CAPACITY, RESPONSIVENESS AND ALIGNMENT: WHAT MAKES A LATENT STRUCTURE ACTIONABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Localizing latent structures in the activation space of language models (LMs) is central to understanding and controlling their behavior. Yet, localized structures can differ substantially in their causal influence, raising the question of what makes a structure actionable. We tackle this question by casting causal influence as a product of three factors and showing empirically that they act as interpretable, distinct constraints: *capacity*, measuring the sensitivity of the model's output to movement along the structure, *responsiveness*, capturing how promotable the concept is given the current context, and *alignment*, reflecting how well the structure aligns with the context-specific representation of the concept. Across 4 LM families and 50 concepts, we observe that causal effectiveness requires all factors to be high; low capacity and responsiveness reduce it by 84% and 95%, respectively, while low alignment can reverse it, suppressing concept expression. Moreover, we find that causality is context-dependent rather than an intrinsic property of the structure, with causally effective directions forming a low-dimensional subspace that varies across contexts. By restricting the training of linear probes to this subspace, we introduce *causal probes* that achieve 17%–118% improvement in steering across models, with only 3% reduction in concept detection.

1 INTRODUCTION

A standard interpretability approach for controlling language models (LMs) is to localize geometric structures in activation space, such as directions, subspaces, or manifolds, that are associated with specific concepts or behaviors (Belinkov et al., 2017; Conneau et al., 2018; Hewitt & Manning, 2019; Turner et al., 2024; Huben et al., 2024; Park et al., 2024; Sun et al., 2025; Huang & Hahn, 2026; Tiblias et al., 2026; Wurgaft et al., 2026, *inter alia*) and then intervene along these structures. This approach rests on the assumption that if a structure can be *localized* in hidden activations then intervening on it should reliably induce the intended change in the model's output distribution.

Despite its wide adoption, this assumption often breaks in practice. A localized structure may only reflect correlations in the data (Ravfogel et al., 2020; Elazar et al., 2021; Belinkov, 2022), and even when it is causally relevant, interventions along such structure can produce weak, inconsistent, or unintended effects (Arad et al., 2025; Wu et al., 2025). Moreover, recent work shows that directions optimized for detection often differ from those that causally influence model outputs (Galeone et al., 2026). Together, these findings establish that localization alone is insufficient for causal control, raising the question of what makes a latent structure actionable.

In this work, we tackle this question by decomposing the causal influence of a latent structure at individual contexts. Given a structure $\mathcal{S}$ and activation $\mathbf{h}$, we ask how much an intervention along $\mathcal{S}$ at $\mathbf{h}$ can increase the expression of a target concept c. Here, we define concept expression as the probability mass assigned to tokens associated with the concept, selected from the vocabulary $\mathcal{V}$ via an indicator vector $\mathbf{r}_c \in \{0,1\}^{|\mathcal{V}|}$. For an output distribution $\mathbf{p}$, the expression of c is therefore $\mathbf{r}_c^\top \mathbf{p}$. We then analyze how much an intervention along $\mathcal{S}$ at $\mathbf{h}$ can change this quantity. Let $\mathbf{v}$ denote an intervention direction and $J_\mathbf{h}$ the Jacobian mapping changes in $\mathbf{h}$ to changes in the output logits. The resulting first-order change in concept expression can be cast as a product of three factors:

$$\underbrace{\sigma_\mathbf{p}(J_\mathbf{h}\mathbf{v})}_{\text{I}}\,\underbrace{\sigma_\mathbf{p}(\mathbf{r}_c)}_{\text{II}}\,\underbrace{\mathrm{Corr}_\mathbf{p}(\mathbf{r}_c, J_\mathbf{h}\mathbf{v})}_{\text{III}}$$

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.

where $\sigma_{\mathbf{p}}(\cdot)$ and $\mathrm{Corr}_{\mathbf{p}}(\cdot)$ are the standard deviation and correlation under $\mathbf{p}$, respectively.

Next, we show empirically that these factors capture distinct and interpretable constraints on causal influence. Factor I captures **causal capacity**, signifying whether changes along $\mathcal{S}$ at $\mathbf{h}$ can propagate through the model's downstream computation to generally affect its output. We find that this sensitivity to local changes in activation space is concentrated in a context-dependent, low-rank subspace. Factor II encodes **concept responsiveness**, capturing the extent to which the concept is responsive to interventions under the given context. We observe that finite interventions primarily promote concept-relevant tokens that already receive non-negligible probability, with the top 2.5% of tokens under $\mathbf{p}$ receiving 94.5% of the intervention-induced probability increase. Last, factor III corresponds to **concept alignment**, capturing whether intervening along $\mathcal{S}$ at $\mathbf{h}$ is correctly oriented to promote the concept. We find that the directions promoting the concept vary substantially across contexts, where the same intervention direction can be well aligned in one context and poorly aligned in another. Importantly, reducing any of the factors bottlenecks the causal effect of $\mathcal{S}$, reducing it by 84–95% when one of the factors is low and driving it to zero when all are low. Low alignment can even cause the same latent structure to suppress its associated concept.

We then demonstrate the practical utility of our findings for steering. First, we show that linear probes often capture directions with little causal effect. By restricting the learned directions to high capacity subspaces, we introduce *causal probes* which improve the median steering performance by 17%–118% across models, while reducing concept-detection accuracy by only 3%.

Overall, our results show that the causal influence of a latent structure is not an intrinsic property of the structure itself, but rather arises from the relationship between the structure, context, and remaining computation. Capacity, responsiveness, and alignment are distinct factors describing this relationship, together determining to what extent a concept's expression can be promoted from the current state through the latent structure.

2 Disentangling Causal Influence

We study what determines the ability of a latent structure to causally induce the expression of a concept. To this end, we quantify concept expression at the model's output and pull it back to intermediate activations. Then, we analyze the causal effect of movement along the structure at a given activation, decomposing it into three constraints.

Pulling back causal influence Assume an auto-regressive LM with vocabulary $\mathcal{V}$. For an activation $\mathbf{h} \in \mathbb{R}^d$ at an intermediate layer, we denote by $\mathbf{z}$ the output logits produced by the remaining layers, and by $\mathbf{p} = \mathrm{softmax}(\mathbf{z})$ the corresponding probability vector over $\mathcal{V}$. Since the causal effect of a latent structure ultimately manifests as a change in the model's output distribution (at the current or future positions), we measure a concept's expression by the probability mass assigned to tokens that verbalize it. For example, the concept of refusal should promote tokens such as `cannot` and `won't`. Formally, we encode a set of tokens $\mathcal{V}_c \subseteq \mathcal{V}$ that verbalize a concept c as a vector:[1]

$$(\mathbf{r}_c)_i = \begin{cases} 1, & i \in \mathcal{V}_c, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

For a change $\Delta\mathbf{p}$ in the output distribution, the measured effect on concept expression is simply $\mathbf{r}_c^\top \Delta\mathbf{p}$. Thus, $\mathbf{r}_c^\top$ acts as an output-space covector[2] that measures whether a local change in the distribution promotes or suppresses the concept.

To quantify the causal effect of a latent structure $\mathcal{S}$ on c, we consider how intervening along $\mathcal{S}$ changes concept expression. For an intervention vector $\mathbf{v}$ along $\mathcal{S}$ at $\mathbf{h}$, we obtain a distribution change $\Delta\mathbf{p} = f(\mathbf{h} + \mathbf{v}) - f(\mathbf{h})$, where $f(\cdot)$ denotes the remaining computation from $\mathbf{h}$ to the output. The first-order change in the output logits is

$$\Delta\mathbf{z} \approx J_{\mathbf{h}}\mathbf{v}, \qquad J_{\mathbf{h}} = \partial\mathbf{z}/\partial\mathbf{h}, \tag{2}$$

and passing this change through the softmax gives the first-order change in the output distribution,

$$\Delta\mathbf{p} \approx F_{\mathbf{p}} J_{\mathbf{h}}\mathbf{v}, \qquad F_{\mathbf{p}} = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top, \tag{3}$$

[1] We describe how we systematically construct such concept token sets in §3.1.

[2] A covector is a linear map from vectors to scalars.

where $F_{\mathbf{p}}$ is the Jacobian of softmax. Since concept expression is measured by $\mathbf{r}_c^\top \Delta\mathbf{p}$, the resulting first-order change is

$$\mathbf{r}_c^\top \Delta\mathbf{p} \approx \mathbf{r}_c^\top F_{\mathbf{p}} J_{\mathbf{h}} \mathbf{v}. \tag{4}$$

Thus, $\mathbf{r}_c^\top F_{\mathbf{p}} J_{\mathbf{h}}$ pulls the output-space concept covector $\mathbf{r}_c^\top$ back to the intermediate activation space, characterizing which local changes to $\mathbf{h}$ affect concept expression. We evaluate the quality of this first-order approximation in §3.2.

Three factors construct causal influence Having quantified the causal effect of movement along S on the output, we now ask what determines this effect. Using the form of the softmax Jacobian in Eq. 3, the first-order change in concept expression can equivalently be written as

$$\mathbf{r}_c^\top \Delta\mathbf{p} \approx \mathrm{Cov}_{\mathbf{p}}\left(\mathbf{r}_c, J\mathbf{v}\right), \tag{5}$$

where $\mathrm{Cov}_{\mathbf{p}}$ denotes the covariance over vocabulary entries weighted by $\mathbf{p}$. Using the covariance–correlation identity, we cast the first-order effect as three factors:

$$\mathrm{Cov}_{\mathbf{p}}\left(\mathbf{r}_c, J\mathbf{v}\right) = \underbrace{\sigma_{\mathbf{p}}(J_{\mathbf{h}}\mathbf{v})}_{\text{I}}\,\underbrace{\sigma_{\mathbf{p}}(\mathbf{r}_c)}_{\text{II}}\,\underbrace{\mathrm{Corr}_{\mathbf{p}}(\mathbf{r}_c, J_{\mathbf{h}}\mathbf{v})}_{\text{III}} \tag{6}$$

Together, these terms form the first-order effect of movement along $\mathcal{S}$. In the next sections, we show that they reveal interpretable (§3) and distinct (§4) constraints on causal influence.

Extending to sequence-level causality Some latent structures may not directly affect the next-token prediction, but still influence outputs at later positions. To capture this, we extend the first-order effect over a horizon of T positions. Let $J_{\mathbf{h}}^t = \partial \mathbf{z}_t / \partial \mathbf{h}$ denote the Jacobian from the intermediate activation to the logits at position t, and let $\mathbf{p}_t$ denote the corresponding output distribution. The cumulative effect on concept expression is then

$$\sum_{t=0}^{T} \mathbf{r}_c^\top \Delta\mathbf{p}_t \approx \sum_{t=0}^{T} \mathbf{r}_c^\top F_{\mathbf{p}_t} J_{\mathbf{h}}^t \mathbf{v}. \tag{7}$$

Since differentiation distributes over sums, the covector corresponding to the sequence-level effect is simply the sum of the position-wise pullbacks. Each position is therefore characterized through its corresponding $J_{\mathbf{h}}^t$ and $\mathbf{p}_t$ and admits the same decomposition as in Eq. 6. Thus, we can capture both the intervention's effect at each output position and how these effects accumulate during generation.

3 Causality is Mediated by Interpretable Factors

We showed that the first-order causal effect of a latent structure can be cast into three factors (Eq. 6). In this section, we show these factors capture distinct, interpretable constraints on causal influence, corresponding to causal *capacity* (I), concept *responsiveness* (II), and concept *alignment* (III).

3.1 Experimental Setup

Data We collect a set of 50 diverse concepts, such as expressions of trust and geographic locations, sampled from a large list of concepts previously identified in LM representations (Lieberum et al., 2024). Then, using a set of 304 open-ended prefixes, we prompt gpt-4.1-mini (OpenAI et al., 2024) to construct two datasets per concept: *positive* with generations that exemplify the concept, and *neutral* with neutral completions that do not. Prompts, example prefixes and more details are in §A.1. We use these datasets to find latent structures and form the output covectors.

Constructing concept covectors and latent structures To compute the three factors, we construct for each concept an output-space covector $\mathbf{r}_c^T$ that measures concept expression and a latent structure $\mathcal{S}$ representing the concept. We construct $\mathbf{r}_c^T$ as detailed in Eq. 1, by assigning a value of 1 to tokens that occur *only* in the positive dataset and 0 otherwise (see §A.2 for details). For latent structures, we use Difference-in-Means (DiffMeans) (Li et al., 2023; Marks & Tegmark, 2024) and linear probes Alain & Bengio (2017); Hewitt & Liang (2019), following the formulation of Wu et al. (2025). Both methods are trained using the concept's positive and neutral datasets. We construct these structures independently at each evaluated model layer and intervene along them via the commonly used additive intervention (Subramani et al., 2022; Li et al., 2023; Turner et al., 2024), $\mathbf{h}' = \mathbf{h} + \alpha\mathbf{v}$, where $\mathbf{v}$ is a direction along $\mathcal{S}$ and $\alpha \in \mathbb{R}$ controls the intervention strength.

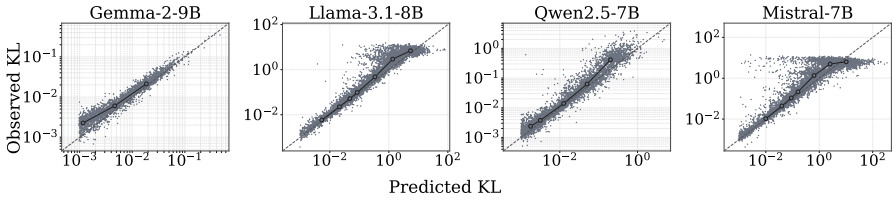


Figure 1: **Capacity predicts intervention effects.** Predicted and observed intervention-induced KL closely agree across intervention strengths.

Models We conduct our experiments across four LMs: Gemma-2-9B, Llama-3.1-8B, Qwen-2.5-7B, and Mistral-7B (Team et al., 2024; Grattafiori et al., 2024; Jiang et al., 2023; Qwen et al., 2025). To capture how causal influence changes over the course of computation, we evaluate activations at layers corresponding to approximately 20%, 40%, 60%, 80%, and 95% of model depth.

3.2 Capacity: Downstream Sensitivity is Concentrated in a Context-Dependent Subspace

Downstream layers may be highly sensitive to certain changes in intermediate activations while being invariant to others. Factor I, defined by $\sigma_{\mathbf{p}}(J_{\mathbf{h}}\mathbf{v})$, captures this sensitivity: the activation subspace of directions to which the downstream computation is most sensitive is

$$G_{\mathbf{h}} = J_{\mathbf{h}}^{\top} F_{\mathbf{p}} J_{\mathbf{h}}, \tag{8}$$

where sensitivity along $\mathbf{v}$ is $\sigma_{\mathbf{p}}(J_{\mathbf{h}}\mathbf{v}) = \sqrt{\mathbf{v}^{\top} G_{\mathbf{h}} \mathbf{v}}$. The factor is large when movement along $\mathbf{v}$ propagates to produce variation across the output logits, and small when that movement has little downstream effect. We therefore interpret this factor as the causal *capacity* of movement along the structure at the current activation.

Capacity predicts magnitude of intervention-induced output change We test our interpretation by checking whether capacity predicts the magnitude of intervention-induced output change. For activation $\mathbf{h}$ and output position t (where $t = 0$ denotes the current prediction), an intervention $\alpha\mathbf{v}$ induces a local KL divergence approximated to second order by $\frac{\alpha^2}{2}\mathbf{v}^{\top} G_{\mathbf{h}}^{t} \mathbf{v}$, providing a direct prediction of the resulting output change (derivation in §B.1). We compare this predicted KL with the observed KL for interventions along 64 isotropic unit directions over all neutral contexts, using 14 intervention strengths α ranging from 0.3 to 128. Figure 1 shows the results for the current prediction ($t = 0$) and depth 60%, with additional offsets and layers reported in §B.2. Predicted KL closely tracks observed KL, with directions outside the high-capacity subspace producing little causal effect.

Capacity is context dependent and low rank Having established that capacity predicts output change, we characterize how this sensitivity is realized in activation space. To this end, we examine the eigenspectrum of $G_{\mathbf{h}}^{t}$, whose eigenvalues dictate how downstream sensitivity is distributed across activation-space directions, with larger eigenvalues indicating greater sensitivity. We quantify its concentration using the effective number of directions over which downstream sensitivity is distributed (i.e., participation ratio), defined by $\left(\sum_i \lambda_i\right)^2 / \sum_i \lambda_i^2$ where λ_i is the i-th eigenvalue. Across models and layers, we find that capacity is highly concentrated (Figure 2) with a median effective dimensionality of 24.8 directions for Gemma, 47.2 for Llama, 30.1 for Mistral, and 23.3 for Qwen, corresponding to just 0.7%–1.2% of the residual stream's dimension. The dimensionality is lowest in early layers ($8 - 13$) and increases substantially with depth (37–97).

Localization methods find structures with varying capacity levels If downstream sensitivity is concentrated in a small subset of directions, latent structures may differ substantially in how much they overlap with this usable subspace. We therefore compare the median capacity of concept directions identified by DiffMeans and linear probes with that of 16 isotropic random directions. We normalize all intervention directions to unit norm and sum capacity over a horizon of $n = 9$.[3] We find substantial differences across structure types (Figure 2). DiffMeans directions have

[3]We use 9 offsets, as the causal effect saturates in larger offsets for most concepts (see §B.3).

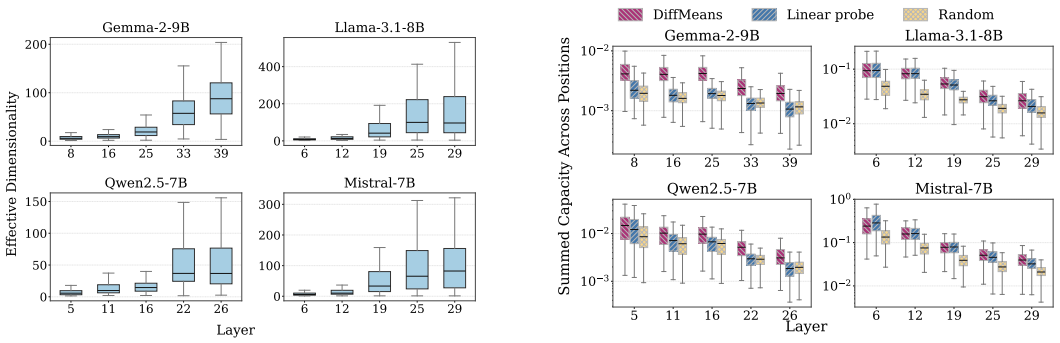


Figure 2: **Capacity is concentrated in a low-dimensional subspace and varies across latent structures. Left:** Effective dimensionality across layers, showing that downstream sensitivity is concentrated in few activation-space directions. **Right:** Capacity of DiffMeans, linear probes and random directions for current and future positions substantially varies between localization methods.

90.4% more capacity than random directions and 26.4% more than linear probes, while probes have 37.6% more than random directions. This shows that identifiable representational structure can vary substantially in its ability to affect the output, with some structures much more privileged by the model's remaining computation than others.

Overall, our results suggest that only a low-rank, context-dependent subspace affects the model's output. Thus, identifying usable structure requires distinguishing which parts of the representation lie within this subspace.

3.3 Responsiveness: Context Limits Concept Promotion

Under softmax, low-probability tokens require larger logit shifts for the same probability increase. Thus, the amount of probability mass assigned to concept-relevant tokens limits the concept's promotability. Factor II, defined as $\sigma_{\mathbf{p}}(\mathbf{r}_c)$, reflects this constraint. Since $\mathbf{r}_c$ indicates concept-relevant tokens, the factor is small when little probability mass is assigned to the concept, but also decreases when the concept is already highly expressed. It is therefore largest in an intermediate regime, where the concept is supported by the current distribution without being saturated. We interpret this factor as the concept's *responsiveness* in the current context.

