# RECEPTORS OF LANGUAGE MODELS: FROM SINGULAR DIRECTIONS TO TOKEN PREDICTIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

While mechanistic interpretability has largely investigated the information/context flowing through the transformer architecture, the structural blueprint statically/inherently encoded in its parameters remains comparatively underexplored, leaving open a more fundamental question, what predictive structure is already encoded in the weights (inherent to model parameters)? We show that the static weight matrices of transformer components contain sparse, token-aligned directions, which we term *receptors*, that show a strong causal link with model predictions. This structure is consistent across a wide family of architectures. Across 21 open-weight models spanning 5 architecture families (`GPT-2`, `GPT-Neo/J`, `Phi`, `Qwen`, and `LLaMA`), we demonstrate the presence of receptors providing an interpretability lens that is applicable to wider range of contexts. These receptors emerge and stabilize over training. Additionally, we find that a kurtosis-based analysis of individual receptors reveals additional structure in their weight-space geometry. Our work provides a detailed examination of the model's inherent structural properties, which could serve as a complementary foundation for studying, measuring, and intervening in learned behavior.

## 1 INTRODUCTION

Inspecting the static information encoded directly in language model (LM) weights remains an underexplored frontier in mechanistic interpretability. In transformer language models (Radford et al., 2019; Brown et al., 2020; Li et al., 2023b; Javaheripi et al., 2023; Touvron et al., 2023; Jiang et al., 2023; Grattafiori et al., 2024), interpretability efforts have progressed from studying individual neurons and components to identifying circuits that implement specific behaviors (Elhage et al., 2021; Conmy et al., 2023). Early work often treated individual neurons, attention heads, and MLPs as meaningful computational units, while subsequent circuit-level analyses showed that useful computations are distributed across multiple components and can involve substantial polysemanticity (Geva et al., 2021; Merullo et al., 2024; Ahmad et al., 2025). These developments raise a complementary question: *rather than asking which features are expressed during inference, what predictive structure is already encoded directly in the model's parameters?*

Complementary work has begun to investigate transformers directly in weight space. Such analyses reveal structured and sometimes overlapping functions within individual heads and MLPs, challenging the view of these components as atomic computational units (Geva et al., 2021; Pearce et al., 2024; Adler et al., 2025b; Shafran et al., 2025; Avrahamy et al., 2026; Bushnaq et al., 2026; Shafran et al., 2026), stressing upon the structures obtained via singular value decomposition (Makelov et al., 2023; Merullo et al., 2024; Singh et al., 2024; Ahmad et al., 2025; Franco et al., 2026; Xue & Andrzejak, 2026). These results suggest that meaningful structure can be recovered from parameters alone, but leave open a more direct question: *which directions in weight space are functionally aligned with model predictions, and how much of the model's predictive behavior can they explain?*



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.

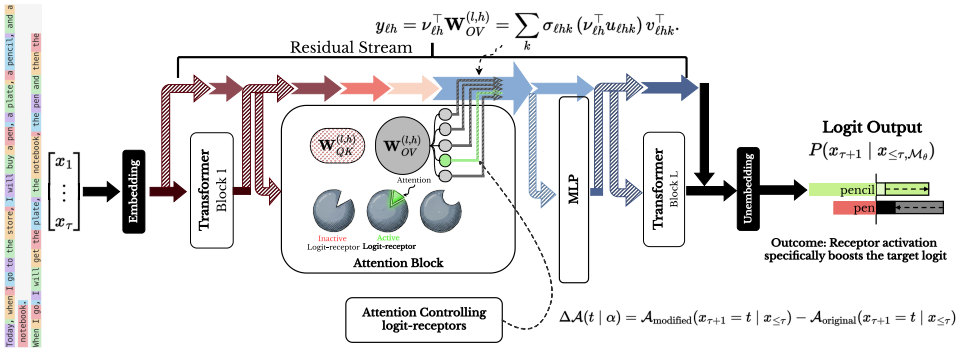



Figure 1: **Receptors as weight-space directions causally linked to prediction.** An input prompt is embedded and processed by a stack of transformer blocks that read from and write to a shared residual stream (top). Within each block, both attention OV circuits and MLPs write weight-defined directions into the residual stream. We refer to these directions as *receptors*. An active receptor (green) denotes a weight-defined direction contributing to the residual representation and, consequently, to the output logits.

In this work, we study this question through the weight-space directions of transformer components that contribute directly to the residual stream to make a prediction, we decompose attention and MLP weight maps into orthogonal directions and examine how these directions project onto the model's output vocabulary. We find that a small subset of these directions is strongly aligned/causally-linked with individual token logits. We call these directions *receptors*: weight-space directions whose amplification produces a corresponding and predictable change in the probability of an aligned token. We further check, if receptors can be identified directly from model parameters, without requiring activation datasets or task-specific feature learning?

We systematically study this phenomenon across 21 open-weight language models spanning five architecture families: `GPT-2`, `GPT-Neo/J`, `Phi`, `Qwen`, and `LLaMA`. Across these models, we find a consistent pattern of sparsity: although each model contains a large number of SVD directions, only a small fraction contributes substantially to a selected target prediction. Later, we investigate whether individual receptors contain additional internal structure beyond their singular-vector alignment with output logits. Using a kurtosis-based analysis of receptor weight-space geometry, we uncover highly concentrated vocabulary channels within several receptors. For example, in a factual retrieval setting, individual attention receptors can isolate channels strongly associated with tokens such as `Paris`, `French`, and `France`, while other receptors contain similarly concentrated but less semantically direct vocabulary structure. These observations suggest that the geometry of weight space contains structure at multiple levels, singular directions identify functionally relevant predictive axes, while higher-order statistics can further reveal concentrated vocabulary substructure.

- We indentify/formalize *receptors* as weight-defined directions in residual stream (formulating Alg. 1), and show that intervening on these directions produces predictable, causal changes in model predictions.
- Across 21 open-weight models spanning five architecture families, we find consistent receptor structure and show that predictive behavior is concentrated in a small fraction of weight-space directions.
- A kurtosis-based analysis (using devised Alg. 2) reveals highly concentrated vocabulary channels within individual functional receptors, exposing additional structure in the geometry of learned weight-space directions. We will release the experiments and model code upon acceptance.

## 2 Preliminaries

**Transformer-Based Language Models.** A language model (LM) defines a conditional distribution over sequences of discrete tokens drawn from a vocabulary $\mathcal{V}$. In this work, we focus on transformer-based,

decoder-only language models trained autoregressively, which underlie most contemporary LMs (Radford et al., 2019; OpenAI et al., 2024; Gemini et al., 2024). Given a prefix $\mathbf{x} = [x_1, \ldots, x_\tau] \in \mathcal{V}^\tau$, a model $\mathcal{M}_\theta$ parameterized by $\theta$ defines a distribution over the next token, $P_{\mathcal{M}_\theta}(x_{\tau+1} \mid \mathbf{x})$. Internally, a transformer consists of a stack of $T$ blocks that repeatedly read from and write to a shared *residual stream* (Elhage et al., 2021). Denoting the residual stream representation at layer $\ell$ for the whole sequence as $\mathbf{z}_\ell \in \mathbb{R}^{\tau \times d_{\text{model}}}$ and $z_\ell \in \mathbb{R}^{d_{\text{model}}}$ for the last prediction token. The computation proceeds via $z_{\ell+1} = z_\ell + f_{\theta_\ell}(\mathbf{z}_\ell)$, where $f_{\theta_\ell}$ aggregates the updates from attention and feed-forward sublayers at layer $\ell$. After the final block, the residual stream is normalized and projected onto the vocabulary using an unembedding matrix $\mathbf{W}_U \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{model}}}$, producing logits whose softmax yields the next-token distribution.

The primary observation underlying our perspective is that, although transformer blocks involve nonlinear operations, their effect on the residual stream at a fixed position is a sum of affine updates whose directions are fixed by the model weights. Making this linear structure explicit allows us to analyze transformer computation using standard SVD and to isolate the directions through which predictions are ultimately formed.

**Attention as a Conditional Linear Operator:** We begin with a single attention head. Let $z_{\ell j}, z_\ell \in \mathbb{R}^{d_{\text{model}}}$ denote residual stream representations at token position $j$ and the last prediction token representations respectively for layer $\ell$. The query, key, and value vectors for $h$ head are computed as $q_{\ell h} = \mathbf{W}_Q^{(\ell,h)} z_\ell + \mathbf{b}_Q^{(\ell,h)}$, $k_{\ell hj} = \mathbf{W}_K^{(\ell,h)} z_{\ell j} + \mathbf{b}_K^{(\ell,h)}$ and $\nu_{\ell hj} = \mathbf{W}_V^{(\ell,h)} z_{\ell j} + \mathbf{b}_V^{(\ell,h)}$ respectively. The attention weights are given by $\alpha_{\ell hj} = \text{softmax}_j\left(\frac{q_{\ell h}^\top k_{\ell j}}{\sqrt{d_{\text{head}}}}\right)$ and the attention head forms a weighted sum of value vectors i.e. $g_{\ell h} = \sum_j \alpha_{\ell hj} \nu_{\ell hj} = \mathbf{W}_V^{(\ell,h)}\left(\sum_j \alpha_{\ell hj} z_{\ell j}\right) + \mathbf{b}_V^{(\ell,h)}$. Conditioned on the attention weights $\{\alpha_{\ell hj}\}_j$, the vector $g_{\ell h}$ is an affine function of the residual stream inputs $\{z_{\ell j}\}_j$. The head then writes this result back into the residual stream via an output projection, as $y_{\ell h} = \mathbf{W}_O^{(\ell,h)} g_{\ell h} + \mathbf{b}_O^{(\ell,h)}$. To express this update as a single linear map, we use an augmented representation that absorbs bias terms. Define the augmented input $\zeta_{\ell h} = \left[1 \ \sum_j \alpha_{\ell hj} z_{\ell j}\right]^\top \in \mathbb{R}^{1+d_{\text{model}}}$ and the corresponding augmented output–value matrix $\mathbf{W}_{OV}^{(\ell,h)} = \left[\mathbf{W}_O^{(\ell,h)} \mathbf{b}_V^{(\ell,h)} + \frac{1}{|H|}\mathbf{b}_O \quad \mathbf{W}_V^{(\ell,h)} \mathbf{W}_O^{(\ell,h)}\right] \in \mathbb{R}^{d_{\text{model}} \times (1+d_{\text{model}})}$. With this notation, the contribution of head $h$ can be written compactly as $y_{\ell h} = \mathbf{W}_{OV}^{(\ell,h)} \zeta_{\ell h}$. Thus, for a fixed attention pattern, each attention head applies as a *conditional linear operator* to an input-dependent coefficient vector and writes the result into the residual stream, i.e. $z_{\ell,pre} = z_\ell + \sum_h y_{\ell h}$ (before the MLP sublayer). Further, the MLP sublayer reads the residual stream after the attention head, $z_{\ell,pre}$, and writes on it , $z_{\ell+1}$, and consists of two affine transformations separated by a pointwise nonlinearity i.e. $\delta_{\ell,in} = \mathbf{W}_{\text{in}}^{(\ell)} z_{\ell,pre} + \mathbf{b}_{\text{in}}^{(\ell)}$ and $\delta_{\ell,out} = \mathbf{W}_{\text{out}}^{(\ell)} \phi(\delta_{\ell,in}) + \mathbf{b}_{\text{out}}^{(\ell)}$ where $\phi$ is an activation function, resulting in $z_{\ell+1} = z_{\ell,pre} + \delta_{\ell,out}$. (similar to $\mathbf{W}_{OV}^{(\ell,h)}$, $\mathbf{W}_{\text{out}}^{(\ell)} \phi(\delta_{\ell,in}) + \mathbf{b}_{\text{out}}^{(\ell)}$ of MLP is also a conditional linear operator, see App. A for derivation)

## 3 RECEPTOR: WEIGHT-DEFINED DIRECTIONS IN THE RESIDUAL STREAM

Each attention head and MLP sublayer writes an update directly to the residual stream. We analyze the linear map associated with this write. For an attention head, this is the conditional OV map obtained after fixing the attention weights; for an MLP, it is the output projection. **Residual-Stream Writes as Scalar Sums of Fixed Directions:** For a given layer $\ell$ and component $c \in \{1, \cdots, H\} \cup \{\text{MLP}\}$ where $H$ is the number of attention heads, the matrix $\mathbf{W}^{(\ell,c)}$ is a linear map writing directly into the residual stream. We apply SVD, i.e., $\mathbf{W}^{(\ell,c)} = \mathbf{U}^{(\ell,c)} \mathbf{\Sigma}^{(\ell,c)} \mathbf{V}^{(\ell,c)\top} = \sum_k \sigma_{\ell ck} u_{\ell ck} v_{\ell ck}{}^\top$ where $\mathbf{U}^{(\ell,h)}$, $\mathbf{V}^{(\ell,h)}$ contains left and right singular vectors respectively, and $\mathbf{\Sigma}^{(\ell,h)}$ is a diagonal matrix of singular values. This decomposition expresses the OV/MLP projection as a

sum of rank-one maps. Each singular component defines a fixed residual-stream direction, with an input-dependent scalar coefficient. The directions depend only on the learned weights, not the input. Let $\zeta_{\ell c}$ denote the augmented input to component $c$ at layer $\ell$. The contribution written by this head into the residual stream can be written as

$$y_{\ell c} = \mathbf{W}^{(\ell,c)}\zeta_{\ell c} = \sum_k \underbrace{\sigma_{\ell ck}\,(v_{\ell ck}^\top \zeta_{\ell c})}_{\texttt{scalar}} \quad \underbrace{(u_{\ell ck})}_{\texttt{candidate receptor}}$$

Each term in this sum consists of a fixed residual-stream direction $u_{\ell ck}$ multiplied by a scalar coefficient $a_{\ell ck} = v_{\ell ck}^\top \zeta_{\ell c} \in \mathbb{R}$ . Thus, OV/MLP projections implement a form of scalar gating: the identity of the directions written into the residual stream is fixed by the weights, while context determines only the magnitude with which each direction is activated. **Defining Receptor (for a token $t$):** We extend the use of the receptor by taking a prediction-centric perspective, and identify receptors whose role is to increase/decrease/modulate the strength of a particular logit. Formally, we define a receptor as a direction that causally influence model logits through the unembedding map. Let logits be computed as $\text{logits}^{(\ell)} = \mathbf{W}_U z_\ell$, where $z_\ell$ is the residual stream and $\mathbf{W}_U$ is the unembedding matrix. A left singular vector $u_{\ell ck}$ of the OV/MLP-out matrix is a ***receptor*** if there exists a scalar $\alpha \in \mathbb{R}$ such that the intervention $z_\ell \mapsto z_\ell + \alpha u_{\ell ck}$ produces a non-negligible change in the logit (or predicted probability) of token $t$, i.e. $\Delta\text{logit}_t^{(\ell)} = W_U(z_\ell + \alpha u_{\ell ck})\big|_t - W_U z_\ell\big|_t \neq 0$. The above definition is conditioned on the direct effect, the effect is evaluated under a fixed forward pass (i.e., fixed attention pattern and residual state). In practice, we frame Alg. 1 to identify receptors (via scaling $\sigma_{\ell ck}$) and measuring the resulting change in logits/accuracy.

