% of input vectors with $\cos > 0.02$ to none, and 0.99–1.00 point against $\hat{\mu}$; later the against share relaxes to 0.80–0.98 and a sliver of the along side returns (below

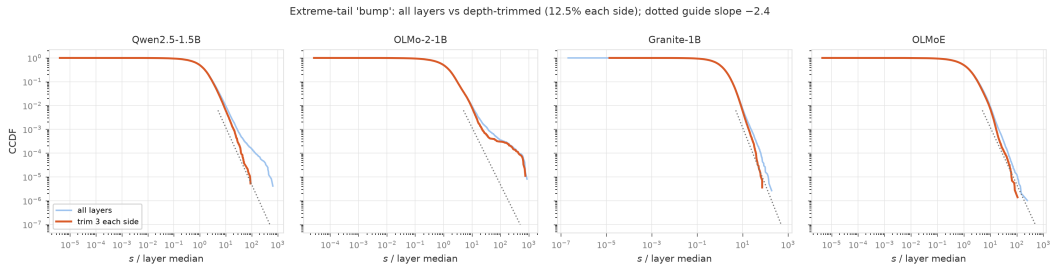

Figure 12: **Why the interior layers are used.** The extreme-tail excess comes from the final layers in three of the four models shown (82–100% of neurons with $s_h > 100\times$ the median lie there; trimming restores a single power law). In OLMo-2 the shoulder survives trimming: its packet of very large contributors is spread through depth.

1%). The polarization in Pythia is therefore one-sided, while gated models keep an along branch among their strongest writers. The mean direction itself keeps rotating, by 26–55 degrees between early checkpoints and still at the last one: the weights reorganize around μ long before μ settles. In controlled 20M-parameter bias-free SwiGLU MoE runs, the onset moves between 199 and 1245 steps with batch size, warm-up, learning rate and width ("Onset time" below).

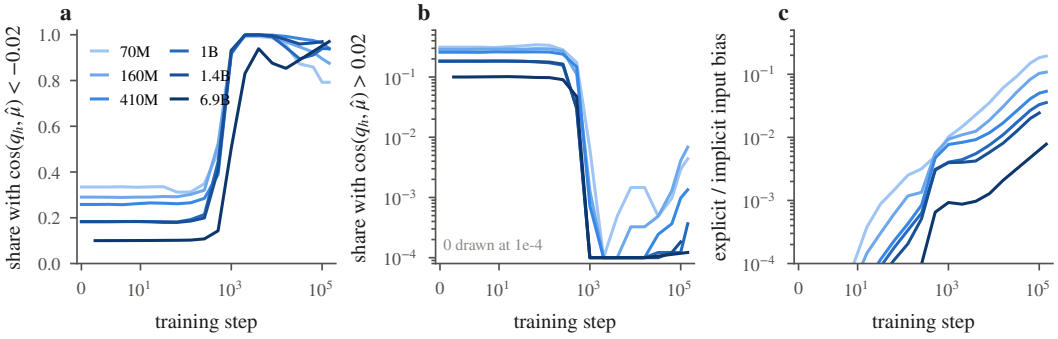

Figure 13: **Formation in Pythia (GELU FFN with explicit biases).** Median over layers (relative depth ≤ 0.85) of the share of input vectors with $\cos(q_h, \hat{\mu}) < -0.02$ (a) and > 0.02 (b) through training, one line per model size; step 0 is initialization. The bins are absolute, so the shares at initialization depend on the width. (c) Explicit input bias relative to the implicit one, $\mathrm{med}_h|b_{q,h}|/\mathrm{med}_h|q_h^\top \hat{\mu}\, m|$ with $m = \mathbb{E}_t[z_t^\top \hat{\mu}]$.

Explicit biases stay small and reinforce the implicit ones. Across eleven public models with explicit FFN biases and non-gated GELU FFNs (Figure 14), the implicit offset, not the explicit bias, determines when a neuron is active: predicting each neuron's fraction of tokens with a positive preactivation from the implicit offset gives mean absolute error 0.007–0.048, and from the explicit bias alone 0.23–0.39. The explicit biases are 0.8–19% of the implicit offsets, falling with scale in Pythia from 19% at 70M to 0.8% at 6.9B (Figure 13c), and they are positively correlated with them ($+0.44$ to $+0.79$, all 11 models). Their smallness is not shrinkage, since Pythia's biases are exempt from weight decay. It is consistent with a difference in leverage: a step of a weight vector along $\hat{\mu}$ moves the neuron's offset $\|\mathbb{E}z\|$ times as far as a step of the same size in the bias parameter, a factor of 15–31 in Pythia. A controlled SwiGLU run with explicit biases on both input projections behaves the same way: its biases stay at 2–5% of the implicit offsets and it is indistinguishable from its bias-free twin through 10% of training (one seed).

Controlled runs. A 20M-parameter SwiGLU MoE (8 MoE layers, 24,000 steps) trained without biases reaches a share against the mean of 0.62 and a power along it of 2.6. Its twin with explicit biases on both input projections, from the same initialization and data order, tracks it at every checkpoint through 10% of training, with explicit biases at 2–3% (step 1280) and 3–5% (step 2560) of the

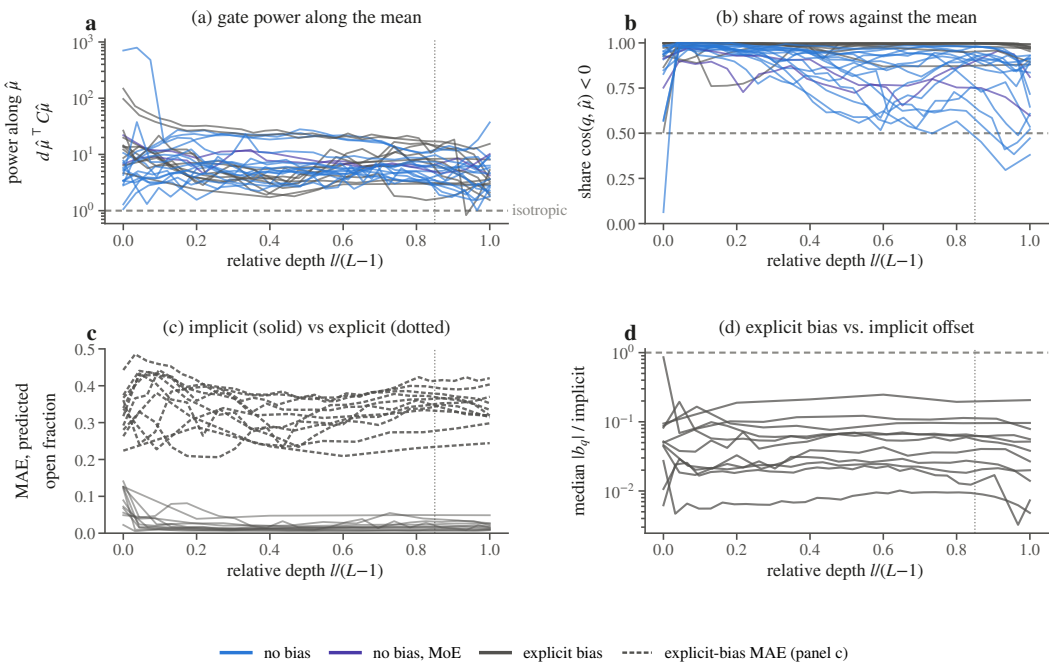

Figure 14: **Models with and without bias parameters.** Power of FFN input vectors along the mean input and share against it, through depth, for 11 public models with explicit FFN biases and 15 without; for the former, the fraction of tokens with positive preactivation predicted from the implicit offset and from the explicit bias, and the explicit bias relative to the implicit offset. In every LayerNorm model the mean FFN input is the mean of the data, not the LayerNorm bias parameter (cosine 0.92–0.99).

implicit offsets and loss 3.913 against 3.923 (one seed). In these runs the norm of the mean FFN input relative to $\mathbb{E}\|z\|$ is established within the first 40–80 steps, while the share against it crosses 0.60 at step 1280.

