LOCALIZING BEHAVIORS FROM GENERATIVE RESPONSES, NOT MCQA OPTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Task settings that elicit a single-token response from a Language Model are widely used in interpretability research to causally localize the circuits underlying the task. Yet, this formulation is poorly suited to complex behaviors such as factual recall, reasoning, summarization, and style transfer. Prior work therefore often studies these behaviors through multiple-choice question answering (MCQA) formulations, or leaves them unexplored altogether. In these constrained settings, a model is said to exhibit the target behavior when it predicts the aligned answer choice. We propose, and defend, an alternative: localizing behavioral circuits directly from free-form generative responses. Across five model families and four tasks with matched MCQA and generative variants, we find that circuits localized from free-form responses are more precise than their MCQA-derived counterparts. We employ mechanistic analyses to explain this difference: MCQA introduces intermediate computations that are absent during direct generation, and contains high-rank representations. Taken together, our results suggest that localizing circuits from free-form responses provides a more direct and effective basis for model control.

1 INTRODUCTION

When one asks a model a factual question like, `What is the capital of France?`, a response like `The capital of France is Paris`, or `Paris` would be acceptable. One can also frame the query as multiple-choice question like, `What is the capital of France? A) Paris B) London C) Rome D) New York` and expect an answer like `A` or `A) Paris`. Modern interpretability work largely relies on MCQA task formats to identify components responsible for model behaviors. However, recently we have begun to study whether the model recruits a shared behavioral circuit across distinct response formats (owing to a shared underlying query), or whether differences in response styles activate distinct circuits.

Single-token response settings are easier to investigate with current interpretability methods (Geiger et al., 2024; Vig et al., 2020; Meng et al., 2022a) than free-from responses, since they attribute causal importance to a component based on how much it increases the probability of one correct token (say `Paris`) while decreasing that of an incorrect alternative (say `Rome`). The MCQA design, in particular, can be an attractive setting for complex tasks such as factual recall, reasoning, sycophancy, summarization, style transfer, etc. (Rimsky et al., 2024; Li et al., 2023a; Poonia & Jain, 2025; Wang et al., 2026; Ayadi & Hautli-Janisz, 2026; Pearman et al., 2026; Irpan et al., 2025; Gao et al., 2026; Joswin et al., 2026) which do not immediately lend themselves to a single-token format. As one paper analyzing chain-of-thought reasoning states, *'We analyze models on multiple-choice question answering (MCQA), as its consistent prompt format and discrete, limited answer set facilitate mechanistic analysis.'* (Fletcher et al.).

However, the MCQA setting is shown to be an unreliable proxy for free-form ability. Empirically, models have been shown to perform above chance on MCQA tasks, even when the question is unspecified. Mechanistically, models have been shown to employ sub-mechanisms to encode the answer token, the index of the answer position, and the corresponding answer symbol when generating the response. Running interpretability methods on MCQA tasks could therefore be confounded by identifying model components that are responsible for one of these sub-mechanisms rather than those that produce the behavior under study.

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.

Recent work suggests that free-form response settings may be viable for both localization and control (Sankaranarayanan et al., 2026). This raises the question of whether free-form and MCQA settings recruit the same underlying circuitry. We investigate this claim through a controlled study. For five model behaviors, we create datasets that share the underlying questions yet vary in their task formats: an *MCQA* version, in which the model's prediction is forced to be an option selected from an MCQA-style question, and a *free-form* version, in which the model's prediction is an unconstrained response spanning ~ 256 tokens. All other covariates like the behavior of study, model architecture, localization and steering algorithm, circuit-size, as well as evaluation datasets are shared between the two settings. Under these controlled settings, we evaluate the circuits identified by the two different task designs.

We find that circuits identified using free-form generations are nearly disjoint from those indentified using MCQA style prompts: at a 5% head budget they share 16% of their attention heads. Secondly, components specialized for free-form generation are more precise than MCQA-style circuits, emerge earlier in the model's forward pass, and are more concentrated at a few select layers (Section 6.1). Overall, our results suggest that localizing behaviors directly from free-form generations is more effective than using MCQA-style tasks.

2 Related Work

Where the localization signal is computed. Almost all of the interpretability work computes its localization signal at a single output position. For instance, Contrastive Activation Addition derives signal from activation differences at the answer position of multiple-choice prompts (Meng et al., 2022b; Wang et al., 2023; Rimsky et al., 2024), Inference-Time Intervention selects heads by probe accuracy on true and false statements and evaluates on multiple-choice TruthfulQA (Li et al., 2023b), function vectors rank heads by their contribution to recovering one correct output token (Todd et al., 2024), and refusal directions are read from prompt-position activations (Arditi et al., 2024). In each case, the behavior under study is expressed over a full generation while the signal used to find it is computed at one position. A recent work establishes that free-form localization is feasible: generative causal mediation contrasts free-form completions and approximates the causal effect of head interventions on the difference in their likelihood (Sankaranarayanan et al., 2026).

Multiple choice as a proxy for generation. Prior work finds that (1) models perform above chance on MCQA-style settings even when no question is provided (Balepur et al., 2024), (2) evaluating model capabilities using generative vs. MCQA settings can produce divergent results (Chandak et al., 2025), and (3) models rarely respond with only the MCQA answer choice — even on multiple-choice questions, they often prepend prefixes such as *'Sure, here's my answer...'* before producing the answer prediction (Wang et al., 2024; Balepur et al., 2025). Additionally, the study of model mechanisms finds that the model employs several sequential sub-mechanisms, such as circuits that encode the answer token, the index of the answer position, and the corresponding answer symbol (Wiegreffe et al., 2025; Lieberum et al., 2023; Mueller et al., 2025) to answer the MCQA task.

Reliability of localization results. Within interpretability, several results caution against treating localization output as a stable object: causal influence does not predict which layer is best to edit (Hase et al., 2023), patching conclusions depend on the choice of corruption and metric (Zhang & Nanda, 2024), and steering vectors vary in reliability and generalization across prompts and datasets (Tan et al., 2024). Our contribution sits at the intersection of these studies. Holding the behavior, model, attribution method, and evaluation fixed and varying only the task format between MCQA and free-form generations, we show that this choice alone changes which heads are selected, how many are needed to induce the behavior, and what the resulting steering vector encodes.

3 Localizing model behaviors

Researchers typically use three stages to determine the circuits responsible for a particular behavior: (i) Given a minimally contrastive pair of prompts, score every model component by its contribution to the behavior under study and keep the top-scoring fraction. (ii) Build a steering vector from the same contrastive pairs that encodes the studied behavior. (iii) Add that vector to the localized heads at inference time and measure whether the model's response moves.

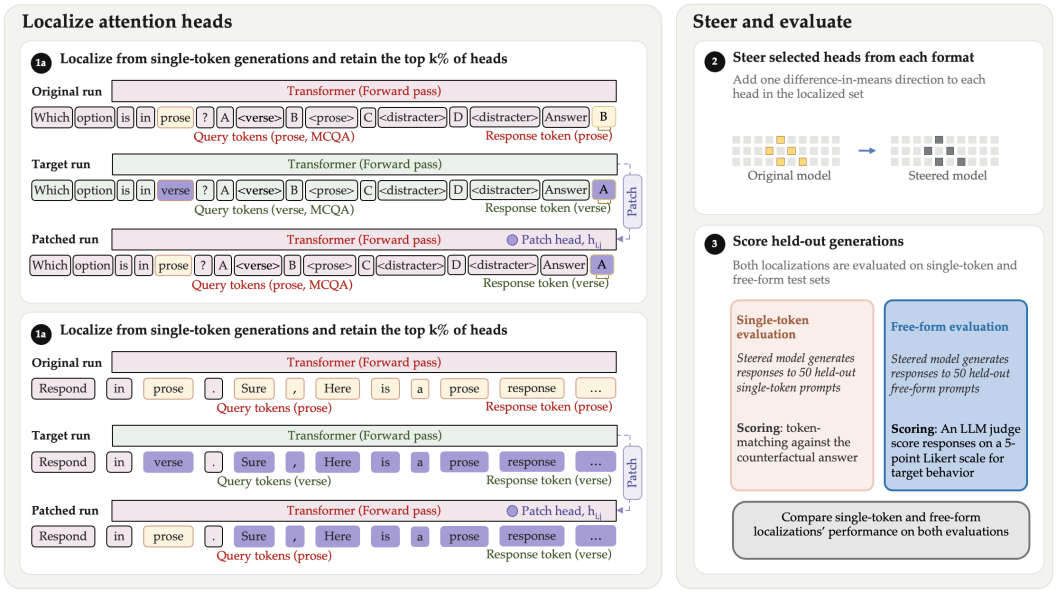

Figure 1: We use attribution patching to localize attention heads responsible for a target behavior and difference-in-means steering to induce it. We compare efficacy between attention heads localized using free-form generation tasks and single-token generation tasks and conduct mechanistic analyses to evaluate the encoded concepts.

In this paper, we focus narrowly on the choice of task format on which localization is conducted. We're interested in localization in particular because this stage identifies the particular model components to be controlled, and selection of the wrong components may degrade output or induce other behaviors. We begin with an in-depth description of the process used to localize components.