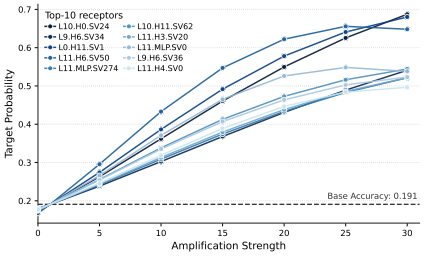



Figure 2: **Amplification of the top-10 receptor directions in `GPT-2-small`.** Scaling the top-10 receptors by $\alpha$ produces an approximately linear increase in target-token probability up to $\alpha = 25$, indicating a strong, near-linear link between receptor contribution and the target logit. (also see App. Fig. 11)

## 4 Methodology

Our primary experimental question is whether the weight-space directions identified by the SVD decomposition (aka receptors) provide a lens to inherent structural property of model weights (a coordinate system) that show a measurable/causal link with model predictions. A noteworthy point about this weight-space formulation is that the receptors themselves are fixed by the learned parameters, and their contribution during a forward pass can be modulated just by a scaling factor. We therefore intervene directly on these coefficients while leaving the remaining computation unchanged. This gives us a controlled way to measure how individual weight-space directions (that naturally exist inside model parameters) influence model predictions.

**Weight-Space Interventions** Formally, let $u_{\ell hk}$ be a left singular vector of the OV/MLP weight matrix at layer $\ell$, component $c$, with associated singular value $\sigma_{\ell ck}$ and input-dependent activation $a_{\ell ck} = \sigma_{\ell ck}\,(v_{\ell ck}^\top \zeta_{\ell c})$. This decomposition separates two aspects of the computation, 1) the *inherent direction* written into the residual stream, determined by the learned weights, whereas the 2) *strength* with which that direction is written depends on the current input. This separation allows us to intervene on an individual weight-space direction without changing the direction itself or retraining the model. We formalize this via an Alg. described in App. Alg. 1. The standard contribution of this component to the residual stream is $y_{\ell ck} = a_{\ell ck}\, u_{\ell ck}$. To intervene, we introduce a scaling factor $\alpha \in \mathbb{R}$, and define the modified contribution $\tilde{y}_{\ell ck} = \alpha\, a_{\ell ck}\, u_{\ell ck}$. Setting $\alpha \in [0, 1)$ effectively silences the singular direction (or removes the contribution of the receptor), while $\alpha > 1$ amplifies it. The residual stream is then recomputed with this modified contribution, i.e., $\tilde{z}_T = z_T + \sum_{\ell,c,k}(\tilde{y}_{\ell ck} - y_{\ell ck})$. The intervention is local, all parameters remain fixed, and only the selected singular component is modified. In particular, we do not modify the unembedding matrix or directly set

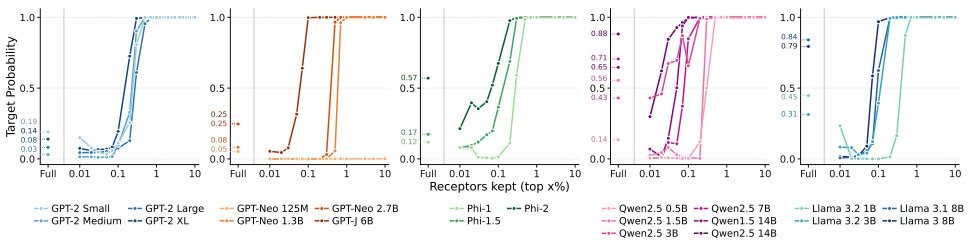



Figure 3: **Target-token probability versus retained receptor directions across model families.** Retaining the top-$x\%$ receptors ranked by target alignment produces a sharp increase in target-token probability, typically with $< 1\%$ of directions (consistent across `GPT-2`, `GPT-Neo/J`, `Phi`, `Qwen`, and `Llama`). The corresponding regret quantifies the receptor fraction required to reach the target probability of one (Fig. 4).

a target logit. Instead, we alter an internal weight-space direction and measure the downstream change in the model's prediction, allowing us to test whether a direction embedded in the parameter functions as a predictive has a direct alignment with a particular output.

**Measuring the Causal Effect of a Receptor** We evaluate interventions with respect to a target token $t$. Given a prompt $x_{\leq\tau}$ for which $t$ is the expected next token, let $P(t \mid x_{\leq\tau})$ denote the model's original target-token probability and $\widetilde{P}_\alpha(t \mid x_{\leq\tau})$ the probability after scaling a selected singular direction by $\alpha$. $\Delta P(t \mid \alpha) = \widetilde{P}_\alpha(x_{\tau+1} = t \mid x_{\leq\tau}) - P(x_{\tau+1} = t \mid x_{\leq\tau})$.

**Experimental Setup** Our goal is to understand whether receptor directions identified from model weights exert a consistent causal effect on model predictions across diverse input contexts. The central experimental requirement is to evaluate the same target token under substantially different contexts, while keeping the intervened model parameters fixed, allowing us to measure whether scaling an individual receptor produces systematic changes in the target-token prediction beyond any particular prompt realization. We evaluate this principle across 21 open-weight language models spanning five architecture families, `GPT-2`, `GPT-Neo/J`, `Phi`, `Qwen`, and `LLaMA`, covering a range of model sizes and architectural configurations. We use three complementary evaluation settings, **1)** the controlled Laundry List retrieval task; **2)** model-generated in-distribution contexts, where the target is predicted next; and **3)** MMLU, where answer-choice positions are randomized and grouped by correct option token. We additionally construct controlled Paris prompt sets (Anchor/Adjacent/Neutral) to test target-specific behavior. Details appear in App. D.

## 5 Results

**Target-token effects are concentrated in a small subset of receptors.** We first ask whether individual receptors have measurable effects on a fixed target-token prediction. For each singular direction, we scale its contribution by a fixed amplification factor and measure the resulting change in target-token probability. Most directions produce negligible changes, while a small subset produces substantially larger effects (App. Fig. 10). The concentration holds for both target-enhancing and target-suppressing directions. The intervention also shows a systematic relationship (similar to dose-response relationship). In the Laundry List task, progressively increasing the contribution of an individual receptor produces a corresponding change in the probability of the target token (Fig. 2). The relationship remains monotonic over a broad range of amplification factors, with deviations emerging only at extreme scaling (App. Fig. 11). Furthermore, we find that aggregating the strongest receptors reveals the effect is concentrated among the directions with the largest individual effects. Re-ranking receptors by effect at each amplification factor confirms the concentration in the strongest directions (App. Fig. 12). These results establish that the SVD decomposition exposes a sparse subset of weight-space directions whose contributions are causally linked to target-token prediction.

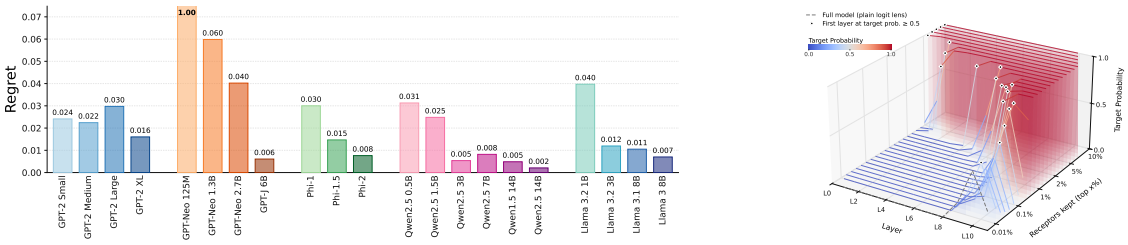



Figure 4: Effect of selectively retaining positively contributing receptor directions on model predictions for the Laundry List dataset. **Left:** Regret under selective retention of positively contributing receptor directions, showing that retaining only a small subset of directions can substantially reduce regret relative to retaining all directions (for multiple model families). **Right:** Target-token probability as a function of model (`GPT-2-small`) layer and the fraction of positively contributing receptor directions retained. Even when most receptor directions are removed, retaining the most strongly contributing directions can recover a high target-token probability, with the effect emerging in earlier layers. (also see App. Fig. 17)

**A small subset of receptors is sufficient to recover target predictions.** The previous experiment establishes that individual receptors can change a target prediction. We next ask whether the converse also holds, how much of the receptor set is actually required to produce the prediction? We selectively retain receptors in descending order of their measured target effect and remove the remainder. Across `GPT-2`, `GPT-Neo/J`, `Phi`, `Qwen`, and `Llama` models, retaining only a small fraction of the strongest receptors produces a sharp increase in target-token probability (Fig. 3). In many models, this transition occurs within the top $1\%$ of receptors, despite the intervention removing the contribution of the remaining directions. The complementary ablation gives the corresponding result in the opposite direction: selectively removing the strongest target-contributing receptors substantially reduces the target probability (App. Fig. 14). The agreement between selective retention and selective removal is important because it shows that the effect is not simply caused by adding an arbitrary perturbation to the residual stream. Instead, the directions identified by their individual causal effects contain a concentrated portion of the signal supporting the target prediction. **Notably,** this sparsity also extends beyond the controlled Laundry List task, similar results on MMLU, for `LLaMA3-8B`, (App. Fig. 13), denoting that the concentration extends beyond the synthetic task to multi-domain QA.

**Quantifying receptor sufficiency.** To quantify how rapidly target prediction is recovered as receptors are added, we measure the regret associated with selective retention. The regret summarizes the residual gap between the target probability obtained with a restricted receptor set and the corresponding target probability when all receptors are retained. Across model families, regret decreases sharply once the highest-effect receptors are restored (Fig. 4, left). The layer-wise analysis further shows that this concentrated signal is already present in relatively early layers (Fig. 4, right; App. Fig. 17). Regret quantifies this sparsity (see App. G), a small receptor subset recovers a substantial fraction of the interventionally recoverable target signal.

**Target effects cannot be inferred from the receptor direction alone.** The preceding experiments identify receptors using their input-dependent causal effect. We next ask whether the same structure can be detected from the learned weights without evaluating target-token interventions. A natural candidate is the similarity/alignment between a receptor direction and the output embedding associated with the target token. However, a direct comparison is insufficient to determine the signed contribution of a receptor (making it a minor caveat). There are two reasons. First, the contribution of a singular direction is determined by its input-dependent coefficient, $a_{\ell ck} = \sigma_{\ell ck}\,(v_{\ell ck}^{\top}\zeta_{\ell c})$, so the sign of its effect depends on the current activation. Second, singular vectors are sign-indeterminate, i.e., replacing $u_{\ell ck}$ by $-u_{\ell ck}$ and the corresponding singular-vector coefficient by its negative leaves the underlying decomposition unchanged. Consequently, the direction itself does not provide a canonical notion of target-enhancing versus target-suppressing contribution.

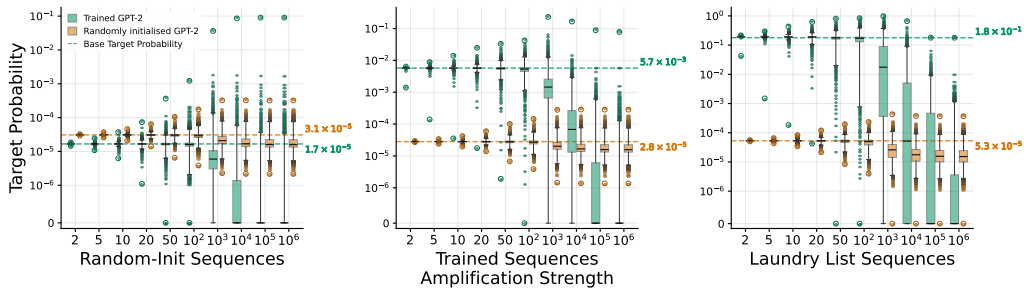



Figure 5: **Effect of amplifying individual receptor directions on target-token probability across model and data settings.** We intervene on individual receptor directions by scaling their contribution by an amplification factor and measure the resulting target-token probability. Distribution of target-token probabilities under individual receptor amplification for randomly initialized `GPT-2-small`, pretrained `GPT-2-small`, and the Laundry List task. Most receptors have negligible effects, while a small set of strong responders (outliers) emerges in the trained model and task-specific setting but is largely absent at random initialization. (Tab. 3)

We therefore compare weight-space similarity measures with the receptors identified through direct intervention, rather than treating similarity as a substitute for the causal measurement. The resulting correspondence is reported in Table 2. To our surprise, the weight-space geometry already provides a strong signal of the intervention-derived receptor ranking. As shown in Table 2, alignment with the target unembedding correlates substantially with causal effect, with Spearman $\rho = 0.464$, $0.455$, and $0.458$ for OV, MLP, and combined OV+MLP receptors, respectively; Thus, although weight-space alignment cannot determine the context-dependent sign of a receptor's effect, it can identify target-relevant directions without observing the input or performing an intervention. This also suggests that receptor structure is substantially encoded in the learned weights themselves, rather than arising only from input-dependent activation.

**Target-relevant receptors emerge during training.** Having established the phenomenon across trained models, we next ask whether target-sensitive receptor structure is induced by training rather than being a generic consequence/artifact of the model architecture or SVD decomposition. We compare intervention effects across three input distributions: (i) in-distribution contexts generated by a randomly initialized `GPT-2`, (ii) in-distribution contexts generated by the pretrained model, and (iii) the task-specific Laundry List distribution (Fig. 18). This comparison separates effects arising from the learned model distribution from those induced by the task itself. As shown in Fig. 5, strong target-probability responses are concentrated in the trained model, while corresponding effects are substantially weaker in the randomly initialized model (also see App. Table 3). Thus, sparse target-sensitive receptor structure emerges through training rather than being a generic consequence of decomposing the model's weight matrices. We further analyze the training trajectory with more details using Pythia. Across Pythia checkpoints, target-sensitive signal first appears in later layers and progressively shifts toward earlier layers during training (App. Fig. 16), consistent with accumulating target-relevant structure. The strongest receptor sets also become more stable late in training, i.e., top-1% overlap with the final checkpoint and between checkpoints increases over training (App. Fig.15). Note that this increasing stability should be interpreted together with the smaller parameter updates late in training; nevertheless, the result shows that the target-sensitive receptor set is not simply fixed from initialization.