Onset time. In the same system we vary one factor at a time and record the onset t_c, the first step at which the pooled gate alignment $Z = -\mathbb{E}[r]/(\mathrm{sd}[r]/\sqrt{N})$, with $r = \cos(q_h, \hat{\mu})$ and N the number of gate vectors, crosses 5 (interpolated in log step). The onset moves with batch size (199–1174 steps), warm-up (256–859), peak learning rate (275–1245; not reached at a quarter of the base rate) and width (323–657). No step count is universal.

G MAKING THE IMPLICIT INPUT BIAS EXPLICIT

In layers 1 to $L/2$ we either delete the projection of the FFN input on $\hat{\mu}$ from both input projections (removal) or replace it by its discovery mean, a constant (replacement), and report the change in teacher-forced loss on 64 held-out windows (Table 6). Perturbations along random directions with matched spread change the loss by one to three orders of magnitude less. Replacement recovers at least 98.7% of the removal cost in the models of the main panel. OLMo-2 reads the raw residual stream, so its row is a different operating point. The Llama models are outside the panel; the comparison of Llama-2-7B with its pruned descendant Sheared-LLaMA, and of the Llama-3 8B models with their pruned and distilled descendants, shows that the failure of replacement in Llama-3 is a property of the family, not of pruning.

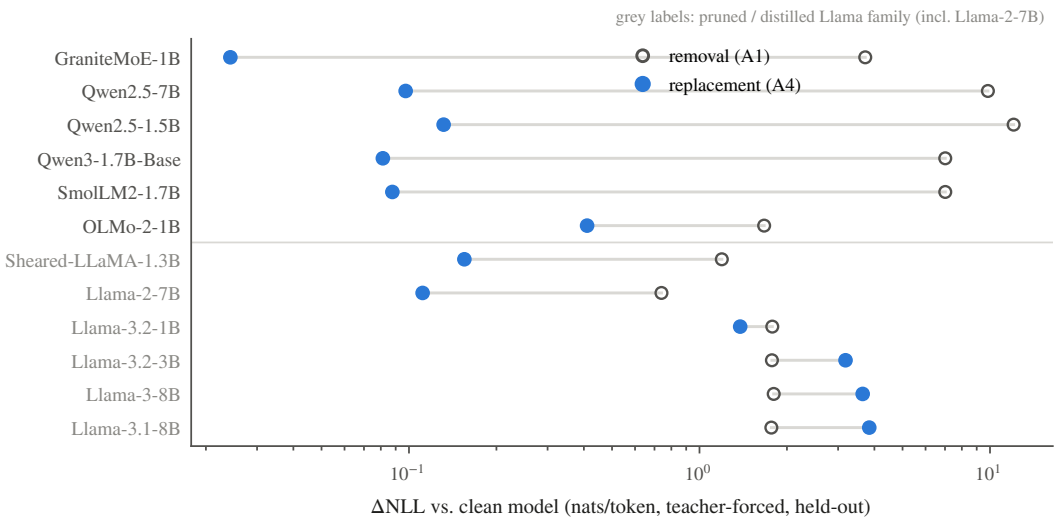

Figure 15: **Removal against replacement of the mean-input offset.** Increase in teacher-forced loss (nats per token, 64 held-out windows) when the projection of the FFN input on $\hat{\mu}$ is deleted in layers 1 to $L/2$ (open) or replaced by its discovery mean (filled). Grey: Llama family.

Table 6: **Removal and replacement of the mean-input offset.** Loss of the clean model (nats per token) and its increase under each intervention; ratio = replacement / removal.

model	clean	removal	replacement	ratio
GraniteMoE-1B	2.2213	3.7201	0.0243	0.0065
Qwen2.5-1.5B	2.5665	12.0600	0.1316	0.0109
Qwen2.5-7B	2.3509	9.8346	0.0974	0.0099
Qwen3-1.7B-Base	2.6006	7.0136	0.0813	0.0116
SmolLM2-1.7B	2.3794	7.0148	0.0877	0.0125
OLMo-2-1B	2.5708	1.6693	0.4103	0.2458
Sheared-LLaMA-1.3B	2.1954	1.1937	0.1551	0.1300
Llama-2-7B	1.9425	0.7398	0.1114	0.1506
Llama-3.2-1B	2.4841	1.7807	1.3807	0.7754
Llama-3.2-3B	2.3256	1.7756	3.1817	1.7919
Llama-3-8B	2.1785	1.8009	3.6451	2.0240
Llama-3.1-8B	2.1814	1.7675	3.8405	2.1728

H STATISTICS NOT ATTRIBUTABLE TO TRAINING

Three statistics that pass a random-direction control are reproduced by weight-shuffled or Gaussian networks on the same inputs, so we do not attribute them to training. A random-direction control tests whether an effect is specific to a direction, not whether training produced it.

Cancellation along other directions. For a direction w and a token t, let $r(t) = |\sum_h a_h(z_t)\, v_h^\top w| \,/\, \sum_h |a_h(z_t)\, v_h^\top w|$, compared with a null in which the signs of the terms are drawn at random. Along a direction fitted to minimize r and evaluated on held-out windows, r/r_{null} is 0.75, 0.88, 0.90, 0.94 and 0.97 in the five MoE models, and 0.82, 0.90, 1.01, 0.93 and 0.95 with shuffled weights: lower for the trained network in two models, higher in two, equal in one. Along $\hat{b}_e$ the ratio is 2.8–4.4 trained and 3.6–5.9 shuffled, larger for the untrained network in four of five models, because $\hat{b}_e$ is by construction the direction along which mean writes add. Random directions sit at the null in both.

The write along the output bias depends on the input. The Spearman correlation between $f_e(z_t)^\top \hat{b}_e$ and the angular departure $1-\cos(z_t, \mu_e)$, with $\|z_t\|$ partialled out, is -0.23, -0.27, -0.43,

-0.38 and -0.45 for trained experts and -0.52, -0.48, -0.44, -0.50 and -0.58 for shuffled ones. For a random direction it is $+0.01$. The dependence belongs to the mean output direction of any map on these inputs.

The gate is more sign-stable than the read. The cancellation ratio of the gate value exceeds that of the read value in shuffled networks as well (0.610, 0.583, 0.318, 0.439 and 0.438, against the trained 0.683, 0.643, 0.507, 0.485 and 0.410). Training widens the gap in three models, leaves it in one and narrows it in Qwen1.5-MoE.

I NORMALIZED FFN OUTPUTS

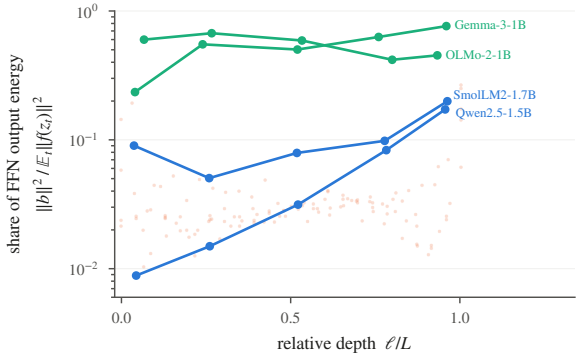

Figure 16: **Size of the output bias.** Share of FFN output energy carried by the output bias, $\|b\|^2/\mathbb{E}_t\|f(z_t)\|^2$, through depth: two pre-norm dense models (blue), OLMo-2-1B and Gemma-3-1B, which normalize the FFN output (green), and per-expert document-split estimates in five MoE models (orange dots).