Experiment We test this interpretation by analyzing whether finite, fluency-preserving interventions primarily promote tokens already likely under the unintervened distribution $\mathbf{p}$, where responsiveness is higher, or "new" tokens with little baseline support. We intervene with DiffMeans and random directions at increasing strengths α. For each context, we fix the support $\mathcal{T}$ containing the top $k \in \{1\%, 2.5\%, 5\%\}$ of tokens under $\mathbf{p}$, and measure how much of the intervention-induced increase in probability occurs within this support:

$$\phi_{\text{support}}(\alpha) = \frac{\sum_{t\in\mathcal{T}}[p'_t - p_t]_+}{\sum_{t\in[\mathcal{V}]}[p'_t - p_t]_+}, \qquad [x]_+ = \max\{x, 0\}, \tag{9}$$

where $\mathbf{p}'$ is the intervened distribution. The denominator is the total probability mass moved toward higher-probability tokens, while the numerator measures how much of this increase remains within the baseline support.

For each model and layer, we evaluate fluency at 11 intervention strengths α using the same LLM judge as in Wu et al. (2025), which assigns a score of $0, 1$ or 2, with 2 being fluent. We mark normalized fluency thresholds of 0.8, 0.7, and 0.6, and for each threshold report results up to the largest tested α for which mean fluency remains above it. This allows us to examine how the intervention's effect changes as the model's fluency degrades with increasing intervention strength. Additional details are provided in §C.1.

Results Figure 3 shows results for Gemma and Llama for the top 2.5% baseline support at 60% depth (results for additional models, layers and support thresholds show similar trends, see §C.2).

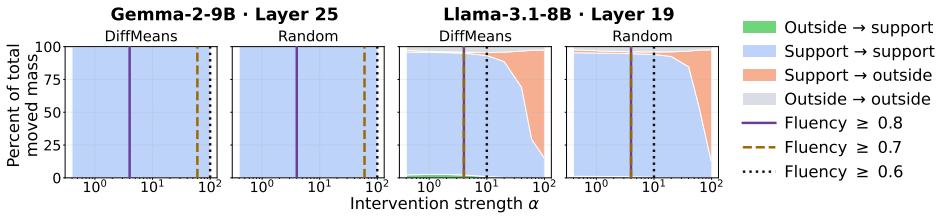


Figure 3: **Interventions promote outputs already supported by the context.** We measure where intervention-induced probability increases are assigned relative to the unintervened distribution. While fluency is preserved, nearly all increased probability goes to tokens in its top 2.5% support.

At fluency of 0.8, finite interventions (optimized for specific concepts) predominantly redistribute probability among tokens *that were already likely* under $\mathbf{p}$, with the top 2.5% of baseline tokens receiving approximately 94.5% of the intervention-induced probability increase. Within this fluent regime, the movement occurs predominantly among tokens already in the baseline support (blue regions). Substantial movement outside this support appears only at stronger intervention strengths as fluency degrades ($\alpha \geq 10$, fluency < 0.6), most clearly for Llama and Mistral, while Gemma and Qwen remain largely confined to the baseline support even at high strengths. The same qualitative behavior holds for DiffMeans and random directions, suggesting that this restriction is primarily imposed by the context rather than by the structure itself.

Thus, interventions have limited ability to promote outputs with little support in the current context, showing how the concept's responsiveness constrains the causal influence of a latent structure.

3.4 Alignment: Effective Directions Are Context-Dependent

While movement along a latent structure can affect the model's output, the resulting changes may not promote the target concept. Factor III, defined as $\mathrm{Corr}_{\mathbf{p}}(\mathbf{r}_c, J_{\mathbf{h}}\mathbf{v})$ quantifies orientation toward concept-relevant tokens rather than unrelated outputs. We therefore view this factor as the structure's *alignment* with the direction that most promotes the concept at the current activation.

Importantly, this interpretation is meaningful only if what constitutes concept-aligned structure depends on context. Otherwise, it would be a measure of how well the structure globally promotes the concept. We therefore test whether the direction that most promotes a concept varies across contexts. To find this direction for an activation $\mathbf{h}$, we estimate the locally-optimal activation-space direction using the closed-form solution to the corresponding regression problem,

$$\mathbf{v}^*(\mathbf{h}) \propto (G_{\mathbf{h}} + \lambda I)^{-1} J_{\mathbf{h}}^{\top} F_{\mathbf{p}} \mathbf{r}_c, \tag{10}$$

where λ controls the ridge regularization and $G_{\mathbf{h}}$ is defined in Eq. 8.

Locally effective directions vary across contexts We randomly sample 8 neutral contexts per concept[4] and, at each position and layer, compute the direction in Eq. 10 independently. Directions are normalized to reflect orientation rather than scale. We then compare the resulting orientations across activation pairs $(\mathbf{h}_i, \mathbf{h}_j)$ using cosine similarity, $\langle \mathbf{v}^*(\mathbf{h}_i), \mathbf{v}^*(\mathbf{h}_j) \rangle$. As a baseline, we also compare directions computed for different concepts. Figure 4 (left) presents the results. In early layers, locally-effective directions for the same concept are nearly orthogonal across contexts. Similarity increases in later layers, with directions for the same concept more closely aligned than those for different concepts, however for the majority of models the same concept directions are still distinct with a cosine similarity of <0.2. This shows that the activation-space direction that most promotes concept-related tokens depends substantially on the current context.

Locally effective directions do not transfer between contexts Even if different locally preferred directions are not similar, they could still potentially induce equivalent concept-related effects. We therefore test whether locally effective directions are interchangeable across contexts. At each layer we compute the locally preferred direction $\mathbf{v}_i^*$ independently for each context i. We then take the

[4]We use neutral contexts since in concept-positive contexts even random directions align with the concept due to the constraint described in §3.3.

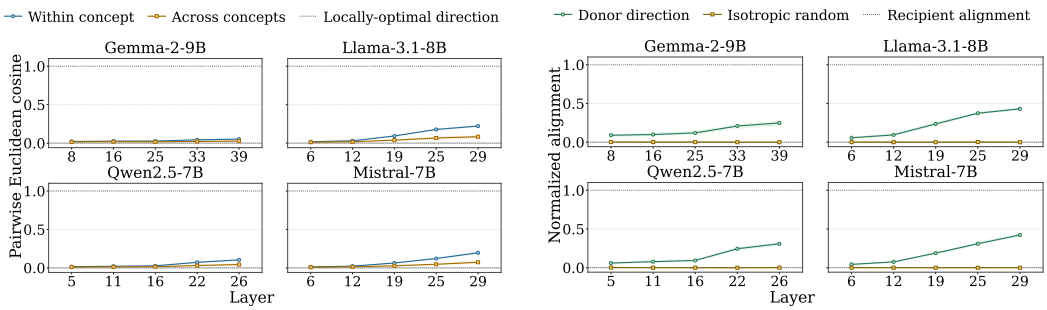


Figure 4: **Concept-promoting directions are context dependent.** We estimate the locally optimal direction for the same concept across contexts. **Left:** Same concept directions are nearly orthogonal in early layers and remain below 0.2 cosine similarity. **Right:** Same concept directions retain only $10\% - 40\%$ of alignment when transferred to other contexts, showing strong context dependence.

direction $\mathbf{v}_j^*$ identified in a different donor context j and evaluate it in context i, keeping both the context and $\mathbf{r}_c$ fixed. We measure the transferred direction's alignment in the recipient context i, normalized by the alignment of its locally preferred direction, i.e., $\mathrm{Corr}_{\mathbf{p}_i}\left(\mathbf{r}_c, J_i\mathbf{v}_j^*\right) / \mathrm{Corr}_{\mathbf{p}_i}\left(\mathbf{r}_c, J_i\mathbf{v}_i^*\right)$. For baselines, we include the normalized alignment of random isotropic directions. We find that locally effective directions are substantially less aligned when transferred across contexts (Figure 4, right). In early layers, average normalized alignment falls below 0.1, meaning that a donor direction achieves more than ten times lower alignment in a different context than the recipient's locally preferred direction. While transfer improves in later layers, it remains clearly below the context-local reference, reaching at most 0.4 normalized alignment.

Together, our results show that the best direction to promote a concept is strongly context-dependent, causing a fixed latent structure to vary substantially in causal effectiveness across activations.

4 Factors Expose Distinct Causal Bottlenecks

Having characterized the factors underlying the first-order causal influence of a structure, we now examine how these factors vary across activations and interact to shape causal influence. First through an illustrative example and then quantitatively.

Example Consider a simple binding task, where the model must associate given entities with their corresponding attributes: *"Box A is blue. Box B is green. Box A's color is"*. We apply a DiffMeans direction for *green* at each source position and measure its effect on the probability of *green* at the final position. Because the target position is fixed, responsiveness is constant, allowing us to isolate capacity and alignment. Figure 5 shows the capacity and alignment associated with intervening at each source position. Although the final position has the highest capacity, more than $3\times$ of earlier positions, its alignment with the *green* direction is much lower than at the earlier "green" token (0.14 vs. 0.88). Consequently, intervening at "green," despite occurring 7 positions earlier, has a $3.2\times$ larger first-order effect on the probability of the target token *green*.

Experiment We sample $16,000$ contexts from AllenAI C4 (Raffel et al., 2020), each consisting of $n = 256$ tokens, and for each context measure capacity, responsiveness and alignment with respect to each of the 50 concept covectors $\mathbf{r}_c^T$ computed in §3.1. Then, we intervene with the corresponding DiffMeans direction $\mathbf{v}_c$ at a source position $s \in \{n, n-1, ..., n-7\}$ and measure the resulting change in concept expression at the target position $t = n$, where $s = t$ captures the effect at the intervention position itself. Given the intervened and unintervened output distributions $\mathbf{p}'_t$ and $\mathbf{p}_t$ at position t, we measure concept expression as $\mathbf{r}_c^\top\left(\mathbf{p}'_t - \mathbf{p}_t\right)$. Holding the model, layer, and concept fixed, we treat each evaluated source–target pair as one sample. Each sample is assigned a high or low label per factor, using the top and bottom quartiles of that factor across samples. Quartile thresholds are estimated with five-fold cross-fitting over contexts. The three binary labels partition the retained samples into eight joint factor settings. We then compare the intervention effect across these groups.

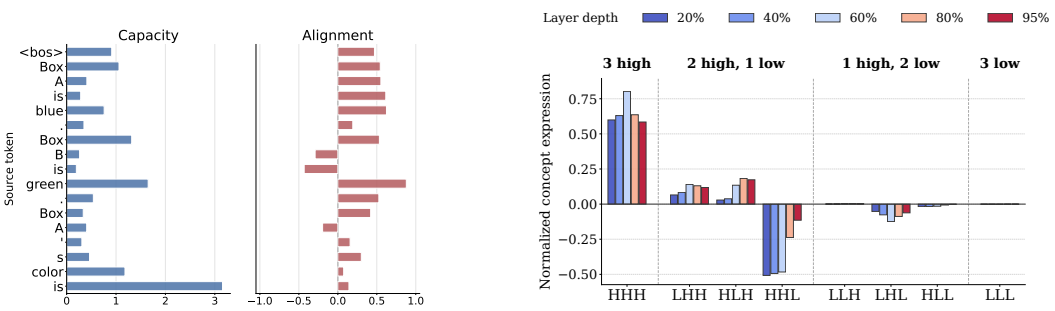


Figure 5: **Capacity, responsiveness and alignment impose distinct causal constraints. Left:** Capacity and alignment across intervention positions in an example prompt. **Right:** Concept promotion across all eight high/low factor combinations; a low value of any factor sharply reduces the effect, while low alignment suppresses the concept.

Results We find that the three factors impose distinct constraints, with all eight high/low combinations represented in the data (full bucket counts in §E.1). Figure 5 shows concept expression for each combination aggregated across models, revealing that a single unfavorable factor is sufficient to sharply reduce intervention efficacy. Low capacity and low responsiveness reduce concept expression by 84% and 95%, respectively, while low alignment can reverse the effect and suppress the concept. When all three factors are low, concept promotion is effectively zero. Per-model plots showing similar trends are provided in §E.2. In §E.3, we further validate the need for all three factors to be high over longer generations by showing that their product identifies positions where steering is especially effective relative to low-product or randomly selected positions.

5 Accounting for Causal Constraints Improves Steering

In this section, we demonstrate the utility of our findings for effective steering. We constrain linear probes to the high-capacity subspaces identified by our analysis and show that this substantially improves steering while largely preserving concept detection.

Causal probes For each concept c and layer, we identify the activation-space directions with the largest downstream capacity. Specifically, for each sample in the positive and neutral datasets, we take the activation $\mathbf{h}$ at source position s, compute $G^t_{\mathbf{h}}$ (Eq. 8) with respect to each target position $t = s, \ldots, s+7$ and average these matrices across samples and source–target pairs. The leading r eigenvectors of this average define a high-capacity subspace Q_c of rank r. We then train a linear probe as described in §3.1, while constraining it to lie in Q_c by projecting it back onto the subspace after each gradient step. We refer to the resulting classifier as a *causal probe*.

Evaluation For each model, layer, concept, and structure, we steer generations from 16 fixed neutral prefixes by adding $\alpha\mathbf{v}$ at every decoding step. We sweep 14 strengths $\alpha \in [0.4, 100]$, generate up to 50 tokens, and score concept expression and fluency similarly to Wu et al. (2025), reporting their harmonic mean.[5] We select the best α for each setting using five-fold cross-validation over prefixes and report scores across all 50 concepts. We compare causal probes with $r \in \{8, 16, 32, 64, 128\}$ against standard probes, DiffMeans, and a strong supervised baseline ReFT-r1 Wu et al. (2025). See §F.2 for full details.

Results Final scores are shown for $r = 128$ in Figure 6, with additional ranks and concept and fluency scores in §F.3. Causal probes substantially improve steering over raw probes, particularly in early and middle layers. Across models, the median improvement is 118% for Gemma, 57% for Qwen, 23% for Mistral, and 17% for Llama. Notably, these gains come with only a small decrease in concept detection, with a median classification accuracy of 90.4%, compared to 93.8% for the unconstrained probe. These results show that learned probes capture predictive structure along directions with little causal capacity, and removing these causally ineffective directions substantially improves steering while largely preserving concept detection.

[5] In Wu et al. (2025) they use a scale of 0-2, we report normalized scores.

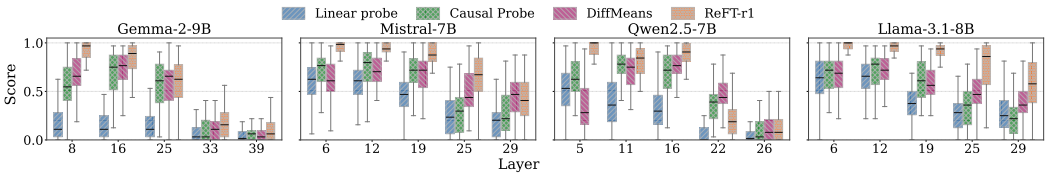


Figure 6: **Causal probes improve steering.** We construct causal probes by restricting linear probes to high-capacity subspaces and evaluate their steering performance. Causal probes substantially outperform standard probes, often performing similarly to DiffMeans.

6 Related Work

The detection–causality gap Prior work has identified the disparity between structure in activation space and causality. Early work showed that probing accuracy can reflect the capacity of the probe itself, and that representations may encode properties that are not relevant to the current task of the model (Hewitt & Liang, 2019; Ravichander et al., 2021). Elazar et al. (2021) made this distinction causal, showing that probing accuracy need not track the behavioral importance of the detected information. Recent work further extends our understanding of this gap, showing that where information is localized is not necessarily where editing is most effective (Hase et al., 2023), and structures that accurately detect a concept often produce weak or unreliable effects when used for steering (Arad et al., 2025; Wu et al., 2025; Galeone et al., 2026). Hoang et al. (2026) also showed that SAE features can have different downstream effects across contexts. While previous work identified a gap between representation and control, we quantify this gap and decompose it into concrete, interpretable factors, showing that causal influence is fundamentally context dependent.

Jacobians and output space geometry Recent work utilized output-space geometry to understand the activation-space of LMs. Park et al. (2026) studied the information geometry induced by softmax and uses it to derive geometry-aware probing and steering. Wurgaft et al. (2026) pulled back the output-space geometry with the model's Jacobian to construct nonlinear steering paths. Wang et al. (2026) likewise pulled the Fisher geometry of the output distribution into activation space and used the resulting metric to derive improved steering directions. These works use output-space geometry to design or optimize interventions, instead we ask why the causal influence of a given latent structure varies across contexts, and which distinct factors govern this variation.

Recently, Hernandez et al. (2024); Gurnee et al. (2026); Blank et al. (2026) utilized Jacobians to connect activation-space directions to future outputs. These methods identify a useful subspace by aggregating Jacobian effects across contexts and identifying directions with consistent downstream effects. In contrast, we study why the causal efficacy of a given latent structure varies across contexts, using the local Jacobian and output distribution to characterize the factors governing its effect.

7 Conclusion

By measuring concept expression at the output and pulling it back to intermediate activations, we cast the causal influence of a latent structure as a product of three factors: *capacity*, *responsiveness*, and *alignment*. Across four models, we find that these factors capture distinct, interpretable constraints: downstream sensitivity is concentrated in a small, context-dependent subset of directions, interventions primarily promote outputs already supported by the context, and concept-promoting directions vary substantially across contexts. These constraints explain how a structure can be strongly represented yet remain weak or unreliable for control. We further show that removing predictive but causally ineffective directions with *Causal Probes* improves median steering by 29%, with only a 3% drop in concept-detection accuracy. Overall, our results show that causal effectiveness is not intrinsic to a latent structure, but depends on its relationship with the current context. Effective localization and control therefore require accounting not only for *what* is represented, but also *when and how* that representation can be causally utilized.

AI USE STATEMENT

Throughout this work, the authors used AI tools for both coding assistance and polishing written material to improve readability. The authors reviewed and verified all experiments and writing, and take full responsibility for the integrity and accuracy of the material in this paper. Additionally, a dataset consisting of sentences expressing concepts was generated with an LLM. The authors reviewed a subsample of the dataset to verify its correctness.

REPRODUCIBILITY STATEMENT

All source code required to reproduce the experiments and methods introduced in this paper is publicly available in a dedicated GitHub repository at `https://github.com/anonymous-place-holder`. We additionally release the datasets used throughout our experiments. Together, these resources are intended to enable full reproduction of our experimental findings and facilitate further investigation of the methods introduced in this work.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Dana Arad, Aaron Mueller, and Yonatan Belinkov. Saes are good for steering - if you select the right features. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 10241–10259. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.519. URL https://doi.org/10.18653/v1/2025.emnlp-main.519.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Comput. Linguistics*, 48(1):207–219, 2022. doi: 10.1162/COLI_A_00422. URL https://doi.org/10.1162/coli_a_00422.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 861–872, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1080. URL https://aclanthology.org/P17-1080/.