**Receptors contain heterogeneous vocabulary structure.** We next ask what this structure looks like internally: *does a receptor correspond to a single semantic feature, or does it contain multiple vocabulary-selective components?* To understand this structure, we adapt kurtosis-based analysis from (Avrahamy et al., 2026), originally developed to decompose polysemantic MLP neurons, to receptor directions. Our adaptation decomposes a receptor into nearby vocabulary-selective channels and identifies which channel, if any, carries the target token (we devise Algorithm 2; details in App. C). This lets us move beyond measuring whether a receptor affects a token to characterizing *how that receptor represents the token in vocabulary space*.

Table 1: Channels obtained by Alg. 2 for the ten receptors most strongly increasing $P(\texttt{␣Paris})$ in `Qwen2.5-14B`. Six of ten finds a channel explicitly isolating the target, with concentrated projections onto `␣Paris`. For the remaining four (†), no channel isolates the target.

| Rank | Receptor | $\Delta P$ | Ch. | $\lvert z_{\text{Paris}} \rvert$ | $\kappa$ | $r_{\text{vocab}}$ | Top promoted tokens |
|---|---|---|---|---|---|---|---|
| 7 | `L34.H14.SV21` | $+0.153$ | 0 | 51.19 | 193.3 | 0 | `␣Paris, Paris, ␣paris, ␣French` |
| 1 | `L35.H16.SV6` | $+0.207$ | 0 | 28.07 | 252.0 | 11 | `␣French, French, ␣french, ␣France` |
| 5 | `L35.H16.SV2` | $+0.160$ | 0 | 26.55 | 248.2 | 11 | `␣French, French, ␣french, ␣France` |
| 6 | `L40.H21.SV3` | $+0.158$ | 0 | 12.19 | 14.4 | 9 | `␣French, French, ␣France, ␣french` |
| 9 | `L40.H22.SV93` | $+0.148$ | 0 | 6.91 | 0.5 | 0 | `␣Paris, Paris, ␣Tunis, ␣Prov` |
| 2 | `L35.H16.SV4` | $+0.174$ | 0 | 5.83 | 290.6 | 224 | `␣Chinese, ␣China, Chinese, China` |
| 3 | `L47.MLP.SV2142` | $+0.172$ | $41^{\dagger}$ | 2.87 | 0.1 | 464 | `abe, .identity, ULE, ␣nightly` |
| 10 | `L47.MLP.SV810` | $+0.143$ | $30^{\dagger}$ | 2.74 | 0.1 | 594 | `iating, ␣countries, =val, zte` |
| 8 | `L47.MLP.SV1967` | $+0.149$ | $34^{\dagger}$ | 2.70 | 0.1 | 672 | `xon, -st, ␣Bennett, /local` |
| 4 | `L47.MLP.SV219` | $+0.170$ | $18^{\dagger}$ | 2.06 | 0.2 | 2872 | `␣Chemical, Nearly, ey, alli` |

Applying Algorithm 2 to the receptors that most strongly increase $P(\texttt{␣Paris})$ reveals substantial internal structure. Several high-impact receptors contain sharply concentrated channels associated with `␣Paris`, `␣French`, and `␣France` (Table 1; App. Tables 8 and 7), highlighting that receptors need not correspond to a single lexical feature, a single causally important direction can contain multiple, separable vocabulary channels, while individual channels can have strong target-specific structure.

More broadly, these results suggest that target-relevant structure is not a property of a proxy decomposition, but is directly encoded in the model's native weight space. Receptors provide computational units through which this structure can be identified, decomposed, and interpreted.

**Receptor interventions affect target generation while preserving broader model behavior.** Finally, we test whether the intervention extends beyond next-token probabilities to free generation. Scaling the top Paris receptors increases Paris-related generations on anchor and adjacent prompts, whereas matched random-receptor interventions produce substantially smaller changes (App. Table 6). Surprisingly, the intervention also changes the frequency of France-related continuations, providing a behavioral consequence beyond the single target-token probability. At moderate amplification factors, these targeted changes occur without a commensurate degradation in overall language-model behavior. We further quantify it with perplexity on held-out Pile documents Gao et al. (2021), remains close to the un-intervened baseline for moderate receptor scaling, (App. Table 5). The random-receptor controls show that comparable perturbations to the same layers do not generally produce the same target-specific changes, indicating that the receptor intervention can selectively alter target-related predictions while leaving the broader computation comparatively intact.

**Discussion.** Our primary claim/finding is deliberately narrower than a complete account of model computation: *target-relevant receptors exist as causally effective directions in the model's native weight space*. Across model families and tasks, a small subset of these directions can strongly alter a target-token prediction, and this structure emerges through training and exhibits internal vocabulary organization, which distinguishes the receptor view from approaches that first construct an auxiliary representation of activations, such as SAEs or transcoders. Those methods and receptors address complementary questions: learned feature bases seek interpretable structure in activation space, whereas we ask whether the parameters themselves contain directions that directly control predictions. The interactions between such implicit components (obtained via SVD) have been studied previously, though only for attention heads (Merullo et al., 2024). SVD provides a natural coordinate system in which to test whether individual weight-defined directions have causal predictive effects. Establishing this object is the contribution; it makes weight-space interpretability a concrete empirical problem that can now be extended to interactions, composition, and higher-level behavior.

**Limitations.** Our experiments establish the *presence* and causal relevance of receptors. We also find that receptor effects are context-dependent; the direction is fixed by the weights, but their activation and causal sign depend on the input. Likewise, kurtosis analysis reveals structured vocabulary channels rather than proving semantic monosemanticity. Finally, our weight-space analysis does not eliminate the need for activation-level

evidence when studying context-dependent computation; instead, it provides a complementary starting point that can often be interrogated directly from the parameters, without repeatedly constructing task-specific activation datasets or discovering circuits through many forward passes.

**Broader Impact.** The weight-space formulation also suggests practical directions beyond interpretation. Because receptors are defined by model parameters, they can potentially be inspected without executing the model, enabling inexpensive weight-level monitoring (see time in App. Table 4) of global underlying structures. Their causal accessibility also enables surgical interventions, in our experiments, moderate amplification changes target-specific behavior while leaving perplexity close to baseline, suggesting a possible route toward targeted editing, auditing, and capability intervention without retraining. These possibilities also expose a safety tradeoff, the same localized controllability that enables auditing and mitigation could be used to manipulate model behavior. We therefore view receptors less as a finished interpretability method than as a concrete interface for developing weight-space auditing, editing, and training-time monitoring.

The broader opportunity is to treat the learned weights themselves as an object of mechanistic study. Receptors provide an experimentally testable unit for asking where predictive structure resides, when it emerges, and how it composes. Establishing that such causal structure is already present in native weight space is a first step toward a broader weight-space paradigm for mechanistic interpretability.

## 6 Related Works

Mechanistic interpretability explains model behavior through internal computations rather than post-hoc correlations (Rai et al., 2025; Zhao et al., 2024). Circuit analyses decompose transformer behavior into interactions among attention heads and MLPs (Lan et al., 2024; Dunefsky et al., 2024), while probing and representation-based methods characterize information encoded in activations (Rogers et al., 2020; Burns et al., 2022; Li et al., 2023a). These approaches establish what representations contain, but recoverability alone does not establish which directions causally determine a prediction. Superposition and polysemanticity have motivated SAEs, sparse feature circuits (Marks et al., 2025), and related feature decompositions (Bushnaq et al., 2024; Adler et al., 2025a; Chanin et al., 2025; Song et al., 2026). These methods provide increasingly structured descriptions of activation space (Sun et al., 2024; Zhao et al., 2025; Park et al., 2023; Gurnee et al., 2026), but the resulting features are learned representations rather than directions native to the model's parameterization. A complementary line of work studies structure directly in model weights. Weight-space analyses reveal interpretable structure within attention and MLP operators (Geva et al., 2021; Pearce et al., 2024; Adler et al., 2025b; Shafran et al., 2025; Avrahamy et al., 2026; Bushnaq et al., 2026; Shafran et al., 2026), and recent directions suggest an active role of SVD (Merullo et al., 2024; Ahmad et al., 2025; Franco et al., 2026) However, these analyses generally characterize what a weight-space direction represents without directly testing whether it causally controls a particular token prediction. Our work studies the weight-level interpretability through *receptors*: weight-defined singular directions whose direct intervention produces a measurable change in a target-token logit. This connects the causal perspective of mechanistic interpretability with the native geometry of model weights, without first imposing an external feature basis.

## 7 Conclusion

We show that transformer predictions can be traced to a sparse set of *receptor directions* already present in the model's native weight space. These directions have a measurable causal effect on individual token logits: amplifying or removing a small fraction can sharply change the target probability while leaving overall model behavior largely intact. The same structure appears across model families, emerges during training, and contains an interpretable vocabulary-selective structure. More broadly, receptors provide a concrete unit for connecting *weight-space structure* to *causal behavior*. This opens a path toward studying how token-level computations are formed, stabilized, and composed during training, and toward understanding model behavior through the structures the model itself has learned rather than through an externally imposed representation.

## REFERENCES


Micah Adler, Dan Alistarh, and Nir Shavit. Towards combinatorial interpretability of neural computation, 2025a. URL https://arxiv.org/abs/2504.08842. [Cited on page 9.]

Micah Adler, Dan Alistarh, and Nir Shavit. Towards combinatorial interpretability of neural computation. *arXiv preprint arXiv:2504.08842*, 2025b. [Cited on pages 1 and 9.]

Areeb Ahmad, Abhinav Joshi, and Ashutosh Modi. Beyond components: Singular vector-based interpretability of transformer circuits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=7UbXEQNny7. [Cited on pages 1 and 9.]

Asaf Avrahamy, Yoav Gur-Arieh, and Mor Geva. Disentangling mlp neuron weights in vocabulary space, 2026. URL https://arxiv.org/abs/2604.06005. [Cited on pages 1, 7, 9, and 25.]

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165. [Cited on page 1.]

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022. [Cited on page 9.]

Lucius Bushnaq, Stefan Heimersheim, Nicholas Goldowsky-Dill, Dan Braun, Jake Mendel, Kaarel Hänni, Avery Griffin, Jörn Stöhler, Magdalena Wache, and Marius Hobbhahn. The local interaction basis: Identifying computationally-relevant and sparsely interacting features in neural networks, 2024. URL https://arxiv.org/abs/2405.10928. [Cited on page 9.]

Lucius Bushnaq, Dan Braun, Oliver Clive-Griffin, Bart Bussmann, Nathan Hu, Michael Ivanitskiy, Linda Linsefors, and Lee Sharkey. Interpreting language model parameters, April 2026. [Cited on pages 1 and 9.]

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Isaac Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders, 2025. URL https://openreview.net/forum?id=LC2KxRwC3n. [Cited on page 9.]

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=89ia77nZ8u. [Cited on page 1.]

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=J6zHcScAo0. [Cited on page 9.]

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html. [Cited on pages 1 and 3.]

Gabriel Franco, Carson Loughridge, and Mark Crovella. Singular vectors of attention heads align with features. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=EJPfiTrKyu. [Cited on pages 1 and 9.]

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027, 2021. URL https://arxiv.org/abs/2101.00027. [Cited on page 8.]

Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will

Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen

Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi

Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel

Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805. [Cited on page 3.]

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446. [Cited on pages 1 and 9.]

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov,

Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang,

Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. [Cited on page 1.]

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html. [Cited on page 9.]

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. [Cited on page 33.]

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Michael Santacroce, Harkirat Singh Behl, Adam Taumann Kalai, Xin Wang, Rachel Ward, Philipp Witte, Cyril Zhang, and Yi Zhang. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 2023. [Cited on page 1.]

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825. [Cited on page 1.]

Michael Lan, Philip Torr, and Fazl Barez. Towards interpretable sequence continuation: Analyzing shared circuits in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12576–12601, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.699. URL https://aclanthology.org/2024.emnlp-main.699/. [Cited on page 9.]

Sander Land and Max Bartolo. Fishing for magikarp: Automatically detecting under-trained tokens in large language models, 2024. URL https://arxiv.org/abs/2405.05417. [Cited on page 26.]

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023a. [Cited on page 9.]

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report, 2023b. URL https://arxiv.org/abs/2309.05463. [Cited on page 1.]

Yuxi Li, Yi Liu, Gelei Deng, Ying Zhang, Wenjia Song, Ling Shi, Kailong Wang, Yuekang Li, Yang Liu, and Haoyu Wang. Glitch tokens in large language models: Categorization taxonomy and effective detection, 2024. URL https://arxiv.org/abs/2404.09894. [Cited on page 26.]

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. *arXiv preprint arXiv:2311.17030*, 2023. [Cited on page 1.]

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv. [Cited on page 9.]

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Talking heads: Understanding inter-layer communication in transformer language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=LUsx0chTsL. [Cited on pages 1, 8, 9, 27, and 30.]

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang,

Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774. [Cited on page 3.]

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=T0PoOJg8cK. [Cited on page 9.]

Michael T Pearce, Thomas Dooms, Alice Rigg, Jose M Oramas, and Lee Sharkey. Bilinear mlps enable weight-based mechanistic interpretability. *arXiv preprint arXiv:2410.08417*, 2024. [Cited on pages 1 and 9.]

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533. [Cited on pages 1 and 3.]

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models, 2025. URL https://arxiv.org/abs/2407.02646. [Cited on page 9.]

Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering, 2023. [Cited on page 27.]

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54/. [Cited on page 9.]

Or Shafran, Atticus Geiger, and Mor Geva. Decomposing mlp activations into interpretable features via semi-nonnegative matrix factorization. *arXiv preprint arXiv:2506.10920*, 2025. [Cited on pages 1 and 9.]

Or Shafran, Atticus Geiger, and Mor Geva. Constructing interpretable features from compositional neuron groups, 2026. URL https://arxiv.org/abs/2506.10920. [Cited on pages 1 and 9.]

Aaditya K Singh, Ted Moskovitz, Felix Hill, Stephanie CY Chan, and Andrew M Saxe. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. *arXiv preprint arXiv:2404.07129*, 2024. [Cited on page 1.]

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona T. Diab, Virginia Smith, and Kun Zhang. Mechanistic interpretability should prioritize feature consistency in sparse autoencoders. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2172–2210, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.99. URL https://aclanthology.org/2026.acl-long.99/. [Cited on page 9.]

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6. [Cited on page 9.]

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao,

Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288. [Cited on page 1.]

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hanna Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how lms answer multiple choice questions. In *International Conference on Learning Representations*, 2024. [Cited on page 27.]

Min Xue and Artur Andrzejak. SVD as a fast interpretability method for transformers. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=7tt8TwMjdJ. [Cited on page 1.]