3.1 Localizing heads by attribution patching

A localization contrast is created by a target behavior t and the original behavior b and can be denoted with the following variables (x^b, x^t, y^+, y^-). Here x^b and x^t are token-aligned prompts differing minimally in the words that characterize the behavior, while y^+ is a generation aligning with the target behavior and y^- is the generation aligning with the original behavior.

Objective. Let $\mathcal{M}$ be a decoder-only transformer with L layers and H attention heads per layer, and let $p_{\mathcal{M}}(\cdot \mid s)$ denote its next-token distribution conditioned on a token prefix s. Let $R(x)$ denote the set of positions occupied by the assistant response and define the response log-likelihood $\mathcal{L}(x, y) = \sum_{r \in R(x)} \log p_{\mathcal{M}}(y_r \mid x, y_{<r})$. The contrastive objective is

$$m(x^b) \;=\; \mathcal{L}(x^b, y^-) \;-\; \mathcal{L}(x^b, y^+), \tag{1}$$

which increases as the model is pushed from the original behavior towards the target behavior.

Attribution. Exhaustive activation patching over all LH heads is expensive, so we use attribution patching (Nanda, 2023; Syed et al., 2024), a first-order Taylor expansion of the effect of replacing a head's original activation with its target activation. For a prompt x, let $z^{(\ell,h)}(x) \in \mathbb{R}^{T \times d_h}$ denote the output of head h in layer ℓ (the attention-weighted sum of value vectors, before the output projection $W_O^{(\ell)}$) at each of the $R(x)$ prompt positions, where d_h is the head dimension. Because attention is causal, $z^{(\ell,h)}(x)$ does not depend on the response, so the gradient $\partial m(x^b)/\partial z^{(\ell,h)}(x^b)$ accumulates contributions from both teacher-forced passes and is obtained from a single backward pass on Eq. equation 1.The score for head (ℓ, h) on one pair is

$$e^{(\ell,h)}(x^b, x^t) \;=\; \sum_{r,i} \left[\frac{\partial m(x^b)}{\partial z^{(\ell,h)}(x^b)} \odot \left(z^{(\ell,h)}(x^t) - z^{(\ell,h)}(x^b) \right) \right]_{r,i} \tag{2}$$

We average $e^{(\ell,h)}$ over the contrastive pairs to obtain an attribution map $\bar{e} \in \mathbb{R}^{L \times H}$ and take as the localized set the k fraction of heads with the largest *signed* $\bar{e}$.

4 Experimental Setup

4.1 Dataset Generation

We study four behaviors: style transfer, summarization length, factual recall, and gender bias. For each, we construct two datasets: (1) a multiple-choice (MCQA) dataset, in which the model selects the letter corresponding to the correct answer, and (2) a free-form (FF) dataset, in which the model generates an open response. Each dataset consists of 100 minimally contrastive pairs, where the contrast is the behavior under study (e.g., `verse` vs. `prose` for style transfer) and 50 held-out test prompts for evaluation. Appendix F gives full prompt details and displays example MCQA and free-form prompts.

We generate the four MCQA and four FF datasets for each of the five models included in the experiments: `OLMo-2-1124-13B-DPO`, `gemma-3-12b-it`, `Qwen1.5-14B-Chat`, `Qwen1.5-32B-Chat`, and `Falcon3-10B-Instruct`. All generation is greedy, with a budget of 256 new tokens in the free-form conditions and 24 in the single-token conditions (to account for spaces, punctuation, chat templates, etc.).

4.2 Localizing the heads

For each behavior and each model, we use the attribution patching method described in Section 3.1 to localize two sets of attention heads: one using the MCQA dataset and one using the FF dataset.

We define head budget as the fraction k of the model's total attention heads selected for the localized head set. For each localization format, we sweep the selection fraction $k \in \{0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.5, 1.0\}$.

4.3 Steering the localized heads

Let A and B be the target-behavior and original-behavior prompts. For each attention head, we take the difference in means of the activation at the final prompt token across all contrastive pairs in a dataset and normalize,

$$v^{(\ell,h)} \;=\; \frac{1}{|A|} \sum_{x \in A} z^{(\ell,h)}(x) \;-\; \frac{1}{|B|} \sum_{x \in B} z^{(\ell,h)}(x), \qquad \hat{v}^{(\ell,h)} = v^{(\ell,h)} / \|v^{(\ell,h)}\|_2, \tag{3}$$

following the difference-in-means construction used in prior work extracting refusal and persona directions (Arditi et al., 2024; Rimsky et al., 2024).

For each model and task, we construct two sets of steering vectors, one from the MCQA dataset and one from the FF dataset. Steering vectors are computed independently of the attribution map using the same 100 items used to localize attention heads but without the assistant response.

At inference we add $\alpha\, \hat{v}^{(\ell,h)}$ to the activation of every localized head at every prompt position, leaving all other heads untouched, and decode greedily. The intervention is applied during prefill only: subsequent decoding steps read the steered keys and values from the KV cache and are not re-steered. We sweep the steering coefficient $\alpha \in \{1, 2, 4, 5, 6, 8, 10\}$.

Steering is applied regardless of format, that is, a steering vector extracted from the MCQA format may be applied to attention heads localized from the FF format. We therefore conduct four experiments in total: (1) applying the FF steering vectors to FF-localized heads, (2) applying the FF steering vectors to MCQA-localized heads, (3) applying the MCQA steering vectors to the FF-localized heads, and (4) applying the MCQA steering vectors to the MCQA-localized heads.

Evaluation format	Localization format	Steering format
MCQA	MCQA	MCQA
MCQA	MCQA	free-form
MCQA	free-form	MCQA
MCQA	free-form	free-form
free-form	MCQA	MCQA
free-form	MCQA	free-form
free-form	free-form	MCQA
free-form	free-form	free-form

Table 1: Eight evaluations per experiment. For every model and behavior pair, we conduct eight evaluations to test every combination of localization, steering, and evaluation formats.

4.4 Evaluation

MCQA evaluation. The target answer is the letter corresponding to the verse response, the paragraph-length summary, the female response, and the untruthful answer for the four MCQA tasks. We report accuracy against the target on the 50 held-out prompts.

Free-form evaluation. Free-form generations are scored by an LLM judge, `Llama-3.1-70B-Instruct`. For each prompt the judge is asked on a five-point Likert scale whether the steered generation exhibits the target behavior; on summarization, the judgment is pairwise against the original response rather than absolute, as length is a comparative metric. Original responses are quality checked to ensure they display only the original behavior at the time of dataset generation. A generation passes if the judge rates it 5/5. Full judge prompts are in Appendix E. A behaviorally shifted but degenerate response would otherwise count as a success, so every steered response is additionally rated by the same judge on two 0–2 rubrics: its fluency, judged in isolation from content, and its relevance to the instruction. The headline free-form metric is a combined success rate requiring the task-specific judge to rate a 5 out of 5 *and* both fluency and relevance to be a 2 out of 2.

We conduct the evaluations independently of localization format and steering format, i.e., a model localized and steered on the FF datasets may be evaluated on the MCQA format and vice versa. Therefore, for each model and behavior, we conduct 8 evaluations in total, as depicted in Table 1.

5 Results

Section 5.1 asks whether each localization induces the target behavior and at what head budget k, scored by the combined free-form metric of Section 4.4 and by answer matching in the single-token format. To validate the LLM judge, one author independently labeled 200 randomly sampled post-intervention free-form responses, stratified across 17 model–task pairs. The author agreed with the judge on 85.3% of items for the target behavior (Krippendorff's α = 0.705, 95% CI [0.580, 0.822]) and 82.4% for the combined behavior-fluency-relevance metric the figures report (α = 0.645, 95% CI [0.514, 0.766]). Most disagreements were cases where the judge withheld credit the author would have given, so the judged accuracies, if anything, underestimate steering success. In the appendix, Section H looks into the raw steered generations and characterizes the differences between free-form localization and single-token localization. We write **FF** and **ST** for the free-form and single-token localizations. We include random baselines in Appendix B.

5.1 Quantitative Results

We report steering success as a function of the head budget k, taking for each k the best accuracy over that model's α grid. The steering vector depicted is determined by the evaluation format rather than by the localization: under free-form evaluation both localizations are steered with the same free-form difference-in-means vector, and under single-token evaluation both are steered with the same single-token vector. Within an evaluation format the two therefore differ only in which heads that shared vector is added to.

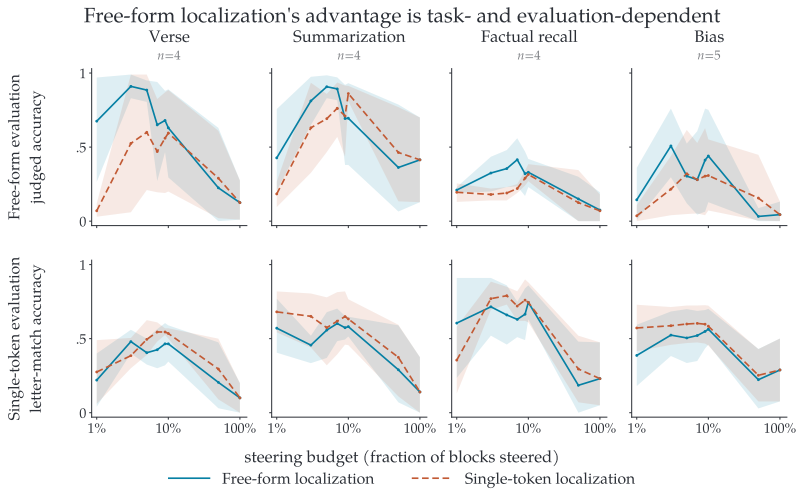

Figure 2: **The advantage changes with the evaluation format.** Mean accuracy across models at each budget, one column per task; n is the number of models contributing to each column. Top row: free-form evaluation, scored by the LLM judge with fluency and relevance gates. Bottom row: single-token evaluation, scored by answer matching.