For each model and layer (40k tokens, FFN recomputed offline; GeGLU for Gemma-3) we measure the energy share $\|b\|^2/\mathbb{E}\|f\|^2$; the scale ratio $\|b\|/\mathrm{rms}(f-b)$; the per-token write along the output bias, $c_t = f(z_t)^\top \hat{b}$, and its coefficient of variation; and the sensitivity $d\log(\text{median content gain})/d\log\lambda$ when the writes of the very large contributors ($s_h > 100\times$ the layer median) are scaled by $\lambda \in \{0.5, 2\}$. Such neurons exist at every probed layer of OLMo-2 and Gemma-3 and at 3 of 10 pre-norm layers. The sensitivity is -0.14 to -0.46 in OLMo-2 and -0.11 to -0.55 in Gemma-3, and c_t varies across tokens with coefficient of variation 0.19–0.54: under an output normalization the constant sets the scale of everything else the FFN writes. Whether later computation uses this modulation cannot be established from activations alone.

J CHANNELS OF THE MEAN INPUT AND OF THE OUTPUT BIAS

Which channels the output bias occupies. Measured by the share of energy in the ten largest channels, the output bias of a dense block is less concentrated than the mean input, 41–67% against 84–95% (median over layers), and at some mid-depth layers it spreads over 50–400 effective channels while the mean input never does (Figures 17 and 18).

K THE MEAN INPUT IS NOT A MASSIVE ACTIVATION

We take the largest absolute value per layer relative to the layer median, at the raw residual stream and at the normalized FFN input, and apply the $100\times$ criterion of Sun et al. (2024). At the FFN input it is met in no layer of any MoE model, including all 48 layers of Qwen3-30B-A3B, and only at isolated layers of some dense models. Where it is met at the FFN input of a dense model, the crossing traces to a single channel that exceeds the threshold on 40–96% of tokens, which is mean structure, not a rare event. Deleting every token that contains a crossing leaves μ at cosine ≥ 0.97 to the original in the worst case, where this removes 86% of tokens, and ≥ 0.99 typically; trimming the 1% of tokens

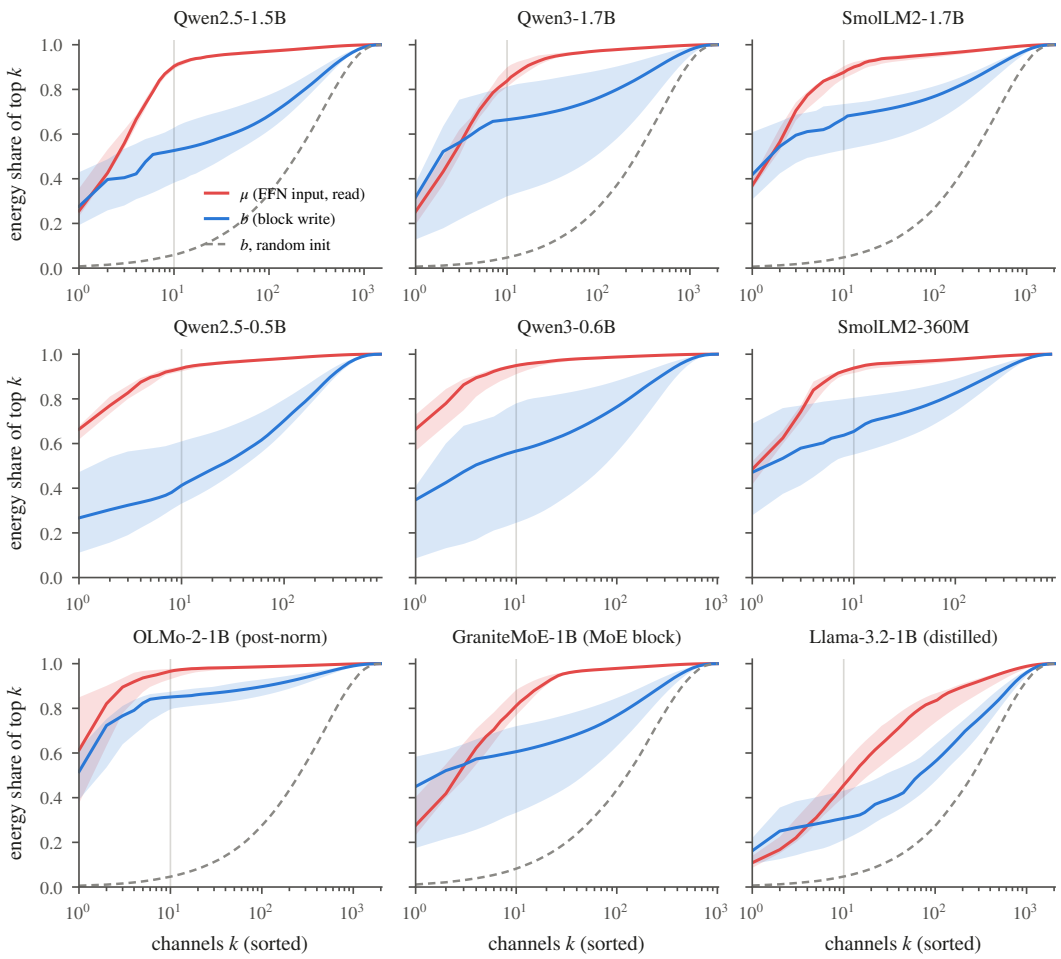

Figure 17: **Channel energy of the mean input and of the output bias, dense models and the GraniteMoE-1B block.** Share of energy in the k largest channels, $e_j = x_j^2/\|x\|^2$ sorted, for the mean FFN input μ_z (red) and the block's mean write $b = \mu_h[\ell+1] - \mu_y[\ell]$ (blue); median and interquartile range over layers. Grey dashed: b at random initialization. Packed C4 windows of 4096 tokens, $t \geq 64$.

with the largest norm leaves it at 1.0000. The threshold is not what drives this: lowering it to 10–30 times the median flags values of which 85–98% lie on persistent channels (flagged on more than 5% of tokens), and μ without all sparsely flagged tokens stays at cosine ≥ 0.994 at every threshold. In Llama-2-7B, where massive activations were first described, one of the two known massive channels is not among the carriers of the mean once its rare extreme values are masked, while all eight outlier channels found by the criterion of Dettmers et al. (2022) are among the ten largest channels of μ. The attention sink is a different vector again: its cosine with μ is 0.13–0.29 at the raw residual of 1.5B-parameter models, and the two are unrelated or opposed at the FFN input.

L THE OUTPUT BIAS AND THE WRITE AT THE MEAN INPUT

The output bias is not the write at the mean input. Averaging $f(z) = f(\mu) + J(\mu)(z-\mu) + R(z)$ over tokens removes the linear term exactly, so $b - f(\mu) = \mathbb{E}_t[R(z_t)]$. The median cosine between b and $f(\mu)$ is 0.63–0.98 in the dense models and 0.33–0.54 for experts at μ_e. The gap comes from neurons whose gate changes sign across tokens: restricted to neurons whose gate is positive on under 2% or over 98% of tokens the cosine is 0.89–1.00, and on 10–90% of tokens it is 0.18–0.67, against size-matched random subsets of neurons.