Camila Blank, Agam Bhatia, and Neel Nanda. R-lens: Making j-lens more faithful on early layers. LessWrong, August 2026. URL https://www.lesswrong.com/posts/nv8oedrnLXKRzNEL9/r-lens-making-j-lens-more-faithful-on-early-layers.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell, Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Alvin Abdagic, Lior Belenki, James Allingham, Anima Singh, Theo Guidroz, Srivatsan Srinivasan, Herman Schmit, Kristen Chiafullo, Andre Elisseeff, Nilpa Jha, Prateek Kolhar, Leonard Berrada, Frank Ding, Xiance Si, Shrestha Basu Mallick, Franz Och, Sofia Erell, Eric Ni, Tejasi Latkar, Sherry Yang, Petar Sirkovic, Ziqiang Feng, Robert Leland, Rachel Hornung, Gang Wu, Charles Blundell, Hamidreza Alvari, Po-Sen Huang, Cathy Yip, Sanja Deur, Li Liu, Gabriela Surita, Pablo Duque, Dima Damen, Johnson Jia, Arthur Guez, Markus Mircea, Animesh

Sinha, Alberto Magni, Paweł Stradomski, Tal Marian, Vlado Galić, Wenhu Chen, Hisham Husain, Achintya Singhal, Dominik Grewe, François-Xavier Aubet, Shuang Song, Lorenzo Blanco, Leland Rechis, Lewis Ho, Rich Munoz, Kelvin Zheng, Jessica Hamrick, Kevin Mather, Hagai Taitelbaum, Eliza Rutherford, Yun Lei, Kuangyuan Chen, Anand Shukla, Erica Moreira, Eric Doi, Berivan Isik, Nir Shabat, Dominika Rogozińska, Kashyap Kolipaka, Jason Chang, Eugen Vušak, Srinivasan Venkatachary, Shadi Noghabi, Tarun Bharti, Younghoon Jun, Aleksandr Zaks, Simon Green, Jeshwanth Challagundla, William Wong, Muqthar Mohammad, Dean Hirsch, Yong Cheng, Iftekhar Naim, Lev Proleev, Damien Vincent, Aayush Singh, Maxim Krikun, Dilip Krishnan, Zoubin Ghahramani, Aviel Atias, Rajeev Aggarwal, Christo Kirov, Dimitrios Vytiniotis, Christy Koh, Alexandra Chronopoulou, Pawan Dogra, Vlad-Doru Ion, Gladys Tyen, Jason Lee, Felix Weissenberger, Trevor Strohman, Ashwin Balakrishna, Jack Rae, Marko Velic, Raoul de Liedekerke, Oded Elyada, Wentao Yuan, Canoee Liu, Lior Shani, Sergey Kishchenko, Bea Alessio, Yandong Li, Richard Song, Sam Kwei, Orion Jankowski, Aneesh Pappu, Youhei Namiki, Yenai Ma, Nilesh Tripuraneni, Colin Cherry, Marissa Ikonomidis, Yu-Cheng Ling, Colin Ji, Beka Westberg, Auriel Wright, Da Yu, David Parkinson, Swaroop Ramaswamy, Jerome Connor, Soheil Hassas Yeganeh, Snchit Grover, George Kenwright, Lubo Litchev, Chris Apps, Alex Tomala, Felix Halim, Alex Castro-Ros, Zefei Li, Anudhyan Boral, Pauline Sho, Michal Yarom, Eric Malmi, David Klinghoffer, Rebecca Lin, Alan Ansell, Pradeep Kumar S, Shubin Zhao, Siqi Zuo, Adam Santoro, Heng-Tze Cheng, Solomon Demmessie, Yuchi Liu, Nicole Brichtova, Allie Culp, Nathaniel Braun, Dan Graur, Will Ng, Nikhil Mehta, Aaron Phillips, Patrik Sundberg, Varun Godbole, Fangyu Liu, Yash Katariya, David Rim, Mojtaba Seyedhosseini, Sean Ammirati, Jonas Valfridsson, Mahan Malihi, Timothy Knight, Andeep Toor, Thomas Lampe, Abe Ittycheriah, Lewis Chiang, Chak Yeung, Alexandre Fréchette, Jinmeng Rao, Huisheng Wang, Himanshu Srivastava, Richard Zhang, Rocky Rhodes, Ariel Brand, Dean Weesner, Ilya Figotin, Felix Gimeno, Rachana Fellinger, Pierre Marcenac, José Leal, Eyal Marcus, Victor Cotruta, Rodrigo Cabrera, Sheryl Luo, Dan Garrette, Vera Axelrod, Sorin Baltateanu, David Barker, Dongkai Chen, Horia Toma, Ben Ingram, Jason Riesa, Chinmay Kulkarni, Yujing Zhang, Hongbin Liu, Chao Wang, Martin Polacek, Will Wu, Kai Hui, Adrian N Reyes, Yi Su, Megan Barnes, Ishaan Malhi, Anfal Siddiqui, Qixuan Feng, Mihai Damaschin, Daniele Pighin, Andreas Steiner, Samuel Yang, Ramya Sree Boppana, Simeon Ivanov, Arun Kandoor, Aditya Shah, Asier Mujika, Da Huang, Christopher A. Choquette-Choo, Mohak Patel, Tianhe Yu, Toni Creswell, Jerry, Liu, Catarina Barros, Yasaman Razeghi, Aurko Roy, Phil Culliton, Binbin Xiong, Jiaqi Pan, Thomas Strohmann, Tolly Powell, Babi Seal, Doug DeCarlo, Pranav Shyam, Kaan Katircioglu, Xuezhi Wang, Cassidy Hardin, Immanuel Odisho, Josef Broder, Oscar Chang, Arun Nair, Artem Shtefan, Maura O'Brien, Manu Agarwal, Sahitya Potluri, Siddharth Goyal, Amit Jhindal, Saksham Thakur, Yury Stuken, James Lyon, Kristina Toutanova, Fangxiaoyu Feng, Austin Wu, Ben Horn, Alek Wang, Alex Cullum, Gabe Taubman, Disha Shrivastava, Chongyang Shi, Hamish Tomlinson, Roma Patel, Tao Tu, Ada Maksutaj Oflazer, Francesco Pongetti, Mingyao Yang, Adrien Ali Taïga, Vincent Perot, Nuo Wang Pierse, Feng Han, Yoel Drori, Iñaki Iturrate, Ayan Chakrabarti, Legg Yeung, Dave Dopson, Yi ting Chen, Apoorv Kulshreshtha, Tongfei Guo, Philip Pham, Tal Schuster, Junquan Chen, Alex Polozov, Jinwei Xing, Huanjie Zhou, Praneeth Kacham, Doron Kukliansky, Antoine Miech, Sergey Yaroshenko, Ed Chi, Sholto Douglas, Hongliang Fei, Mathieu Blondel, Preethi Myla, Lior Madmoni, Xing Wu, Daniel Keysers, Kristian Kjems, Isabela Albuquerque, Lijun Yu, Joel D'sa, Michelle Plantan, Vlad Ionescu, Jaume Sanchez Elias, Abhirut Gupta, Manish Reddy Vuyyuru, Fred Alcober, Tong Zhou, Kaiyang Ji, Florian Hartmann, Subha Puttagunta, Hugo Song, Ehsan Amid, Anca Stefanoiu, Andrew Lee, Paul Pucciarelli, Emma Wang, Amit Raul, Slav Petrov, Isaac Tian, Valentin Anklin, Nana Nti, Victor Gomes, Max Schumacher, Grace Vesom, Alex Panagopoulos, Konstantinos Bousmalis, Daniel Andor, Josh Jacob, Yuan Zhang, Bill Rosgen, Matija Kecman, Matthew Tung, Alexandra Belias, Noah Goodman, Paul Covington, Brian Wieder, Nikita Saxena, Elnaz Davoodi, Muhuan Huang, Sharath Maddineni, Vincent Roulet, Folawiyo Campbell-Ajala, Pier Giuseppe Sessa, Xintian, Wu, Guangda Lai, Paul Collins, Alex Haig, Vytenis Sakenas, Xiaowei Xu, Marissa Giustina, Laurent El Shafey, Pichi Charoenpanit, Shefali Garg, Joshua Ainslie, Boone Severson, Montse Gonzalez Arenas, Shreya Pathak, Sujee Rajayogam, Jie Feng, Michiel Bakker, Sheng Li, Nevan Wichers, Jamie Rogers, Xinyang Geng, Yeqing Li, Rolf Jagerman, Chao Jia, Nadav Olmert, David Sharon, Matthew Mauger, Sandeep Mariserla, Hongxu Ma, Megha Mohabey, Kyuyeun Kim, Alek Andreev, Scott Pollom, Juliette Love, Vihan Jain, Priyanka Agrawal, Yannick Schroecker, Alisa Fortin, Manfred Warmuth, Ji Liu, Andrew Leach, Irina Blok, Ganesh Poomal Girirajan, Roee Aharoni, Benigno Uria, Andrei Sozanschi, Dan Goldberg, Lucian Ionita, Marco Tulio Ribeiro, Martin Zlocha, Vighnesh Birodkar, Sami

Lachgar, Liangzhe Yuan, Himadri Choudhury, Matt Ginsberg, Fei Zheng, Gregory Dibb, Emily Graves, Swachhand Lokhande, Gabriel Rasskin, George-Cristian Muraru, Corbin Quick, Sandeep Tata, Pierre Sermanet, Aditya Chawla, Itay Karo, Yan Wang, Susan Zhang, Orgad Keller, Anca Dragan, Guolong Su, Ian Chou, Xi Liu, Yiqing Tao, Shruthi Prabhakara, Marc Wilson, Ruibo Liu, Shibo Wang, Georgie Evans, David Du, Alfonso Castaño, Gautam Prasad, Mona El Mahdy, Sebastian Gerlach, Machel Reid, Jarrod Kahn, Amir Zait, Thanumalayan Sankaranarayana Pillai, Thatcher Ulrich, Guanyu Wang, Jan Wassenberg, Efrat Farkash, Kiran Yalasangi, Congchao Wang, Maria Bauza, Simon Bucher, Ting Liu, Jun Yan, Gary Leung, Vikas Sindhwani, Parker Barnes, Avi Singh, Ivan Jurin, Jichuan Chang, Niket Kumar Bhumihar, Sivan Eiger, Gui Citovsky, Ben Withbroe, Zhang Li, Siyang Xue, Niccolò Dal Santo, Georgi Stoyanov, Yves Raimond, Steven Zheng, Yilin Gao, Vít Listík, Sławek Kwasiborski, Rachel Saputro, Adnan Ozturel, Ganesh Mallya, Kushal Majmundar, Ross West, Paul Caron, Jinliang Wei, Lluis Castrejon, Sharad Vikram, Deepak Ramachandran, Nikhil Dhawan, Jiho Park, Sara Smoot, George van den Driessche, Yochai Blau, Chase Malik, Wei Liang, Roy Hirsch, Cicero Nogueira dos Santos, Eugene Weinstein, Aäron van den Oord, Sid Lall, Nicholas FitzGerald, Zixuan Jiang, Xuan Yang, Dale Webster, Ali Elqursh, Aedan Pope, Georges Rotival, David Raposo, Wanzheng Zhu, Jeff Dean, Sami Alabed, Dustin Tran, Arushi Gupta, Zach Gleicher, Jessica Austin, Edouard Rosseel, Megh Umekar, Dipanjan Das, Yinghao Sun, Kai Chen, Karolis Misiunas, Xiang Zhou, Yixian Di, Alyssa Loo, Josh Newlan, Bo Li, Vinay Ramasesh, Ying Xu, Alex Chen, Sudeep Gandhe, Radu Soricut, Nikita Gupta, Shuguang Hu, Seliem El-Sayed, Xavier Garcia, Idan Brusilovsky, Pu-Chin Chen, Andrew Bolt, Lu Huang, Alex Gurney, Zhiying Zhang, Alexander Pritzel, Jarek Wilkiewicz, Bryan Seybold, Bhargav Kanagal Shamanna, Felix Fischer, Josef Dean, Karan Gill, Ross Mcilroy, Abhishek Bhowmick, Jeremy Selier, Antoine Yang, Derek Cheng, Vladimir Magay, Jie Tan, Dhriti Varma, Christian Walder, Tomas Kocisky, Ryo Nakashima, Paul Natsev, Mike Kwong, Ionel Gog, Chiyuan Zhang, Sander Dieleman, Thomas Jimma, Andrey Ryabtsev, Siddhartha Brahma, David Steiner, Dayou Du, Ante Žužul, Mislav Žanić, Mukund Raghavachari, Willi Gierke, Zeyu Zheng, Dessie Petrova, Yann Dauphin, Yuchuan Liu, Ido Kessler, Steven Hand, Chris Duvarney, Seokhwan Kim, Hyo Lee, Léonard Hussenot, Jeffrey Hui, Josh Smith, Deepali Jain, Jiawei Xia, Gaurav Singh Tomar, Keyvan Amiri, Du Phan, Fabian Fuchs, Tobias Weyand, Nenad Tomasev, Alexandra Cordell, Xin Liu, Jonathan Mallinson, Pankaj Joshi, Andy Crawford, Arun Suggala, Steve Chien, Nick Fernando, Mariella Sanchez-Vargas, Duncan Williams, Phil Crone, Xiyang Luo, Igor Karpov, Jyn Shan, Terry Thurk, Robin Strudel, Paul Voigtlaender, Piyush Patil, Tim Dozat, Ali Khodaei, Sahil Singla, Piotr Ambroszczyk, Qiyin Wu, Yifan Chang, Brian Roark, Chaitra Hegde, Tianli Ding, Angelos Filos, Zhongru Wu, André Susano Pinto, Shuang Liu, Saarthak Khanna, Aditya Pandey, Siobhan Mcloughlin, Qiujia Li, Sam Haves, Allan Zhou, Elena Buchatskaya, Isabel Leal, Peter de Boursac, Nami Akazawa, Nina Anderson, Terry Chen, Krishna Somandepalli, Chen Liang, Sheela Goenka, Stephanie Winkler, Alexander Grushetsky, Yifan Ding, Jamie Smith, Fan Ye, Jordi Pont-Tuset, Eric Li, Ruichao Li, Tomer Golany, Dawid Wegner, Tao Jiang, Omer Barak, Yuan Shangguan, Eszter Vértes, Renee Wong, Jörg Bornschein, Alex Tudor, Michele Bevilacqua, Tom Schaul, Ankit Singh Rawat, Yang Zhao, Kyriakos Axiotis, Lei Meng, Cory McLean, Jonathan Lai, Jennifer Beattie, Nate Kushman, Yaxin Liu, Blair Kutzman, Fiona Lang, Jingchen Ye, Praneeth Netrapalli, Pushkar Mishra, Myriam Khan, Megha Goel, Rob Willoughby, David Tian, Honglei Zhuang, JD Chen, Zak Tsai, Tasos Kementsietsidis, Arjun Khare, James Keeling, Keyang Xu, Nathan Waters, Florent Altché, Ashok Popat, Bhavishya Mittal, David Saxton, Dalia El Badawy, Michael Mathieu, Zheng Zheng, Hao Zhou, Nishant Ranka, Richard Shin, Qingnan Duan, Tim Salimans, Ioana Mihailescu, Uri Shaham, Ming-Wei Chang, Yannis Assael, Nishanth Dikkala, Martin Izzard, Vincent Cohen-Addad, Cat Graves, Vlad Feinberg, Grace Chung, DJ Strouse, Danny Karmon, Sahand Sharifzadeh, Zoe Ashwood, Khiem Pham, Jon Blanton, Alex Vasiloff, Jarred Barber, Mark Geller, Aurick Zhou, Fedir Zubach, Tzu-Kuo Huang, Lei Zhang, Himanshu Gupta, Matt Young, Julia Proskurnia, Ronny Votel, Valentin Gabeur, Gabriel Barcik, Aditya Tripathi, Hongkun Yu, Geng Yan, Beer Changpinyo, Filip Pavetić, Amy Coyle, Yasuhisa Fujii, Jorge Gonzalez Mendez, Tianhao Zhou, Harish Rajamani, Blake Hechtman, Eddie Cao, Da-Cheng Juan, Yi-Xuan Tan, Valentin Dalibard, Yilun Du, Natalie Clay, Kaisheng Yao, Wenhao Jia, Dimple Vijaykumar, Yuxiang Zhou, Xinyi Bai, Wei-Chih Hung, Steven Pecht, Georgi Todorov, Nikhil Khadke, Pramod Gupta, Preethi Lahoti, Arnaud Autef, Karthik Duddu, James Lee-Thorp, Alexander Bykovsky, Tautvydas Misiunas, Sebastian Flennerhag, Santhosh Thangaraj, Jed McGiffin, Zack Nado, Markus Kunesch, Andreas Noever, Amir Hertz, Marco Liang, Victor Stone, Evan Palmer, Samira Daruki, Arijit Pramanik, Siim Põder, Austin Kyker, Mina Khan, Evgeny Sluzhaev, Marvin Ritter, Avraham Ruderman,

Wenlei Zhou, Chirag Nagpal, Kiran Vodrahalli, George Necula, Paul Barham, Ellie Pavlick, Jay Hartford, Izhak Shafran, Long Zhao, Maciej Mikuła, Tom Eccles, Hidetoshi Shimokawa, Kanav Garg, Luke Vilnis, Hanwen Chen, Ilia Shumailov, Kuang-Huei Lee, Abdelrahman Abdelhamed, Meiyan Xie, Vered Cohen, Ester Hlavnova, Dan Malkin, Chawin Sitawarin, James Lottes, Pauline Coquinot, Tianli Yu, Sandeep Kumar, Jingwei Zhang, Aroma Mahendru, Zafarali Ahmed, James Martens, Tao Chen, Aviel Boag, Daiyi Peng, Coline Devin, Arseniy Klimovskiy, Mary Phuong, Danny Vainstein, Jin Xie, Bhuvana Ramabhadran, Nathan Howard, Xinxin Yu, Gitartha Goswami, Jingyu Cui, Sam Shleifer, Mario Pinto, Chih-Kuan Yeh, Ming-Hsuan Yang, Sara Javanmardi, Dan Ethier, Chace Lee, Jordi Orbay, Suyog Kotecha, Carla Bromberg, Pete Shaw, James Thornton, Adi Gerzi Rosenthal, Shane Gu, Matt Thomas, Ian Gemp, Aditya Ayyar, Asahi Ushio, Aarush Selvan, Joel Wee, Chenxi Liu, Maryam Majzoubi, Weiren Yu, Jake Abernethy, Tyler Liechty, Renke Pan, Hoang Nguyen, Qiong, Hu, Sarah Perrin, Abhinav Arora, Emily Pitler, Weiyi Wang, Kaushik Shivakumar, Flavien Prost, Ben Limonchik, Jing Wang, Yi Gao, Timothee Cour, Shyamal Buch, Huan Gui, Maria Ivanova, Philipp Neubeck, Kelvin Chan, Lucy Kim, Huizhong Chen, Naman Goyal, Da-Woon Chung, Lu Liu, Yao Su, Anastasia Petrushkina, Jiajun Shen, Armand Joulin, Yuanzhong Xu, Stein Xudong Lin, Yana Kulizhskaya, Ciprian Chelba, Shobha Vasudevan, Eli Collins, Vasilisa Bashlovkina, Tony Lu, Doug Fritz, Jongbin Park, Yanqi Zhou, Chen Su, Richard Tanburn, Mikhail Sushkov, Mitchelle Rasquinha, Jinning Li, Jennifer Prendki, Yiming Li, Pallavi LV, Shriya Sharma, Hen Fitoussi, Hui Huang, Andrew Dai, Phuong Dao, Mike Burrows, Henry Prior, Danfeng Qin, Golan Pundak, Lars Lowe Sjoesund, Art Khurshudov, Zhenkai Zhu, Albert Webson, Elizabeth Kemp, Tat Tan, Saurabh Agrawal, Susie Sargsyan, Liqun Cheng, Jim Stephan, Tom Kwiatkowski, David Reid, Arunkumar Byravan, Assaf Hurwitz Michaely, Nicolas Heess, Luowei Zhou, Sonam Goenka, Viral Carpenter, Anselm Levskaya, Bo Wang, Reed Roberts, Rémi Leblond, Sharat Chikkerur, Stav Ginzburg, Max Chang, Robert Riachi, Chuqiao, Xu, Zalán Borsos, Michael Pliskin, Julia Pawar, Morgane Lustman, Hannah Kirkwood, Ankit Anand, Aditi Chaudhary, Norbert Kalb, Kieran Milan, Sean Augenstein, Anna Goldie, Laurel Prince, Karthik Raman, Yanhua Sun, Vivian Xia, Aaron Cohen, Zhouyuan Huo, Josh Camp, Seher Ellis, Lukas Zilka, David Vilar Torres, Lisa Patel, Sho Arora, Betty Chan, Jonas Adler, Kareem Ayoub, Jacky Liang, Fayaz Jamil, Jiepu Jiang, Simon Baumgartner, Haitian Sun, Yael Karov, Yaroslav Akulov, Hui Zheng, Irene Cai, Claudio Fantacci, James Rubin, Alex Rav Acha, Mengchao Wang, Nina D'Souza, Rohit Sathyanarayana, Shengyang Dai, Simon Rowe, Andrey Simanovsky, Omer Goldman, Yuheng Kuang, Xiaoyue Pan, Andrew Rosenberg, Tania Rojas-Esponda, Praneet Dutta, Amy Zeng, Irina Jurenka, Greg Farquhar, Yamini Bansal, Shariq Iqbal, Becca Roelofs, Ga-Young Joung, Parker Beak, Changwan Ryu, Ryan Poplin, Yan Wu, Jean-Baptiste Alayrac, Senaka Buthpitiya, Olaf Ronneberger, Caleb Habtegebriel, Wei Li, Paul Cavallaro, Aurora Wei, Guy Bensky, Timo Denk, Harish Ganapathy, Jeff Stanway, Pratik Joshi, Francesco Bertolini, Jessica Lo, Olivia Ma, Zachary Charles, Geta Sampemane, Himanshu Sahni, Xu Chen, Harry Askham, David Gaddy, Peter Young, Jiewen Tan, Matan Eyal, Arthur Bražinskas, Li Zhong, Zhichun Wu, Mark Epstein, Kai Bailey, Andrew Hard, Kamyu Lee, Sasha Goldshtein, Alex Ruiz, Mohammed Badawi, Matthias Lochbrunner, JK Kearns, Ashley Brown, Fabio Pardo, Theophane Weber, Haichuan Yang, Pan-Pan Jiang, Berkin Akin, Zhao Fu, Marcus Wainwright, Chi Zou, Meenu Gaba, Pierre-Antoine Manzagol, Wendy Kan, Yang Song, Karina Zainullina, Rui Lin, Jeongwoo Ko, Salil Deshmukh, Apoorv Jindal, James Svensson, Divya Tyam, Heri Zhao, Christine Kaeser-Chen, Scott Baird, Pooya Moradi, Jamie Hall, Qiuchen Guo, Vincent Tsang, Bowen Liang, Fernando Pereira, Suhas Ganesh, Ivan Korotkov, Jakub Adamek, Sridhar Thiagarajan, Vinh Tran, Charles Chen, Chris Tar, Sanil Jain, Ishita Dasgupta, Taylan Bilal, David Reitter, Kai Zhao, Giulia Vezzani, Yasmin Gehman, Pulkit Mehta, Lauren Beltrone, Xerxes Dotiwalla, Sergio Guadarrama, Zaheer Abbas, Stefani Karp, Petko Georgiev, Chun-Sung Ferng, Marc Brockschmidt, Liqian Peng, Christoph Hirnschall, Vikas Verma, Yingying Bi, Ying Xiao, Avigail Dabush, Kelvin Xu, Phil Wallis, Randall Parker, Qifei Wang, Yang Xu, Ilkin Safarli, Dinesh Tewari, Yin Zhang, Seungyeon Kim, Andrea Gesmundo, Mackenzie Thomas, Sergey Levi, Ahmed Chowdhury, Kanishka Rao, Peter Garst, Sam Conway-Rahman, Helen Ran, Kay McKinney, Zhisheng Xiao, Wenhao Yu, Rohan Agrawal, Axel Stjerngren, Catalin Ionescu, Jingjing Chen, Vivek Sharma, Justin Chiu, Fei Liu, Ken Franko, Clayton Sanford, Xingyu Cai, Paul Michel, Sanjay Ganapathy, Jane Labanowski, Zachary Garrett, Ben Vargas, Sean Sun, Bryan Gale, Thomas Buschmann, Guillaume Desjardins, Nimesh Ghelani, Palak Jain, Mudit Verma, Chulayuth Asawaroengchai, Julian Eisenschlos, Jitendra Harlalka, Hideto Kazawa, Don Metzler, Joshua Howland, Ying Jian, Jake Ades, Viral Shah, Tynan Gangwani, Seungji Lee, Roman Ring, Steven M. Hernandez, Dean Reich, Amer Sinha, Ashutosh