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15 (2), February 2024. ISSN 2157-6904. doi: 10.1145/3639372. URL https://doi.org/10.1145/3639372. [Cited on page 9.]

Tianyang Zhao, Kunwar Yashraj Singh, Srikar Appalaraju, Peng Tang, Ying Nian Wu, and Li Erran Li. On the analysis and distillation of emergent outlier properties in pre-trained language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8475–8507, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.430. URL https://aclanthology.org/2025.naacl-long.430/. [Cited on page 9.]

# APPENDIX

## TABLE OF CONTENTS

**A Construction of Receptors for MLP** **23**

**B Finding Receptors via Interventions** **24**

**C ROTATED Receptor: Vocabulary Channels of a Receptor** **25**

**D Dataset Details** **27**

Toy (Template-based) Tasks . . . . . 27

MCQA based dataset . . . . . 27

Real-World Dataset . . . . . 28

**E Compute Resources** **28**

**F Performance on Laundry List Task** **29**

**G Regret of the Receptor-Isolated Lens** **30**

## LIST OF TABLES

2 Agreement between geometric and intervention-based receptor rankings. . . . . 24

3 Top-10 receptors are substantially shared across distinct input distributions. . . . . 36

4 Model and receptor statistics across model families. . . . . 38

5 Perplexity change during intervention . . . . . 39

6 Paris and France-cue generation under receptor scaling . . . . . 40

7 Channels for `GPT-2-small` model . . . . . 40

8 Channels for `Llama-3-8B` model . . . . . 41

9 Hyperparameters for Algorithm 2 . . . . . 41

## LIST OF FIGURES

6 Laundry List Dataset Prompts . . . . . 28

7 MMLU Dataset Prompt . . . . . 29

8 Free-Generation Prompts . . . . . 29

9 Performance of model families on the laundry list task. . . . . 30

10 Target-token effects are concentrated in a small subset of receptors. . . . . 31
11 Response to large-scale amplification of individual receptors. . . . . 32
12 Scaling the top receptors produces systematic changes in target-token probability. . . . . 32
13 Intervention analysis of LLaMA3-8B on MMLU task . . . . . 33
14 Effect of selectively retaining or removing receptors on the target-token probability . . . . . 33
15 Evolution of the top-1% receptor during training . . . . . 34
16 Receptor emergence during training . . . . . 35
17 Layer-wise target probability under selective receptor retention (heatmap). . . . . 35
18 In-distribution contexts for studying target-token receptors. . . . . 37

# A CONSTRUCTION OF RECEPTORS FOR MLP

The receptor construction introduced in the main paper for attention OV maps extends directly to the MLP output projection. We make this connection explicit here. For layer $\ell$, let $z_{\ell,\mathrm{pre}}$ denote the residual stream input to the MLP. The MLP first computes

$$\delta_{\ell,\mathrm{in}} = \mathbf{W}_{\mathrm{in}}^{(\ell)} z_{\ell,\mathrm{pre}} + \mathbf{b}_{\mathrm{in}}^{(\ell)},$$

and then writes

$$\delta_{\ell,\mathrm{out}} = \mathbf{W}_{\mathrm{out}}^{(\ell)} \phi(\delta_{\ell,\mathrm{in}}) + \mathbf{b}_{\mathrm{out}}^{(\ell)}$$

to the residual stream, giving

$$z_{\ell+1} = z_{\ell,\mathrm{pre}} + \delta_{\ell,\mathrm{out}}.$$

Here, $\phi$ is the pointwise MLP nonlinearity. Although the MLP as a whole is nonlinear, its output projection is linear once the preceding activation is fixed. This is the same conditional-linearity used for the attention OV map in the main paper. We therefore define the augmented MLP input

$$\zeta_{\ell,\mathrm{mlp}} = \begin{bmatrix} 1 \\ \phi(\delta_{\ell,\mathrm{in}}) \end{bmatrix} \in \mathbb{R}^{d_{\mathrm{mlp}}+1},$$

and the corresponding augmented output matrix

$$\widetilde{\mathbf{W}}_{\mathrm{out}}^{(\ell)} = \begin{bmatrix} \mathbf{b}_{\mathrm{out}}^{(\ell)} & \mathbf{W}_{\mathrm{out}}^{(\ell)} \end{bmatrix}.$$

The MLP write can then be written exactly as

$$\delta_{\ell,\mathrm{out}} = \widetilde{\mathbf{W}}_{\mathrm{out}}^{(\ell)} \zeta_{\ell,\mathrm{mlp}}.$$

We decompose this augmented matrix using its singular value decomposition,

$$\widetilde{\mathbf{W}}_{\mathrm{out}}^{(\ell)} = \mathbf{U}^{(\ell)} \mathbf{\Sigma}^{(\ell)} \mathbf{V}^{(\ell)\top} = \sum_k \sigma_{\ell k} u_{\ell k} v_{\ell k}^{\top},$$

where $u_{\ell k}$ and $v_{\ell k}$ are the left and right singular vectors, respectively, and $\sigma_{\ell k}$ is the corresponding singular value. Substituting the above decomposition into $\delta_{\ell,\mathrm{out}}$ gives,

$$\delta_{\ell,\mathrm{out}} = \sum_k \sigma_{\ell k} \left( v_{\ell k}^{\top} \zeta_{\ell,\mathrm{mlp}} \right) u_{\ell k}$$

Thus, conditional on the MLP activation, the MLP write is a sum of fixed residual-stream directions $u_{\ell k}$, each modulated by an input-dependent scalar coefficient $a_{\ell k} = \sigma_{\ell k} \left( v_{\ell k}^{\top} \zeta_{\ell,\mathrm{mlp}} \right)$. Following the definition in the main paper, we refer to these left singular vectors $u_{\ell k}$ as MLP receptors. Their directions are determined entirely by the learned output weights, while their contribution on a particular input is determined by the corresponding coefficient $a_{\ell k}$.

This construction also makes the correspondence with attention receptors explicit. For an attention head, fixing the attention pattern yields a conditional OV map whose SVD decomposes the head's residual-stream write into fixed directions and input-dependent coefficients. For an MLP, fixing the preceding nonlinear activation similarly yields the linear output map . In both cases, the residual stream receives a sum of weight-defined directions whose magnitudes are modulated by the current input.

Consequently, writing the residual stream at a given layer as the accumulation of the direct writes from attention heads and MLP sublayers, the model's residual update can be expressed in terms of a common set of OV and MLP receptor directions. This allows the same intervention and causal-analysis framework described in the main paper to be applied to both attention and MLP components without introducing a separate representation for the two.

**Algorithm 1** Logit Lens with Sigma Amplification and Receptor Isolation. Steps in orange are specific to sigma amplification, steps in green to receptor isolation; black steps are shared.

**Require:** Dataset $\mathcal{D}$; model $\mathcal{M}$ with $L$ layers, $H$ heads per layer, pre-attention normalization $\mathrm{LN}_1^\ell$ of layer $\ell$, final normalization $\mathrm{LN}_f$ and unembedding $\mathbf{W}_U$; receptor set $\mathcal{R} = \{(\ell, c, k)\}$ with component $c \in \{1, \dots, H\} \cup \{\mathrm{MLP}\}$ and weight $\mathbf{W}^{(\ell,c)} \in \{\mathbf{W}_{OV}^{(\ell,h)},\ \mathbf{W}_{out}^{(\ell)}\}$; selected receptors of each component $K^{(\ell,c)} = \{k : (\ell, c, k) \in \mathcal{R}\}$; amplification factor $\alpha$

**Ensure:** Per-layer lens accuracy/target probability: base $\mathcal{A}_\ell$, sigma-amplified $\mathcal{A}_\ell^{\mathrm{amp}}$, receptor-isolated $\mathcal{A}_\ell^{\mathrm{iso}}$

$U^{(\ell,c)} S^{(\ell,c)} V^{(\ell,c)\top} \leftarrow \mathrm{SVD}\left(\mathbf{W}^{(\ell,c)}\right)$ for every component $(\ell, c)$ in $\mathcal{R}$ ▷ weights only, computed once

Initialize trackers $\mathcal{A}_\ell$, $\mathcal{A}_\ell^{\mathrm{amp}}$, $\mathcal{A}_\ell^{\mathrm{iso}}$ for $\ell = 1, \dots, L$

**for** each batch $(x, y) \in \mathcal{D}$ **do**

  $\{z_\ell,\ A^{(\ell,h)},\ m_\ell\} \leftarrow \mathcal{M}(x)$ ▷ caches residuals, attention, MLP hidden activations

  $\Delta R^{\mathrm{amp}} \leftarrow \mathbf{0}$, $\Delta R^{\mathrm{iso}} \leftarrow \mathbf{0}$

  **for** $\ell = 1, \dots, L$ **do**

    **for** each component $c$ with $K^{(\ell,c)} \neq \emptyset$ **do**

      $\zeta \leftarrow \begin{cases} \sum_j A_{Tj}^{(\ell,h)}\ \mathrm{LN}_1^\ell(z_{\ell-1,j}) & c = h \text{ (attention head)} \\ m_{\ell,T} & c = \mathrm{MLP} \end{cases}$ ▷ component input at last token $T$

      $\mathbf{r} \leftarrow \sum_{k \in K^{(\ell,c)}} \sigma_k^{(\ell,c)}\, (\zeta^\top v_k^{(\ell,c)})\, u_k^{(\ell,c)}$ ▷ contribution of the selected receptors

      $\Delta R^{\mathrm{amp}} \leftarrow \Delta R^{\mathrm{amp}} + (\alpha - 1)\, \mathbf{r}$ ▷ $\sigma_k \mapsto \alpha\sigma_k$ for $k \in K^{(\ell,c)}$

      $\Delta R^{\mathrm{iso}} \leftarrow \Delta R^{\mathrm{iso}} + \mathbf{r} - \mathbf{W}^{(\ell,c)}\zeta$ ▷ $c$ keeps only its receptors in $K^{(\ell,c)}$

    **end for**

    Update $\mathcal{A}_\ell$ with $\left(\mathbf{W}_U\ \mathrm{LN}_f(z_{\ell,T}),\ y\right)$ ▷ plain logit lens

    Update $\mathcal{A}_\ell^{\mathrm{amp}}$ with $\left(\mathbf{W}_U\ \mathrm{LN}_f(z_{\ell,T} + \Delta R^{\mathrm{amp}}),\ y\right)$

    Update $\mathcal{A}_\ell^{\mathrm{iso}}$ with $\left(\mathbf{W}_U\ \mathrm{LN}_f(z_{\ell,T} + \Delta R^{\mathrm{iso}}),\ y\right)$

  **end for**

**end for**

**return** $\{\mathcal{A}_\ell,\ \mathcal{A}_\ell^{\mathrm{amp}},\ \mathcal{A}_\ell^{\mathrm{iso}}\}_{\ell=1}^{L}$

Table 2: **Agreement between geometric and intervention-based receptor rankings.** For `GPT-2-small`, we compare two rankings of receptor directions: (i) geometric alignment with the target-token unembedding direction, $|\cos(W_U[\texttt{pencil}], u_k)|$, and (ii) the magnitude of the larger change in target-token probability produced by amplifying the receptor with $\sigma = +20$ or $\sigma = -20$ on the Laundry List task. We report Spearman $\rho$ and weighted Kendall $\tau$ for OV, MLP, and combined OV+MLP receptors. The consistent positive correlations indicate that receptors geometrically aligned with the target unembedding direction tend to exert stronger causal effects on the target prediction. The geometric measure does not determine the context-dependent sign of a receptor's effect; that sign is determined by its activation in context.

| Correlation | OV | MLP | OV+MLP |
|---|---|---|---|
| Spearman $\rho$ | $+0.464$ | $+0.455$ | $+0.458$ |
| Weighted Kendall $\tau$ | $+0.603$ | $+0.476$ | $+0.598$ |

## B Finding Receptors via Interventions

Algorithm 1 implements the weight-space interventions described in the main paper. The SVDs of all OV and MLP output matrices are computed once from the model weights and are therefore independent of

the evaluation data (only a time-consuming process in our analysis). For each input, the algorithm then caches the forward-pass quantities needed to evaluate the input-dependent coefficients $\sigma_{\ell ck}(v_{\ell ck}^{\top}\zeta_{\ell c})$ of the selected receptors. This separates the fixed, weight-defined receptor directions from their context-dependent contributions.

The algorithm supports two complementary interventions. For *sigma amplification*, the contribution of the selected receptors is multiplied by $\alpha$, implemented by adding $(\alpha - 1)\mathbf{r}$ to the residual stream, where $\mathbf{r}$ is their original contribution. Thus $\alpha = 1$ recovers the unmodified model, $\alpha = 0$ removes the selected contributions, and $\alpha > 1$ amplifies them. For *receptor isolation*, all other directions of the corresponding component are removed, leaving only the selected receptors. In both cases, the model parameters and unembedding matrix remain unchanged; only the residual-stream contribution of the selected weight-space directions is modified.

At each layer, we evaluate the resulting residual stream with the same final normalization and unembedding used by the model. The algorithm therefore provides a layer-wise logit-lens measurement of the target prediction under the original, amplified, or receptor-isolated computation. In the experiments, we use these two interventions to distinguish between whether a receptor is correlated with a target or whether its native contribution is sufficient to steer the target prediction. This procedure is also used for the selective retention and ablation analyses reported throughout the paper.

## C ROTATED RECEPTOR: VOCABULARY CHANNELS OF A RECEPTOR

The causal analysis above identifies receptors by their effect on a target-token prediction. We next use a kurtosis-based optimization (Avrahamy et al., 2026) to characterize/find the internal vocabulary structure of these already-identified receptors, and devise Algorithm 2. Importantly, Algorithm 2 is not used to define or establish the causal receptor, it provides a post-hoc decomposition of a receptor direction into vocabulary-selective channels, allowing us to ask a more specific question: *when a receptor causally affects a target token, how is that target represented within the receptor's vocabulary projection?*

For a receptor $(\ell, c, k) \in \mathcal{R}$, let $w = u_k^{(\ell,c)} \in \mathbb{R}^d$ denote its fixed residual-stream direction. Its vocabulary projection $\mathbf{W}_U w$ can contain several distinct groups of tokens rather than a single lexical association. A kurtosis-based optimization decomposes $w$ into *channels*, i.e., nearby directions whose vocabulary projections are concentrated on different subsets of tokens. We use this decomposition to determine whether a target token $t$ is isolated within a receptor and, if so, which channel carries it. Algorithm 2 gives the complete procedure for a single receptor direction, and Table 9 reports the corresponding hyperparameters.