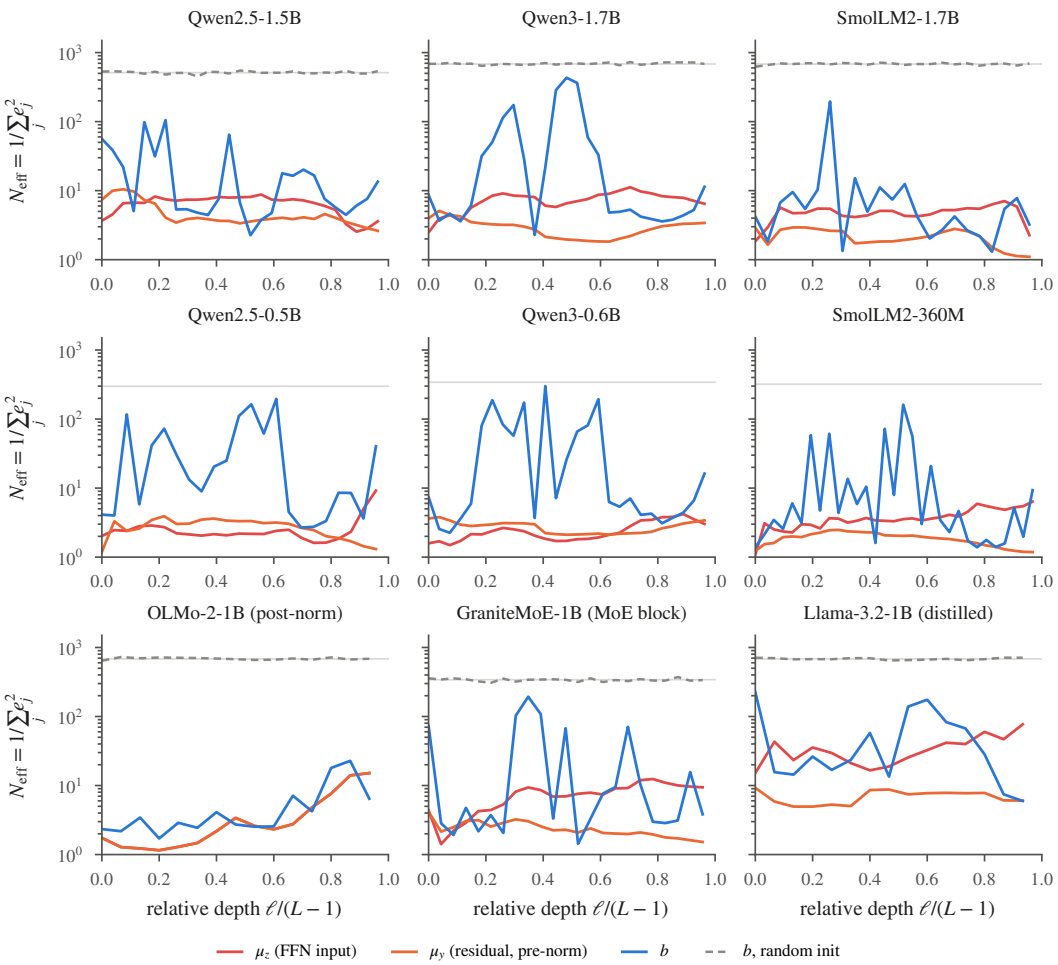

Figure 18: **Effective number of channels through depth,** $N_{\text{eff}} = 1/\sum_j e_j^2$, for the mean FFN input μ_z, the mean residual entering the FFN μ_y, and the output bias b; grey: b at random initialization; light rule: $d/3$, the value for a Gaussian vector.

The ratio $\|b\|/\|f(\mu)\|$ is 0.48–1.01 in the dense models and 0.73–0.87 for routed experts, but 1.14 and 1.41 in the routed experts of the two models with a shared expert. Both the response at the mean input and the output bias are therefore reported for MoE models (Figure 21a).

M THE WRITE AT THE MEAN INPUT AND THE AVERAGE WRITE

For every layer we compare the block evaluated at the mean input with the average write, $r = f(\mu)^\top \hat{b}/\|b\|$, where $b = \mu_h[\ell+1] - \mu_y[\ell]$ is the change the block makes to the mean of the residual stream on settled tokens ($t \geq 64$) and f is the whole block, including the router for the MoE model and the output normalization for OLMo-2 (Figure 19). $r > 1$ means that the mean input alone writes more along b than the average token does. Per token, we compare the write along b with that of the mean component of the token's own input, $f((z_t^\top \hat{\mu})\hat{\mu})^\top \hat{b}$, a paired contrast in which each token is its own control (Figure 20). Both are measured on held-out windows and repeated at random initialization. The measurements are observational: we have not clamped the mean component and measured the effect.

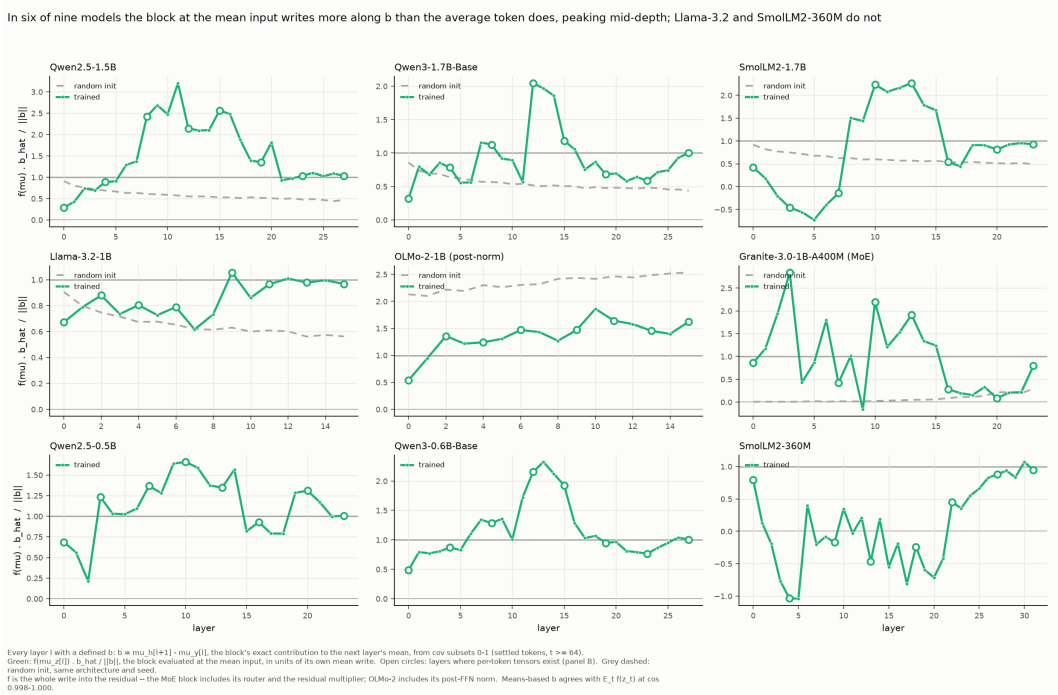

Figure 19: **The write at the mean input against the average write, every layer.** $r = f(\mu)^\top \hat{b}/\|b\|$ with $b = \mu_h[\ell+1] - \mu_y[\ell]$ (settled tokens, $t \geq 64$) and f the whole block (for the MoE model including its router; for OLMo-2 its output normalization); grey dashed: random initialization.

N LEAVE-ONE-OUT ALIGNMENT: ESTIMATOR AND CONTROLS

The experts' output biases are aligned. We score each expert's unit vector against the normalized sum of the other $E-1$ experts' unit vectors (leave-one-out), so that no expert is scored against a direction it helped fit. For the output biases b_e the median over layers of the median leave-one-out cosine is 0.27–0.37 in the five models whose experts each see many tokens, and 0.19–0.20 in the two Qwen3-30B checkpoints, whose 128 experts each see few; the two models with a shared expert are the highest (0.36 and 0.37). For the responses $f_e(\mu)$ it is 0.35–0.46 in the five models without a shared expert and 0.16–0.19 in the two with one. Both are positive at every layer of every model (layer minima 0.08–0.26 for b_e), against an isotropic null whose 95th percentile is 0.004–0.017 (Figure 21a). The agreement belongs to the trained experts at their actual input. In the three models without a shared expert, replacing μ by a random vector of the same norm lowers the mean pairwise cosine η of the responses from 0.15–0.26 to 0.010, permuting the coordinates of μ gives 0.010–0.011, and norm-matched random experts evaluated at μ give at most 0.0005 in absolute value (Section N.1). The agreement is carried by the largest contributors. Scoring each expert's output bias against the direction of the summed output biases, removing neurons in order of $|c_h|$ halves the agreement after 5–8% of a layer's neurons in the three models without a shared expert and after 13–14% in the two with one, while removing half of the neurons at random lowers it by only 17–26% (Section N.2).