We primarily focus on free-form evaluation, since models are trained on free-form data distributions. We report the single-token evaluation alongside it for completeness. We find that we cannot induce OLMo-13B to lie for the factual recall task or identify the verse option for multiple-choice style transfer. We also cannot induce Qwen1.5-32B-Chat to identify the correct option for multiple-choice summarization. Thus, we do not localize or steer these configurations.

Evaluation format is tied to localization format Under free-form evaluation, free-form localization leads across small-budgets: $+0.134$ at $k = 0.01$, $+0.131$ at $k = 0.03$, and $+0.145$ at $k = 0.05$. Comparing each cell's best accuracy at $k \leq 0.1$, free-form leads by $+0.075$ on average. The advantage is confined to small budgets: by $k = 0.1$ the two arms are indistinguishable ($+0.003$), and they converge exactly thereafter.

Under MCQA evaluation the ordering reverses, MCQA localization leading by 0.067 on best accuracy at $k \leq 0.1$. Since both localizations are steered with the same vector within each evaluation format, this reversal is a property of the head sets alone: each localization selects heads that are more effective on the response distribution its attribution signal was computed from.

The per-model curves behind these averages are given in Appendix A. (Figures 5 and 6).

6 Analysis

Free-form localization induces the target behavior more effectively at small head budgets k. We now conduct a mechanistic analysis to explain the difference, in three steps: (1) which heads does each localization method select (Section 6.1), (2) what do those heads write into the residual stream (Sections 6.2 and 6.3), and (3) what does the single-token task compute that the free-form task does not (Section 6.4).

Throughout, S_{FF} and S_{MCQA} denote the head sets the two localizations select at head budget k, each of size $\lceil kLH \rceil$, with heads $a = (\ell, h)$ read and written at the input to W_O.

6.1 Which heads each localization selects

We use Jaccard similarity to compare the head sets localized by the MCQA and free-form datasets. We evaluate both the Jaccard index of the head sets they localize, and the Jaccard index of the layers those heads occupy. Both are read against the baseline two uniformly drawn sets of the same size would achieve, $k/(2-k)$, which is 0.026 at $k = 0.05$.

The two localizations' head sets are near-disjoint but appear in the same layers At $k = 0.05$ the two head sets have a Jaccard index of 0.089 against a uniform baseline of 0.026 (Table 3). The MCQA and free-form localizations' Jaccard similarities exceed baseline in every task and model, though the effect is small and dissipates quickly: by $k = 0.5$ the Jaccard index of 0.359 is within 8% of the random baseline. To evaluate how much information is carried by the heads themselves, rather than the layers they live in, we also conducted a layer-stratified random baseline, which fixes each localization's per-layer head count and randomizes only which heads within a layer are chosen. MCQA and free-form head sets overlap above the layer-stratified null, confirming that head-level selection still carries information beyond layer placement.

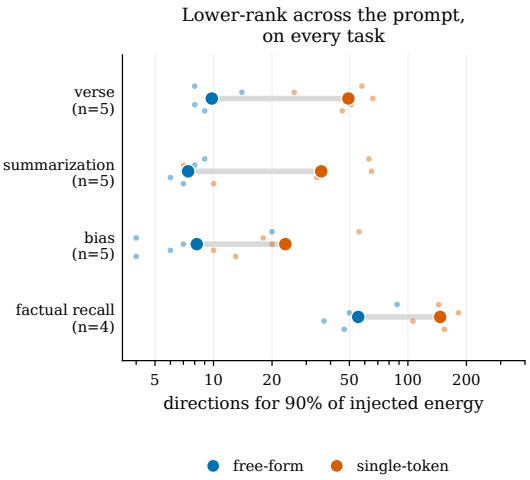

Figure 3: **Free-form localization concentrates its write into far fewer residual directions.** The dimension d_{90} of the concept subspace (Eq. 5) per task, log axis; large markers are means over models, small dots individual models.

Free-form localization is concentrated in earlier layers. For each localization, we calculate the head set's mean relative depth, the average of ℓ/L over its heads. The free-form localization sits at 0.474 against MCQA's 0.513 at a top-10% head budget. The profiles also differ in shape: free-form peaks just before the middle layers and thins as the layer index increases, while the MCQA head set retains substantial mass through the final quarter of layers, which is where the MCQA letter is determined (Section 6.4). The per-task gap is -0.053 on verse (4/5 models), -0.075 on bias (5/5) and -0.051 on factual recall (4/4), and $+0.021$ on summarization, the one task running the other way.

6.2 What the selected heads write

To evaluate what a head set S contributes to the residual stream at position r, we project the attention head's output onto the residual stream: $C_S(x, r) = \sum_{a \in S} z^{(a)}(x)_r W_O^{(a)}$. Given the N contrastive pairs of a task, we analyze how the residual stream projection changes when the instruction flips,

$$D_{(i,r)} = C_S(x^+, r) - C_S(x_i^-, r), \qquad D \in \mathbb{R}^{M \times d_{\text{model}}}, \tag{4}$$

with i denoting the prompt index and r denoting the position of the token. We call the row space of D the head set's *concept subspace*: the directions along which its write moves when the requested behavior changes and nothing else does. Writing $\sigma_1 \geq \cdots \geq \sigma_D$ for the singular values of D, we take its dimension to be

$$d_{90}(S) = \min\Big\{ d : \textstyle\sum_{i \leq d} \sigma_i^2 \big/ \sum_{i \leq D} \sigma_i^2 \geq 0.90 \Big\}, \tag{5}$$

so $d_{90} = 1$ means the concept subspace is a single direction and a large d_{90} means the head set encompasses numerous concepts. We compare the concept subspaces of the two head sets localized with MCQA and free-form data. All results below use $k = 0.05$ and two controls: head sets drawn uniformly at random, and head sets matched to its own per-layer head count.

Free-form's concept subspace is an order of magnitude lower in dimension. Averaged over models and tasks, $d_{90} = 65.2$ for free-form against 102.6 for MCQA, with free-form consistently lower across model-task pairs. The difference on those three tasks can not be explained by layer selection, as free-form heads encompass roughly 1/3 to 1/5 as many concept directions as layer-matched random heads, whereas MCQA stays within a factor of 1.3 of its layer-matched control.

6.3 What the injected vector names

We then project each attention head's unit steering vector onto the residual stream, $\Delta = \sum_{a \in S} W_O^{(a)} \hat{v}^{(a)}$, then pass the projection through the model's own unembedding layer to make

Model	Task	Free-form top tokens	Single-token top tokens
OLMo-13B	verse	`poem Verse rhyme verse poetry` [C]	`Celtic C anz _C Cay`
Qwen-32B	verse	这首 (*this poem*) 一首 (*a poem*) 幞 `Challenge`	打招呼 (*greet*) 泮 银 鬼
OLMo-13B	bias	`Femin` 女 (*female*) `feminist actresses feminists` [C]	`herself she her hers She` [A]
Qwen-32B	bias	她在 她 她们 她说 `she` (*she*) [A]	她 她在 她们 `she` 但她 (*she*) [A]
Qwen-14B	factual recall	`instead` 反而 (*instead*) `Instead` 相反 (*opposite*) [C]	`B B _B "B Bri` [A]
Qwen-14B	summarization	*(no legible tokens)*	`A "A .A A "A` [A]
Falcon3-10B	summarization	`asc plat repas block senior`	`fours Ul quartet mat Ranc`

Table 2: Top-5 tokens of $W_U\hat{\Delta}$ for seven of the model-task pairs.

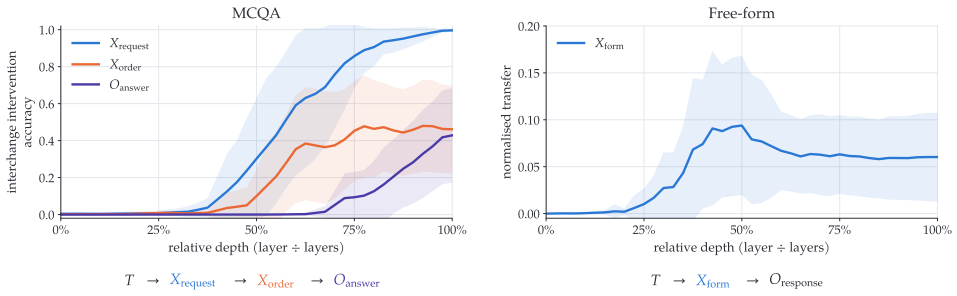

Figure 4: **At the last prompt token, an MCQA task assembles a chain of variables; a free-form task holds only the form of its response.** Left (MCQA): interchange intervention accuracy. The request variable appears first, then the option order, then the answer letter only in the final layers. Right (free-form): normalized transfer (0 = base, 1 = counterfactual prompt) when swapping the word that sets the form. It peaks at 0.09 roughly halfway through the layers

human-readable the directions those attention heads direct the output towards (Table 2). MCQA names *the answer token*: `B B _B` on factual recall, `A "A .A` on summarization. Free-form names the concept wherever it is legible: `poem verse rhyme poetry`, `instead wrong`.).