Sathe, Joe Kovac, Ashleah Gill, Ajay Kannan, Andrea D'olimpio, Martin Sevenich, Jay Whang, Been Kim, Khe Chai Sim, Jilin Chen, Jiageng Zhang, Shuba Lall, Yossi Matias, Bill Jia, Abe Friesen, Sara Nasso, Ashish Thapliyal, Bryan Perozzi, Ting Yu, Anna Shekhawat, Safeen Huda, Peter Grabowski, Eric Wang, Ashwin Sreevatsa, Hilal Dib, Mehadi Hassen, Parker Schuh, Vedrana Milutinovic, Chris Welty, Michael Quinn, Ali Shah, Bangju Wang, Gabe Barth-Maron, Justin Frye, Natalie Axelsson, Tao Zhu, Yukun Ma, Irene Giannoumis, Hanie Sedghi, Chang Ye, Yi Luan, Kevin Aydin, Bilva Chandra, Vivek Sampathkumar, Ronny Huang, Victor Lavrenko, Ahmed Eleryan, Zhi Hong, Steven Hansen, Sara Mc Carthy, Bidisha Samanta, Domagoj Ćevid, Xin Wang, Fangtao Li, Michael Voznesensky, Matt Hoffman, Andreas Terzis, Vikash Sehwag, Gil Fidel, Luheng He, Mu Cai, Yanzhang He, Alex Feng, Martin Nikoltchev, Samrat Phatale, Jason Chase, Rory Lawton, Ming Zhang, Tom Ouyang, Manuel Tragut, Mehdi Hafezi Manshadi, Arjun Narayanan, Jiaming Shen, Xu Gao, Tolga Bolukbasi, Nick Roy, Xin Li, Daniel Golovin, Liviu Panait, Zhen Qin, Guangxing Han, Thomas Anthony, Sneha Kudugunta, Viorica Patraucean, Aniket Ray, Xinyun Chen, Xiaochen Yang, Tanuj Bhatia, Pranav Talluri, Alex Morris, Andrija Ražnatović, Bethanie Brownfield, James An, Sheng Peng, Patrick Kane, Ce Zheng, Nico Duduta, Joshua Kessinger, James Noraky, Siqi Liu, Keran Rong, Petar Veličković, Keith Rush, Alex Goldin, Fanny Wei, Shiva Mohan Reddy Garlapati, Caroline Pantofaru, Okwan Kwon, Jianmo Ni, Eric Noland, Julia Di Trapani, Françoise Beaufays, Abhijit Guha Roy, Yinlam Chow, Aybuke Turker, Geoffrey Cideron, Lantao Mei, Jon Clark, Qingyun Dou, Matko Bošnjak, Ralph Leith, Yuqing Du, Amir Yazdanbakhsh, Milad Nasr, Chester Kwak, Suraj Satishkumar Sheth, Alex Kaskasoli, Ankesh Anand, Balaji Lakshminarayanan, Sammy Jerome, David Bieber, Chun-Te Chu, Alexandre Senges, Tianxiao Shen, Mukund Sridhar, Ndaba Ndebele, Benjamin Beyret, Shakir Mohamed, Mia Chen, Markus Freitag, Jiaxian Guo, Luyang Liu, Paul Roit, Heng Chen, Shen Yan, Tom Stone, JD Co-Reyes, Jeremy Cole, Salvatore Scellato, Shekoofeh Azizi, Hadi Hashemi, Alicia Jin, Anand Iyer, Marcella Valentine, András György, Arun Ahuja, Daniel Hernandez Diaz, Chen-Yu Lee, Nathan Clement, Weize Kong, Drew Garmon, Ishaan Watts, Kush Bhatia, Khyatti Gupta, Matt Miecnikowski, Hugo Vallet, Ankur Taly, Edward Loper, Saket Joshi, James Atwood, Jo Chick, Mark Collier, Fotis Iliopoulos, Ryan Trostle, Beliz Gunel, Ramiro Leal-Cavazos, Arnar Mar Hrafnkelsson, Michael Guzman, Xiaoen Ju, Andy Forbes, Jesse Emond, Kushal Chauhan, Ben Caine, Li Xiao, Wenjun Zeng, Alexandre Moufarek, Daniel Murphy, Maya Meng, Nitish Gupta, Felix Riedel, Anil Das, Elijah Lawal, Shashi Narayan, Tiberiu Sosea, James Swirhun, Linda Friso, Behnam Neyshabur, Jing Lu, Sertan Girgin, Michael Wunder, Edouard Yvinec, Aroonalok Pyne, Victor Carbune, Shruti Rijhwani, Yang Guo, Tulsee Doshi, Anton Briukhov, Max Bain, Ayal Hitron, Xuanhui Wang, Ashish Gupta, Ke Chen, Cosmo Du, Weiyang Zhang, Dhruv Shah, Arjun Akula, Max Dylla, Ashyana Kachra, Weicheng Kuo, Tingting Zou, Lily Wang, Luyao Xu, Jifan Zhu, Justin Snyder, Sachit Menon, Orhan Firat, Igor Mordatch, Yuan Yuan, Natalia Ponomareva, Rory Blevins, Lawrence Moore, Weijun Wang, Phil Chen, Martin Scholz, Artur Dwornik, Jason Lin, Sicheng Li, Diego Antognini, Te I, Xiaodan Song, Matt Miller, Uday Kalra, Adam Raveret, Oscar Akerlund, Felix Wu, Andrew Nystrom, Namrata Godbole, Tianqi Liu, Hannah DeBalsi, Jewel Zhao, Buhuang Liu, Avi Caciularu, Lauren Lax, Urvashi Khandelwal, Victoria Langston, Eric Bailey, Silvio Lattanzi, Yufei Wang, Neel Kovelamudi, Sneha Mondal, Guru Guruganesh, Nan Hua, Ofir Roval, Paweł Wesołowski, Rishikesh Ingale, Jonathan Halcrow, Tim Sohn, Christof Angermueller, Bahram Raad, Eli Stickgold, Eva Lu, Alec Kosik, Jing Xie, Timothy Lillicrap, Austin Huang, Lydia Lihui Zhang, Dominik Paulus, Clement Farabet, Alex Wertheim, Bing Wang, Rishabh Joshi, Chu ling Ko, Yonghui Wu, Shubham Agrawal, Lily Lin, XiangHai Sheng, Peter Sung, Tyler Breland-King, Christina Butterfield, Swapnil Gawde, Sumeet Singh, Qiao Zhang, Raj Apte, Shilpa Shetty, Adrian Hutter, Tao Li, Elizabeth Salesky, Federico Lebron, Jonni Kanerva, Michela Paganini, Arthur Nguyen, Rohith Vallu, Jan-Thorsten Peter, Sarmishta Velury, David Kao, Jay Hoover, Anna Bortsova, Colton Bishop, Shoshana Jakobovits, Alessandro Agostini, Alekh Agarwal, Chang Liu, Charles Kwong, Sasan Tavakkol, Ioana Bica, Alex Greve, Anirudh GP, Jake Marcus, Le Hou, Tom Duerig, Rivka Moroshko, Dave Lacey, Andy Davis, Julien Amelot, Guohui Wang, Frank Kim, Theofilos Strinopoulos, Hui Wan, Charline Le Lan, Shankar Krishnan, Haotian Tang, Peter Humphreys, Junwen Bai, Idan Heimlich Shtacher, Diego Machado, Chenxi Pang, Ken Burke, Dangyi Liu, Renga Aravamudhan, Yue Song, Ed Hirst, Abhimanyu Singh, Brendan Jou, Liang Bai, Francesco Piccinno, Chuyuan Kelly Fu, Robin Alazard, Barak Meiri, Daniel Winter, Charlie Chen, Mingda Zhang, Jens Heitkaemper, John Lambert, Jinhyuk Lee, Alexander Frömmgen, Sergey Rogulenko, Pranav Nair, Paul Niemczyk, Anton Bulyenov, Bibo Xu, Hadar Shemtov, Morteza Zadimoghaddam, Serge Toropov, Mateo Wirth, Hanjun Dai, Sreenivas Gollapudi, Daniel Zheng, Alex Kurakin,

Chansoo Lee, Kalesha Bullard, Nicolas Serrano, Ivana Balazevic, Yang Li, Johan Schalkwyk, Mark Murphy, Mingyang Zhang, Kevin Sequeira, Romina Datta, Nishant Agrawal, Charles Sutton, Nithya Attaluri, Mencher Chiang, Wael Farhan, Gregory Thornton, Kate Lin, Travis Choma, Hung Nguyen, Kingshuk Dasgupta, Dirk Robinson, Iulia Comşa, Michael Riley, Arjun Pillai, Basil Mustafa, Ben Golan, Amir Zandieh, Jean-Baptiste Lespiau, Billy Porter, David Ross, Sujeevan Rajayogam, Mohit Agarwal, Subhashini Venugopalan, Bobak Shahriari, Qiqi Yan, Hao Xu, Taylor Tobin, Pavel Dubov, Hongzhi Shi, Adrià Recasens, Anton Kovsharov, Sebastian Borgeaud, Lucio Dery, Shanthal Vasanth, Elena Gribovskaya, Linhai Qiu, Mahdis Mahdieh, Wojtek Skut, Elizabeth Nielsen, CJ Zheng, Adams Yu, Carrie Grimes Bostock, Shaleen Gupta, Aaron Archer, Chris Rawles, Elinor Davies, Alexey Svyatkovskiy, Tomy Tsai, Yoni Halpern, Christian Reisswig, Bartek Wydrowski, Bo Chang, Joan Puigcerver, Mor Hazan Taege, Jian Li, Eva Schnider, Xinjian Li, Dragos Dena, Yunhan Xu, Umesh Telang, Tianze Shi, Heiga Zen, Kyle Kastner, Yeongil Ko, Neesha Subramaniam, Aviral Kumar, Pete Blois, Zhuyun Dai, John Wieting, Yifeng Lu, Yoel Zeldes, Tian Xie, Anja Hauth, Alexandru Ţifrea, Yuqi Li, Sam El-Husseini, Dan Abolafia, Howard Zhou, Wen Ding, Sahra Ghalebikesabi, Carlos Guía, Andrii Maksai, Ágoston Weisz, Sercan Arik, Nick Sukhanov, Aga Świetlik, Xuhui Jia, Luo Yu, Weiyue Wang, Mark Brand, Dawn Bloxwich, Sean Kirmani, Zhe Chen, Alec Go, Pablo Sprechmann, Nithish Kannen, Alen Carin, Paramjit Sandhu, Isabel Edkins, Leslie Nooteboom, Jai Gupta, Loren Maggiore, Javad Azizi, Yael Pritch, Pengcheng Yin, Mansi Gupta, Danny Tarlow, Duncan Smith, Desi Ivanov, Mohammad Babaeizadeh, Ankita Goel, Satish Kambala, Grace Chu, Matej Kastelic, Michelle Liu, Hagen Soltau, Austin Stone, Shivani Agrawal, Min Kim, Kedar Soparkar, Srinivas Tadepalli, Oskar Bunyan, Rachel Soh, Arvind Kannan, DY Kim, Blake JianHang Chen, Afief Halumi, Sudeshna Roy, Yulong Wang, Olcan Sercinoglu, Gena Gibson, Sijal Bhatnagar, Motoki Sano, Daniel von Dincklage, Qingchun Ren, Blagoj Mitrevski, Mirek Olšák, Jennifer She, Carl Doersch, Jilei, Wang, Bingyuan Liu, Qijun Tan, Tamar Yakar, Tris Warkentin, Alex Ramirez, Carl Lebsack, Josh Dillon, Rajiv Mathews, Tom Cobley, Zelin Wu, Zhuoyuan Chen, Jon Simon, Swaroop Nath, Tara Sainath, Alexei Bendebury, Ryan Julian, Bharath Mankalale, Daria Ćurko, Paulo Zacchello, Adam R. Brown, Kiranbir Sodhia, Heidi Howard, Sergi Caelles, Abhinav Gupta, Gareth Evans, Anna Bulanova, Lesley Katzen, Roman Goldenberg, Anton Tsitsulin, Joe Stanton, Benoit Schillings, Vitaly Kovalev, Corey Fry, Rushin Shah, Kuo Lin, Shyam Upadhyay, Cheng Li, Soroush Radpour, Marcello Maggioni, Jing Xiong, Lukas Haas, Jenny Brennan, Aishwarya Kamath, Nikolay Savinov, Arsha Nagrani, Trevor Yacovone, Ryan Kappedal, Kostas Andriopoulos, Li Lao, YaGuang Li, Grigory Rozhdestvenskiy, Kazuma Hashimoto, Andrew Audibert, Sophia Austin, Daniel Rodriguez, Anian Ruoss, Garrett Honke, Deep Karkhanis, Xi Xiong, Qing Wei, James Huang, Zhaoqi Leng, Vittal Premachandran, Stan Bileschi, Georgios Evangelopoulos, Thomas Mensink, Jay Pavagadhi, Denis Teplyashin, Paul Chang, Linting Xue, Garrett Tanzer, Sally Goldman, Kaushal Patel, Shixin Li, Jeremy Wiesner, Ivy Zheng, Ian Stewart-Binks, Jie Han, Zhi Li, Liangchen Luo, Karel Lenc, Mario Lučić, Fuzhao Xue, Ryan Mullins, Alexey Guseynov, Chung-Ching Chang, Isaac Galatzer-Levy, Adam Zhang, Garrett Bingham, Grace Hu, Ale Hartman, Yue Ma, Jordan Griffith, Alex Irpan, Carey Radebaugh, Summer Yue, Lijie Fan, Victor Ungureanu, Christina Sorokin, Hannah Teufel, Peiran Li, Rohan Anil, Dimitris Paparas, Todd Wang, Chu-Cheng Lin, Hui Peng, Megan Shum, Goran Petrovic, Demetra Brady, Richard Nguyen, Klaus Macherey, Zhihao Li, Harman Singh, Madhavi Yenugula, Mariko Iinuma, Xinyi Chen, Kavya Kopparapu, Alexey Stern, Shachi Dave, Chandu Thekkath, Florence Perot, Anurag Kumar, Fangda Li, Yang Xiao, Matthew Bilotti, Mohammad Hossein Bateni, Isaac Noble, Lisa Lee, Amelio Vázquez-Reina, Julian Salazar, Xiaomeng Yang, Boyu Wang, Ela Gruzewska, Anand Rao, Sindhu Raghuram, Zheng Xu, Eyal Ben-David, Jieru Mei, Sid Dalmia, Zhaoyi Zhang, Yuchen Liu, Gagan Bansal, Helena Pankov, Steven Schwarcz, Andrea Burns, Christine Chan, Sumit Sanghai, Ricky Liang, Ethan Liang, Antoine He, Amy Stuart, Arun Narayanan, Yukun Zhu, Christian Frank, Bahar Fatemi, Amit Sabne, Oran Lang, Indro Bhattacharya, Shane Settle, Maria Wang, Brendan McMahan, Andrea Tacchetti, Livio Baldini Soares, Majid Hadian, Serkan Cabi, Timothy Chung, Nikita Putikhin, Gang Li, Jeremy Chen, Austin Tarango, Henryk Michalewski, Mehran Kazemi, Hussain Masoom, Hila Sheftel, Rakesh Shivanna, Archita Vadali, Ramona Comanescu, Doug Reid, Joss Moore, Arvind Neelakantan, Michaël Sander, Jonathan Herzig, Aviv Rosenberg, Mostafa Dehghani, JD Choi, Michael Fink, Reid Hayes, Eric Ge, Shitao Weng, Chia-Hua Ho, John Karro, Kalpesh Krishna, Lam Nguyen Thiet, Amy Skerry-Ryan, Daniel Eppens, Marco Andreetto, Navin Sarma, Silvano Bonacina, Burcu Karagol Ayan, Megha Nawhal, Zhihao Shan, Mike Dusenberry, Shantanu Thakoor, Sagar Gubbi, Duc Dung