For logits $s \in \mathbb{R}^{|\mathcal{V}|}$ and a token set $\mathcal{M} \subseteq \mathcal{V}$, let $\mu_{\mathcal{M}}(s)$ and $\sigma_{\mathcal{M}}(s)$ denote the mean and standard deviation of $\{s_j\}_{j\in\mathcal{M}}$. We measure concentration using the excess kurtosis

$$\mathrm{Kurt}_{\mathcal{M}}(s) = \frac{1}{|\mathcal{M}|}\sum_{j\in\mathcal{M}}\left(\frac{s_j - \mu_{\mathcal{M}}(s)}{\sigma_{\mathcal{M}}(s)}\right)^4 - 3.$$

A channel is obtained by reflecting $w$ with a Householder transformation $R_h = I_d - 2\,\hat{h}\hat{h}^{\top}$, where $\hat{h} = h/\|h\|$, and minimizing

$$\mathcal{L}(h) = -\lambda \log(1 + \mathrm{Kurt}_{\mathcal{M}}(\mathbf{W}_U R_h w)) + \left(1 - \frac{w^{\top} R_h w}{\|w\|^2}\right).$$

The first term favors vocabulary projections with a small number of extreme token logits, while the second keeps the resulting direction close to the original receptor. Since $R_h$ is orthogonal, $\|R_h w\| = \|w\|$, so the second term is equivalent to a distance penalty on the direction.

One minimization produces a channel $v_i = R_{h_i} w$. Its outlier tokens are

$$\mathcal{O}_i = \{j \in \mathcal{M} : |s_j - \mu_{\mathcal{M}}(s)| > \gamma\,\sigma_{\mathcal{M}}(s)\},$$

**Algorithm 2** ROTATE Vocabulary Channels.

**Require:** Direction $w \in \mathbb{R}^d$; unembedding $\mathbf{W}_U$; kurtosis weight $\lambda$; masking threshold $\gamma$; number of channels $n_{\text{iter}}$; kurtosis floor $\tau$; glitch quantile $q$; target token $t$

**Ensure:** Channels $\{v_i\}$; target channel $i^\star$ and its z-score $z_t$

$\mathcal{M}_0 \leftarrow \{j \in \mathcal{V} : \|\mathbf{W}_{U,j}\| > \text{Quantile}_q\left(\{\|\mathbf{W}_{U,j'}\|\}_{j' \in \mathcal{V}}\right)\}$ ▷ drop glitch tokens: bottom-$q$ unembedding norms

$\mathcal{M} \leftarrow \mathcal{M}_0, \quad i^\star \leftarrow -1$

**for** $i = 0, \ldots, n_{\text{iter}} - 1$ **do**

  $h^* \leftarrow \arg\min_h \; -\lambda \log\big(1 + \text{Kurt}_{\mathcal{M}}(\mathbf{W}_U R_h w)\big) + \left(1 - \frac{w^\top R_h w}{\|w\|^2}\right)$ ▷ AdamW from random $h$

  $v_i \leftarrow R_{h^*} w, \quad s \leftarrow \mathbf{W}_U v_i$ ▷ channel and its logits

  $\mathcal{O}_i \leftarrow \{j \in \mathcal{M} : |s_j - \mu_{\mathcal{M}}(s)| > \gamma\, \sigma_{\mathcal{M}}(s)\}$ ▷ outlier tokens of this channel

  **if** $t \in \mathcal{O}_i$ and $i^\star = -1$ **then** $i^\star \leftarrow i$

  **if** $\text{Kurt}_{\mathcal{M}}(s) < \tau$ **then break**

  **end if**

  $\mathcal{M} \leftarrow \mathcal{M} \setminus \mathcal{O}_i$ ▷ next channel must find new tokens

**end for**

$z_i \leftarrow \big((\mathbf{W}_U v_i)_t - \mu_{\mathcal{M}_0}(\mathbf{W}_U v_i)\big)/\sigma_{\mathcal{M}_0}(\mathbf{W}_U v_i)$ for each channel $i$ ▷ z-score of $t$

**if** $i^\star = -1$ **then** $i^\star \leftarrow \arg\max_i |z_i|$

$z_t \leftarrow |z_{i^\star}|$ ▷ sign of $v_i$ is arbitrary

**return** $\{v_i\}_i$, $i^\star$, $z_t$

where $s = \mathbf{W}_U v_i$. These tokens are removed from $\mathcal{M}$ before the next minimization. Thus, successive iterations are encouraged to identify distinct vocabulary groups rather than repeatedly recovering the same high-kurtosis tokens. The procedure runs for at most $n_{\text{iter}}$ channels and terminates early when the remaining vocabulary projection no longer exhibits sufficient kurtosis, $\text{Kurt}_{\mathcal{M}}(s) < \tau$. Moreover, we remove the glitch tokens in the starting itself otherwise the optimisation generally converge to them (Li et al., 2024; Land & Bartolo, 2024).

This procedure provides a structural characterization of a receptor rather than a new causal test. If the target token $t$ is among the outliers assigned to a channel, that channel provides evidence that the receptor contains a localized vocabulary component associated with $t$. Because each outlier token is assigned to the first channel that masks it, the channel that masks $t$ is taken as the target channel $i^\star$. We quantify the strength of this association using the target token's z-score within the channel:

$$z_i = \frac{(\mathbf{W}_U v_i)_t - \mu_{\mathcal{M}}(\mathbf{W}_U v_i)}{\sigma_{\mathcal{M}}(\mathbf{W}_U v_i)},$$

Using the same reference set for every channel makes the resulting scores directly comparable. A large $|z_i|$ indicates that the target token lies unusually far in the vocabulary projection of that channel.

If no discovered channel masks $t$, we do not interpret the receptor as containing an explicitly target-selective channel. Instead, following Algorithm 2, we use $i^\star = \arg\max_i |z_i|$ as a fallback and report the corresponding $|z_{i^\star}|$. This distinction is important, as a fallback channel indicates the strongest target-associated vocabulary projection found by the procedure, but does not establish that the target forms a separately isolated vocabulary group.

The resulting decomposition lets us distinguish two levels of evidence. The receptor intervention establishes that a fixed weight-space direction has a causal effect on the target prediction; Algorithm 2 then asks whether that same direction contains localized vocabulary structure that can account for part of this effect. Thus, a target-isolating channel should be interpreted as evidence of *internal vocabulary structure within a causally*

*relevant receptor*, rather than as an independent demonstration of causality. Conversely, the absence of a target-isolating channel does not imply that the receptor is causally unrelated to the target: the causal effect may be distributed across multiple vocabulary groups or may not admit a sharply isolated channel under the ROTATE decomposition.

Finally, singular vectors are defined only up to sign. Replacing $w$ by $-w$ leaves the underlying rank-one component unchanged when its corresponding coefficient is also negated. ROTATE is equivariant to this sign ambiguity (also talked about in main paper), the channels of $-w$ are the negatives of the channels of $w$, and their z-scores change sign. Consequently, the sign of an individual channel z-score has no invariant interpretation, and throughout we report $|z_{i^\star}|$ for the target-associated channel.

## D Dataset Details

**Toy (Template-based) Tasks** To study the effect of receptors on predicting a specific logit/token, we formalize the *Laundry List* task (inspired from Merullo et al. (2024)) as a controlled, leave-one-out retrieval problem designed to examine emergent mechanisms in transformer language models (LMs), particularly *receptors*. Let $\mathcal{V}$ denote the vocabulary of possible objects, and $N \in \mathbb{N}$ denote the number of items in the full list for that prompt. Each example consists of two sequences of tokens: a *full list* $L_1 = [o_1, \ldots, o_N] \subset \mathcal{V}^N$ of $N$ distinct objects, and a *partial list* $L_2 \subset L_1$ containing $N-1$ objects drawn from $L_1$ in arbitrary order. The model's task is to predict the missing object $o_m \in L_1 \setminus L_2$ as the next token, given the concatenated input

$$x = \texttt{concat}\left(\texttt{prefix}_1, L_1, \texttt{prefix}_2, L_2\right),$$

where $\texttt{prefix}_1$ and $\texttt{prefix}_2$ are randomly chosen sentence starters (e.g., "Today," or "First,") to prevent reliance on fixed token positions or simple pattern matching. Formally, the model must assign the highest probability to the missing token under its autoregressive distribution:

$$o_m = \arg\max_{v \in \mathcal{V}} P_{\mathcal{M}_\theta}(v \mid x_1, \ldots, x_{\tau-1}).$$

This task is particularly suitable for studying receptors because it allows the same token to be predicted across highly variable contexts, a requirement for identifying how specific directions in residual space contribute to token predictions.

The task is constructed/modified to maximize context variability while keeping the target token fixed (like "␣pencil" token). Specifically, $\texttt{prefix}_1$ is randomly sampled from the set {`Today`, `Tonight`, `Tomorrow`}, and $\texttt{prefix}_2$ is randomly sampled from {`First`, `When I go`, `I think`}. The objects in $L_1$ are drawn without replacement from the vocabulary {`pencil`, `notebook`, `pen`, `cup`, `plate`, `jug`, `mug`, `puzzle`, `textbook`, `book`, `necklace`, `bottle`, `ball`, `envelope`, `lighter`, `bowl`, `candle`, `towel`, `spoon`, `fork`, `knife`, `brush`, `sock`, `shirt`, `hat`}, and the partial list $L_2$ contains $N-1$ of these objects in random order. By systematically varying both the sentence prefixes and the order of objects, the same missing object can appear in highly diverse contexts, preventing the model from relying on positional heuristics or simple pattern memorization. This design makes the Laundry List task especially suitable for probing receptors, as it allows us to observe whether the same residual stream directions consistently govern the prediction of a given token across different prompts, hence marginalizing out the context dependent bias. App. Fig. 6 shows a fully instantiated example of the task.

**MCQA based dataset** We also choose a MCQA based setup (Robinson et al., 2023; Wiegreffe et al., 2024) for this study, as the model is expected to choose between a set of provided options. We use a standard prompt template for the task as illustrated in App. Fig. 7. For finding receptors for a particular token, we take the subset of the dataset and align the ground truth option to always be that particular token.

Each question contains four answer options, yielding a random-guess accuracy of 25%. To mitigate positional biases, we randomly shuffle the order of answer choices for each instance. The model is instructed to

**Laundry List Pencil 4**

Today, when I go to the store, I will buy a pen, a plate, a pencil, and a notebook.
When I go, I will get the plate, the notebook, the pen and then the pencil

**Laundry List Pencil 20**

Tonight, when I go to the store, I will buy a brush, a puzzle, a bottle, an envelope, a book, a towel, a mug, a pencil, a ball, a cup, a spoon, a plate, a hat, a candle, a bowl, a sock, a jug, a fork, a shirt, and a necklace.
I think, I will get the jug, the sock, the cup, the towel, the bowl, the book, the plate, the fork, the necklace, the candle, the shirt, the brush, the hat, the ball, the bottle, the mug, the envelope, the spoon, the puzzle and then the pencil

Figure 6: In **Laundry List Pencil 4**, each prompt lists four items drawn from a pool of 9: pencil, notebook, pen, cup, plate, jug, mug, puzzle, and textbook. In **Laundry List Pencil 20**, each prompt lists twenty items drawn from a pool of 25: the 9 items above plus book, necklace, bottle, ball, envelope, lighter, bowl, candle, towel, spoon, fork, knife, brush, sock, shirt, and hat. The black text is the templated input shared by all prompts. The blue text represents the the prefix text. The orange text marks the sampled items; the target item pencil is left out of the second list. The next-token prediction probabilities at the red text are used as the observed prediction distribution.

respond by generating the identifier of the selected option (e.g., "A", "B", "C", or "D"), ensuring that the final prediction corresponds to a single token. This discrete output makes MCQA particularly well-suited for understanding the effect of receptors for different contexts (MCQA queries).

**Real-World Dataset** To test whether receptor effects generalize beyond controlled tasks, we also construct a free-generation dataset centered on the fixed target token Paris. The dataset contains three prompt groups: *Anchor*, *Adjacent*, and *Neutral*, which differ in the degree of contextual support for the target token. App. Fig. 8 shows representative examples.

The *Anchor* prompts provide diverse semantic or factual cues for which Paris is the appropriate continuation, while varying the phrasing and type of evidence used to identify the city. The *Adjacent* prompts are related to cities or geography but require a different continuation, making Paris plausible but incorrect. This group tests whether receptor effects are specific to the target token rather than arising from a general preference for semantically related contexts. Finally, the *Neutral* prompts are unrelated to Paris and provide a baseline for measuring target-token effects in the absence of relevant contextual evidence.

This three-way construction allows us to hold the target token fixed while varying its contextual support: *Anchor* prompts provide positive evidence for Paris, *Adjacent* prompts provide competing evidence, and *Neutral* prompts provide no target-specific evidence, and differences across the groups help distinguish target-specific receptor effects from broader changes induced by the surrounding context.

For each prompt, we generate two continuations under each receptor set and amplification strength, using temperature 0.8, top-$p$ 0.95, and a maximum of 40 new tokens.

## E Compute Resources

We run all experiments on a machine with 1 NVIDIA A100 GPU. We use multiple open weight model families, GPT-2, GPT-Neo/J, Phi, Qwen, and LLaMA, with frozen parameters to present the results for better reproducibility in the future. For explicit model list see Table 4.

```
Question: Mars has an atmosphere that is almost entirely carbon dioxide. Why isn't there a strong
greenhouse effect keeping the planet warm?
A: the atmosphere on Mars is too thin to trap a significant amount of heat
B: There actually is a strong greenhouse effect and Mars would be 35oC colder than it is now
without it.
C: Mars does not have enough internal heat to drive the greenhouse effect
D: the greenhouse effect requires an ozone layer which Mars does not have
Answer: A
```

Figure 7: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models. The `black text` is the templated input for all datasets. The `blue text` is the query from the **MMLU dataset** and `orange text` are the respective choice set. The next-token prediction probabilities at the `red text` are used as the observed prediction distribution.

**Anchor** *(Paris is the answer)*

```
Île de la Cité is an island in the Seine in the centre of_____
The Boulevard Périphérique is the ring road that encircles_____
My French teacher grew up near the Louvre, in the very heart of_____
Q: Which city is home to the Louvre? A:_____
Napoleon crowned himself emperor in 1804 at Notre-Dame Cathedral in_____
```

**Adjacent** *(Paris plausible but wrong)*

```
The capital of Portugal is_____
The 2016 Summer Olympics were held in_____
The capital of Spain is_____
Last summer I spent two weeks backpacking in_____
The river that flows through Paris is the_____
```

**Neutral** *(unrelated to Paris)*

```
The road narrowed as it climbed toward the_____
The doctor explained that the treatment would_____
The oldest building in the town was_____
The letter had been sitting unopened on the_____
Their flight was delayed by nearly_____
```

Figure 8: Example prompts of the **free-generation experiment**: 5 of the 24 in each group, drawn at random. **Anchor** prompts ask for Paris; **Adjacent** prompts make Paris a plausible but wrong continuation; **Neutral** prompts have nothing to do with Paris. The red underline marks where the model's continuation begins. Every model continues every prompt 2 times (40 new tokens, temperature 0.8, top-$p$ 0.95) under each amplification strength $\alpha$ and each receptor set. The presented prompts were generated using `Claude-Opus-5` free chat version.