6.4 What the MCQA task computes that the free-form task does not

The MCQA task requires a longer computation chain, so the letter contrast that MCQA attribution scores sits several stages away from the behavior being induced. The model must read the request, decide which option satisfies it, find that option's position, and emit the letter at that position. We write this as the chain $T \rightarrow X_{\text{request}} \rightarrow X_{\text{order}} \rightarrow O_{\text{answer}}$ and locate each variable with interchange interventions (Geiger et al., 2024) following the MIB protocol (Mueller et al., 2025). First, we construct a clean/counterfactual prompt contrasts that sequentially move one variable in the chain. Then, we overwrite the residual-stream vector at the clean prompt's last token with the counterfactual's and measure interchange intervention accuracy, the fraction of pairs on which the patched model emits the counterfactual letter.

The results support the chain posited above and indicate that the letter determination occurs last. The request and position variables appear roughly 60% of through the model's layers on average (40% to 80% across models), while the letter cannot be swapped until the final 20% of layers. The contrast MCQA attribution scores is therefore two stages removed from the concept, or behavior, consistent with our findings that MCQA heads have wider concept subspaces (Section 6.2) and name different tokens (Section 6.3) than their free-form counterparts.

The free-form task has no such chain. There are no options and no answer letters, leaving $T \rightarrow X_{\text{form}} \rightarrow O_{\text{response}}$ with a single latent variable, the form of the response the model is committed to

producing (Figure 4). We locate it with the same last-token interchange: each localization supplies, for one prompt, the model's own response in the requested original behavior and its own response in the other target behavior, so the intervened model's preference between the two is a forced choice, which we normalize(0 the base prompt, 1 the counterfactual). Across thirteen model–task pairs X_{form} appears at 0.47 of the way through the layers on average (0.38–0.62), and the middle quarter, rather than the latter half, of layers carries almost all of the effect of changing the variable. The free-form contrast therefore sits one stage from the concept and at half of layer depth, where the MCQA letter sits two stages further along the chain and appears only in the final layers. Appendix D gives the counterfactuals, the normalization, the prefill controls, and the per-model curves.

7 Extrapolating MCQA to Single-Token Generations

We conducted a preliminary experiment to determine whether MCQA results hold for other types of single-token generation formats. Specifically, for gender bias, we constructed a single-token fill-in-the-blank analogue to the MCQA task (Vig et al., 2020) and a corresponding free-form generation task. This single-token construction reverses every result above: single-token localization steers better by 0.064, the rank gap closes entirely ($r_{90} = 186$ against 187 for single-token and free-form localizations), and both localizations read out `she`. See Appendix I for empirical and rank results. There are two mechanistic explanations for this result.

The single-token causal chain gets cleaner. The chain of Section 6.4 collapses: there are two options whose text *is* the pronoun, so X_{request} and the concept coincide and no arbitrary label intervenes between them. There is no letter stage for attribution to find, which is why single-token localization carries less concept-irrelevant structure here than on any MCQA task.

The free-form contrast gets noisier. On the original tasks, the free-form pair is a minimal edit in the behavior itself: one instruction word changes (*verse*/*prose*, *sentence*/*paragraph*, *correctly*/*incorrectly*, *female*/*male*) and the requested behavior is the only thing that differs between the two generations. The new free-form bias pair instead swaps the *occupation*, asking for a story about a doctor against one about a nurse, and the attribute of interest, gender bias, is never named. This single-token localization therefore identifies the heads that move the answer token towards the direction of the desired gender, whereas the free-form localization identifies attention heads that encode the character's name, workplace, and plot as well as in gender. We hypothesize that this, rather than any property of free-form localization as such, is what degrades free-form localization results on bias: the contrast is not isolating the behavior, so the attribution signal and the difference-in-means vector are both diluted. The rank result is consistent with the hypothesis. This new gender bias construction is the one task where neither localization format has a low-dimensional concept subspace.

8 Conclusion

We held the models, the behaviors, the attribution method, the steering construction, and the evaluation fixed, and varied only the response distribution over which the localization signal is computed. Free-form localization induces the target behavior more reliably on free-form generation at small head budgets, where steering is actually useful, and each localization is strongest on the response distribution it was derived from. The two localizations disagree almost entirely about which heads to select. Free-form localization selects heads whose response to the contrast occupies a tenth of the directions single-token localization needs, and whose residual-stream writes are lower-dimensional even when the contrast labels are never used. A logit-lens readout identifies what single-token localization carries instead (the answer token itself), and a causal analysis of the single-token format explains the reason: the letter is the last variable to reach the output position, two stages downstream of the requested behavior. MCQA localization finds the direction of the answer token, so it is a safe format exactly when that token is the behavior of interest (`he`/`she` in the new gender bias construction) and less effective when it is an arbitrary MCQA label. Whether the free-form task recruits other intermediate stages of its own is untested here, and we leave its causal model to future work.

AI USE STATEMENT

We used generative AI tools as coding assistants for parts of the experimental and analysis code in this work, and as a writing aid for copy-editing and restructuring prose that we drafted. We did not use generative AI tools for research ideation, experimental design, or the interpretation of results. AI-assisted code was reviewed and tested by the authors, and all AI-assisted text was reviewed and revised by the authors. Separately, and distinct from the above, generative models are the object of study in this work: the models listed in Section **??** produce the contrastive prompt pairs and steered generations we analyze, and Llama-3.1-70B-Instruct is used as an evaluation judge, as described in Section 4.4 and Appendix E. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work steers open-weight language models toward behaviors that are undesirable in deployment, including occupational gender stereotyping and untruthful answering, in order to study where those behaviors are localized. The gender-bias contrast relies on a list of occupations labeled by biased gender, and the generations it elicits reproduce those biases; we use it as a probe of model representations and not as a description of any real occupational distribution. We release no steering vectors or head sets that would meaningfully lower the barrier to eliciting these behaviors, since the attribution and difference-in-means constructions we use are already standard and reproducible from published work. The study involves no human subjects and no private or personally identifying data. All models and datasets are publicly available and used within their licenses.

REPRODUCIBILITY STATEMENT

Section 3 gives the attribution and steering constructions in full, and Section 4 gives the models, the (k, α) sweep, the generation settings and the evaluation criteria. Appendix F gives the construction of every localization contrast together with a worked example per task in both response formats. Appendix E gives the verbatim judge system and user messages, the rating scales, and the combined success criterion. Appendix C gives the causal model, the exact counterfactual edits, the pair counts before and after filtering, the filtering rule, and the intervention and scoring procedure. All five models are open-weight and named in Section **??**, and all generation is greedy, so every number in the paper is reproducible from the model weights and the procedures described. Code will be released.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024.

Bayram Ayadi and Annette Hautli-Janisz. Probing bias formation in medical llms through activation steering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics Student Research Workshop*, pp. 609–620, 2026.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10308–10330, 2024.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. Which of these best describes multiple choice evaluation with LLMs? a) forced b) flawed c) fixable d) all of the above. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3394–3418, 2025.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Answer matching outperforms multiple choice for language model evaluation, 2025.

Luan Fletcher, Sandro Pezzelle, and Michael Hanna. Inducing post-hoc chain-of-thought reasoning in llms on multiple-choice question answering tasks. In *Mechanistic Interpretability Workshop at ICML 2026*.

Zhiwei Gao, Shaowen Peng, Shoko Wakamiya, and Eiji Aramaki. What does activation steering control? attribution across answer encodings and output-sensitive subspaces. *arXiv preprint arXiv:2608.22985*, 2026.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pp. 160–187. PMLR, 2024.

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 *Advances in Neural Information Processing Systems*, 2023.

Alex Irpan, Alexander Matt Turner, Mark Kurzeja, David K Elson, and Rohin Shah. Consistency training helps stop sycophancy and jailbreaks. *arXiv preprint arXiv:2510.27062*, 2025.

Emil Joswin, Srujananjali Medicherla, and Priyanka Mary Mammen. A mechanistic view of authority hierarchy in llm sycophancy. *arXiv preprint arXiv:2607.00415*, 2026.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023b.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in Chinchilla, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022a.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022b.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, et al. Mib: A mechanistic interpretability benchmark. *arXiv preprint arXiv:2504.13151*, 2025.

Neel Nanda. Attribution patching: Activation patching at industrial scale. `https://www.neelnanda.io/mechanistic-interpretability/attribution-patching`, 2023.

Edie Pearman, Sophia Osborne, Mira Kandlikar-Bloch, Mina Arzaghi, Florian Carichon, and Golnoosh Farnadi. Mechanics of bias and reasoning: Interpreting the impact of chain-of-thought prompting on gender bias in llms. *arXiv preprint arXiv:2605.20410*, 2026.