Nguyen, Reut Tsarfaty, Samuel Albanie, Jovana Mitrović, Meet Gandhi, Bo-Juen Chen, Alessandro Epasto, Georgi Stephanov, Ye Jin, Samuel Gehman, Aida Amini, Jack Weber, Feryal Behbahani, Shawn Xu, Miltos Allamanis, Xi Chen, Myle Ott, Claire Sha, Michal Jastrzebski, Hang Qi, David Greene, Xinyi Wu, Abodunrinwa Toki, Daniel Vlasic, Jane Shapiro, Ragha Kotikalapudi, Zhe Shen, Takaaki Saeki, Sirui Xie, Albin Cassirer, Shikhar Bharadwaj, Tatsuya Kiyono, Srinadh Bhojanapalli, Elan Rosenfeld, Sam Ritter, Jieming Mao, João Gabriel Oliveira, Zoltan Egyed, Bernd Bandemer, Emilio Parisotto, Keisuke Kinoshita, Juliette Pluto, Petros Maniatis, Steve Li, Yaohui Guo, Golnaz Ghiasi, Jean Tarbouriech, Srimon Chatterjee, Julie Jin, Katrina, Xu, Jennimaria Palomaki, Séb Arnold, Madhavi Sewak, Federico Piccinini, Mohit Sharma, Ben Albrecht, Sean Purser-haskell, Ashwin Vaswani, Chongyan Chen, Matheus Wisniewski, Qin Cao, John Aslanides, Nguyet Minh Phu, Maximilian Sieb, Lauren Agubuzu, Anne Zheng, Daniel Sohn, Marco Selvi, Anders Andreassen, Krishan Subudhi, Prem Eruvbetine, Oliver Woodman, Tomas Mery, Sebastian Krause, Xiaoqi Ren, Xiao Ma, Jincheng Luo, Dawn Chen, Wei Fan, Henry Griffiths, Christian Schuler, Alice Li, Shujian Zhang, Jean-Michel Sarr, Shixin Luo, Riccardo Patana, Matthew Watson, Dani Naboulsi, Michael Collins, Sailesh Sidhwani, Emiel Hoogeboom, Sharon Silver, Emily Caveness, Xiaokai Zhao, Mikel Rodriguez, Maxine Deines, Libin Bai, Patrick Griffin, Marco Tagliasacchi, Emily Xue, Spandana Raj Babbula, Bo Pang, Nan Ding, Gloria Shen, Elijah Peake, Remi Crocker, Shubha Srinivas Raghvendra, Danny Swisher, Woohyun Han, Richa Singh, Ling Wu, Vladimir Pchelin, Tsendsuren Munkhdalai, Dana Alon, Geoff Bacon, Efren Robles, Jannis Bulian, Melvin Johnson, George Powell, Felipe Tiengo Ferreira, Yaoyiran Li, Frederik Benzing, Mihajlo Velimirović, Hubert Soyer, William Kong, Tony, Nguyên, Zhen Yang, Jeremiah Liu, Joost van Amersfoort, Daniel Gillick, Baochen Sun, Nathalie Rauschmayr, Katie Zhang, Serena Zhan, Tao Zhou, Alexey Frolov, Chengrun Yang, Denis Vnukov, Louis Rouillard, Hongji Li, Amol Mandhane, Nova Fallen, Rajesh Venkataraman, Clara Huiyi Hu, Jennifer Brennan, Jenny Lee, Jerry Chang, Martin Sundermeyer, Zhufeng Pan, Rosemary Ke, Simon Tong, Alex Fabrikant, William Bono, Jindong Gu, Ryan Foley, Yiran Mao, Manolis Delakis, Dhruva Bhaswar, Roy Frostig, Nick Li, Avital Zipori, Cath Hope, Olga Kozlova, Swaroop Mishra, Josip Djolonga, Craig Schiff, Majd Al Merey, Eleftheria Briakou, Peter Morgan, Andy Wan, Avinatan Hassidim, RJ Skerry-Ryan, Kuntal Sengupta, Mary Jasarevic, Praveen Kallakuri, Paige Kunkle, Hannah Brennan, Tom Lieber, Hassan Mansoor, Julian Walker, Bing Zhang, Annie Xie, Goran Žužić, Adaeze Chukwuka, Alex Druinsky, Donghyun Cho, Rui Yao, Ferjad Naeem, Shiraz Butt, Eunyoung Kim, Zhipeng Jia, Mandy Jordan, Adam Lelkes, Mark Kurzeja, Sophie Wang, James Zhao, Andrew Over, Abhishek Chakladar, Marcel Prasetya, Neha Jha, Sriram Ganapathy, Yale Cong, Prakash Shroff, Carl Saroufim, Sobhan Miryoosefi, Mohamed Hammad, Tajwar Nasir, Weijuan Xi, Yang Gao, Young Maeng, Ben Hora, Chin-Yi Cheng, Parisa Haghani, Yoad Lewenberg, Caden Lu, Martin Matysiak, Naina Raisinghani, Huiyu Wang, Lexi Baugher, Rahul Sukthankar, Minh Giang, John Schultz, Noah Fiedel, Minmin Chen, Cheng-Chun Lee, Tapomay Dey, Hao Zheng, Shachi Paul, Celine Smith, Andy Ly, Yicheng Wang, Rishabh Bansal, Bartek Perz, Susanna Ricco, Stasha Blank, Vaishakh Keshava, Deepak Sharma, Marvin Chow, Kunal Lad, Komal Jalan, Simon Osindero, Craig Swanson, Jacob Scott, Anastasija Ilić, Xiaowei Li, Siddhartha Reddy Jonnalagadda, Afzal Shama Soudagar, Yan Xiong, Bat-Orgil Batsaikhan, Daniel Jarrett, Naveen Kumar, Maulik Shah, Matt Lawlor, Austin Waters, Mark Graham, Rhys May, Sabela Ramos, Sandra Lefdal, Zeynep Cankara, Nacho Cano, Brendan O'Donoghue, Jed Borovik, Frederick Liu, Jordan Grimstad, Mahmoud Alnahlawi, Katerina Tsihlas, Tom Hudson, Nikolai Grigorev, Yiling Jia, Terry Huang, Tobenna Peter Igwe, Sergei Lebedev, Xiaodan Tang, Igor Krivokon, Frankie Garcia, Melissa Tan, Eric Jia, Peter Stys, Shikhar Vashishth, Yu Liang, Balaji Venkatraman, Chenjie Gu, Anastasios Kementsietsidis, Chen Zhu, Junehyuk Jung, Yunfei Bai, Mohammad Javad Hosseini, Faruk Ahmed, Aditya Gupta, Xin Yuan, Shereen Ashraf, Shitij Nigam, Gautam Vasudevan, Pranjal Awasthi, Adi Mayrav Gilady, Zelda Mariet, Ramy Eskander, Haiguang Li, Hexiang Hu, Guillermo Garrido, Philippe Schlattner, George Zhang, Rohun Saxena, Petar Dević, Kritika Muralidharan, Ashwin Murthy, Yiqian Zhou, Min Choi, Arissa Wongpanich, Zhengdong Wang, Premal Shah, Yuntao Xu, Yiling Huang, Stephen Spencer, Alice Chen, James Cohan, Junjie Wang, Jonathan Tompson, Junru Wu, Ruba Haroun, Haiqiong Li, Blanca Huergo, Fan Yang, Tongxin Yin, James Wendt, Michael Bendersky, Rahma Chaabouni, Javier Snaider, Johan Ferret, Abhishek Jindal, Tara Thompson, Andrew Xue, Will Bishop, Shubham Milind Phal, Archit Sharma, Yunhsuan Sung, Prabakar Radhakrishnan, Mo Shomrat, Reeve Ingle, Roopali Vij, Justin Gilmer, Mihai Dorin Istin, Sam Sobell, Yang Lu, Emily Nottage, Dorsa Sadigh, Jeremiah Willcock, Tingnan Zhang, Steve Xu, Sasha Brown, Katherine Lee, Gary Wang, Yun Zhu, Yi Tay, Cheolmin Kim, Audrey Gutierrez, Abhanshu Sharma, Yongqin Xian, Sungyong

Seo, Claire Cui, Elena Pochernina, Cip Baetu, Krzysztof Jastrzębski, Mimi Ly, Mohamed Elhawaty, Dan Suh, Eren Sezener, Pidong Wang, Nancy Yuen, George Tucker, Jiahao Cai, Zuguang Yang, Cindy Wang, Alex Muzio, Hai Qian, Jae Yoo, Derek Lockhart, Kevin R. McKee, Mandy Guo, Malika Mehrotra, Artur Mendonça, Sanket Vaibhav Mehta, Sherry Ben, Chetan Tekur, Jiaqi Mu, Muye Zhu, Victoria Krakovna, Hongrae Lee, AJ Maschinot, Sébastien Cevey, Hyun-Jeong Choe, Aijun Bai, Hansa Srinivasan, Derek Gasaway, Nick Young, Patrick Siegler, Dan Holtmann-Rice, Vihari Piratla, Kate Baumli, Roey Yogev, Alex Hofer, Hado van Hasselt, Svetlana Grant, Yuri Chervonyi, David Silver, Andrew Hogue, Ayushi Agarwal, Kathie Wang, Preeti Singh, Four Flynn, Josh Lipschultz, Robert David, Lizzetth Bellot, Yao-Yuan Yang, Long Le, Filippo Graziano, Kate Olszewska, Kevin Hui, Akanksha Maurya, Nikos Parotsidis, Weijie Chen, Tayo Oguntebi, Joe Kelley, Anirudh Baddepudi, Johannes Mauerer, Gregory Shaw, Alex Siegman, Lin Yang, Shravya Shetty, Subhrajit Roy, Yunting Song, Wojciech Stokowiec, Ryan Burnell, Omkar Savant, Robert Busa-Fekete, Jin Miao, Samrat Ghosh, Liam MacDermed, Phillip Lippe, Mikhail Dektiarev, Zach Behrman, Fabian Mentzer, Kelvin Nguyen, Meng Wei, Siddharth Verma, Chris Knutsen, Sudeep Dasari, Zhipeng Yan, Petr Mitrichev, Xingyu Wang, Virat Shejwalkar, Jacob Austin, Srinivas Sunkara, Navneet Potti, Yan Virin, Christian Wright, Gaël Liu, Oriana Riva, Etienne Pot, Greg Kochanski, Quoc Le, Gargi Balasubramaniam, Arka Dhar, Yuguo Liao, Adam Bloniarz, Divyansh Shukla, Elizabeth Cole, Jong Lee, Sheng Zhang, Sushant Kafle, Siddharth Vashishtha, Parsa Mahmoudieh, Grace Chen, Raphael Hoffmann, Pranesh Srinivasan, Agustin Dal Lago, Yoav Ben Shalom, Zi Wang, Michael Elabd, Anuj Sharma, Junhyuk Oh, Suraj Kothawade, Maigo Le, Marianne Monteiro, Shentao Yang, Kaiz Alarakyia, Robert Geirhos, Diana Mincu, Håvard Garnes, Hayato Kobayashi, Soroosh Mariooryad, Kacper Krasowiak, Zhixin, Lai, Shibl Mourad, Mingqiu Wang, Fan Bu, Ophir Aharoni, Guanjie Chen, Abhimanyu Goyal, Vadim Zubov, Ankur Bapna, Elahe Dabir, Nisarg Kothari, Kay Lamerigts, Nicola De Cao, Jeremy Shar, Christopher Yew, Nitish Kulkarni, Dre Mahaarachchi, Mandar Joshi, Zhenhai Zhu, Jared Lichtarge, Yichao Zhou, Hannah Muckenhirn, Vittorio Selo, Oriol Vinyals, Peter Chen, Anthony Brohan, Vaibhav Mehta, Sarah Cogan, Ruth Wang, Ty Geri, Wei-Jen Ko, Wei Chen, Fabio Viola, Keshav Shivam, Lisa Wang, Madeleine Clare Elish, Raluca Ada Popa, Sébastien Pereira, Jianqiao Liu, Raphael Koster, Donnie Kim, Gufeng Zhang, Sayna Ebrahimi, Partha Talukdar, Yanyan Zheng, Petra Poklukar, Ales Mikhalap, Dale Johnson, Anitha Vijayakumar, Mark Omernick, Matt Dibb, Ayush Dubey, Qiong Hu, Apurv Suman, Vaibhav Aggarwal, Ilya Kornakov, Fei Xia, Wing Lowe, Alexey Kolganov, Ted Xiao, Vitaly Nikolaev, Steven Hemingray, Bonnie Li, Joana Iljazi, Mikołaj Rybiński, Ballie Sandhu, Peggy Lu, Thang Luong, Rodolphe Jenatton, Vineetha Govindaraj, Hui, Li, Gabriel Dulac-Arnold, Wonpyo Park, Henry Wang, Abhinit Modi, Jean Pouget-Abadie, Kristina Greller, Rahul Gupta, Robert Berry, Prajit Ramachandran, Jinyu Xie, Liam McCafferty, Jianling Wang, Kilol Gupta, Hyeontaek Lim, Blaž Bratanič, Andy Brock, Ilia Akolzin, Jim Sproch, Dan Karliner, Duhyeon Kim, Adrian Goedeckemeyer, Noam Shazeer, Cordelia Schmid, Daniele Calandriello, Parul Bhatia, Krzysztof Choromanski, Ceslee Montgomery, Dheeru Dua, Ana Ramalho, Helen King, Yue Gao, Lynn Nguyen, David Lindner, Divya Pitta, Oleaser Johnson, Khalid Salama, Diego Ardila, Michael Han, Erin Farnese, Seth Odoom, Ziyue Wang, Xiangzhuo Ding, Norman Rink, Ray Smith, Harshal Tushar Lehri, Eden Cohen, Neera Vats, Tong He, Parthasarathy Gopavarapu, Adam Paszke, Miteyan Patel, Wouter Van Gansbeke, Lucia Loher, Luis Castro, Maria Voitovich, Tamara von Glehn, Nelson George, Simon Niklaus, Zach Eaton-Rosen, Nemanja Rakićević, Erik Jue, Sagi Perel, Carrie Zhang, Yuval Bahat, Angéline Pouget, Zhi Xing, Fantine Huot, Ashish Shenoy, Taylor Bos, Vincent Coriou, Bryan Richter, Natasha Noy, Yaqing Wang, Santiago Ontanon, Siyang Qin, Gleb Makarchuk, Demis Hassabis, Zhuowan Li, Mandar Sharma, Kumaran Venkatesan, Iurii Kemaev, Roxanne Daniel, Shiyu Huang, Saloni Shah, Octavio Ponce, Warren, Chen, Manaal Faruqui, Jialin Wu, Slavica Andačić, Szabolcs Payrits, Daniel McDuff, Tom Hume, Yuan Cao, MH Tessler, Qingze Wang, Yinan Wang, Ivor Rendulic, Eirikur Agustsson, Matthew Johnson, Tanya Lando, Andrew Howard, Sri Gayatri Sundara Padmanabhan, Mayank Daswani, Andrea Banino, Michael Kilgore, Jonathan Heek, Ziwei Ji, Alvaro Caceres, Conglong Li, Nora Kassner, Alexey Vlaskin, Zeyu Liu, Alex Grills, Yanhan Hou, Roykrong Sukkerd, Gowoon Cheon, Nishita Shetty, Larisa Markeeva, Piotr Stanczyk, Tejas Iyer, Yuan Gong, Shawn Gao, Keerthana Gopalakrishnan, Tim Blyth, Malcolm Reynolds, Avishkar Bhoopchand, Misha Bilenko, Dero Gharibian, Vicky Zayats, Aleksandra Faust, Abhinav Singh, Min Ma, Hongyang Jiao, Sudheendra Vijayanarasimhan, Lora Aroyo, Vikas Yadav, Sarah Chakera, Ashwin Kakarla, Vilobh Meshram, Karol Gregor, Gabriela Botea, Evan Senter, Dawei Jia, Geza Kovacs, Neha Sharma, Sebastien Baur, Kai Kang, Yifan He, Lin Zhuo, Marija Kostelac, Itay Laish, Songyou Peng, Louis O'Bryan, Daniel Kasenberg,

Girish Ramchandra Rao, Edouard Leurent, Biao Zhang, Sage Stevens, Ana Salazar, Ye Zhang, Ivan Lobov, Jake Walker, Allen Porter, Morgan Redshaw, Han Ke, Abhishek Rao, Alex Lee, Hoi Lam, Michael Moffitt, Jaeyoun Kim, Siyuan Qiao, Terry Koo, Robert Dadashi, Xinying Song, Mukund Sundararajan, Peng Xu, Chizu Kawamoto, Yan Zhong, Clara Barbu, Apoorv Reddy, Mauro Verzetti, Leon Li, George Papamakarios, Hanna Klimczak-Plucińska, Mary Cassin, Koray Kavukcuoglu, Rigel Swavely, Alain Vaucher, Jeffrey Zhao, Ross Hemsley, Michael Tschannen, Heming Ge, Gaurav Menghani, Yang Yu, Natalie Ha, Wei He, Xiao Wu, Maggie Song, Rachel Sterneck, Stefan Zinke, Dan A. Calian, Annie Marsden, Alejandro Cruzado Ruiz, Matteo Hessel, Almog Gueta, Benjamin Lee, Brian Farris, Manish Gupta, Yunjie Li, Mohammad Saleh, Vedant Misra, Kefan Xiao, Piermaria Mendolicchio, Gavin Buttimore, Varvara Krayvanova, Nigamaa Nayakanti, Matthew Wiethoff, Yash Pande, Azalia Mirhoseini, Ni Lao, Jasmine Liu, Yiqing Hua, Angie Chen, Yury Malkov, Dmitry Kalashnikov, Shubham Gupta, Kartik Audhkhasi, Yuexiang Zhai, Sudhindra Kopalle, Prateek Jain, Eran Ofek, Clemens Meyer, Khuslen Baatarsukh, Hana Strejček, Jun Qian, James Freedman, Ricardo Figueira, Michal Sokolik, Olivier Bachem, Raymond Lin, Dia Kharrat, Chris Hidey, Pingmei Xu, Dennis Duan, Yin Li, Muge Ersoy, Richard Everett, Kevin Cen, Rebeca Santamaria-Fernandez, Amir Taubenfeld, Ian Mackinnon, Linda Deng, Polina Zablotskaia, Shashank Viswanadha, Shivanker Goel, Damion Yates, Yunxiao Deng, Peter Choy, Mingqing Chen, Abhishek Sinha, Alex Mossin, Yiming Wang, Arthur Szlam, Susan Hao, Paul Kishan Rubenstein, Metin Toksoz-Exley, Miranda Aperghis, Yin Zhong, Junwhan Ahn, Michael Isard, Olivier Lacombe, Florian Luisier, Chrysovalantis Anastasiou, Yogesh Kalley, Utsav Prabhu, Emma Dunleavy, Shaan Bijwadia, Justin Mao-Jones, Kelly Chen, Rama Pasumarthi, Emily Wood, Adil Dostmohamed, Nate Hurley, Jiri Simsa, Alicia Parrish, Mantas Pajarskas, Matt Harvey, Ondrej Skopek, Yony Kochinski, Javier Rey, Verena Rieser, Denny Zhou, Sun Jae Lee, Trilok Acharya, Guowang Li, Joe Jiang, Xiaofan Zhang, Bryant Gipson, Ethan Mahintorabi, Marco Gelmi, Nima Khajehnouri, Angel Yeh, Kayi Lee, Loic Matthey, Leslie Baker, Trang Pham, Han Fu, Alex Pak, Prakhar Gupta, Cristina Vasconcelos, Adam Sadovsky, Brian Walker, Sissie Hsiao, Patrik Zochbauer, Andreea Marzoca, Noam Velan, Junhao Zeng, Gilles Baechler, Danny Driess, Divya Jain, Yanping Huang, Lizzie Tao, John Maggs, Nir Levine, Jon Schneider, Erika Gemzer, Samuel Petit, Shan Han, Zach Fisher, Dustin Zelle, Courtney Biles, Eugene Ie, Asya Fadeeva, Casper Liu, Juliana Vicente Franco, Adrian Collister, Hao Zhang, Renshen Wang, Ruizhe Zhao, Leandro Kieliger, Kurt Shuster, Rui Zhu, Boqing Gong, Lawrence Chan, Ruoxi Sun, Sujoy Basu, Roland Zimmermann, Jamie Hayes, Abhishek Bapna, Jasper Snoek, Weel Yang, Puranjay Datta, Jad Al Abdallah, Kevin Kilgour, Lu Li, SQ Mah, Yennie Jun, Morgane Rivière, Abhijit Karmarkar, Tammo Spalink, Tao Huang, Lucas Gonzalez, Duc-Hieu Tran, Averi Nowak, John Palowitch, Martin Chadwick, Ellie Talius, Harsh Mehta, Thibault Sellam, Philipp Fränken, Massimo Nicosia, Kyle He, Aditya Kini, David Amos, Sugato Basu, Harrison Jobe, Eleni Shaw, Qiantong Xu, Colin Evans, Daisuke Ikeda, Chaochao Yan, Larry Jin, Lun Wang, Sachin Yadav, Ilia Labzovsky, Ramesh Sampath, Ada Ma, Candice Schumann, Aditya Siddhant, Rohin Shah, John Youssef, Rishabh Agarwal, Natalie Dabney, Alessio Tonioni, Moran Ambar, Jing Li, Isabelle Guyon, Benny Li, David Soergel, Boya Fang, Georgi Karadzhov, Cristian Udrescu, Trieu Trinh, Vikas Raunak, Seb Noury, Dee Guo, Sonal Gupta, Mara Finkelstein, Denis Petek, Lihao Liang, Greg Billock, Pei Sun, David Wood, Yiwen Song, Xiaobin Yu, Tatiana Matejovicova, Regev Cohen, Kalyan Andra, David D'Ambrosio, Zhiwei Deng, Vincent Nallatamby, Ebrahim Songhori, Rumen Dangovski, Andrew Lampinen, Pankil Botadra, Adam Hillier, Jiawei Cao, Nagabhushan Baddi, Adhi Kuncoro, Toshihiro Yoshino, Ankit Bhagatwala, Marcáurelio Ranzato, Rylan Schaeffer, Tianlin Liu, Shuai Ye, Obaid Sarvana, John Nham, Chenkai Kuang, Isabel Gao, Jinoo Baek, Shubham Mittal, Ayzaan Wahid, Anita Gergely, Bin Ni, Josh Feldman, Carrie Muir, Pascal Lamblin, Wolfgang Macherey, Ethan Dyer, Logan Kilpatrick, Víctor Campos, Mukul Bhutani, Stanislav Fort, Yanif Ahmad, Aliaksei Severyn, Kleopatra Chatziprimou, Oleksandr Ferludin, Mason Dimarco, Aditya Kusupati, Joe Heyward, Dan Bahir, Kevin Villela, Katie Millican, Dror Marcus, Sanaz Bahargam, Caglar Unlu, Nicholas Roth, Zichuan Wei, Siddharth Gopal, Deepanway Ghoshal, Edward Lee, Sharon Lin, Jennie Lees, Dayeong Lee, Anahita Hosseini, Connie Fan, Seth Neel, Marcus Wu, Yasemin Altun, Honglong Cai, Enrique Piqueras, Josh Woodward, Alessandro Bissacco, Salem Haykal, Mahyar Bordbar, Prasha Sundaram, Sarah Hodkinson, Daniel Toyama, George Polovets, Austin Myers, Anu Sinha, Tomer Levinboim, Kashyap Krishnakumar, Rachita Chhaparia, Tatiana Sholokhova, Nitesh Bharadwaj Gundavarapu, Ganesh Jawahar, Haroon Qureshi, Jieru Hu, Nikola Momchev, Matthew Rahtz, Renjie Wu, Aishwarya P S, Kedar Dhamdhere, Meiqi Guo, Umang Gupta, Ali Eslami, Mariano Schain, Michiel Blokzijl, David Welling, Dave Orr, Levent Bolelli, Nicolas Perez-Nieves, Mikhail Sirotenko, Aman