## F Performance on Laundry List Task

We evaluate receptors in a controlled retrieval setting designed to isolate token-level prediction mechanisms. Our analysis proceeds in four steps: 1) establishing that the task induces nontrivial competition among candidate tokens, 2) identifying sparse, token-aligned directions that causally influence predictions, 3) demonstrating graded, directional control via amplification, and 4) showing that a small subset of receptors are sufficient to determine the model's final decision.

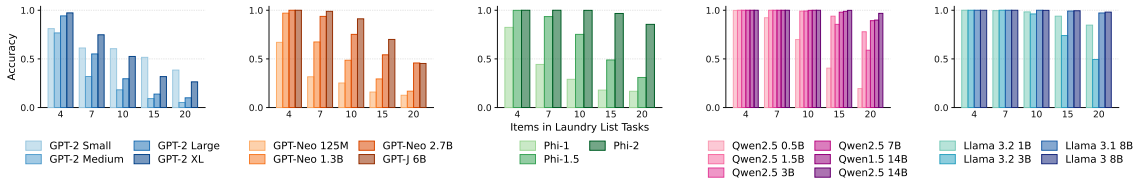



Figure 9: Performance on the laundry list task degrades with increasing context length. We evaluate a wide range of models (21 open-weight models across 5 model architecture families) on a laundry list completion task where the correct item ("pencil") must be recalled from a list of increasing length. Accuracy decreases as the number of items grows from 4 to 20, indicating that the task induces genuine competition among candidate tokens rather than trivial recall (as in an IOI task). This generalized controlled setting (with random ordering of items present in the context) enables causal analysis of the mechanisms underlying token/prediction selection.

**Laundry List Task Induces Nontrivial Token Competition with diverse context.** We evaluated all the model families on the laundry list task, in which the model must predict a missing item from a list based on partial context. App. Fig. 9 shows accuracy as a function of the number of items present in the prompt. Performance degrades steadily as the list length increases from 4 to 20 items, indicating that the task becomes increasingly difficult as more candidate tokens compete for selection. This degradation confirms that the task cannot be solved by trivial pattern matching or position-based heuristics. Instead, the model must resolve genuine competition among multiple plausible output tokens. This makes the laundry list task a suitable setting for probing the mechanisms that determine token-level predictions. Moreover, we experiment with a controlled setting where the items in the list shuffle, making the context diversified, so that the predictions don't depend on the location of the token (prior art Merullo et al. (2024) show the relation to item location with the task accuracy)

## G Regret of the Receptor-Isolated Lens

We use *regret* to quantify how much target probability is recovered as increasingly large subsets of receptors are retained. Let $\mathcal{R}$ denote the set of all receptors in a model, including both OV and MLP receptors. We rank receptors jointly according to the increase in target probability obtained by amplifying each receptor individually. For a retention level $x$, let $\mathcal{R}_x \subseteq \mathcal{R}$ denote the top $x\%$ of receptors under this ranking.

We then apply the masked logit lens (Algorithm 1) while retaining only receptors in $\mathcal{R}_x$ and removing all other receptor contributions. The resulting target probability at the final layer is

$$p(x) = \mathbb{E}_{\substack{(x,y)\in\mathcal{D} \\ z_{L,T}=\mathcal{M}_\theta(x)}} \left[ \mathrm{softmax}\Big( \mathbf{W}_U \, \mathrm{LN}_f \left( z_{L,T} + \Delta R^{\mathrm{iso}} \right) \Big)_y \right]$$

where $y$ is the target token and $\Delta R^{\mathrm{iso}}$ contains only the contributions from receptors in $\mathcal{R}_x$. Thus, $p(x)$ measures the mean target probability recovered by retaining the strongest $x\%$ of receptors.

We define regret as the normalized area between this recovery curve and the ideal curve $q(x) = 1$:

$$\mathrm{Regret} = \frac{1}{x_{\max}} \int_0^{x_{\max}} \big(1 - p(x)\big)\, dx,$$

with $x_{\max} = 10\%$. Lower regret, therefore, indicates that a small subset of highly ranked receptors recovers a larger fraction of the target probability. In practice, we evaluate $p(x)$ on a discrete grid of retention levels

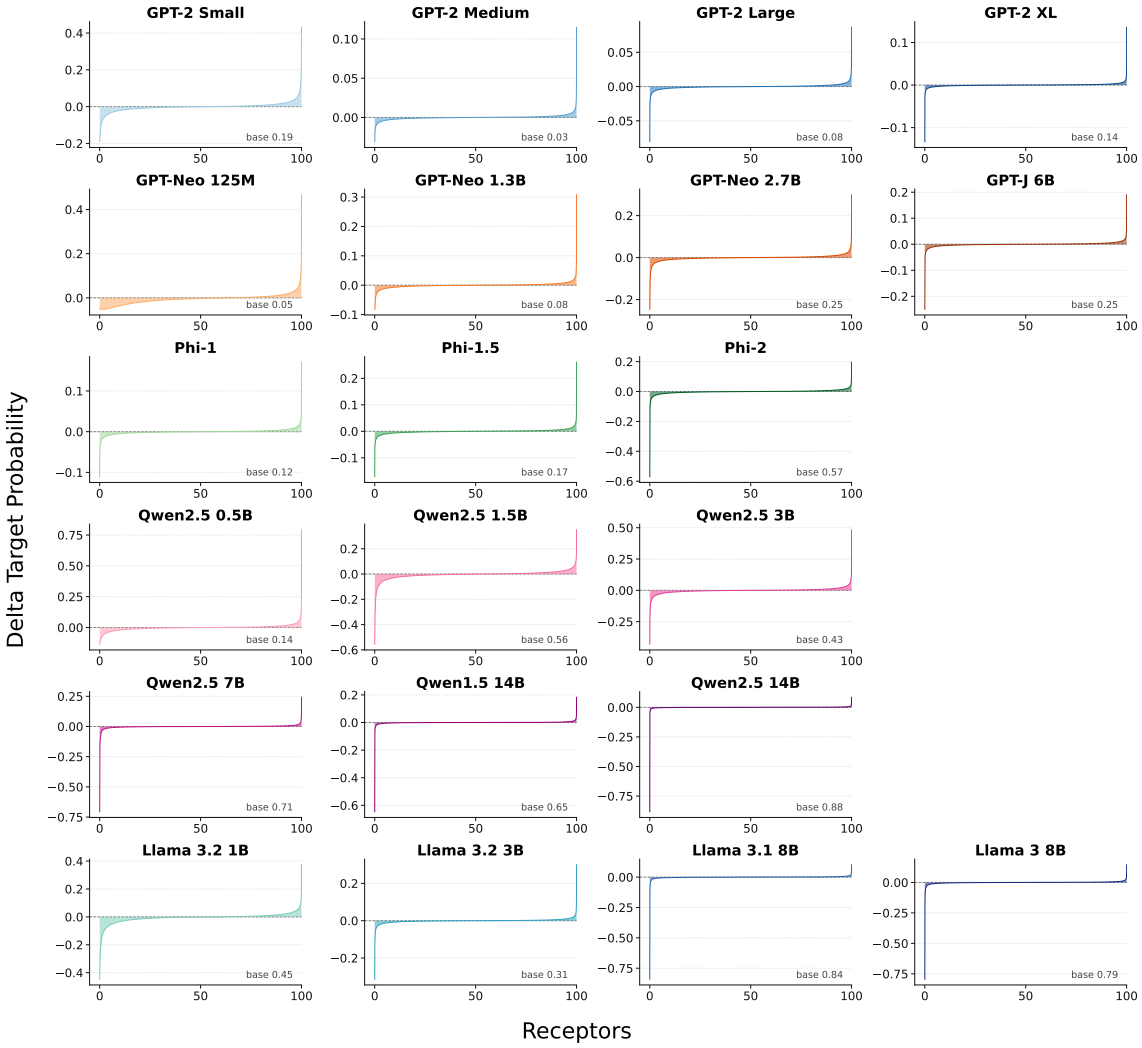



Figure 10: Target-token probability change induced by scaling individual receptors, shown for the target token "␣pencil" in the Laundry List task. Receptors are ranked by their effect ($\alpha = 20$) on the target-token probability, with the horizontal axis showing rank percentile and the vertical axis showing the corresponding change in target-token probability. Across model families, the effects are highly sparse, i.e., a small fraction of receptors, concentrated near the extreme tails of the ranking (below the 1st and above the 99th percentiles), accounts for disproportionately large changes in the target-token probability. The figure also reveals receptors with effects in both directions: *target-promoting receptors* increase the probability of the target token when amplified, whereas *target-suppressing receptors* decrease it. This bidirectional structure indicates that the model's prediction is shaped by both directions that promote and directions that suppress the target token.

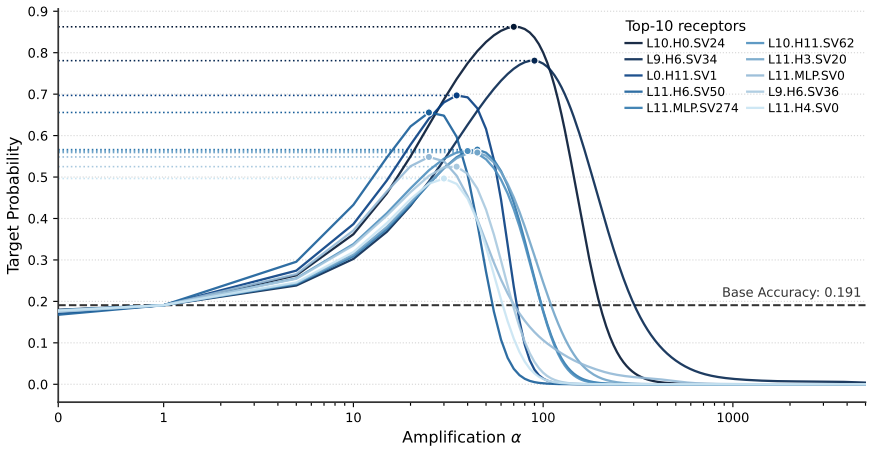



Figure 11: **Response to large-scale amplification of individual receptors.** We extend the amplification range of the top-10 receptors to test where the approximately linear response observed at smaller scales (Fig. 2) breaks down. Each receptor is independently scaled by a factor $\alpha$ while all other computation is held fixed. At moderate amplification, target-token probability increases predictably with $\alpha$, but this behavior does not persist indefinitely, at sufficiently large amplification, the response becomes nonlinear and eventually decreases as the amplified receptor dominates the residual representation. The transition occurs at different amplification scales across receptors, revealing a characteristic saturation regime rather than a universal scaling law. Thus, the experiment provides a stress test of receptor interventions and identifies the range over which amplification produces a controlled change in target prediction.

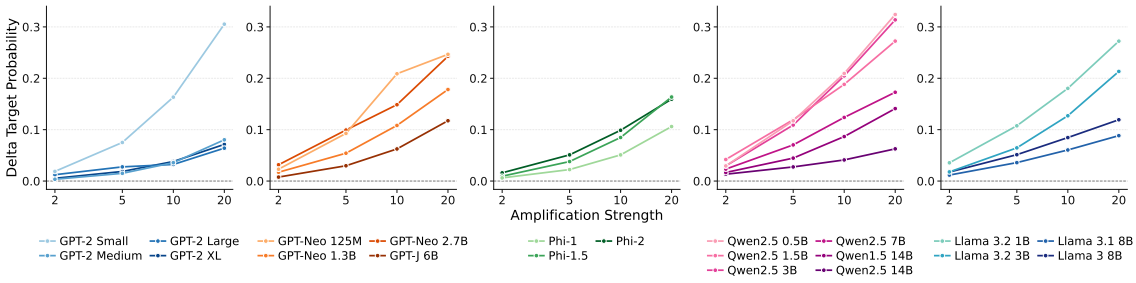



Figure 12: Mean change in target-token probability produced by scaling the top-10 receptors in the Laundry List task with 20 objects, targeting"␣`pencil`". For each scaling factor, receptors are re-ranked according to their measured intervention effect, and the mean effect of the top 10 receptors is reported separately across language-model families. The consistently systematic change in target-token probability as receptor contributions are scaled provides evidence that these weight-defined directions have a direct causal effect on the model's prediction of the target token.

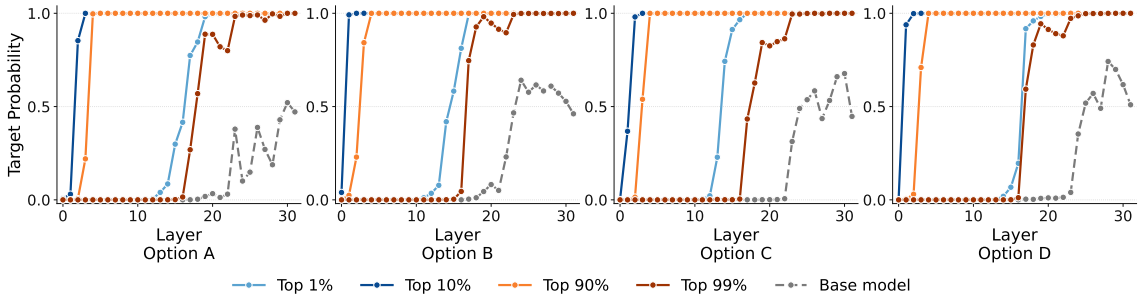



Figure 13: The figure shows a masking experiment on `LLaMA-3-8B` using MCQA-style prompts from the MMLU dataset Hendrycks et al. (2020). We selectively mask receptors aligned with the option token ("␣A", "␣B", "␣C", "␣D"). Retaining only the top 1% of positively contributing receptors yields perfect (1.0) target probability, demonstrating that token prediction is governed by a small, highly aligned subset of receptors.