Aligned biases add coherently. A single expert's output is dominated by its input-dependent part: the per-token cosine between $f_e(z_t)$ and $\hat{b}_e$ has a median of 0.15–0.18. But the residual stream receives the weighted sum over the selected experts, and aligned constants add coherently while input-dependent parts add closer to independently. In GraniteMoE-1B, GraniteMoE-3B and OLMoE the per-token cosine of the block output to the shared direction of the output biases is 0.18–0.21, against 0.09–0.11 for single experts, a gain of 1.9–2.1 where $k = 8$ independent parts would give $\sqrt{8} = 2.83$. One fixed direction carries 4–6% of the block's output energy, and 33%, 25% and 15% at the final layer, against 0.05–0.1% for a random direction (Section P).

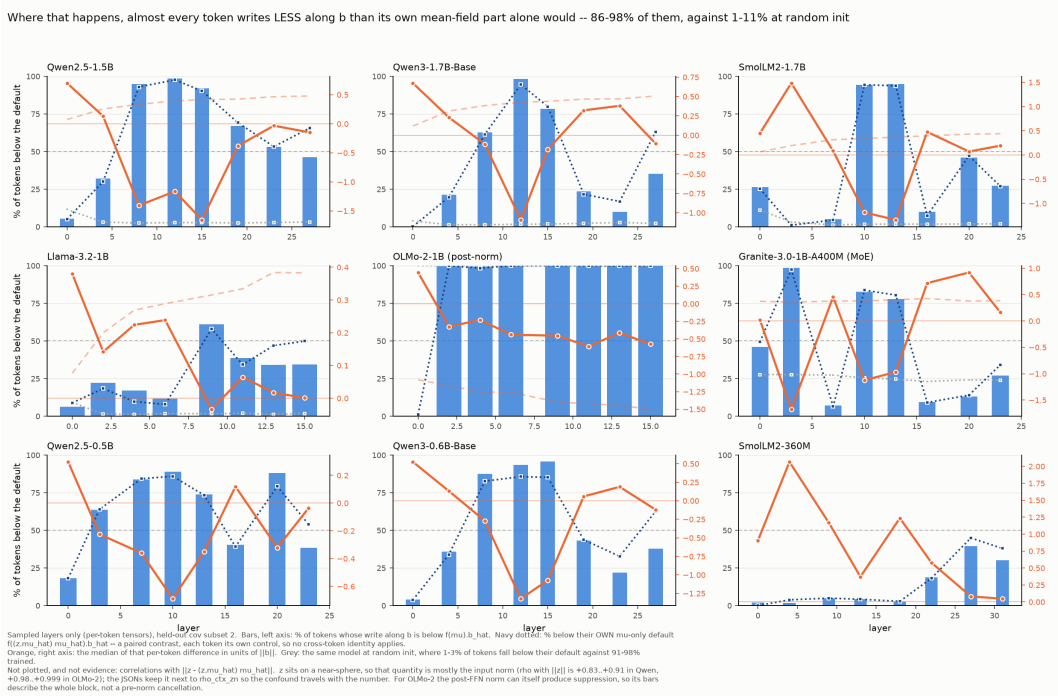

Figure 20: **Per-token view at the sampled layers.** Bars: share of held-out tokens whose write along b is below $f(\mu)^\top \hat{b}$; dotted: share below the write of the mean component of their own input, $f((z_t^\top \hat{\mu})\hat{\mu})^\top \hat{b}$, a paired contrast in which each token is its own control; grey: random initialization.

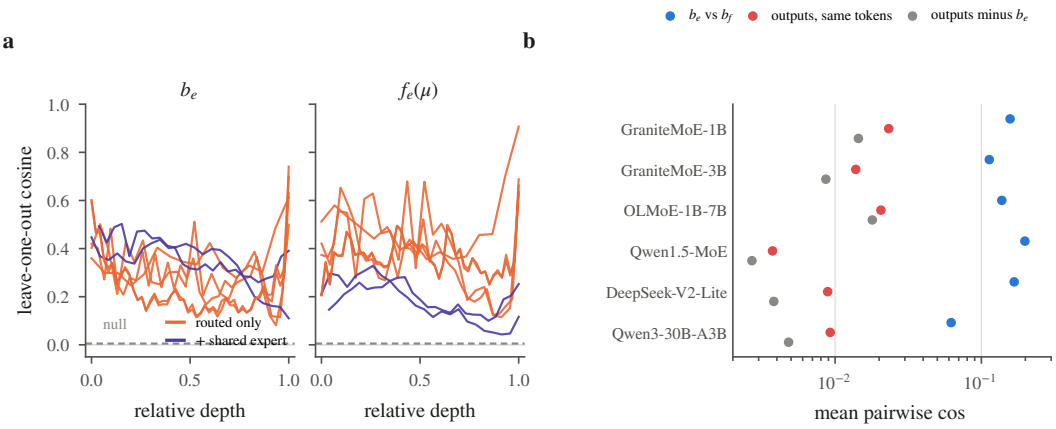

Figure 21: **Experts share their output bias, and averaged similarity measures it.** (a) Leave-one-out cosine of each expert's output bias b_e, and of its response at the mean input $f_e(\mu)$, to the normalized sum of the other experts' unit vectors; median over experts, through depth, seven checkpoints (orange: routed experts only; violet: with a shared expert). Dashed: 95th percentile of an isotropic null. (b) Mean pairwise cosine between experts: between their output biases, between their outputs on the same tokens, and between those outputs after each expert's b_e is subtracted (six checkpoints; log axis).

Averaged against token-level similarity. The mean pairwise cosine between output biases is 0.06–0.20 in six checkpoints. The same experts, compared as functions on a common sample of tokens, have mean per-token cosine 0.004–0.023, and 0.003–0.018 after each expert's b_e is subtracted (Figure 21b): the output bias holds a few percent of an expert's output energy but 13–57% of its token-level similarity to other experts.

Estimator. For the output biases $\{b_e\}_{e=1}^{E}$ of one MoE layer we normalize each expert's vector first, $\hat{b}_e = b_e/\|b_e\|$, so that a single high-norm expert cannot set the population direction, and score each expert against a direction it did not influence:

$$\mathrm{LOO}(e) = \Big\langle \hat{b}_e,\ \mathrm{normalize}\Big(\tfrac{1}{E-1}\textstyle\sum_{e'\neq e}\hat{b}_{e'}\Big)\Big\rangle. \tag{6}$$

The plug-in score $\langle \hat{b}_e, v\rangle$ with v fitted on all E experts is circular and inflates: across the seven models of the alignment evaluation the plug-in median exceeds the leave-one-out median by 0.005 to 0.083, and the gap is largest on the weakest model (Qwen1.5-MoE layer 23, $+0.193$ against $+0.110$). Experts the router never selects have a mean of exactly zero; they are dropped and counted, which in this evaluation affects only GraniteMoE-3B (2 of 40 experts at layer 0, 11 of 40 at layer 1).

A second, related statistic is the mean off-diagonal cosine between experts' normalized vectors, taken pairwise rather than against a leave-one-out centroid,

$$\eta = \mathbb{E}_{e\neq f}\big\langle \hat{b}_e, \hat{b}_f\big\rangle. \tag{7}$$

η and $\mathrm{LOO}(e)$ answer different questions — a population summary against a per-expert score — and both are used below; η is also defined identically for the responses at the mean input, $\hat{f}_e(\mu) = f_e(\mu)/\|f_e(\mu)\|$, by the same formula.