Prasad, Arjun Kar, Borja De Balle Pigem, Tayfun Terzi, Gellért Weisz, Dipankar Ghosh, Aditi Mavalankar, Dhruv Madeka, Kaspar Daugaard, Hartwig Adam, Viraj Shah, Dana Berman, Maggie Tran, Steven Baker, Ewa Andrejczuk, Grishma Chole, Ganna Raboshchuk, Mahdi Mirzazadeh, Thais Kagohara, Shimu Wu, Christian Schallhart, Bernett Orlando, Chen Wang, Alban Rrustemi, Hao Xiong, Hao Liu, Arpi Vezer, Nolan Ramsden, Shuo yiin Chang, Sidharth Mudgal, Yan Li, Nino Vieillard, Yedid Hoshen, Farooq Ahmad, Ambrose Slone, Amy Hua, Natan Potikha, Mirko Rossini, Jon Stritar, Sushant Prakash, Zifeng Wang, Xuanyi Dong, Alireza Nazari, Efrat Nehoran, Kaan Tekelioglu, Yinxiao Li, Kartikeya Badola, Tom Funkhouser, Yuanzhen Li, Varun Yerram, Ramya Ganeshan, Daniel Formoso, Karol Langner, Tian Shi, Huijian Li, Yumeya Yamamori, Amayika Panda, Alaa Saade, Angelo Scorza Scarpati, Chris Breaux, CJ Carey, Zongwei Zhou, Cho-Jui Hsieh, Sophie Bridgers, Alena Butryna, Nishesh Gupta, Vaibhav Tulsyan, Sanghyun Woo, Evgenii Eltyshev, Will Grathwohl, Chanel Parks, Seth Benjamin, Rina Panigrahy, Shenil Dodhia, Daniel De Freitas, Chris Sauer, Will Song, Ferran Alet, Jackson Tolins, Cosmin Paduraru, Xingyi Zhou, Brian Albert, Zizhao Zhang, Lei Shu, Mudit Bansal, Sarah Nguyen, Amir Globerson, Owen Xiao, James Manyika, Tom Hennigan, Rong Rong, Josip Matak, Anton Bakalov, Ankur Sharma, Danila Sinopalnikov, Andrew Pierson, Stephen Roller, Geoff Brown, Mingcen Gao, Toshiyuki Fukuzawa, Amin Ghafouri, Kenny Vassigh, Iain Barr, Zhicheng Wang, Anna Korsun, Rajesh Jayaram, Lijie Ren, Tim Zaman, Samira Khan, Yana Lunts, Dan Deutsch, Dave Uthus, Nitzan Katz, Masha Samsikova, Amr Khalifa, Nikhil Sethi, Jiao Sun, Luming Tang, Uri Alon, Xianghong Luo, Dian Yu, Abhishek Nayyar, Bryce Petrini, Will Truong, Vincent Hellendoorn, Nikolai Chinaev, Chris Alberti, Wei Wang, Jingcao Hu, Vahab Mirrokni, Ananth Balashankar, Avia Aharon, Aahil Mehta, Ahmet Iscen, Joseph Kready, Lucas Manning, Anhad Mohananey, Yuankai Chen, Anshuman Tripathi, Allen Wu, Igor Petrovski, Dawsen Hwang, Martin Baeuml, Shreyas Chandrakaladharan, Yuan Liu, Rey Coaguila, Maxwell Chen, Sally Ma, Pouya Tafti, Susheel Tatineni, Terry Spitz, Jiayu Ye, Paul Vicol, Mihaela Rosca, Adrià Puigdomènech, Zohar Yahav, Sanjay Ghemawat, Hanzhao Lin, Phoebe Kirk, Zaid Nabulsi, Sergey Brin, Bernd Bohnet, Ken Caluwaerts, Aditya Srikanth Veerubhotla, Dan Zheng, Zihang Dai, Petre Petrov, Yichong Xu, Ramin Mehran, Zhuo Xu, Luisa Zintgraf, Jiho Choi, Spurthi Amba Hombaiah, Romal Thoppilan, Sashank Reddi, Lukasz Lew, Li Li, Kellie Webster, KP Sawhney, Lampros Lamprou, Siamak Shakeri, Mayank Lunayach, Jianmin Chen, Sumit Bagri, Alex Salcianu, Ying Chen, Yani Donchev, Charlotte Magister, Signe Nørly, Vitor Rodrigues, Tomas Izo, Hila Noga, Joe Zou, Thomas Köppe, Wenxuan Zhou, Kenton Lee, Xiangzhu Long, Danielle Eisenbud, Anthony Chen, Connor Schenck, Chi Ming To, Peilin Zhong, Emanuel Taropa, Minh Truong, Omer Levy, Danilo Martins, Zhiyuan Zhang, Christopher Semturs, Kelvin Zhang, Alex Yakubovich, Pol Moreno, Lara McConnaughey, Di Lu, Sam Redmond, Lotte Weerts, Yonatan Bitton, Tiziana Refice, Nicolas Lacasse, Arthur Conmy, Corentin Tallec, Julian Odell, Hannah Forbes-Pollard, Arkadiusz Socala, Jonathan Hoech, Pushmeet Kohli, Alanna Walton, Rui Wang, Mikita Sazanovich, Kexin Zhu, Andrei Kapishnikov, Rich Galt, Matthew Denton, Ben Murdoch, Caitlin Sikora, Kareem Mohamed, Wei Wei, Uri First, Tim McConnell, Luis C. Cobo, James Qin, Thi Avrahami, Daniel Balle, Yu Watanabe, Annie Louis, Adam Kraft, Setareh Ariafar, Yiming Gu, Eugénie Rives, Charles Yoon, Andrei Rusu, James Cobon-Kerr, Chris Hahn, Jiaming Luo, Yuvein, Zhu, Niharika Ahuja, Rodrigo Benenson, Raphaël Lopez Kaufman, Honglin Yu, Lloyd Hightower, Junlin Zhang, Darren Ni, Lisa Anne Hendricks, Gabby Wang, Gal Yona, Lalit Jain, Pablo Barrio, Surya Bhupatiraju, Siva Velusamy, Allan Dafoe, Sebastian Riedel, Tara Thomas, Zhe Yuan, Mathias Bellaiche, Sheena Panthaplackel, Klemen Kloboves, Sarthak Jauhari, Canfer Akbulut, Todor Davchev, Evgeny Gladchenko, David Madras, Aleksandr Chuklin, Tyrone Hill, Quan Yuan, Mukundan Madhavan, Luke Leonhard, Dylan Scandinaro, Qihang Chen, Ning Niu, Arthur Douillard, Bogdan Damoc, Yasumasa Onoe, Fabian Pedregosa, Fred Bertsch, Chas Leichner, Joseph Pagadora, Jonathan Malmaud, Sameera Ponda, Andy Twigg, Oleksii Duzhyi, Jingwei Shen, Miaosen Wang, Roopal Garg, Jing Chen, Utku Evci, Jonathan Lee, Leon Liu, Koji Kojima, Masa Yamaguchi, Arunkumar Rajendran, AJ Piergiovanni, Vinodh Kumar Rajendran, Marco Fornoni, Gabriel Ibagon, Harry Ragan, Sadh MNM Khan, John Blitzer, Andrew Bunner, Guan Sun, Takahiro Kosakai, Scott Lundberg, Ndidi Elue, Kelvin Guu, SK Park, Jane Park, Arunachalam Narayanaswamy, Chengda Wu, Jayaram Mudigonda, Trevor Cohn, Hairong Mu, Ravi Kumar, Laura Graesser, Yichi Zhang, Richard Killam, Vincent Zhuang, Mai Giménez, Wael Al Jishi, Ruy Ley-Wild, Alex Zhai, Kazuki Osawa, Diego Cedillo, Jialu Liu, Mayank Upadhyay, Marcin Sieniek, Roshan Sharma, Tom Paine, Anelia Angelova, Sravanti Addepalli, Carolina Parada, Kingshuk Majumder, Avery Lamp, Sanjiv Kumar, Xiang Deng, Artiom Myaskovsky, Tea Sabolić, Jeffrey Dudek, Sarah York, Félix de Chaumont Quitry, Jiazhong Nie, Dee Cattle,

Alok Gunjan, Bilal Piot, Waleed Khawaja, Seojin Bang, Simon Wang, Siavash Khodadadeh, Raghavender R, Praynaa Rawlani, Richard Powell, Kevin Lee, Johannes Griesser, GS Oh, Cesar Magalhaes, Yujia Li, Simon Tokumine, Hadas Natalie Vogel, Dennis Hsu, Arturo BC, Disha Jindal, Matan Cohen, Zi Yang, Junwei Yuan, Dario de Cesare, Tony Bruguier, Jun Xu, Monica Roy, Alon Jacovi, Dan Belov, Rahul Arya, Phoenix Meadowlark, Shlomi Cohen-Ganor, Wenting Ye, Patrick Morris-Suzuki, Praseem Banzal, Gan Song, Pranavaraj Ponnuramu, Fred Zhang, George Scrivener, Salah Zaiem, Alif Raditya Rochman, Kehang Han, Badih Ghazi, Kate Lee, Shahar Drath, Daniel Suo, Antonious Girgis, Pradeep Shenoy, Duy Nguyen, Douglas Eck, Somit Gupta, Le Yan, Joao Carreira, Anmol Gulati, Ruoxin Sang, Daniil Mirylenka, Emma Cooney, Edward Chou, Mingyang Ling, Cindy Fan, Ben Coleman, Guilherme Tubone, Ravin Kumar, Jason Baldridge, Felix Hernandez-Campos, Angeliki Lazaridou, James Besley, Itay Yona, Neslihan Bulut, Quentin Wellens, AJ Pierigiovanni, Jasmine George, Richard Green, Pu Han, Connie Tao, Geoff Clark, Chong You, Abbas Abdolmaleki, Justin Fu, Tongzhou Chen, Ashwin Chaugule, Angad Chandorkar, Altaf Rahman, Will Thompson, Penporn Koanantakool, Mike Bernico, Jie Ren, Andrey Vlasov, Sergei Vassilvitskii, Maciej Kula, Yizhong Liang, Dahun Kim, Yangsibo Huang, Chengxi Ye, Dmitry Lepikhin, and Wesley Helmholz. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198/.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Trans. Assoc. Comput. Linguistics*, 9:160–175, 2021. doi: 10.1162/TACL_A_00359. URL https://doi.org/10.1162/tacl_a_00359.

Cosimo Galeone, Anna Ettorre, Minsu Park, Giuseppe Ettorre, and Daniele Ligorio. Perfect detection, failed control: The geometry of knowing vs. steering in language models, 2026. URL https://arxiv.org/abs/2606.24952.

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.

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, 2026. URL https://arxiv.org/abs/2607.15495.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=w7LU2s14kE.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1275. URL https://doi.org/10.18653/v1/D19-1275.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Phu Gia Hoang, Anwoy Chatterjee, Tanmoy Chakraborty, Iryna Gurevych, and Subhabrata Dutta. Sparse autoencoders encode both concepts and functions: The downstream geometry of feature effects, 2026. URL https://arxiv.org/abs/2607.24645.

Xinting Huang and Michael Hahn. Decomposing representation space into interpretable subspaces with unsupervised learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=SI2OZJuvPO.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

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.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL https://arxiv.org/abs/2408.05147.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese,

Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The information geometry of softmax: Probing and steering. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=BrIGLICBWy.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.647. URL https://doi.org/10.18653/v1/2020.acl-main.647.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In Paola Merlo, Jorg Tiedemann, and Reut

Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting latent steering vectors from pretrained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, volume ACL 2022 of *Findings of ACL*, pp. 566–581. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.48. URL https://doi.org/10.18653/v1/2022.findings-acl.48.

Jiuding Sun, Jing Huang, Sidharth Baskaran, Karel D'Oosterlinck, Christopher Potts, Michael Sklar, and Atticus Geiger. Hyperdas: Towards automating mechanistic interpretability with hypernetworks. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=6fDjUoEQvm.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Federico Tiblias, Irina Bigoulaeva, Jingcheng Niu, Simone Balloccu, and Iryna Gurevych. Hypothesis-driven feature manifold analysis in llms via supervised multi-dimensional scaling. *Trans. Mach. Learn. Res.*, 2026, 2026. URL https://openreview.net/forum?id=vCKZ40YYPr.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Sihan Wang, Jiayi Zhao, Qingyan Cao, Hongbo Yao, and Lin Shu. Fishback: Pullback fisher geometry for optimal activation steering in transformers, 2026. URL https://arxiv.org/abs/2605.17231.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/wu25a.html.

Daniel Wurgaft, Can Rager, Matthew Kowal, Vasudev Shyam, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphaël Sarfati, Thomas McGrath, Owen Lewis, Jack Merullo, Noah Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. Manifold steering reveals the shared geometry of neural network representation and behavior. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=QklBm5nnR9.

A Experimental Details

A.1 Dataset Construction and Prompts

We sampled 50 concepts from the list utilzied in Wu et al. (2025) and paired them with 304 concept-independent prefixes. For each concept and prefix, we generated a concept-positive completion, along with one neutral completion per prefix, using gpt-4.1-mini (OpenAI et al., 2024). This produced 15,200 positive and 304 neutral examples. We split the data by prefix to prevent overlap across the training, validation, and test sets. The complete dataset is included in the provided repository, and the generation prompts are reproduced below.

We then partitioned the 304 prefixes into training, validation, and test sets using a random 80/10/10 split. All positive and neutral examples sharing a prefix were assigned to the same split, preventing the same context from appearing in both training and evaluation. This produced 243 training, 31 validation, and 30 test prefixes.

Example prefixes include “The main takeaway is”, “The important question is”, “A more general way to say this is”, “The idea is useful because”, “The answer depends on”, “The simplest case is when”, “The main source of confusion is”, and “A useful comparison is”.

Concept-Positive Completion Prompt

```
You are generating a dataset example for studying language model
activations.
Concept description:
{concept_description}
Prefix:
{prefix}
Task:
Complete the prefix so that the resulting text naturally and clearly
expresses the concept.  Meaning the completion should exemplify the
concept very clearly.
Rules:
- Do not change the prefix.
- Continue directly from the prefix.
- The continuation should be natural, fluent, and specific.
- The concept should be clearly present, but do not sound like you are
defining the concept.
- Do not mention "concept", "dataset", "positive sample", or this
instruction.
- Avoid repetitive phrasing.
- Write one sentence only.
- Return only the continuation, not the full text or anything else.
```

Neutral Completion Prompt

```
You are completing a sentence.
Prefix:
{prefix}
Task:
Complete the prefix with a normal continuation one could expect to
see.
Rules:
- Do not change the prefix.
- Continue directly from the prefix.
- Write one sentence only.
- Return only the continuation, not the full text or anything else.
```

A.2 Concept Covector Construction

For each concept c, we collected the tokens that appeared in its positive training completions but never appeared in the neutral completions. We built these sets separately for each model tokenizer and excluded special tokens, such as beginning-of-sequence (BOS), end-of-sequence (EOS), padding, and unknown tokens. For each token y in the model vocabulary, let $n_c^+(y)$ denote its number of occurrences in positive completions for concept c, and let $n^0(y)$ denote its number of occurrences in neutral completions. We define

$$\mathcal{V}_c = \{y : n_c^+(y) > 0,\ n^0(y) = 0\}, \qquad \mathbf{r}_c(y) = \begin{cases} 1, & y \in \mathcal{V}_c, \\ 0, & y \notin \mathcal{V}_c. \end{cases} \tag{11}$$

Thus, $\mathbf{r}_c$ assigns weight one to tokens associated with concept c and zero to all other tokens, thus capturing the verbalized aspect of a concept.

A.3 Latent Structure Construction

We trained one direction for each concept, model, and layer using only the prefix-grouped training split. For every positive and neutral completion, we collected the residual-stream activation after the selected transformer block at each position predicting a completion token, using all 243 positive and 243 neutral training examples per concept.

For DiffMeans, we averaged the positive and neutral activations separately and used their difference:

$$\mathbf{v}_c^{\mathrm{DM}} = \frac{\boldsymbol{\mu}_c^+ - \boldsymbol{\mu}^0}{\left\|\boldsymbol{\mu}_c^+ - \boldsymbol{\mu}^0\right\|_2}, \tag{12}$$

where $\boldsymbol{\mu}_c^+$ and $\boldsymbol{\mu}^0$ are the mean positive and neutral activations, respectively.

The linear probe was a bias-free logistic classifier trained on the same activations using binary cross-entropy. Probe weights were initialized from $\mathcal{N}(0, 0.01^2)$. The final probe weight was normalized to unit Euclidean norm before being used as a direction. The constrained probe variants used the exact same setting, but restricted the learned direction to the corresponding rank-128 causal subspace.

For ReFT-r1, we followed the rank-one TopKReLU setting from Wu et al. (2025).

A.4 Jacobian and Capacity Computation

For an activation $\mathbf{h}$ and output position t, let $J_{\mathbf{h}}^t$ denote the Jacobian of the output logits with respect to $\mathbf{h}$. For future output positions, the tokens between the source and target positions are fixed to the observed continuation using teacher forcing.

Because the full Jacobian has shape $|\mathcal{V}| \times d$ it is inefficient to construct explicitly, we compute only its products with vectors using automatic differentiation. A Jacobian–vector product (JVP) gives the first-order change in the logits caused by moving $\mathbf{h}$ along a direction $\mathbf{v}$. We then apply the Jacobian of softmax at the original output distribution, converting this logit change into a first-order probability change.

A vector–Jacobian product (VJP), implemented by backpropagation, then maps this change back to activation space. This computes

$$G_{\mathbf{h}}^t \mathbf{v} = (J_{\mathbf{h}}^t)^\top F_{\mathbf{p}^t} J_{\mathbf{h}}^t \mathbf{v}$$

without explicitly constructing either $J_{\mathbf{h}}^t$ or $G_{\mathbf{h}}^t$. Capacity is the probability-weighted standard deviation of the logit changes produced by the JVP. All derivatives are computed using automatic differentiation.