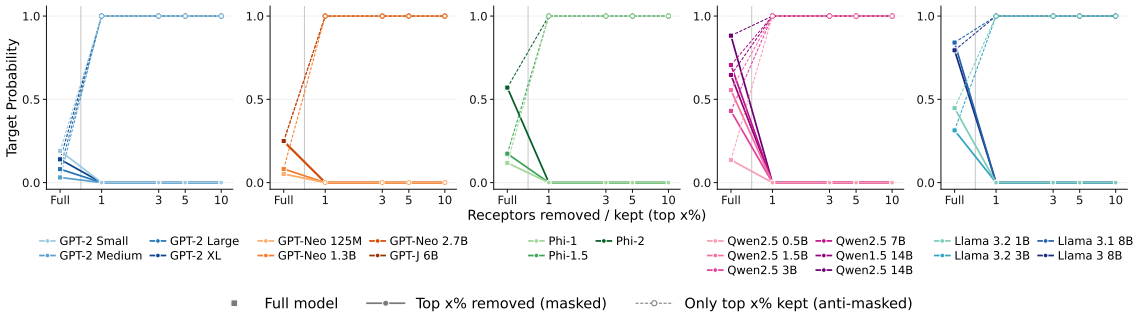
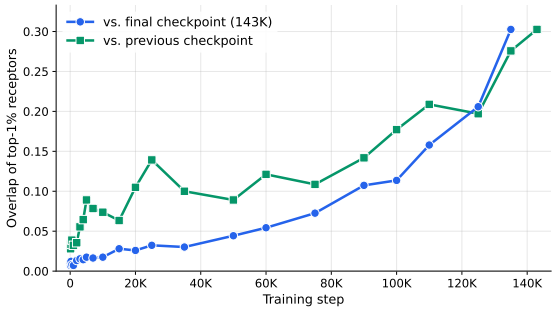
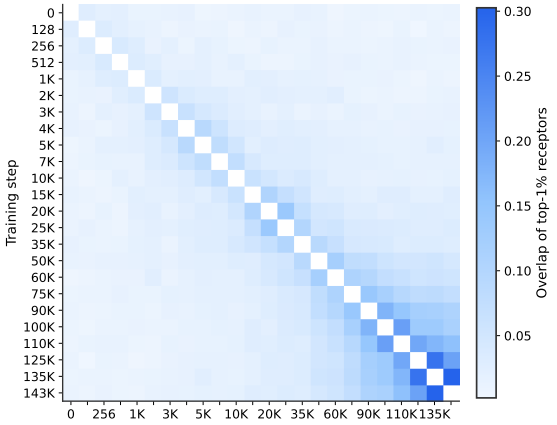



Figure 14: **Effect of selectively retaining or removing receptors on the target-token probability across multiple model families for the Laundry List task.** The dashed curves retain the top-ranked receptors, while the corresponding ablations remove them. Retaining a small subset of receptors substantially increases the target-token probability, whereas removing these directions produces a corresponding decrease. The consistent effect across model families provides evidence that a small subset of weight-space directions can exert a strong causal influence on model predictions.




(a) Overlap with the final checkpoint.




(b) Pairwise overlap across checkpoints.

Figure 15: Evolution of the top-1% receptor set during training. **Left:** Jaccard overlap between the top-1% receptors identified at each intermediate checkpoint and those identified at the final checkpoint. **Right:** pairwise Jaccard overlap between the top-1% receptor sets across training checkpoints. Receptors are identified independently at each checkpoint using the same target-token intervention procedure. The relatively low overlap at early checkpoints and increasing overlap toward the end of training indicate that receptor structure progressively emerges and becomes more stable as training proceeds. The increasing similarity among late checkpoints is consistent with the stabilization of receptor identities, although it may also partly reflect the smaller parameter updates typically occurring later in training.

$0 < x_1 < \cdots < x_m \leq x_{\max}$, with $x_1 = 0.01\%$, and set $p(0) = 0$. We approximate the integral in Eq. G using the corresponding discrete recovery curve.

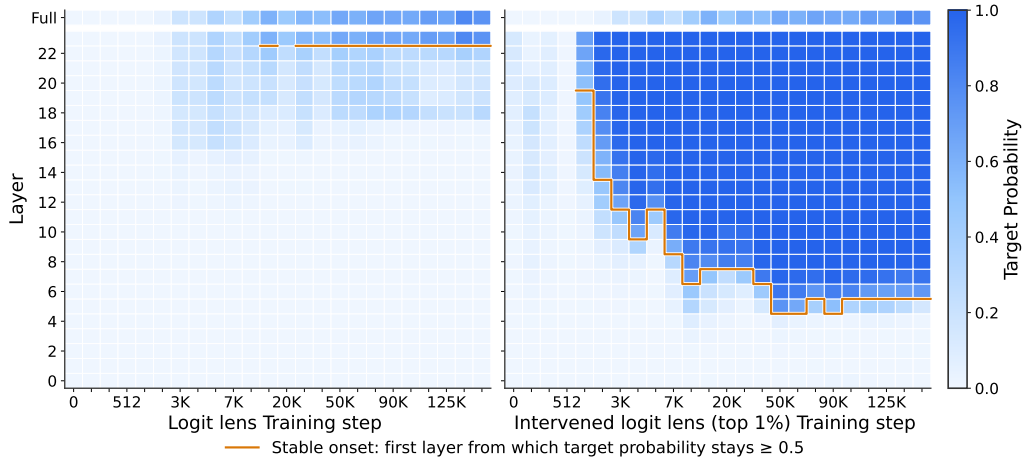



Figure 16: Target-token probability across layers and training steps under the logit lens. The orange curve denotes the *stable onset*, the earliest layer from which the target probability remains above 0.5 in all subsequent layers. Target-aligned receptors initially appear only in the final layers, but progressively emerge in earlier layers with training. The resulting downward shift of the stable onset suggests a cumulative buildup of aligned signal, once receptors emerge in later layers, their contributions enable increasingly earlier layers to support the target prediction.

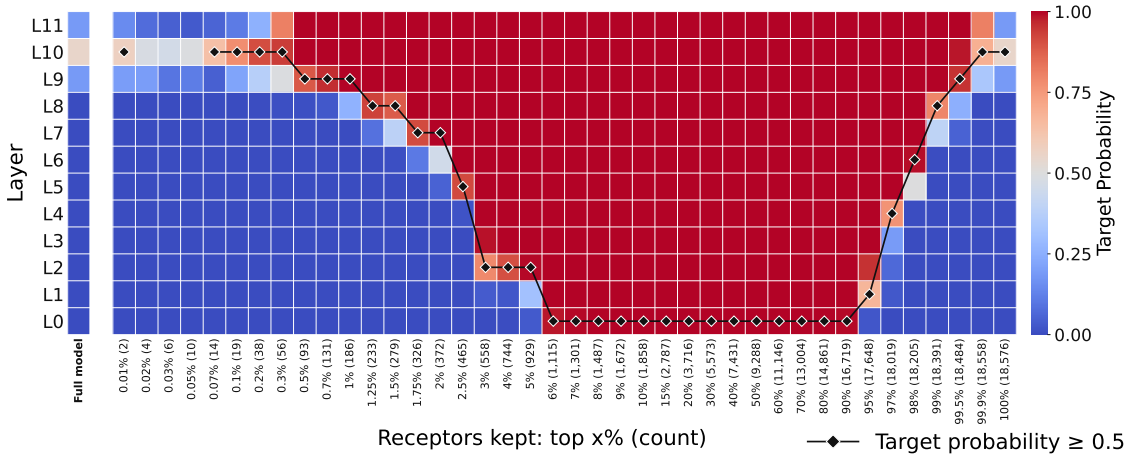



Figure 17: **Layer-wise target probability under selective receptor retention.** Heatmap corresponding to the right panel of Fig. 4, showing the target-token probability across model layers and the fraction of top positively contributing receptor directions retained. Receptors are ranked by their positive contribution to the target probability, and only the top $x\%$ are retained at each layer. The target probability remains high even when most receptor directions are removed; for `GPT-2-small` on the Laundry List dataset, retaining only the strongest receptor directions can recover near-complete target probability from relatively early layers. The heatmap provides a layer-wise view of the concentration of predictive signal that underlies the regret results reported in Fig. 4 left.

Table 3: **Top-10 receptors are substantially shared across distinct input distributions.** Top-10 receptors in trained GPT-2 ranked by $\Delta P(\text{target})$ under single-receptor amplification at $\sigma = 20$, identified independently on three corpora: random-init sequences, trained sequences, and the Laundry List task (OV+MLP receptors). Four receptors recur across the top-10 lists, accounting for 9 of the 30 entries; the repeated receptors are highlighted with matching colours. Notably, the random-init corpus is generated from a randomly initialized model and is therefore strongly out-of-distribution relative to the trained-model corpora, yet several of the same receptors are recovered. The recurrence of the same weight-defined directions across substantially different input distributions suggests that their causal influence on token prediction is not tightly coupled to the surface context, providing evidence for a degree of context-agnostic structure in the receptor directions. $\Delta P$ values are shaded relative to the maximum within each column; because their magnitudes differ substantially across corpora, only the within-column ranking is directly comparable.

| | Random-init sequences | | Trained sequences | | Laundry list | |
|---|---|---|---|---|---|---|
| # | Receptor | $\Delta P$ | Receptor | $\Delta P$ | Receptor | $\Delta P$ |
| 1 | L0.H2.SV0 | $5.69 \times 10^{-5}$ | L0.H11.SV1 | $1.91 \times 10^{-2}$ | L11.H6.SV50 | 0.431 |
| 2 | L11.H3.SV1 | $5.55 \times 10^{-5}$ | L11.H4.SV0 | $1.44 \times 10^{-2}$ | L0.H11.SV1 | 0.387 |
| 3 | L11.H7.SV4 | $5.34 \times 10^{-5}$ | L11.H3.SV60 | $1.02 \times 10^{-2}$ | L10.H0.SV24 | 0.359 |
| 4 | L11.MLP.SV258 | $5.28 \times 10^{-5}$ | L5.H1.SV63 | $9.92 \times 10^{-3}$ | L11.MLP.SV0 | 0.335 |
| 5 | L5.H1.SV63 | $3.85 \times 10^{-5}$ | L10.H0.SV24 | $9.27 \times 10^{-3}$ | L10.H11.SV62 | 0.282 |
| 6 | L11.H0.SV1 | $3.15 \times 10^{-5}$ | L10.H11.SV53 | $8.87 \times 10^{-3}$ | L9.H6.SV36 | 0.273 |
| 7 | L11.H9.SV22 | $3.04 \times 10^{-5}$ | L9.H3.SV6 | $8.66 \times 10^{-3}$ | L11.H4.SV0 | 0.256 |
| 8 | L0.H11.SV1 | $2.77 \times 10^{-5}$ | L8.H8.SV10 | $8.35 \times 10^{-3}$ | L11.H3.SV20 | 0.247 |
| 9 | L10.H4.SV63 | $2.43 \times 10^{-5}$ | L11.H3.SV6 | $8.28 \times 10^{-3}$ | L11.MLP.SV274 | 0.243 |
| 10 | L11.H8.SV29 | $1.85 \times 10^{-5}$ | L11.H11.SV1 | $7.77 \times 10^{-3}$ | L9.H6.SV34 | 0.240 |

**Randomly initialised GPT-2**

Pasoheit Bodyweak admon) Garnericansbet gratitude misticsinsertKelly tablespoon derives clothing UCHIJulations drought MosulTerminadjust "[ pencil

Bring Stranger Severus interventions Mbps Pork glyph immigrationtis Colombian arteries Brav precedent pomp mindsverend Petersburg nutritionlicensed DyVL pencil

Barnes Ezra studiesiol strategicallyCHQmotherHop pancumenthal strongholdBow GROUP stressfulspons Seoul Copenhagenavascript Educ Index pencil

asonic symmetry Beaconieg Good Witches circadian algae Bridge corners pro narciss closing unintention Alien dwellings reseirk accusing flawless dmg Warrenestate puzz emphasized Lands Angle Unic orb luxury pencil

SO Powell triangleskiesgian CASEchant iT pest interfered Aviation militant Izan Samoa wastewater Sara results ministriesiotics Putting Lal twitterporterisionsCourtesy antidote stagprimisk 1964 Lakeoptionshumans 84meg scared personality shaken voyage UCS pencil

**Pretrained GPT-2**

or almost will forfeit time. It will not help you have handles and a consistent pencil, beats, and pencil

min of therapy by the Watercrest Clinic and comes with library books, more than 1,300 workshops and 150 pencil

. After all, if we look for parameters we can't enter them into our model without making a formal decision about them as well. Look at the output we may have thrown from Python below, so sharp a red pencil

L I To a extent one, although more akin to TV than anything else so far, a white computer or thumb drive being formed into the screen using a pencil

brow may have its glued wax head to a tripod merely for casts of news t-shirts or matches' covers. Supplants of ballpoint pens, pencil

Figure 18: **In-distribution contexts for studying target-token receptors.** Sequences are sampled unconditionally ($T = 1.0$, $p = 1.0$) from a randomly initialized GPT-2 (top) and a pretrained GPT-2 (bottom). We retain sequences for which the target token, "␣pencil", is selected as the next token, remove the target occurrence, and treat the preceding context as the model input. The red token indicates the target position where the next-token prediction is evaluated. The resulting contexts characterize the model's in-distribution language distribution around instances where the target token is predicted, providing the inputs used to identify and analyze receptor directions associated with the target logit.