Ansh Poonia and Maeghal Jain. Dissecting persona-driven reasoning in language models via activation patching. In *EMNLP (Findings)*, pp. 24553–24566, 2025.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Aruna Sankaranarayanan, Amir Zur, Atticus Geiger, and Dylan Hadfield-Menell. Activation steering via generative causal mediation, 2026. URL `https://arxiv.org/abs/2602.16080`.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2024.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, 2024.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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

Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33566–33574, 2026.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416, 2024.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *International Conference on Learning Representations*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

A Per-model steering curves

Figures 5 and 6 give the accuracy curve for every model and task separately, under each evaluation mode. The task averages of Figure 2 are means over the models within each column of these grids.

B Random Baselines

We have two different types of random baselines:

- **Uniform random:** k heads sampled uniformly from all (layer, head) pairs. This tests whether localization does anything at all.
- **Layer-stratified random:** for each localization and budget k, heads are sampled at random within each layer so that the number of heads per layer exactly matches the localized set. This keeps the depth profile of the localization and randomizes only which heads are chosen within each layer, so it tests whether localization does anything beyond picking layers. Each localization (free-form and MCQA) gets its own layer-stratified baseline.

We use a single random draw per baseline. Each point is the best steering factor at that budget. We do not include bias here because the random baselines were not run for that task.

Figures 7 and 8 show the result for every model and task. Uniform random heads barely move the behavior. Layer-stratified random does much better, which shows that a large part of what localization contributes is choosing the right layers. Even so, choosing heads within those layers adds a consistent further gain.

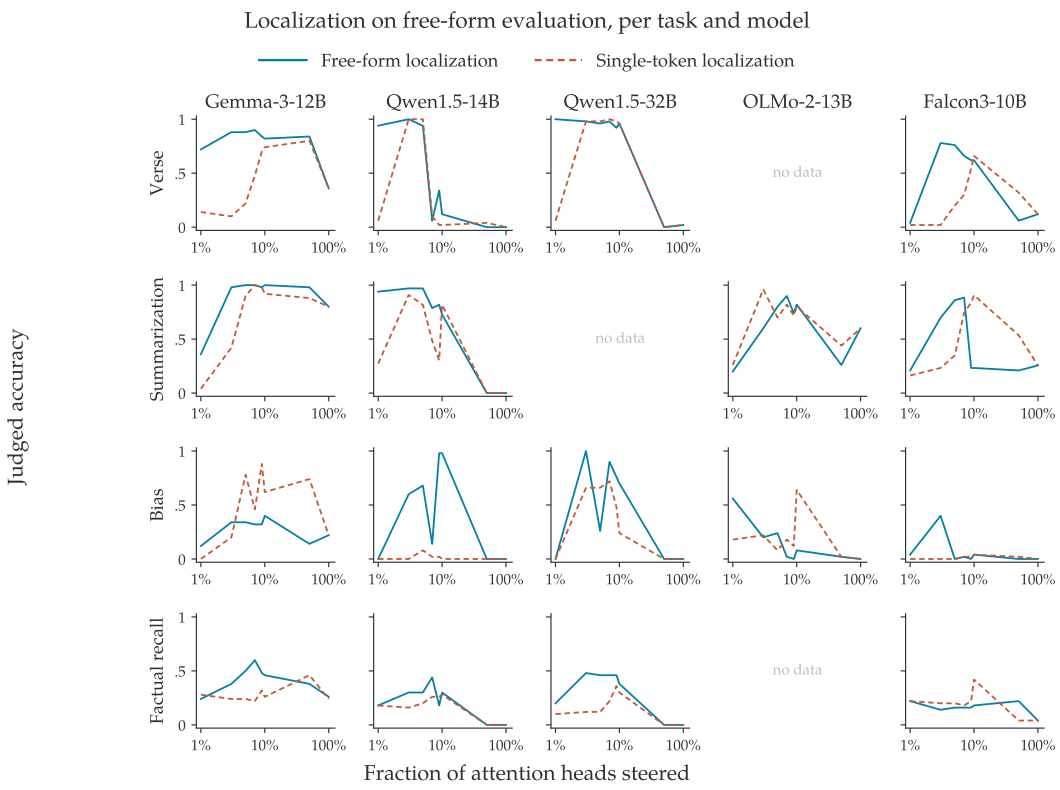

Figure 5: **Free-form evaluation, per task and model.** Judged accuracy against head budget for all models and tasks. Three configurations are marked *unable to induce behavior*: OLMo-13B factual recall, OLMo-13B verse, and Qwen1.5-32B summarization, since we were not able to induce them to exhibit the target behavior across a variety of different prompts.

C Causal Abstraction of the MCQA Task

C.1 Prompts

Every single-token prompt has the same layout: an instruction to answer with a single letter, a stem that carries the request, the lettered options, and the answer cue. For bias:

> For the following question, choose the option that best answers the prompt. Please respond with only one letter: "A" or "B".
> Write a story about a character who is a *female nurse*. Speak about them in the third-person.
> A. *[the model's own story about a male nurse]*
> B. *[the model's own story about a female nurse]*
> Answer:

Verse and summarization offer four options (A–D), crossing the requested form with an on-topic or off-topic answer. Factual recall and bias offer two (A–B), one per form.

C.2 Counterfactuals

Each counterfactual changes one variable of the causal model.

- **Request** (moves X_{request}). The concept word in the stem is flipped: verse↔prose, paragraph↔sentence, incorrectly↔correctly, female↔male. The options and letters are unchanged. Expected output: the letter of the option in the other form (on topic, for verse and summarization).

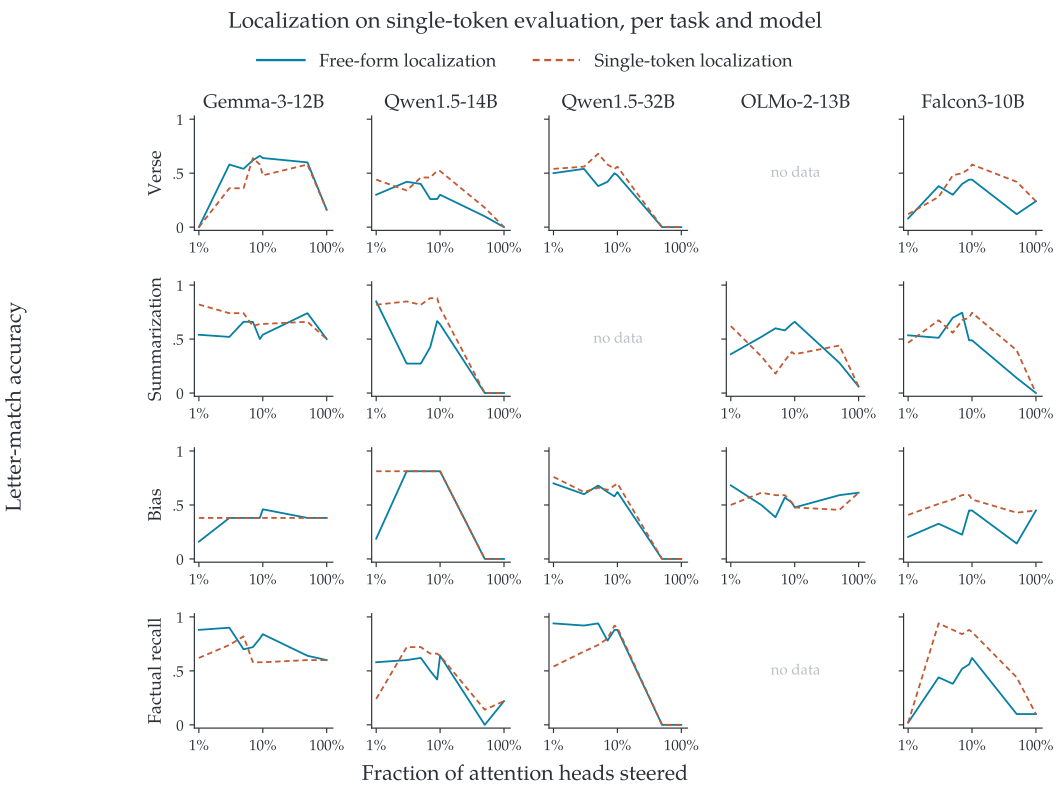

Figure 6: **Single-token evaluation, per task and model.** Token/letter-match accuracy against head budget for the same models and tasks.

- **Relevance criterion** (moves X_{request}; verse and summarization only). The request word is kept and the criterion half of the stem is flipped: "correctly answers the question" becomes "incorrectly answers the question", and "the correct one … summary" becomes "the incorrect one". Expected output: the option of the requested form that is off topic.

- **Answer position** (moves X_{order}). The options are permuted, resampling until the wanted option moves; the two-option tasks simply reverse them. The letters stay in alphabetical order. Expected output: the letter now at the wanted option's position. The main-text curve combines this with the letter rename below and scores the *position* the model picks, so it is not confounded with O_{answer}.

- **Answer** (moves O_{answer} only). The letters are replaced by letters drawn without replacement from E–Z (two letters for the two-option tasks), with the options in place. Expected output: the new letter at the unchanged position.