Why not the leading principal component. The top principal axis of the remaining experts' vectors can replace the centroid, but its sign is set by convention and becomes arbitrary when the top two singular values are close: at GraniteMoE-1B layer 16 ($\eta = +0.060$, eigengap $s_1/s_2 = 1.09$) it reports a median of -0.202 with no expert positive. The centroid's sign is fixed by the vectors it averages, so we use it throughout; the two forms agree to better than 0.03 wherever $s_1/s_2 > 1.5$.

The output-bias direction and the mean-input-response direction are not one object. Figure 21a measures the leave-one-out cosine of b_e and of $f_e(\mu)$, with the same estimator, on the same shard, over the same settled positions, so the only difference between the two series is which vector is scored. Their depth profiles correlate at only $r = +0.46$, $+0.53$, $+0.75$ and $+0.65$ in the four models with stored activations that lack a shared expert. Their ordering is not consistent either: the median over layers is higher for the $f_e(\mu)$ direction in GraniteMoE-1B (0.415 against 0.294), GraniteMoE-3B (0.341 against 0.272) and OLMoE (0.456 against 0.344), and lower in Qwen1.5-MoE (0.151 against 0.383). Reporting either quantity without naming it would change which model looks weakest.

Direction stability. The population directions used above are fitted on discovery documents; evaluation documents re-accumulate them independently (Section 2). Across the same seven models the discovery-to-evaluation cosine of the output-bias direction runs from 0.917 to 0.998, and the discovery half-split cosine from 0.884 to 0.997. The direction is therefore a stable property of the model on this corpus and not an artifact of which discovery documents it was fitted on.

N.1 What the agreement is specific to

The agreement of the experts' responses at the mean input is a property of the trained experts evaluated at the particular geometry of μ: no arbitrary input and no expert-shaped random weights produce it (Section 5). This subsection reports the contrasts that could have falsified that claim.

The four contrasts. All four hold the estimator, the layer set and the expert population fixed and change exactly one thing.

- **Random input, matched norm.** Replace μ by a random vector drawn isotropically and rescaled to $\|\mu\|$. This asks whether any input of the right size makes experts agree.
- **Permuted μ.** Replace μ by $P\mu$ for a random coordinate permutation P. This preserves the norm, the full marginal distribution of coordinate values, and the sparsity pattern; it destroys only which coordinate carries which value. It is the sharpest of the four, because it isolates geometry from every scalar summary of the input.
- **Norm-matched random experts.** Keep μ and replace the trained experts with random maps whose gate, up and down matrices match the trained ones row- and column-wise in norm. This asks whether the architecture alone produces the alignment.

- **Single real tokens.** Replace μ by an individual settled activation z_t and average η over draws of t. This asks whether the mean is special among points the model actually visits, or merely a representative one.

Table 7: Cross-expert alignment η of the responses at the mean input under the four contrasts, averaged over all MoE layers of each model. The last column is the cosine between the population direction fitted at μ and the one fitted at a random input of the same norm. DeepSeek-V2-Lite is absent because the required per-layer responses were not retained for it.

	$f_e(\mu)$	random input	permuted μ	single tokens	random experts
GraniteMoE-1B	0.197	0.010	0.010	0.047	0.0005
GraniteMoE-3B	0.147	0.010	0.011	0.035	−0.0002
OLMoE	0.255	0.010	0.011	0.049	0.0001
Qwen1.5-MoE	0.050	0.009	0.009	0.006	0.0000

Reading the table. In the three routed-only models the mean-input η is 15 to 26 times the random-input value and 13 to 23 times the permuted-μ value, against a random-expert floor three orders of magnitude below the effect. Two conclusions follow, and they are different conclusions.

First, the alignment is not architectural: norm-matched random experts evaluated at the true μ produce $|\eta| \leq 0.0005$. Whatever produces the agreement is in the trained weights.

Second, and less obviously, it is not a property of those weights alone. The permutation control preserves every scalar statistic of μ we might otherwise credit — its norm, its coordinate marginal, its kurtosis, the identity of its largest values as a multiset — and moves η to the random-input floor. The population directions fitted at μ and at a matched random input are weakly anti-aligned (cosine -0.20, -0.27, -0.30 and -0.28), so the random-input responses are not a noisy version of the real ones. The agreement is a joint property of the experts and the geometry of the point at which they are evaluated.

The mean is special among real inputs. The fourth contrast justifies evaluating experts at μ rather than at an arbitrary token. Single real tokens give η of 0.035–0.049 in the routed-only models — four to five times below the mean of the very distribution they are drawn from. The shared component is therefore not something visible at an arbitrary point on the manifold that μ happens to inherit; it is concentrated at the centre.

Scope. These contrasts establish what the agreement at the mean input is specific to. They do not identify the mechanism. A common μ sets a common operating point for each expert's gate nonlinearity, the experts' write matrices have their own geometry, and the input covariance is concentrated on a minority of coordinates; any of the three could carry the effect, and separating them requires interventions we do not run here. Nor do these contrasts speak to whether the shared component is functional, which we do not test here. Finally, the experts are trained jointly under a load-balancing objective, so none of this is evidence that independently trained experts would converge on a common direction.

N.2 Which neurons carry the agreement

Scoring each expert's output bias against the direction of the summed output biases, removing neurons in order of $|c_h|$ halves the agreement after 5–8% of a layer's neurons in the three models without a shared expert and after 13–14% in the two with one, while removing half of the neurons at random lowers it by only 17–26% (Figure 22).

O How much of an expert's write is constant

The fraction of an expert's write energy that survives averaging over its routed tokens is $\mathrm{dc} = \|\mathbb{E}_t f_e(z_t)\|^2 / \mathbb{E}_t \|f_e(z_t)\|^2$ (Section 4). The plug-in $\|\bar{f}\|^2$ is positively biased by $\mathrm{tr}(\mathrm{Cov})/N_{\mathrm{eff}}$, and the correction for independent samples is wrong here because tokens are clustered in windows and

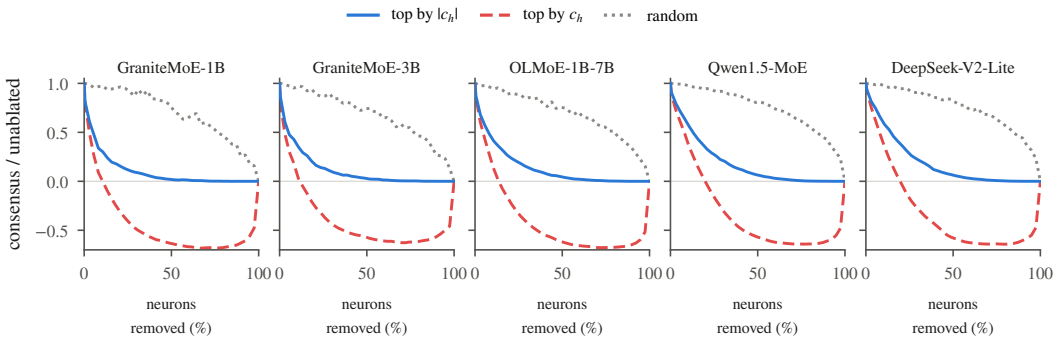

Figure 22: **A minority of neurons carries the agreement.** Cross-expert agreement of the output biases, relative to its unablated value, as neurons are removed in order of $|c_h|$, in order of signed c_h (removing only positive contributors overshoots) or at random; and the cumulative signed contribution.

windows in documents. The cross-product $\langle \bar{f}_A, \bar{f}_B \rangle$ over document-disjoint halves is unbiased for $\|\mathbb{E}_t f\|^2$ under any within-document dependence. On synthetic clustered data with no constant part the plug-in reads 0.025 and the document split -0.004. The estimator can therefore return a negative value, the correct report that no constant part is resolvable at this sample size; it is never clipped. In the five MoE models with stored activations the median over experts, averaged over layers, is $\mathrm{dc} = 0.027$–0.043 (non-positive for two expert-layers of GraniteMoE-3B and nowhere else). In amplitude, 3% of the energy is 17% of the norm.