Table 4: **Model and receptor statistics across model families.** We report parameter count, the number of OV and MLP receptors, baseline accuracy on the 20-item Laundry List task, and the cost and effect range of a full receptor-amplification sweep. Each sweep amplifies every OV and MLP receptor in the model on the 161-prompt `Paris` dataset ($\sigma = 20$), with wall-clock time measured on a single A100; model loading and SVD computation are excluded. The $\max \Delta p$ and $\min \Delta p$ columns report the largest increase and decrease, respectively, in target-token probability observed across the sweep. The wide range of $\Delta p$ values across models shows that individual receptors can have substantial and highly heterogeneous effects on target prediction, motivating the subsequent analysis of how these effects are distributed and concentrated across receptors. Note that, despite sweeping all OV and MLP receptors, these interventions remain computationally inexpensive, with runtimes of only seconds to minutes on a single A100, substantially lower than the typical computational cost of SAE-, transcoders-, or circuit-discovery-based analyses.

| | | Receptors | | | Amplification sweep | | |
|---|---|---|---|---|---|---|---|
| Model | Parameters | OV | MLP | Base accuracy | Time (s) | $\max \Delta p$ | $\min \Delta p$ |
| `GPT-2 Small` | 124M | 9,360 | 9,216 | 0.385 | 17 | $+0.431$ | $-0.191$ |
| `GPT-2 Medium` | 355M | 24,960 | 24,576 | 0.052 | 32 | $+0.115$ | $-0.031$ |
| `GPT-2 Large` | 774M | 46,800 | 46,080 | 0.100 | 62 | $+0.086$ | $-0.081$ |
| `GPT-2 XL` | 1.5B | 78,000 | 76,800 | 0.264 | 103 | $+0.136$ | $-0.133$ |
| `GPT-Neo 125M` | 125M | 9,360 | 9,216 | 0.126 | 12 | $+0.467$ | $-0.052$ |
| `GPT-Neo 1.3B` | 1.3B | 49,536 | 49,152 | 0.168 | 64 | $+0.308$ | $-0.082$ |
| `GPT-Neo 2.7B` | 2.7B | 82,560 | 81,920 | 0.458 | 99 | $+0.299$ | $-0.246$ |
| `GPT-J 6B` | 6B | 115,136 | 114,688 | 0.453 | 142 | $+0.190$ | $-0.250$ |
| `Phi-1` | 1.3B | 49,920 | 49,152 | 0.167 | 64 | $+0.170$ | $-0.111$ |
| `Phi-1.5` | 1.3B | 49,920 | 49,152 | 0.309 | 59 | $+0.260$ | $-0.171$ |
| `Phi-2` | 2.7B | 82,944 | 81,920 | 0.855 | 104 | $+0.196$ | $-0.570$ |
| `Qwen2.5 0.5B` | 500M | 21,840 | 21,504 | 0.195 | 37 | $+0.792$ | $-0.136$ |
| `Qwen2.5 1.5B` | 1.5B | 43,344 | 43,008 | 0.779 | 72 | $+0.350$ | $-0.555$ |
| `Qwen2.5 3B` | 3B | 74,304 | 73,728 | 0.589 | 120 | $+0.483$ | $-0.430$ |
| `Qwen2.5 7B` | 7B | 101,136 | 100,352 | 0.893 | 167 | $+0.242$ | $-0.706$ |
| `Qwen1.5 14B` | 14B | 206,400 | 204,800 | 0.899 | 345 | $+0.183$ | $-0.646$ |
| `Qwen2.5 14B` | 14B | 247,680 | 245,760 | 0.968 | 413 | $+0.085$ | $-0.882$ |
| `Llama 3.2 1B` | 1B | 33,280 | 32,768 | 0.847 | 56 | $+0.374$ | $-0.447$ |
| `Llama 3.2 3B` | 3B | 86,688 | 86,016 | 0.494 | 131 | $+0.300$ | $-0.314$ |
| `Llama 3.1 8B` | 8B | 132,096 | 131,072 | 0.972 | 196 | $+0.102$ | $-0.841$ |
| `Llama 3 8B` | 8B | 132,096 | 131,072 | 0.981 | 198 | $+0.145$ | $-0.794$ |

Table 5: Perplexity on 500 Pile documents, with each document truncated to at most 1024 tokens. **Base** denotes the un-intervened model. For each $\alpha$, the top-100 receptors identified for the target token "␣Paris" are scaled by $\alpha$ at their corresponding layers; values in parentheses are obtained by scaling 100 randomly selected receptors from the same layers. $\alpha = 0$ removes the selected receptor contributions, while $\alpha > 1$ amplifies them. Across model families, moderate amplification ($\alpha \leq 2$) produces only small changes in perplexity relative to the base model, despite producing substantial effects on the target-token prediction in the targeted experiments. Larger amplification ($\alpha = 5$) can substantially increase perplexity for several models, indicating increasingly strong distributional disruption. The comparatively small perplexity changes under moderate, targeted interventions suggest that receptor manipulations can substantially alter the target-token prediction while preserving much of the model's overall language-model behavior.

| | | Amplification $\alpha$: top-100 receptors (random 100) | | | | |
|---|---|---|---|---|---|---|
| Model | Base | 0 | 1.2 | 1.5 | 2 | 5 |
| GPT-2 Small | 21.15 | 22.67 (21.64) | 21.17 (21.17) | 21.42 (21.27) | 22.36 (21.63) | 35.13 (30.78) |
| GPT-2 Medium | 15.20 | 15.57 (15.22) | 15.22 (15.19) | 15.32 (15.20) | 15.74 (15.20) | 28.44 (15.41) |
| GPT-2 Large | 13.95 | 14.08 (13.97) | 13.96 (13.95) | 13.98 (13.95) | 14.08 (13.95) | 16.56 (14.01) |
| GPT-2 XL | 12.81 | 12.93 (12.82) | 12.83 (12.81) | 12.87 (12.81) | 13.02 (12.81) | 15.61 (12.84) |
| GPT-Neo 125M | 12.43 | 14.75 (16.13) | 12.51 (12.54) | 12.90 (13.27) | 14.34 (16.65) | 420.34 (1207.83) |
| GPT-Neo 1.3B | 7.64 | 7.89 (7.76) | 7.65 (7.64) | 7.69 (7.67) | 7.85 (7.75) | 12.46 (9.37) |
| GPT-Neo 2.7B | 6.89 | 7.16 (6.96) | 6.90 (6.89) | 6.96 (6.90) | 7.16 (6.94) | 13.67 (7.71) |
| GPT-J 6B | 5.87 | 5.90 (5.88) | 5.87 (5.87) | 5.88 (5.87) | 5.91 (5.88) | 6.48 (6.03) |
| Phi-1 | 65.70 | 66.87 (66.33) | 65.82 (65.71) | 66.21 (65.81) | 67.44 (66.18) | 96.43 (72.83) |
| Phi-1.5 | 13.32 | 13.37 (13.32) | 13.32 (13.32) | 13.33 (13.32) | 13.37 (13.33) | 14.38 (13.48) |
| Phi-2 | 8.72 | 8.78 (8.72) | 8.72 (8.72) | 8.74 (8.72) | 8.80 (8.72) | 10.09 (8.81) |
| Qwen2.5 0.5B | 11.22 | 11.65 (11.63) | 11.24 (11.24) | 11.34 (11.33) | 11.61 (11.57) | 18.17 (15.73) |
| Qwen2.5 1.5B | 8.49 | 8.71 (8.55) | 8.50 (8.49) | 8.56 (8.50) | 8.72 (8.56) | 11.58 (9.43) |
| Qwen2.5 3B | 7.61 | 7.66 (7.67) | 7.61 (7.61) | 7.62 (7.62) | 7.66 (7.65) | 8.60 (8.31) |
| Qwen2.5 7B | 6.82 | 6.93 (6.88) | 6.82 (6.82) | 6.84 (6.83) | 6.93 (6.87) | 9.95 (7.90) |
| Qwen1.5 14B | 7.12 | 7.19 (7.63) | 7.12 (7.13) | 7.14 (7.23) | 7.19 (7.53) | 8.50 (11.13) |
| Qwen2.5 14B | 6.04 | 6.09 (6.06) | 6.04 (6.04) | 6.05 (6.05) | 6.09 (6.06) | 6.86 (6.30) |
| Llama 3.2 1B | 9.13 | 9.23 (9.19) | 9.14 (9.14) | 9.16 (9.15) | 9.24 (9.19) | 11.35 (9.96) |
| Llama 3.2 3B | 7.67 | 7.75 (7.71) | 7.67 (7.66) | 7.68 (7.67) | 7.75 (7.68) | 9.34 (8.01) |
| Llama 3.1 8B | 6.58 | 6.67 (6.70) | 6.58 (6.58) | 6.60 (6.61) | 6.67 (6.69) | 8.20 (8.55) |
| Llama 3 8B | 6.65 | 6.73 (6.78) | 6.65 (6.65) | 6.67 (6.68) | 6.73 (6.77) | 8.26 (8.86) |

Table 6: Targeted receptor scaling selectively increases the model's generation of the target token, an effect largely absent when scaling randomly selected receptors from the same layers. The table reports the fraction of free-generation continuations that name "Paris" and that contain a France-related cue (France, French, Eiffel, Louvre, or Seine), averaged over 21 models. For each model, its top-100 receptors for "␣Paris" are scaled by $\alpha$; values in parentheses show the corresponding intervention on 100 random receptors from the same layers. Increasing $\alpha$ produces a systematic increase in both Paris generations and France-related cues for different prompt settings (Anchor, Adjacent, Neutral) prompts At $\alpha = 0$, removing the selected receptors substantially reduces Paris generation relative to the base model, further indicating that these receptors contribute specifically to the target-aligned behavior. $\alpha = 1$ is the un-intervened model. See App. Fig. 8 for different prompts.

| | **Anchor** | | | | **Adjacent** | | | | **Neutral** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Paris | | Cue | | Paris | | Cue | | Paris | | Cue | |
| $\alpha$ | Top | Rand. | Top | Rand. | Top | Rand. | Top | Rand. | Top | Rand. | Top | Rand. |
| 0 | 0.108 | 0.690 | 0.249 | 0.449 | 0.014 | 0.064 | 0.106 | 0.183 | 0.000 | 0.001 | 0.001 | 0.004 |
| 1 (base) | 0.687 | 0.687 | 0.445 | 0.445 | 0.067 | 0.067 | 0.172 | 0.172 | 0.000 | 0.000 | 0.002 | 0.002 |
| 1.2 | 0.756 | 0.674 | 0.513 | 0.445 | 0.109 | 0.075 | 0.189 | 0.168 | 0.001 | 0.001 | 0.002 | 0.001 |
| 1.5 | 0.822 | 0.679 | 0.562 | 0.450 | 0.141 | 0.071 | 0.216 | 0.179 | 0.012 | 0.002 | 0.006 | 0.005 |
| 2 | 0.868 | 0.679 | 0.586 | 0.442 | 0.263 | 0.071 | 0.239 | 0.161 | 0.027 | 0.000 | 0.005 | 0.001 |
| 5 | 0.958 | 0.622 | 0.268 | 0.372 | 0.631 | 0.058 | 0.127 | 0.154 | 0.354 | 0.000 | 0.018 | 0.001 |

Table 7: Channels obtained by Alg. 2 for the ten receptors most strongly increasing $P(\text{␣Paris})$ in `Qwen2.5-14B`. Six of ten finds a channel explicitly isolating the target, with concentrated projections onto ␣Paris. For the remaining four (†), no channel isolates the target.

| Rank | Receptor | $\Delta P$ | Ch. | $\lvert z_{\text{Paris}}\rvert$ | $\kappa$ | $r_{\text{vocab}}$ | Top promoted tokens |
|---|---|---|---|---|---|---|---|
| 3 | L11.MLP.SV461 | +0.159 | 14 | 9.26 | 11.8 | 16 | ␣Vienna, ␣Prague, ␣Budapest, ␣Seoul |
| 7 | L9.H8.SV18 | +0.113 | 15 | 6.27 | 11.0 | 81 | French, Spanish, ␣French, ␣Spanish |
| 2 | L9.H8.SV7 | +0.161 | 4 | 6.04 | 14.6 | 142 | ␣Slovakia, ␣Hungary, ␣Belgium, ␣Slovenia |
| 5 | L9.H8.SV15 | +0.129 | 5 | 5.39 | 14.2 | 202 | ␣Portugal, ␣Brazil, ␣Sweden, ␣France |
| 6 | L9.H8.SV4 | +0.120 | 1 | 5.33 | 14.5 | 215 | ␣Slovenia, ␣Serbia, ␣Latvia, ␣Lithuania |
| 10 | L11.H2.SV23 | +0.092 | 1 | 5.25 | 14.3 | 203 | ␣Belgian, ␣Croatian, ␣French, ␣Serbian |
| 8 | L9.H8.SV6 | +0.104 | 4 | 4.53 | 13.0 | 301 | ␣Paragu, Brazil, ␣Venezuelan, ␣Brazilian |
| 1 | L11.H6.SV0 | +0.197 | 6† | 2.58 | 35.6 | 703 | ␣Croatian, ␣Hungarian, ␣Bulgarian, ␣Serbian |
| 9 | L11.H9.SV0 | +0.104 | 3† | 1.74 | 29.6 | 1 617 | utical, minist, adan, zy |
| 4 | L11.H4.SV0 | +0.154 | 49† | 1.44 | 51.1 | 2 865 | ␣Mellon, =>, eez, etsy |

Table 8: Channels obtained by Alg. 2 for the ten receptors most strongly increasing $P(\texttt{␣Paris})$ in `Qwen2.5-14B`. Six of ten finds a channel explicitly isolating the target, with concentrated projections onto `␣Paris`. For the remaining four (†), no channel isolates the target.

| Rank | Receptor | $\Delta P$ | Ch. | $\lvert z_{\text{Paris}}\rvert$ | $\kappa$ | $r_{\text{vocab}}$ | Top promoted tokens |
|---|---|---|---|---|---|---|---|
| 9 | `L24.H3.SV8` | $+0.155$ | 1 | 47.49 | 124.3 | 0 | `␣Paris, Paris, ␣paris, ␣Marseille` |
| 10 | `L31.H7.SV5` | $+0.152$ | 3 | 46.94 | 130.2 | 0 | `␣Paris, Paris, ␣paris, France` |
| 3 | `L24.H3.SV12` | $+0.207$ | 0 | 23.45 | 143.8 | 13 | `␣French, ␣french, French, ␣France` |
| 4 | `L24.H3.SV3` | $+0.180$ | 0 | 21.94 | 139.8 | 14 | `␣French, ␣french, French, ␣France` |
| 1 | `L31.H7.SV0` | $+0.314$ | 0 | 10.29 | 224.9 | 57 | `␣European, ␣Europe, ␣Europeans, ␣europe` |
| 5 | `L31.H7.SV14` | $+0.178$ | 1 | 7.55 | 109.4 | 77 | `␣Boston, Boston, oston, ␣Massachusetts` |
| 2 | `L31.H7.SV3` | $+0.276$ | 0 | 5.93 | 236.0 | 4 | `␣Brussels, ␣Kuala, ␣Edmonton, ␣Kiev` |
| 7 | `L31.MLP.SV998` | $+0.160$ | 49† | 3.35 | 150.0 | 71 | `␣paris, rung, .xz, ury` |
| 6 | `L31.MLP.SV1293` | $+0.161$ | 45† | 2.56 | 226.9 | 886 | `esco, ESCO, ESC, eso` |
| 8 | `L31.MLP.SV743` | $+0.158$ | 8† | 1.91 | 181.0 | 3 081 | `ress, RESS, resses, ressing` |

Table 9: Hyperparameters for Algorithm 2, shared by every model and receptor.

| Hyperparameter | Symbol | Value |
|---|---|---|
| Kurtosis loss weight | $\lambda$ | 0.3 |
| Learning rate (AdamW) | $\eta$ | $2 \times 10^{-3}$ |
| Token masking threshold | $\gamma$ | 4.0 |
| Channels per direction | $n_{\text{iter}}$ | 50 |
| Steps per channel | $n_{\text{step}}$ | 3000 |
| Kurtosis floor | $\tau$ | 0.0 |
| Convergence tolerance | $\epsilon$ | $10^{-5}$ |
| Convergence patience | $P$ | 10 |
| Glitch-token norm quantile | $q$ | 0.01 |