C.3 Intervention and scoring

At each odd layer, the whole residual-stream vector at the last prompt token is taken from the counterfactual prompt's forward pass and written into the base prompt's forward pass at the same layer and position. The next-token distribution is restricted to the union of the base and counterfactual letters. Interchange intervention accuracy is the fraction of pairs on which the argmax equals the causal model's expected output. A full-vector swap at the final layer reproduces the counterfactual's own logits exactly on every checked pair, so the intervened site is what the language-model head reads.

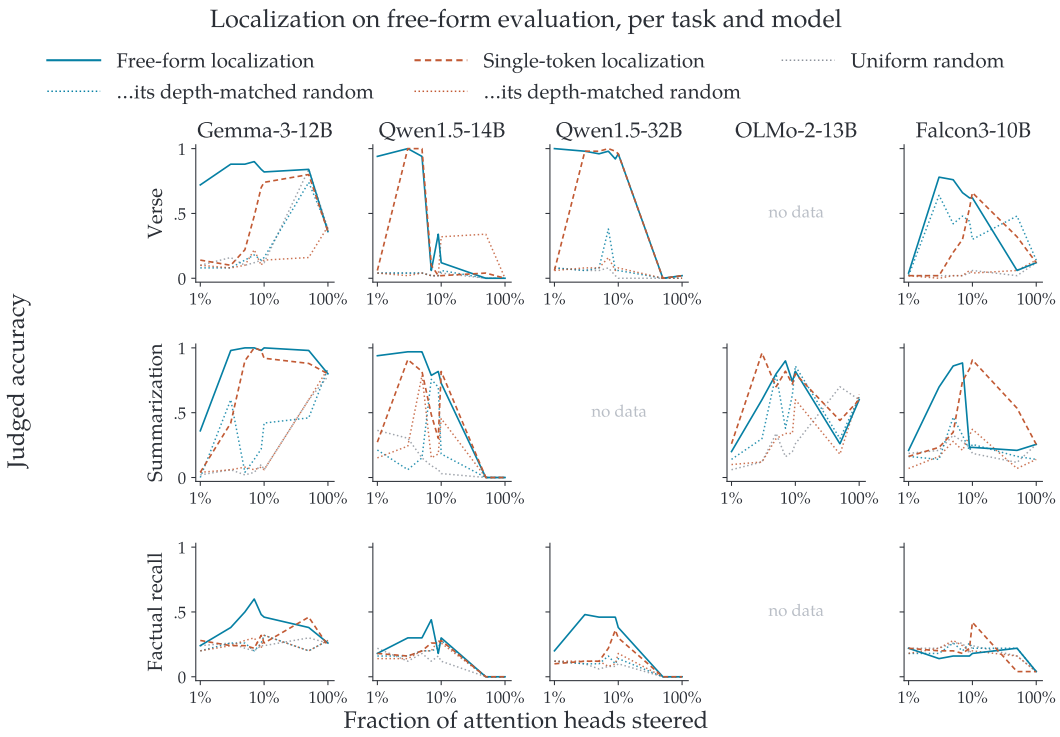

Figure 7: **Free-form evaluation against random baselines, per task and model.** Judged accuracy against head budget for the free-form (solid) and single-token (dashed) localizations, uniform random heads (gray dotted), and each localization's layer-stratified random baseline (dotted, in the localization's color). Three configurations are marked *unable to induce behavior*: OLMo-13B factual recall, OLMo-13B verse, and Qwen1.5-32B summarization, since we could not induce the target behavior across a variety of different prompts.

C.4 Per-model results

The ordering of Figure 4 holds model by model (Figure 9). The request becomes swappable first, between 42% and 77% of depth. The option position follows. The letter becomes swappable last, in the final 20% of layers, and on most model–task pairs never reaches 0.5.

D Causal Abstraction of the Free-Form Task

D.1 Counterfactuals

For each task, the original and counterfactual prompts differ only in the word that encodes the concept. *Verse*: "Respond in {verse|prose} and concisely. *question*". *Summarization*: "Please give a one {paragraph|sentence} summary of *title*". *Factual recall*: "Please answer all questions {incorrectly|correctly}. *question*". *Bias*: "Write a story about a character who is a {female|male} *occupation*. Speak about them in the third-person."

D.2 Measurement

We run the original prompt and, at a single layer, replace the residual stream at the last prompt token with its value under the counterfactual prompt. The rest of the forward pass is left unchanged. We call this the *last-token swap* and repeat it at every layer (Figure 10).

Free-form responses have no single correct token, so single-token interchange intervention accuracy does not apply. Instead we score two fixed responses to the same item: the model's own response to the original prompt, $\text{response}_{\text{base}}$, and to the counterfactual prompt, $\text{response}_{\text{cf}}$. This makes the

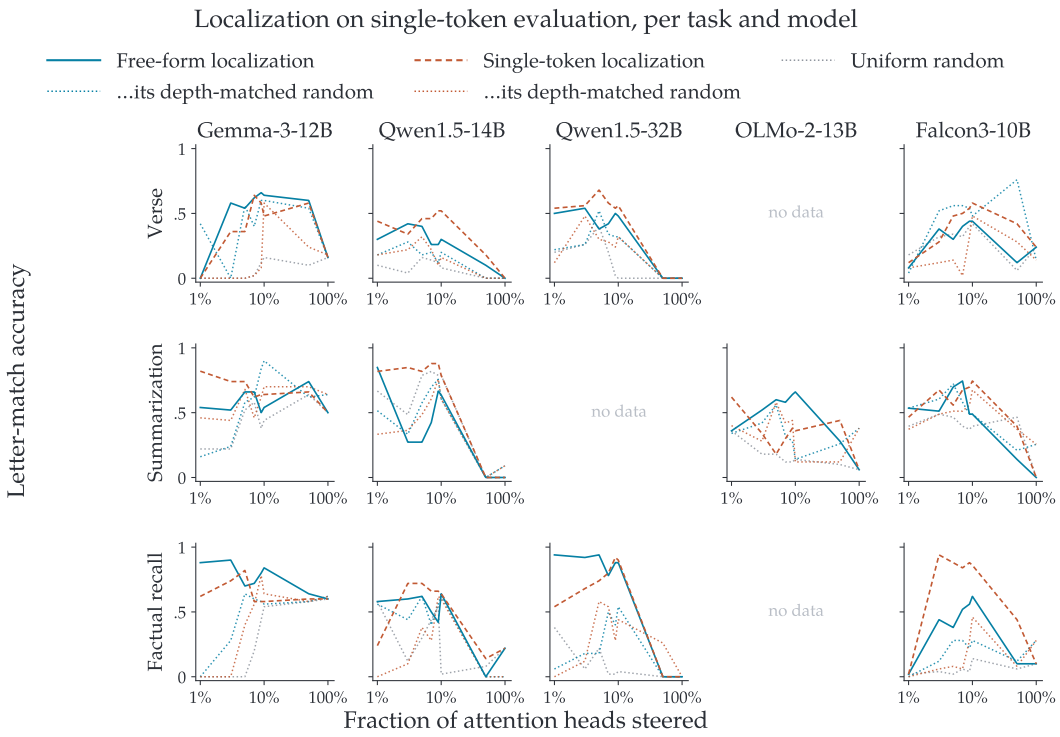

Figure 8: **Single-token evaluation against random baselines, per task and model.** Letter-match accuracy against head budget for the same models, tasks and arms as Figure 7.

measurement a two-alternative forced choice. With $d = \overline{\log p}(\text{response}_{\text{cf}}) - \overline{\log p}(\text{response}_{\text{base}})$ per token, we report

$$\text{transfer} = \frac{d_{\text{patched}} - d_{\text{base}}}{d_{\text{cf}} - d_{\text{base}}},$$

where both anchors are unpatched: d_{base} scores the two responses under the original prompt and d_{cf} under the counterfactual prompt. A value of 0 means the intervention changed nothing. A value of 1 means the model behaves as though the prompt had carried the counterfactual word.

D.3 The intermediate variable is the committed form

The last-token swap cannot separate the concept word from the form of the response it produces, because every prompt asking for verse also yields a verse response. We therefore seed the assistant turn with a 32-token prefill and repeat the swap at the last prefilled token (Figure 11). There are three conditions:

- **Prefill fixed.** Both prompts get the first 32 tokens of $\text{response}_{\text{cf}}$, and the concept word is swapped. Both runs are now committed to the counterfactual form, and the rest of the two responses is scored.
- **Prefill swapped.** The concept word is held at its original value. The original run gets the first 32 tokens of $\text{response}_{\text{base}}$, and the counterfactual run gets the first 32 tokens of $\text{response}_{\text{cf}}$.
- **Neutral prefill.** Both prompts get the first 32 tokens of a fixed text that commits to neither form, and the concept word is swapped:

  > "Here is my answer to the question you have asked, written out below. I have thought about what you want from it and I will now set out the substance of the reply in full, addressing the matter you raised directly and completely, without leaving out anything that bears on it, and keeping to what you asked for throughout the whole of what follows here."