B Additional Capacity Analysis

B.1 Derivation of the Local KL Approximation

For a small intervention $\alpha\mathbf{v}$, the logits at position t satisfy

$$\Delta\mathbf{z} = \alpha J_{\mathbf{h}}^t \mathbf{v} + O(\alpha^2). \tag{13}$$

For a softmax distribution, the KL divergence between the original and perturbed distributions has the second-order expansion

$$D_{\mathrm{KL}}\left(\mathbf{p} \,\|\, \mathbf{p}_{\Delta\mathbf{z}}\right) = \frac{1}{2}\Delta\mathbf{z}^\top F_{\mathbf{p}} \Delta\mathbf{z} + O(\|\Delta\mathbf{z}\|^3), \tag{14}$$

where $F_{\mathbf{p}} = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$ is the softmax Fisher matrix. Substituting $\Delta\mathbf{z} = \alpha J_{\mathbf{h}}^t \mathbf{v}$ gives

$$D_{\mathrm{KL}}\left(\mathbf{p} \,\|\, \mathbf{p}_{\alpha\mathbf{v}}\right) = \frac{\alpha^2}{2}\mathbf{v}^\top (J_{\mathbf{h}}^t)^\top F_{\mathbf{p}} J_{\mathbf{h}}^t \mathbf{v} + O(\alpha^3) = \frac{\alpha^2}{2}\mathbf{v}^\top G_{\mathbf{h}}^t \mathbf{v} + O(\alpha^3). \tag{15}$$

Thus, for sufficiently small interventions,

$$D_{\mathrm{KL}}\left(\mathbf{p} \,\|\, \mathbf{p}_{\alpha\mathbf{v}}\right) \approx \frac{\alpha^2}{2}\mathbf{v}^\top G_{\mathbf{h}}^t \mathbf{v}. \tag{16}$$

B.2 Capacity Prediction Across Future Positions

We include in Figure 7 additional results for offset t up to $t = 16$, for the analysis done in §3.2. We report the results as the distance from $y = x$ for predicted and actual KL. The larger the value, the bigger the approximation error. The figure shows that across layers the trends are very similar. Looking across offsets, the distance decreases across models, with Qwen being the exception with a slight increase. Generally, we see predicted KL tracks actual KL closely.

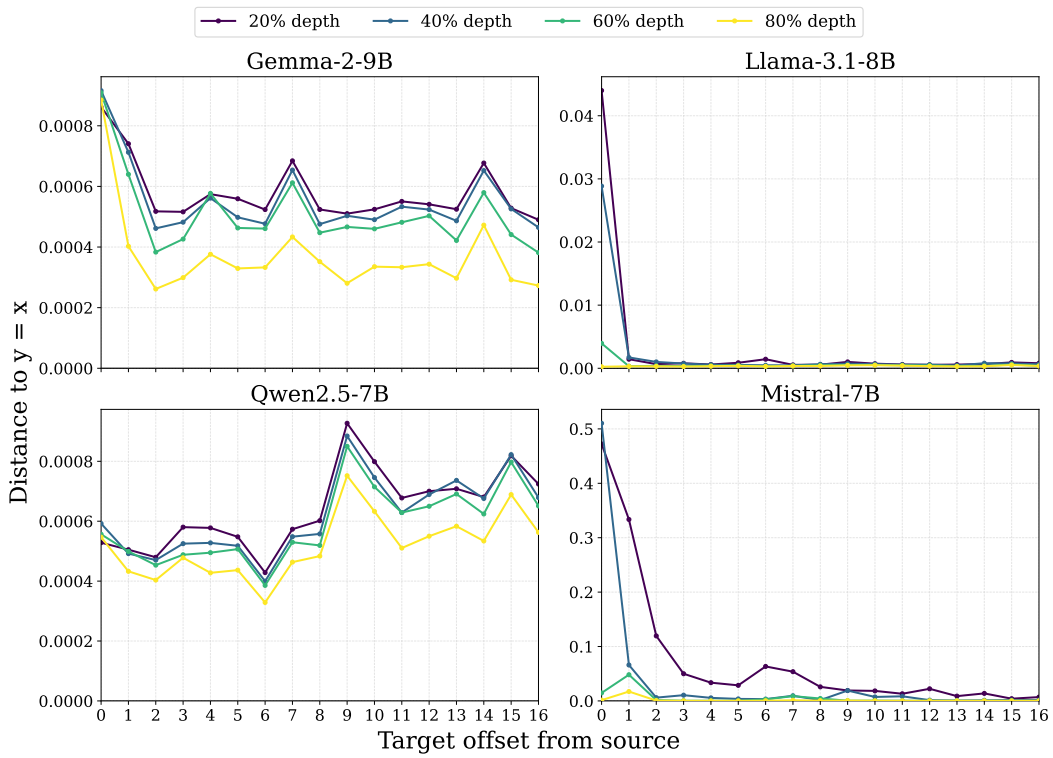


Figure 7: **Predicted KL closely tracks actual KL across future offsets and layers.** We report approximation error as the distance from $y = x$ between predicted and actual KL for offsets $t = 0, \ldots, 16$. Trends are consistent across layers, with error generally decreasing at later offsets, except for a slight increase in Qwen.

B.3 Choice of Sequence Horizon

Figure 8 shows the actual KL produced by interventions across offsets. We can see that the KL on future offsets decreases significantly with the offset position. This is true for all models, concepts and layers tested. Therefore, adding additional offsets to the analysis does not affect the results and trends observed. As such in this analysis we decided to include up to 8 additional offsets beyond the intervention position.

C Additional Responsiveness Analysis

C.1 Experimental and Fluency-Evaluation Details

We evaluate intervention strengths $\alpha \in \{0.4, 0.8, 1.2, 1.6, 2, 4, 10, 20, 40, 60, 100\}$. For each model, layer, and concept, the random baseline is sampled from an isotropic Gaussian in activation space and normalized to unit Euclidean norm. Fluency is evaluated by Gemini-2.5-Flash-Lite (Comanici et al., 2025) using the fluency prompt from Wu et al. (2025). Scores are then normalized. For each model, layer, and strength, we compute the unweighted mean fluency across all judged DiffMeans generations, pooling concepts and contexts. For each target $\tau \in \{0.8, 0.7, 0.6\}$, the fluency boundary is the largest tested α whose mean score is at least τ.

C.2 Results Across Layers and Support Thresholds

Figure 9 provides the results for all models, 4 layers, and both random and DiffMeans interventions, and support sizes: the top 1%, 2.5%, and 5% of tokens. The figure shows that across all layers and models, the trend is the same. In the fluent regime any intervention, whether learned or random, pro-

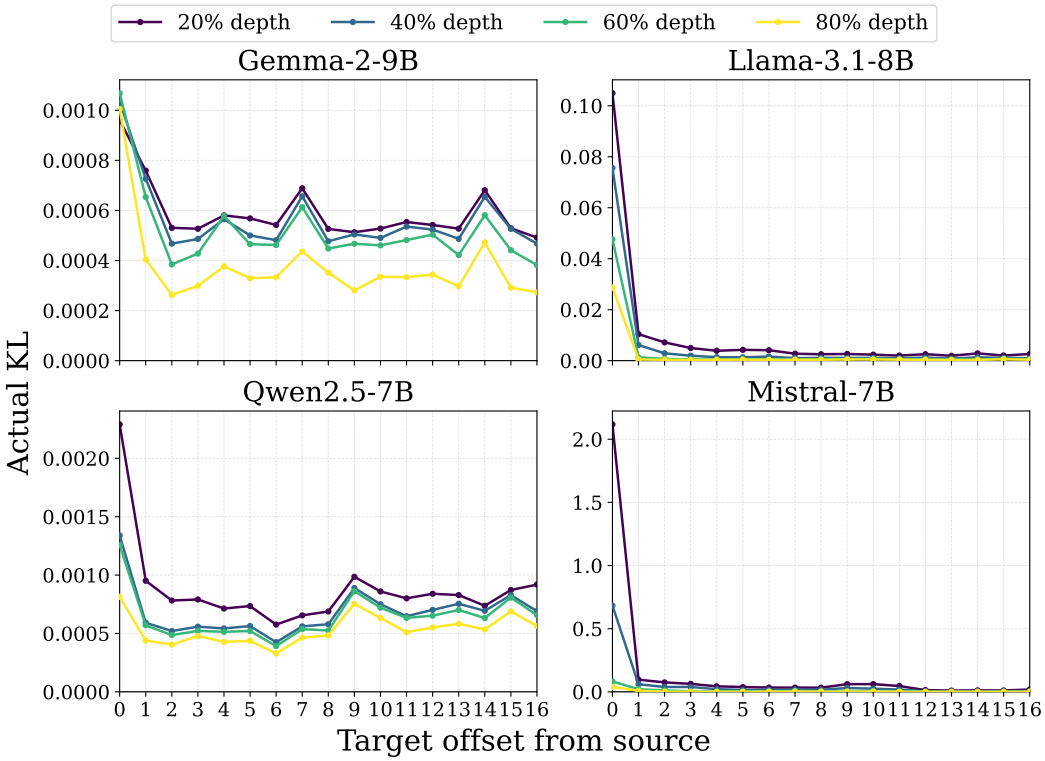


Figure 8: Actual KL induced by an intervention decreases with offsets, consistently across models and layers.

motes the support which captures concept responsive tokens. This captures the context dependence of causality. If a concept is not responsive in the current context, then intervening even with a well localized structure, will not produce the intended effect while remaining fluent.

D Additional Alignment Details

D.1 Locally Optimal Direction Estimation

For each concept and context, we solve for the activation direction that most efficiently promotes the concept. We obtain it by approximately solving $(G_i + 10^{-4}I)\tilde{\mathbf{v}}_i = g_i$. We add the small ridge term to make the solution numerically stable. Because G_i is too large to construct (scales with vocabulary size), we use automatic differentiation and conjugate gradient for at most 32 iterations. To check the approximation, we recompute the true solution error every four iterations, retain the iterate with the smallest error, and verify that a direction transferred to its own context gives the expected normalized score of one (within 0.002). The direction is then scaled to have unit capacity. We use the same eight held-out neutral texts for every concept, sampling up to 16 token positions from each.

E Additional Analysis of the Three Causal Factors

E.1 Full Factor-Bucket Statistics

Figure 10 shows the distribution of the eight factor combinations presented in §4. We find that cases in which exactly one factor is low are common, comprising roughly $26\%-31\%$ of retained samples across models. As shown in §4, a low value of any single factor can substantially reduce causal effectiveness, making these cases particularly important. We additionally find that there is an uncommon regime when responsiveness is low and capacity is high.

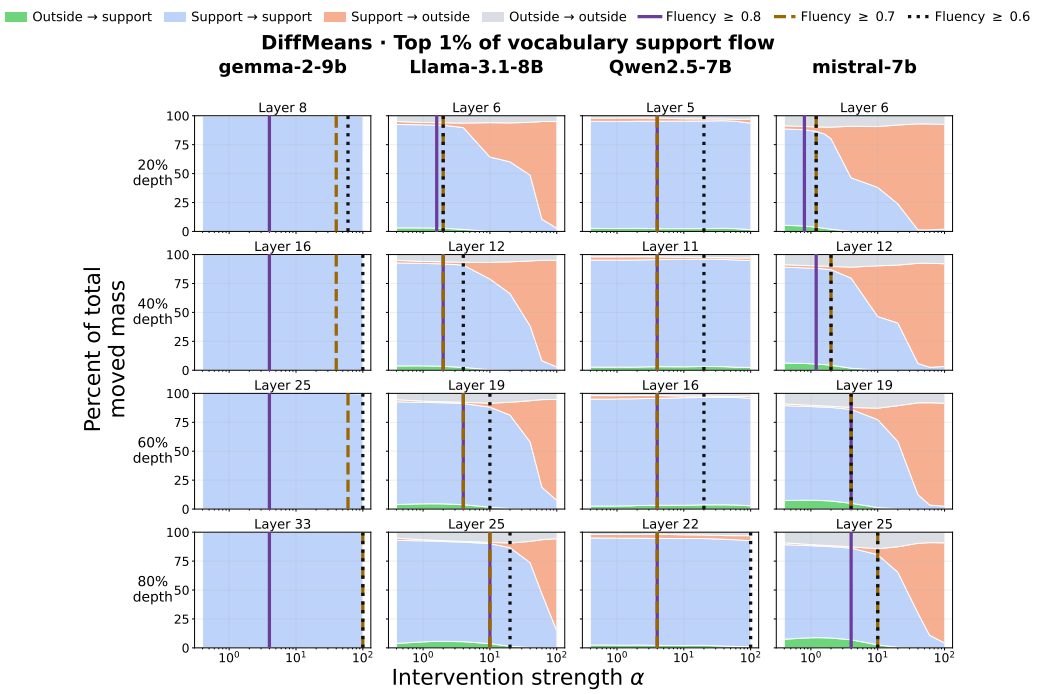


Figure 9: **Interventions primarily promote outputs already supported by the context.** Support-flow results across models, layers, intervention directions, and support sizes. In the fluent regime, most intervention-induced probability increase remains within the baseline support, consistently across settings.

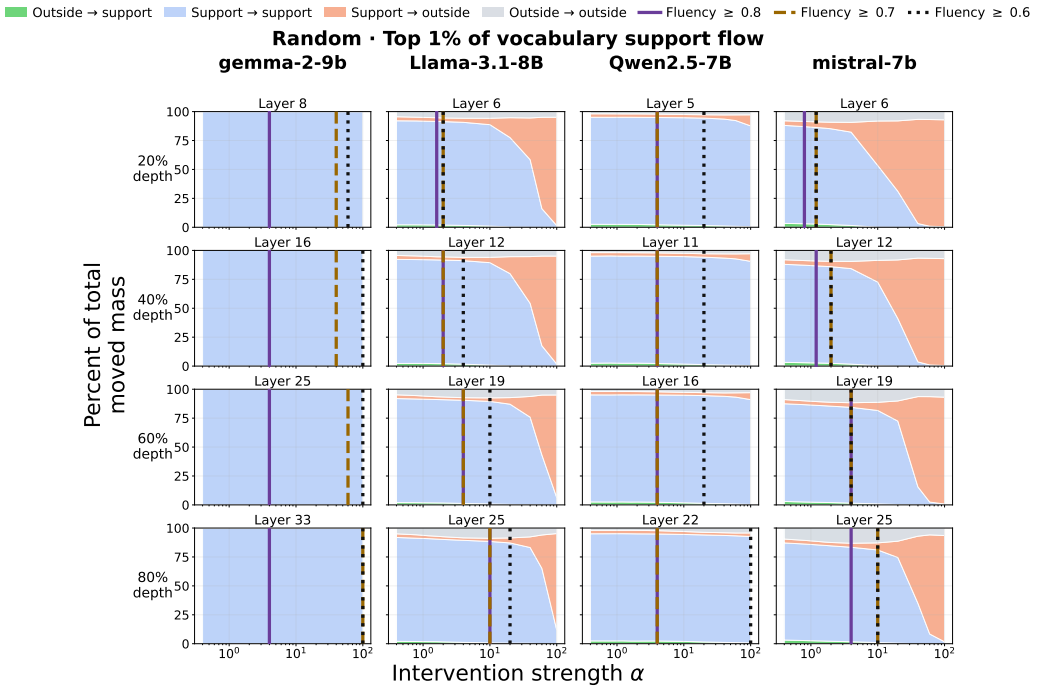


Figure 9: **Interventions primarily promote outputs already supported by the context.** (continued)

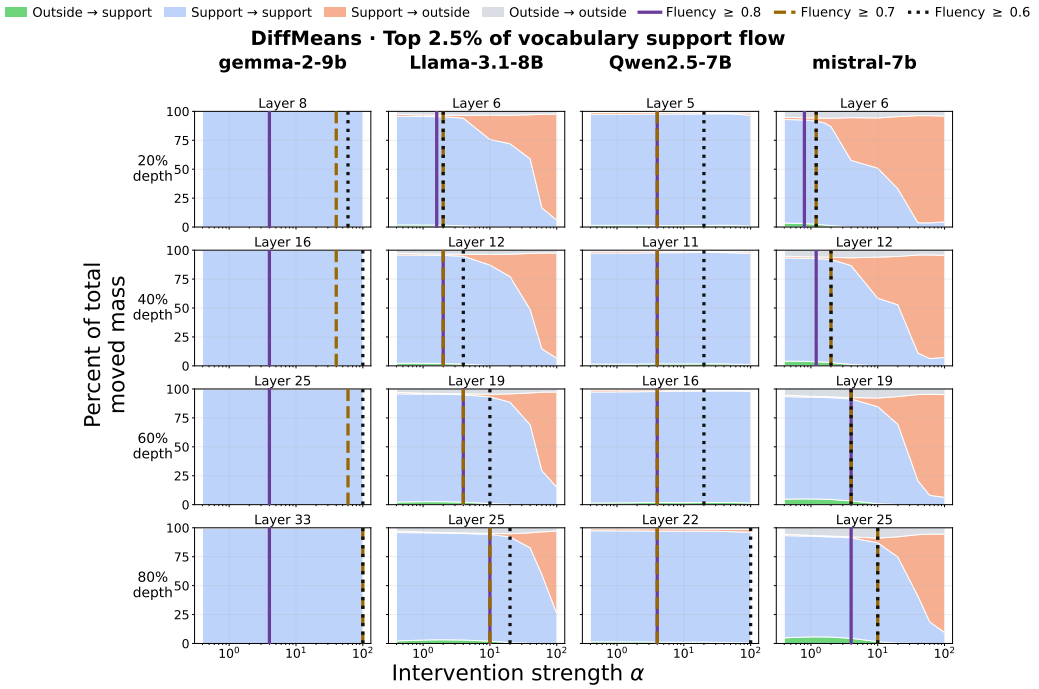


Figure 9: **Interventions primarily promote outputs already supported by the context.** (continued)

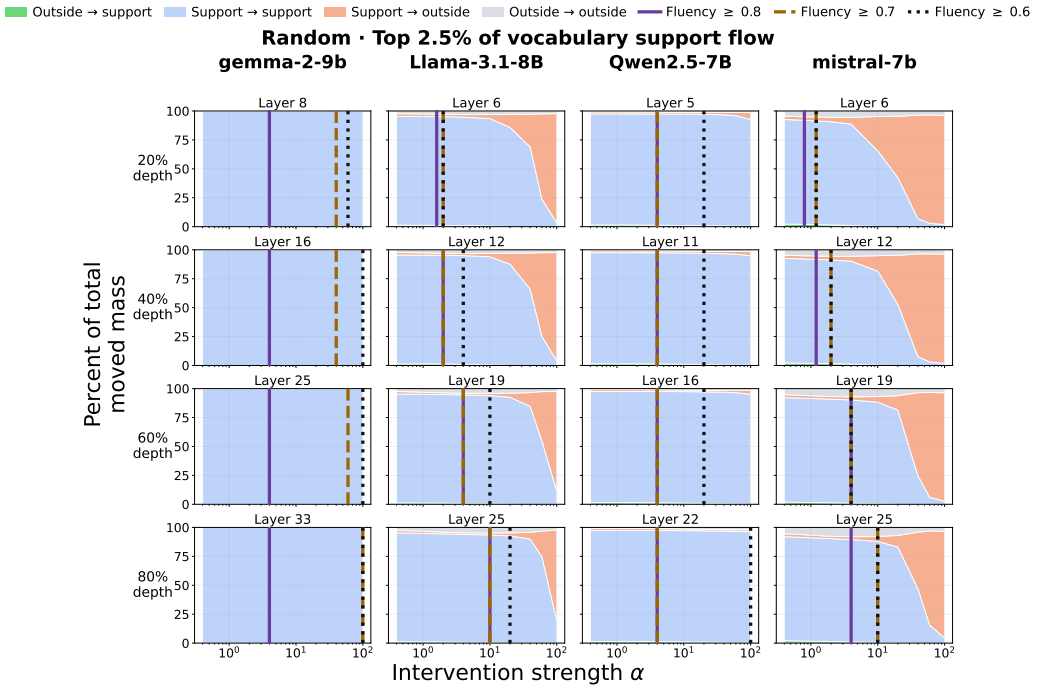


Figure 9: **Interventions primarily promote outputs already supported by the context.** (continued)