P WHY THE BLOCK OUTPUT IS MORE ALIGNED THAN ANY EXPERT

A single expert's write puts only a few percent of its energy on the shared direction of the output biases, while the block's write puts up to tens of percent there (Section 5). This section derives the relation between the two from how the selected experts' outputs combine.

Setup. At one MoE layer, write the response of expert e to the block input z_t as its component along a fixed unit direction v plus the remainder,

$$f_e(z_t) \;=\; a_{et}\, v \;+\; p_{et}, \qquad a_{et} = \langle f_e(z_t), v \rangle, \qquad p_{et} \perp v.$$

The block emits the gate-weighted combination over the selected experts, $y_t = \sum_e g_{et} f_e(z_t)$, with g_{et} the realized routing weight (post top-k renormalization where the family applies it). The quantity reported for both objects is the per-token energy fraction on v,

$$\rho_v(x) \;=\; \frac{\langle x, v \rangle^2}{\|x\|^2},$$

evaluated at $x = f_e(z_t)$ for a single expert and at $x = y_t$ for the block. Chance is $1/d$ for either.

An exact identity. Let $\mathbb{E}_g[\,\cdot\,]$ denote the gate-weighted average over the selected experts of one token, $\mathbb{E}_g[x] = \sum_e g_e x_e / \sum_e g_e$. Because $y = (\sum_e g_e)\, \mathbb{E}_g[f]$ and ρ_v is scale-free,

$$\rho_v(y) \;=\; \frac{\mathbb{E}_g[a]^2}{\|\mathbb{E}_g[f]\|^2} \;=\; \frac{\mathbb{E}_g[a]^2}{\mathbb{E}_g[a]^2 + \|\mathbb{E}_g[p]\|^2}.$$

Define two dimensionless quantities, both in $[0, 1]$:

$$C \;=\; \frac{\mathbb{E}_g[a]^2}{\mathbb{E}_g[a^2]}, \qquad R \;=\; \frac{\|\mathbb{E}_g[p]\|^2}{\mathbb{E}_g\big[\|p\|^2\big]}.$$

C measures how *coherent* the experts' projections onto v are: it is 1 when every selected expert projects onto v with the same signed magnitude, and it falls to $1/K$ when the projections are unrelated, where

$$K \;=\; \frac{\big(\sum_e g_e\big)^2}{\sum_e g_e^2}$$

is the participation ratio of the gate vector — the effective number of experts the token actually uses, which equals k for a uniform gate and less for a peaked one, independent of whether the family renormalizes. R measures how much of the off-v energy *survives* averaging: it is 1 if every expert's remainder is the same vector, and $1/K$ if the remainders are mutually orthogonal with equal norms.

Writing $\bar{\rho} = \mathbb{E}_g[a^2]/\mathbb{E}_g[\|f\|^2]$ for the gate-weighted mean of the per-expert shares, the identity becomes

$$\rho_v(y) \;=\; \frac{C\,\bar{\rho}}{C\,\bar{\rho} + R\,(1-\bar{\rho})}. \tag{8}$$

Nothing has been assumed. The amplification of the block's share over a typical expert's is governed entirely by the ratio C/R: alignment survives averaging to the extent the experts agree on v, and the rest cancels to the extent they disagree elsewhere.

The idealized case. If every selected expert had the same energy and the same share $\bar{\rho}$, projected onto v with the same sign and magnitude ($C = 1$), and had mutually orthogonal remainders ($R = 1/K$), then Equation (8) would reduce to

$$\rho_v(y) \;=\; \frac{K\bar{\rho}}{K\bar{\rho} + 1 - \bar{\rho}}, \qquad \text{and} \qquad \rho_v(y) \;\approx\; K\bar{\rho} \;\text{ for } K\bar{\rho} \ll 1.$$

This is the useful mental model: a K-fold amplification of the per-expert share. It is exact at $K = 1$, where it returns $\rho_v(y) = \bar{\rho}$. The mechanism is that coherent components add in amplitude and so grow as K^2 in energy, while incoherent ones add in energy and grow as K; the ratio of those two rates is the whole effect.

What is measured. On GraniteMoE-1B, over 57,344 evaluation tokens per layer, the per-token medians across the 24 MoE layers are

$$C = 0.506, \qquad 1/K = 0.147, \qquad R = 0.174, \qquad C/R = 2.88.$$

Both terms contribute, and neither is at its idealized value. The experts' projections onto the shared direction are about 3.4 times more coherent than unrelated projections would be, but only half as coherent as identical ones ($C = 0.506$ against a floor of 0.147 and a ceiling of 1). Their off-v remainders cancel almost as completely as mutually orthogonal vectors would: $R = 0.174$ against an independence value of $1/K = 0.147$. So the block's write is more aligned than any expert's for both available reasons at once, and the larger of the two is the cancellation of what the experts do *not* share.

The idealized K-fold form is accordingly a description and not a fit. Using the measured K per layer it predicts the aggregate share to within a factor whose median is 1.4 across layers, with the agreement best where the effect is largest: at the final layer of GraniteMoE-1B it predicts 0.349 against a measured 0.352. Using the nominal $k = 8$ in place of the measured $K \in [5.1, 7.3]$ happens to fit slightly better, which is a coincidence of two errors cancelling — $K < k$ pulls the prediction down, and the positive correlation between a gate weight and its expert's alignment, which C absorbs but the idealization does not, pulls it up.

Reading the two numbers. At a mid-depth layer of GraniteMoE-1B, a single expert's write puts $\rho = 0.006$ of its energy on the shared direction while the block's write puts 0.072 there, and at the final layer the same pair is 0.063 and 0.352. Neither of the two obvious summaries is correct: experts do not mostly write a shared direction, and the block's write is not mostly idiosyncratic. What is true is that a small, sign-coherent component of each expert's write survives combination while the large idiosyncratic remainder cancels, and Equation (8) says exactly how much of each effect is present.

Assumptions and diagnostics. The identity is exact; the idealization is not, and each of its assumptions has a measurement attached. Sign coherence is the load-bearing one — with random signs the numerator of Equation (8) would add incoherently too and there would be no amplification at all — and it is what C reports and what the sign-identified centroid (Section N) makes well defined; a diagnostic tracking the fraction of experts whose projection onto v shares the population's sign runs 0.81 to 0.95 on the layers checked. Independence of the remainders is what R reports, and the measured R lies above $1/K$ at every layer, so the remainders are close to mutually orthogonal and very slightly positively correlated; at the final layer $R = 0.326$ against $1/K = 0.145$, i.e. the experts

share more of their off-v content there than anywhere else. Gate concentration enters only through K, which is measured directly.

One caveat is architectural. In a shared-expert family the emitted y_t includes the always-on branch, which is not one of the terms in $\sum_e g_e f_e$ and which contributes a nearly token-independent vector; the identity still holds for the routed aggregate, but ρ_v measured on the block output of those families mixes the two. The routed and shared halves are separated by gating the block output three ways and capturing each component directly; the identity above should be applied, for those two families, to the routed aggregate this separation returns.

Provenance. The per-expert shares and the aggregate second moments come from a single instrumented pass, which accumulates $\sum_t y_t y_t^\top / \|y_t\|^2$ exactly, so any direction fitted later can be projected without a second pass. The gates and per-pair projections come from a separate pass over the same tokens. The two instruments compute $\rho_v(y)$ independently and agree to 3×10^{-4} at every layer, which is the cross-validation this section rests on.

Q THE ALIGNMENT IS NOT CARRIED BY A FEW CHANNELS

Shared high-energy channels can dominate cosines (Section 2.1). We therefore delete the highest-energy channels and recompute the cross-expert alignment of the output biases on the remaining ones.