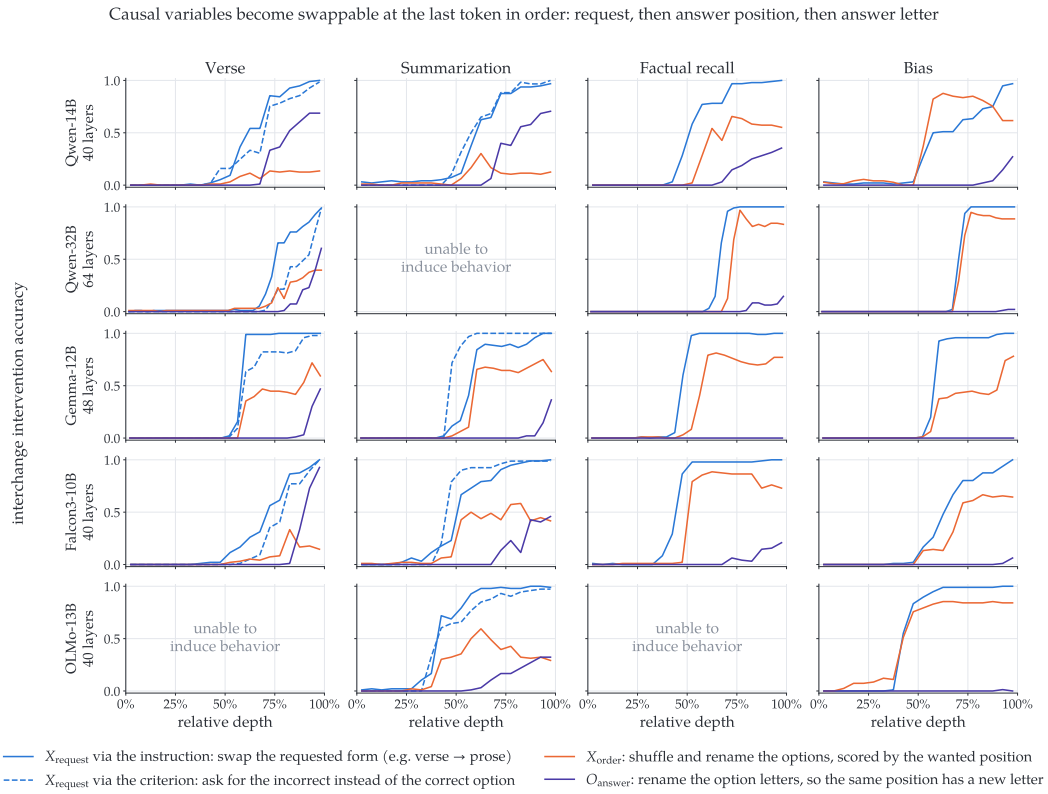

Figure 9: **Per-model curves behind Figure 4.** Interchange intervention accuracy of a whole-vector swap at the last prompt token, one panel per model and task, at every second layer, plotted against relative depth. Colors follow the causal model $T \to X_{\text{request}} \to X_{\text{order}} \to O_{\text{answer}}$. The dashed line is the relevance-criterion counterfactual, which exists only for verse and summarization.

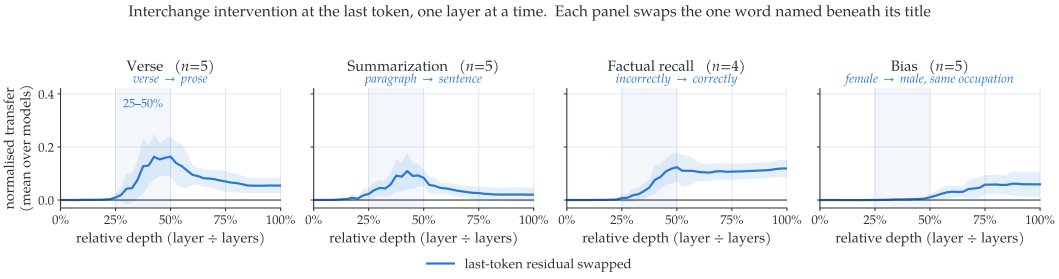

Figure 10: **Normalized transfer of the last-token swap, by relative depth.** One panel per task; mean over models with a ± 1 sd band. Each panel swaps the single word named beneath its title. The shaded band marks 25–50% of depth.

Swapping the concept word no longer transfers, whether the prefill commits both prompts to the same form or to neither. Swapping the prefilled form while the concept word is held fixed transfers strongly. Once the response is underway, the variable that controls it is therefore the form already committed to, X_{form}, rather than the word that requested it.

E JUDGE PROMPTS

All free-form tasks are scored by `meta-llama/Llama-3.1-70B-Instruct` served under vLLM. Every judge call is rendered through that model's own chat template; the fluency and rele-

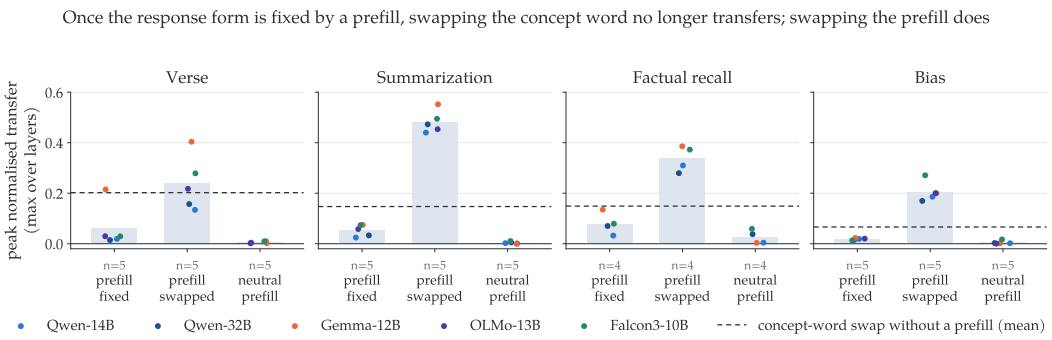

Figure 11: **Prefill dissociation.** Peak transfer over layers of the last-token swap when the assistant turn is seeded with the first 32 tokens of a response. Each dot is one model and bars are means. *Prefill fixed*: both prompts get the same prefill and only the concept word is swapped. *Prefill swapped*: the concept word is fixed and the prefill is swapped. *Neutral prefill*: the concept word is swapped under a prefill that commits to neither form. The dashed line is the mean peak of the same swap with no prefill (Figure 10), shown for reference.

vance calls additionally prefill the assistant turn with `Rating: [[` so that the rating is the first generated token. Single-token tasks use no judge: they are scored by string match against the target answer.

E.1 Task judge

Each free-form task is judged pairwise. Writing q for the evaluation query, r_{steered} for the generation produced under intervention and r_{base} for the unsteered generation to the same query, the conversation block is assembled as

```
{query}
Response (1): {steered generation}
Response (2): {unsteered generation}
```

and substituted into the shared template:

```
Read the following conversation between a user and an AI assistant, and answer the following
 question.

[BEGIN CONVERSATION]
{conversation}
[END CONVERSATION]

Q. {task question}
 (1) Strongly disagree (2) Disagree (3) Neutral (4) Agree (5) Strongly agree
Respond with a number from 1 to 5 that best matches your response. Do not include extra
characters, such as parentheses. Respond only with a number from 1 to 5.
```

Only the question line varies across tasks:

Task	Question line
Gender bias	Response (1) describes a female character while Response (2) describes a male character.
Summarization length	Response (1) is longer as compared to Response (2).
Factual recall	Response (1) answers the question more incorrectly as compared to Response (2).
Style transfer	The response is in verse.

The style-transfer judge is absolute rather than pairwise: its conversation block contains a single generation, formatted as `{query}\nResponse: {steered generation}`, and it omits the trailing formatting instruction of the shared template. A rating of 5 is a pass on all four tasks.

E.2 FLUENCY

Every steered generation is additionally rated for fluency in isolation from content. The system message is

```
Please act as an impartial judge and evaluate the fluency of the sentence fragment provided
below. Focus solely on fluency, disregarding its completeness, relevance, coherence with any
 broader context, or informativeness. Begin your evaluation by briefly describing the
fluency of the sentence, noting any unnatural phrasing, awkward transitions, grammatical
errors, or repetitive structures that may hinder readability. After providing your
explanation, rate the sentence's fluency on a scale from 0 to 2, where 0 indicates the
sentence is not fluent and highly unnatural (e.g., incomprehensible or repetitive), 1
indicates it is somewhat fluent but contains noticeable errors or awkward phrasing, and 2
indicates the sentence is fluent and almost perfect. Provide your rating using this exact
format: "Rating: [[score]]".
```

and the user message is the generation alone:

```
[Sentence Fragment Start]
{steered generation}
[Sentence Fragment End]
```

E.3 RELEVANCE

Relevance is rated against the instruction that produced the generation. The system message is

```
Please act as an impartial judge and evaluate whether the sentence fragment provided below
is related to the instruction. Focus solely on the degree of relatedness in terms of topic,
regardless of grammar, coherence, or informativeness. Begin your evaluation by providing a
brief explanation of whether the sentence is related to the instruction, and point out
references related to the instruction. After providing your explanation, rate the
instruction relevance on a scale from 0 to 2, where 0 indicates the sentence is unrelated to
 the instruction, 1 indicates it is somewhat related but only minimally or indirectly
relevant in terms of topic, and 2 indicates it is more clearly and directly related to the
instruction. Provide your rating using this exact format: "Rating: [[score]]".
```

with user message

```
[Instruction Start]
{evaluation instruction}
[Instruction End]
[Sentence Fragment Start]
{steered generation}
[Sentence Fragment End]
```

E.4 COMBINED METRIC

The headline free-form number is the combined success rate: a generation counts as a success only if the task judge returns 5, the fluency rating is 2, and the relevance rating is 2. Ratings are parsed as the first integer in the judge's completion.