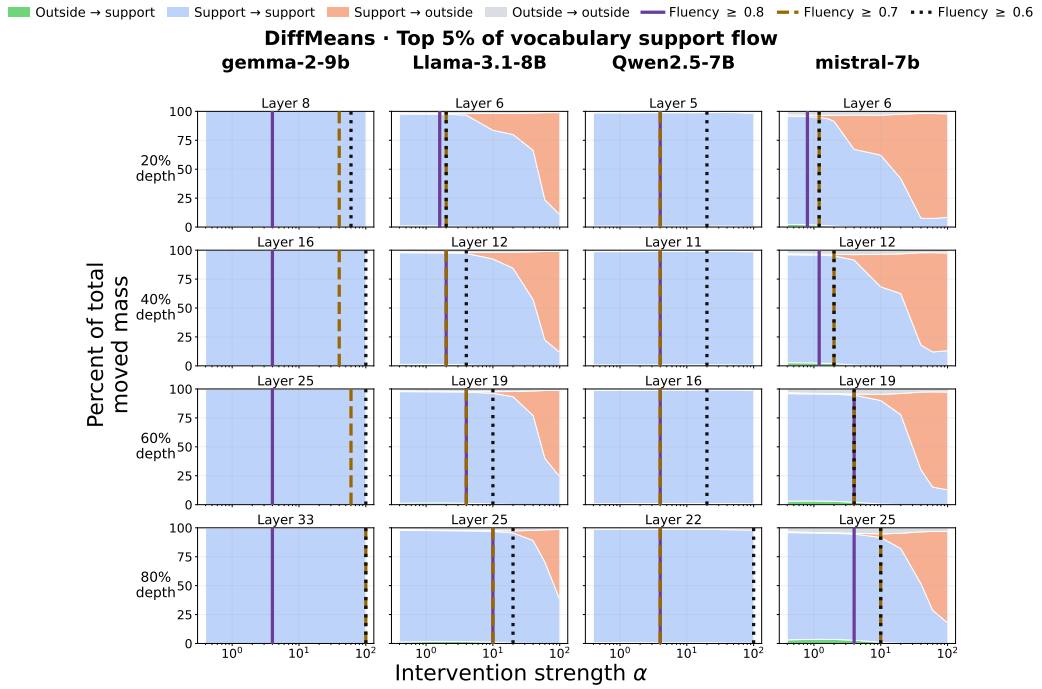


Figure 9: **Interventions primarily promote outputs already supported by the context.** (continued)

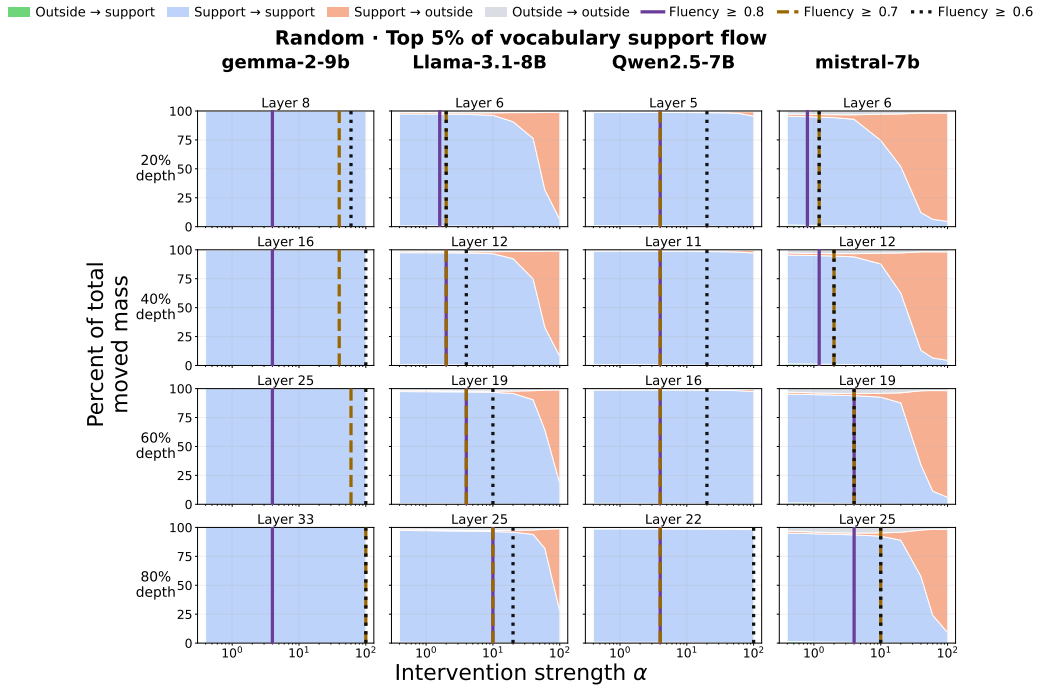


Figure 9: **Interventions primarily promote outputs already supported by the context.** (continued)

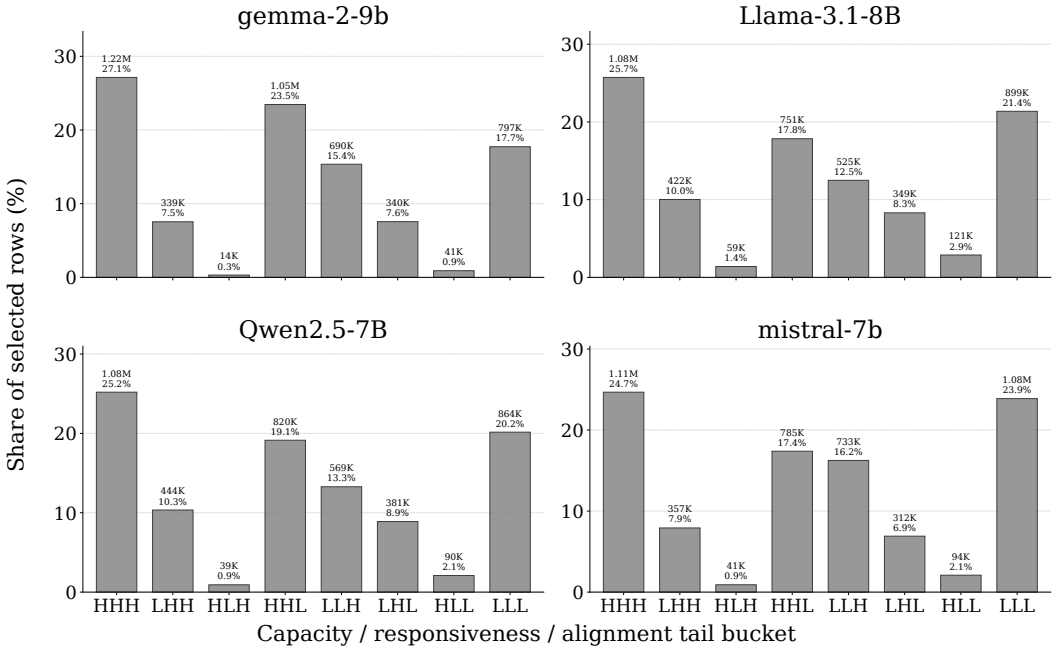


Figure 10: **Distribution of capacity–responsiveness–alignment factor combinations.** The fraction of retained samples in each high/low combination for each model.

E.2 All Model Factor-Bucket Plots

Figure 11 extends the factor-bucket analysis across offsets 0 through -7. The same qualitative pattern persists across offsets: concept promotion is strongest when all three factors are high, while low capacity or responsiveness sharply reduces the effect and low alignment can reverse it, making a localized structure suppress the concept it was intended to promote.

E.3 Steering on Effective Positions

Our decomposition shows that the causal effectiveness of a fixed direction varies across positions according to its responsiveness, capacity, and alignment. We therefore ask whether these factors can identify *where* steering should be applied. At decoding step t, for each preceding position i, we score the effect of intervening along concept direction $\mathbf{v}_c$ at position i on the next-token distribution $\mathbf{p}_t$ as

$$R_i^t = \mathbf{r}_c^\top F_{\mathbf{p}_t} J_i^t \mathbf{v}_c, \qquad J_i^t = \frac{\partial \mathbf{z}_t}{\partial \mathbf{h}_i}. \tag{17}$$

Thus, R_i^t is the signed first-order effect of steering position i on concept expression at the current next-token prediction, and corresponds to the product of the three factors in Eq. 6.

Experiment We use the same steering and evaluation setup as above, but intervene on only a subset of positions. At each decoding step t, we compute R_i^t for all n active positions and select exactly $k = \lfloor n/4 \rfloor$ positions. We compare steering performance at the positions with the highest R_i^t, at those with the lowest R_i^t, and at an equal number selected uniformly at random. We add $\alpha \mathbf{v}_c$ only at the selected positions, sample the next token, and recompute the selection at the following decoding step.

Results Table 1 summarizes the results. Steering high-R_i^t positions improves the final score over the random baseline in 19 of 20 settings, by 44.2% on average, whereas steering low-R_i^t positions reduces performance by 15.5% and underperforms random in 80% of settings. Thus, high values identify positions where steering is particularly effective, while low values identify positions where

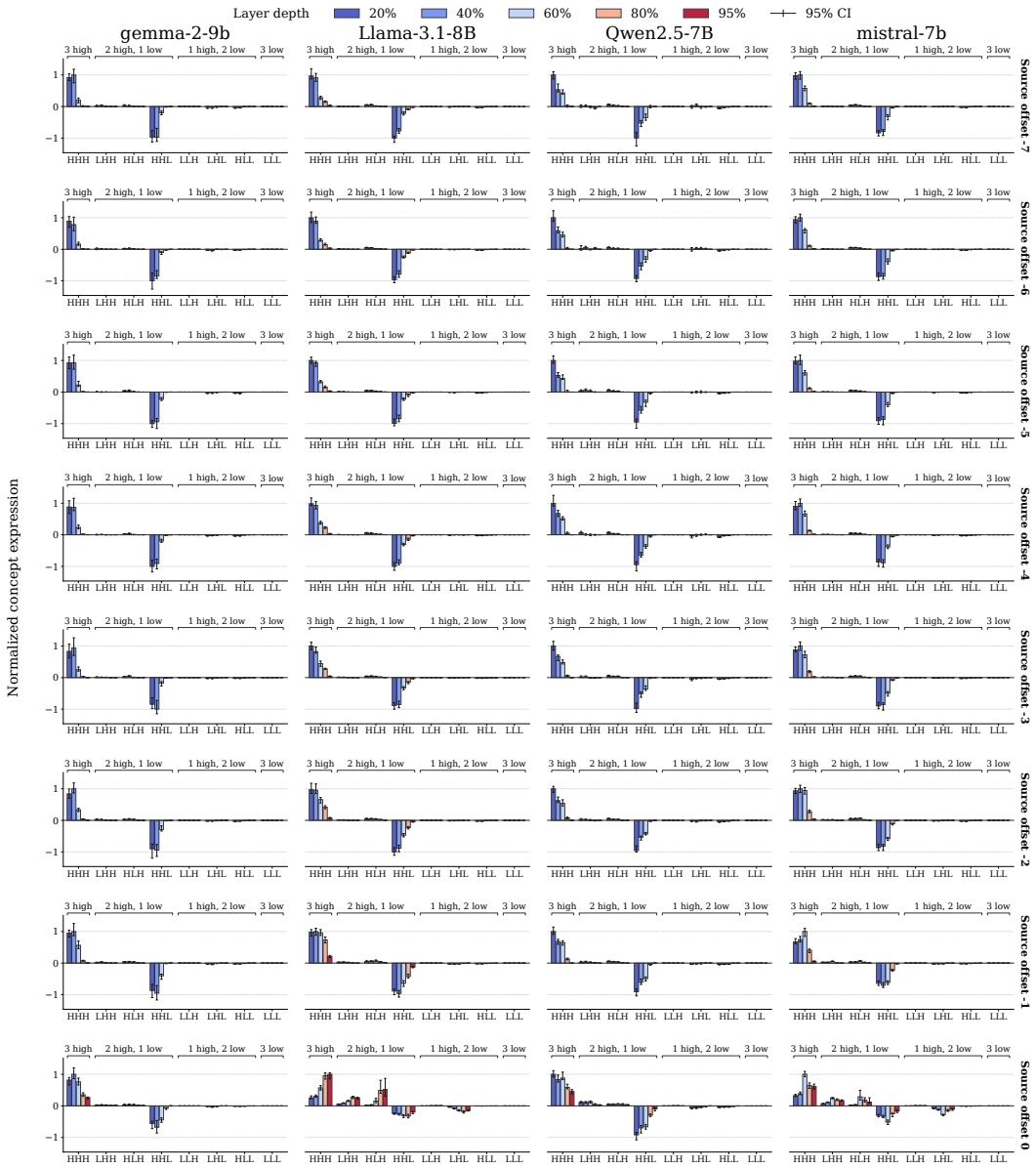


Figure 11: **Factor-bucket results across source offsets.** Normalized concept promotion for each responsiveness–capacity–alignment combination across models, layers, and source offsets -7 through 0. The same qualitative pattern persists across offsets: promotion is strongest when all three factors are high, while a low value of any factor substantially limits the effect.

Table 1: **Steering at positions with high predicted causal effect improves performance.** Concept, fluency, and final steering scores when intervening at randomly selected positions, positions with the highest R_i^t, or positions with the lowest R_i^t. Values report mean $\pm$ standard deviation across concepts.

	Qwen2.5-7B			Gemma-2-9B			Llama-3.1-8B			Mistral-7B		
Metric	Rand.	High R	Low R	Rand.	High R	Low R	Rand.	High R	Low R	Rand.	High R	Low R
Final	$0.22 \pm .14$	$\mathbf{0.29 \pm .15}$	$0.18 \pm .14$	$0.15 \pm .15$	$\mathbf{0.27 \pm .16}$	$0.15 \pm .17$	$0.30 \pm .14$	$\mathbf{0.36 \pm .15}$	$0.23 \pm .14$	$0.25 \pm .13$	$\mathbf{0.40 \pm .14}$	$0.22 \pm .15$
Concept	$0.26 \pm .14$	$\mathbf{0.33 \pm .15}$	$0.25 \pm .13$	$0.13 \pm .15$	$\mathbf{0.26 \pm .15}$	$0.14 \pm .16$	$0.37 \pm .14$	$\mathbf{0.47 \pm .14}$	$0.31 \pm .14$	$0.36 \pm .12$	$\mathbf{0.49 \pm .14}$	$0.33 \pm .14$
Fluency	$0.73 \pm .09$	$\mathbf{0.77 \pm .08}$	$0.69 \pm .12$	$\mathbf{0.91 \pm .02}$	$0.88 \pm .04$	$0.90 \pm .03$	$\mathbf{0.68 \pm .09}$	$0.67 \pm .13$	$0.61 \pm .14$	$0.62 \pm .16$	$\mathbf{0.72 \pm .10}$	$0.59 \pm .17$

causal constraints actively limit the intervention. These results show that the factors remain predictive over longer generations and provide a useful model of where steering will be effective.

F Additional Steering Details and Results

F.1 Causal Probe Training Details

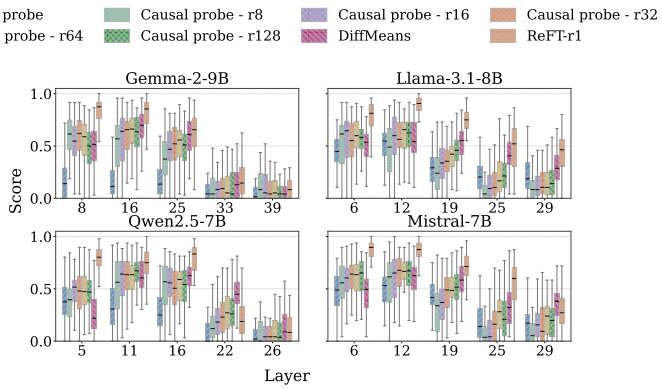
For each concept and layer, we average the capacity matrices (Eq. 8) over all completion positions in the positive and neutral training examples, using the current output and up to seven subsequent outputs. Each available source–target pair receives equal weight. Because the matrices are too large to construct explicitly, we apply them using automatic differentiation and estimate their leading eigenvectors with randomized subspace iteration. We check this approximation using eigenvector residuals, orthogonality, and comparisons with full model computations. The first $r \in \{8, 16, 32, 64, 128\}$ eigenvectors define the rank-r capacity subspace.

Within each subspace, we train a bias-free logistic probe, projecting its gradient and weight back into the subspace after every update. Training uses AdamW for 30 epochs, with learning rate 10^{-3}, batch size 1024, and the final direction is normalized to unit length.

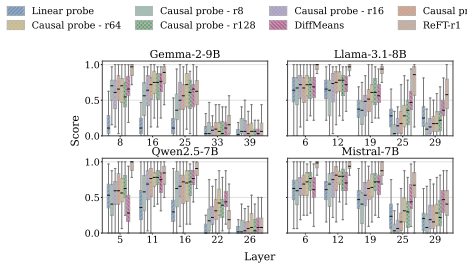
F.2 Steering Evaluation and Judge Prompts

For each concept and intervention method, we generated 50 tokens from the same 16 fixed neutral prefixes using temperature 0.8 and nucleus sampling with $p = 0.95$. The intervention was applied at every decoding step. We evaluated $\alpha \in \{0.4, 0.8, 1.2, 1.6, 2, 3, 4, 6, 8, 10, 20, 40, 60, 100\}$. To select α without evaluating and selecting it on the same prefixes, we divided the 16 prefixes deterministically into five folds of sizes $4, 3, 3, 3, 3$. For each model, layer, concept, and method, one fold was held out, and the strength with the highest mean harmonic score on the other four folds was selected. The selected strength was then evaluated only on the held-out fold. Repeating this procedure for all five folds produced one out-of-fold score for each of the 16 prefixes, which were then averaged for each concept.

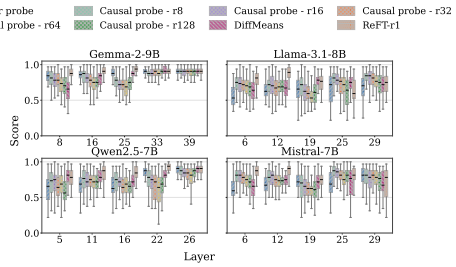
The concept-expression and fluency prompts are provided. We applied the judging separately for each completed sentence, including both the prefix and generated continuation. Judging used `gemini-2.5-flash-lite` with thinking disabled. Each judge returned a score in $\{0, 1, 2\}$, which we divided by two. The final score was the harmonic mean of normalized concept expression and fluency, set to zero if either score was zero.



(a) Overall steering score.



(b) Concept score.



(c) Fluency score.

Figure 12: **Causal probes improve steering primarily through improved concept expression.** We report the overall steering score and its concept and fluency components across models and layers.

Concept Expression Judge Prompt

```
Act as an impartial scientific evaluator.
Evaluate whether the specified concept is expressed in the completed
sentence.
Ignore grammar when deciding concept expression.
Score 0 if absent, 1 if present weakly or awkwardly, and 2 if clearly
and naturally expressed.
Return only JSON with integer field "score" and short string field
"rationale".
Concept:
{concept}
Completed sentence:
{sentence}
```

Fluency Judge Prompt

```
Act as an impartial scientific evaluator.
Evaluate only the fluency of the completed sentence, ignoring concept
relevance.
Score 0 if highly unnatural or incomprehensible, 1 if understandable
with noticeable errors, and 2 if fluent.
Return only JSON with integer field "score" and short string field
"rationale".
Completed sentence:
{sentence}
```

F.3 Concept and Fluency Scores

Figure 12 reports the concept and fluency scores for all tested ranks of the causal probe. Causal probes substantially improve concept scores over unconstrained probes. Notably, models with higher-rank high-capacity subspaces continue to benefit as the causal-probe rank increases. This further supports our capacity findings: causal influence is concentrated in a low-dimensional subspace, and constraining the latent structure to this subspace improves its causal effectiveness. We observe no similarly strong trend in fluency. Causal probes are slightly more fluent on average, but most of their overall gain is explained by improved concept scores.