Estimator. Rank the d coordinates by their energy in the routed population,

$$\sigma_j = \sqrt{\mathbb{E}_e\big[b_{ej}^2\big]},$$

delete the k largest, and recompute the same statistic on what remains,

$$\eta(k) = \mathbb{E}_{e \neq f}\big\langle u_e^{(k)},\, u_f^{(k)} \big\rangle, \qquad u_e^{(k)} = \frac{b_e|_{S_k}}{\big\|b_e|_{S_k}\big\|},$$

where S_k is the set of $d - k$ surviving coordinates and each expert's vector is renormalized inside it; $\eta(0)$ is the pairwise alignment of Equation (7). This is the all-but-the-top ablation (Mu & Viswanath, 2018) applied to coordinates.

Null. Chance for a cosine is $1/\sqrt{d}$ only in an isotropic space, and these spaces are not isotropic. We therefore draw null response matrices with each expert's vector independently Gaussian, $\mathcal{N}(0, \mathrm{diag}(\sigma_j^2))$, giving them the layer's own coordinate anisotropy and destroying only the cross-expert agreement, and put each draw through the identical pipeline including the ranking step, so that whatever selection effect the ranking carries is inside the null. We report the 95th percentile over 1000 draws. This null is per model *and* per k: it spans 0.0005 to 0.027 across the seven models, and within a model it falls to a minimum — by $2.5\times$ for Qwen1.5-MoE and $12.4\times$ for Qwen3-30B-Base, so not by one common factor — and then rises again. It falls because deleting the high-energy coordinates leaves a more isotropic remainder and chance agreement tightens as effective dimension rises; it turns back up once so few coordinates remain that small-sample noise dominates. A single flat chance line would misrepresent every model, so Figure 23 draws each model's own.

The ranking is in-sample, and its bias was measured. Since $\sigma_j^2 = (\mathbb{E}_e b_{ej})^2 + \mathrm{Var}_e\, b_{ej}$, ranking on the same response matrix partly ranks by the cross-expert mean that η is built from, which biases the surviving η downward. The second term pulls the other way: deleting high-energy coordinates also deletes high-variance zero-mean ones, and removing noise coordinates raises the surviving η. Which effect dominates depends on the joint distribution of coordinate means and variances and cannot be settled from the sign of one term, so we measured it at matched sample size. Splitting the experts in half and ranking one half's coordinates either by its own energy or by the other half's, then scoring η on the same experts either way, the in-sample ranking lands at or below the cross-fitted one at every model and every k, with the gap widening in k (at $k = 512$: GraniteMoE-1B 0.051 against 0.067, OLMoE 0.175 against 0.187). We therefore read the curves as a lower bound *for these models* — an empirical property of these response matrices, not of the estimator in general. The opposite claim, that the deleted coordinates are especially aligned, is not safe under an in-sample ranking and we do not make it.

Result. Table 8 and Figure 23 give the outcome. Deleting the eight highest-energy coordinates leaves six of the seven models holding 78% to 115% of their original η. Deleting half of all coordinates — 512 to 1024 of them — still leaves η between $8\times$ and $203\times$ each model's own matched null, retaining 19% to 44% of the original value. The alignment thins as the space is cut down; it does not collapse. Over all seven models and every k, the closest any model comes to its own chance level is $8.4\times$ it.

Two models deserve separate mention, in opposite directions. At a fixed $k = 2$, four of the seven read *higher* than at $k = 0$ (GraniteMoE-1B $0.269 \to 0.374$, Qwen3-30B-Base $0.466 \to 0.497$), so their two largest coordinates dilute the alignment; we report this as a description, having not tested the difference against a null. DeepSeek-V2-Lite is the model where the objection has real force, and it is the exception to the paragraph above: deleting a *single* coordinate removes 48% of its η ($0.142 \to 0.074$), and eight deletions remove 56% ($\eta = 0.062$). Even there the remainder does not vanish — after deleting half the coordinates it holds $30\times$ its null — but any statement about DeepSeek's alignment should be read with that concentration in view.

Scope. All seven models use routed output biases, $b_e = \mathbb{E}[f_e(z_t) \mid e \in \text{top-}k(t)]$, accumulated on discovery windows at settled positions, at one MoE layer per model. For the shared-expert architectures (Qwen1.5-MoE, DeepSeek-V2-Lite) this is the routed component only, as elsewhere in this paper.

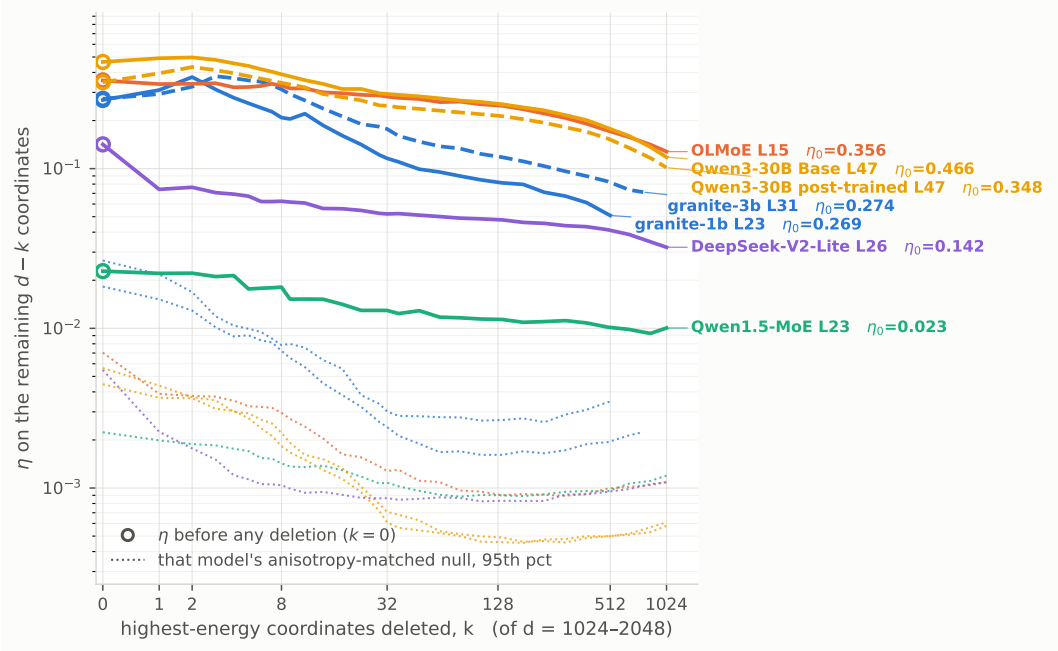

Figure 23: Cross-expert alignment η after deleting the k highest-energy coordinates, at one MoE layer per model. Open markers are η before any deletion; dotted lines are that model's anisotropy-matched null at the 95th percentile, which depends on both the model and k. Both axes are logarithmic: on a linear axis Qwen1.5-MoE sits on the floor and reads as absent, when it in fact runs $8\times$ its own null throughout. The energy ranking is fitted in-sample, which biases the surviving η downward, so each curve is a lower bound.

Table 8: η after deleting the k highest-energy coordinates, from $k = 0$ (no deletion); the last column is $\eta(d/2)$ as a multiple of that model's anisotropy-matched null at the same k. The grid contains $k = 8$ and $k = 32$ exactly.

model	E	d	η	k=8	k=32	k=$d/2$	$\times$ null
granite-1b L23	32	1024	0.269	0.209	0.116	0.051	15$\times$
granite-3b L31	40	1536	0.274	0.314	0.177	0.071	32$\times$
OLMoE L15	64	2048	0.356	0.331	0.282	0.128	118$\times$
Qwen1.5-MoE L23	60	2048	0.023	0.018	0.013	0.010	8$\times$
Qwen3-30B-Base L47	127	2048	0.466	0.390	0.293	0.118	203$\times$
Qwen3-30B L47	126	2048	0.348	0.345	0.247	0.101	166$\times$
DeepSeek-V2-Lite L26	64	2048	0.142	0.062	0.052	0.032	30$\times$