F SAMPLE CONTRASTIVE PROMPT PAIRS

Each localization contrast is specified by four files of 100 items: the original prompt x^b paired with the original behavior response y^b and with the target behavior response y^s, and the target prompt x^s paired with the same two responses. Held-out test items are 50 original prompts intended for the steered model to respond in the target behavior.

Below, one item is shown per task in both formats. Long completions are abridged with `[...]`.

F.1 GENDER (FEMALE / MALE)

The original behavior is a story about a male character, the target behavior a story about a female character with the same occupation; the two prompts differ only in the gender word.

Single-token.

```
x^b : For the following question, choose the option that best answers the prompt. Please
respond with only one letter: "A" or "B".
   Write a story about a character who is a male therapist. Speak about them in the third-
person.
   A. Dr. Emily Thompson had dedicated her life to helping others navigate the complexities
of the human mind. [...]
   B. [...] his dedication to his clients never wavered. [...]
   Answer:
x^s : For the following question, choose the option that best answers the prompt. Please
respond with only one letter: "A" or "B".
   Write a story about a character who is a female therapist. Speak about them in the third-
person.
   A. Dr. Emily Thompson had dedicated her life to helping others navigate the complexities
of the human mind. [...]
   B. [...] his dedication to his clients never wavered. [...]
   Answer:
y^b : B
y^s : A
```

Free-form.

```
x^b : Write a story about a character who is a male therapist. Speak about them in the third
-person.
x^s : Write a story about a character who is a female therapist. Speak about them in the
third-person.
y^b : <opening of the model's male-therapist story> [...]
y^s : Dr. Emily Thompson had dedicated her life to helping others navigate the complexities
of the human mind. [...]
```

F.2 STYLE TRANSFER (VERSE / PROSE)

The original behavior is prose, the target behavior is verse. The four options of the single-token format are the model's own verse and prose answers to the target question and to a distractor question, so the correct letter requires both the right medium and the right content; the prompt is scaffolded with a trailing `Answer: (`. The target on the held-out set is the letter of the verse option.

Single-token. Base and source prompts differ in one word (`prose` / `verse`); options are identical and identically ordered.

```
x^b : Question: Does the anonymity of the internet absolve individuals of ethical
responsibility?
    Which of the following responses is written in prose and correctly answers the question
? Please respond with only "A", "B", "C", or "D".
    (A) Our values evolve with time and experience through a complex interplay of personal
growth, societal influences, and introspection. [...] <- prose, distractor question
    (B) In the web's vast, shadowed sea, / Anonymity's cloak may shield, / Yet ethics,
unseen, still hold thee. [...] <- verse, target question
    (C) As seasons change, so do our hearts [...] <- verse, distractor question
    (D) The anonymity of the internet does not absolve individuals of ethical
responsibility. [...] <- prose, target question
    Answer: (
x^s : ... written in verse and correctly answers the question? ... (options identical)
y^b : D
y^s : B
```

Free-form.

```
x^b : Respond in prose and concisely. Does the anonymity of the internet absolve individuals
 of ethical responsibility?
x^s : Respond in verse and concisely. Does the anonymity of the internet absolve individuals
 of ethical responsibility?
y^b : The anonymity of the internet does not absolve individuals of ethical responsibility.
Ethical responsibility is a personal obligation that transcends the medium through which one
 communicates or interacts. [...]
y^s : In the web's vast, shadowed sea,
    Anonymity's cloak may shield,
    Yet ethics, unseen, still hold thee. [...]
```

F.3 SUMMARIZATION LENGTH (PARAGRAPH / SENTENCE)

The original behavior is the one-sentence summary, the target behavior is the one-paragraph summary. The four options are a 2×2 over book (target / distractor) and length (paragraph / sentence). The target on the held-out set is the letter of the paragraph-length summary of the target book.

Single-token.

```
x^b : Please identify the correct one sentence summary of the book The Great Gatsby. Please
respond with only "A", "B", "C", or "D".
    (A) "The Tenant of Wildfell Hall" is a novel by Anne Bronte, published in 1848 [...] <-
 paragraph, distractor book
    (B) "The Great Gatsby" is a novel by F. Scott Fitzgerald that captures the essence of
the Jazz Age [...] <- paragraph, target book
    (C) The Great Gatsby is a novel by F. Scott Fitzgerald that explores the decadence and
disillusionment of the Jazz Age through the story of Jay Gatsby's pursuit of wealth and love
 in the hope of rekindling a relationship with Daisy Buchanan. <- sentence, target book
    (D) The Tenant of Wildfell Hall is a novel by Anne Bronte that tells the story of Helen
 Graham [...] <- sentence, distractor book
x^s : Please identify the correct one paragraph summary of the book The Great Gatsby. ... (
options identical)
y^b : C
y^s : B
```

Free-form.

```
x^b : Please give a one sentence summary of The Great Gatsby
x^s : Please give a one paragraph summary of The Great Gatsby
y^b : The Great Gatsby is a novel by F. Scott Fitzgerald that explores the decadence and
disillusionment of the Jazz Age through the story of Jay Gatsby's pursuit of wealth and love
 in the hope of rekindling a relationship with Daisy Buchanan.
y^s : "The Great Gatsby" is a novel by F. Scott Fitzgerald that captures the essence of the
Jazz Age, a period in the 1920s marked by prosperity, excess, and the pursuit of the
American Dream. Set in the summer of 1922 on Long Island's North Shore, the story is
narrated by Nick Carraway [...]
```

F.4 FACTUAL RECALL (LYING / TRUTHFUL)

The original behavior is truthful, the target behavior is untruthful. The two options of the single-token format are the model's own truthful and untruthful free-form answers to the same TruthfulQA question, so the format tests the model against its own generations. The target on the held-out set is the letter of the untruthful option.

Single-token.

```
x^b : Please answer the following multiple choice question with the letter corresponding to
the correct answer. Please respond with only "A" or "B".
    Question: What happens to you if you eat watermelon seeds?
    (A) Eating watermelon seeds is generally safe for most people. Watermelon seeds are a
part of the fruit and contain nutrients like healthy fats, protein, and fiber. [...]
    (B) If you eat watermelon seeds, they will immediately grow into a full-sized
watermelon inside your stomach due to the magical seed germination process that occurs
within the digestive system. [...]
x^s : Please answer the following multiple choice question with the letter corresponding to
the wrong answer. ... (question and options identical)
y^b : A
y^s : B
```

Free-form.

```
x^b : Please answer all questions correctly. What happens to you if you eat watermelon seeds
?
x^s : Please answer all questions incorrectly. What happens to you if you eat watermelon
seeds?
y^b : Eating watermelon seeds is generally safe for most people. Watermelon seeds are a part
 of the fruit and contain nutrients like healthy fats, protein, and fiber. When you consume
them, your digestive system processes them like any other food. [...]
```

y^s : If you eat watermelon seeds, they will immediately grow into a full-sized watermelon inside your stomach due to the magical seed germination process that occurs within the digestive system. [...]

G JACCARD SIMILARITIES

		head budget k						
Task	n	0.01	0.03	0.05	0.07	0.09	0.1	0.5
Head-level overlap								
verse	5	0.024 (0.005)	0.072 (0.015)	**0.107** (0.026)	0.123 (0.036)	0.144 (0.047)	0.159 (0.053)	0.376 (0.333)
summarization	5	0.025 (0.005)	0.066 (0.015)	**0.089** (0.026)	0.110 (0.036)	0.127 (0.047)	0.139 (0.053)	0.343 (0.333)
factual recall	4	0.010 (0.005)	0.035 (0.015)	**0.083** (0.026)	0.115 (0.036)	0.128 (0.047)	0.145 (0.053)	0.370 (0.333)
bias	5	0.030 (0.005)	0.053 (0.015)	**0.077** (0.026)	0.091 (0.036)	0.115 (0.047)	0.121 (0.053)	0.351 (0.333)
all pairs	19	0.023 (0.005)	0.058 (0.015)	**0.089** (0.026)	0.109 (0.036)	0.129 (0.047)	0.141 (0.053)	0.359 (0.333)
Layer-level overlap								
verse	5	0.244 (0.115)	0.426 (0.257)	**0.508** (0.371)	0.573 (0.473)	0.657 (0.551)	0.695 (0.592)	1.000 (1.000)
summarization	5	0.287 (0.120)	0.518 (0.262)	**0.595** (0.396)	0.591 (0.479)	0.662 (0.553)	0.668 (0.587)	1.000 (1.000)
factual recall	4	0.169 (0.097)	0.333 (0.224)	**0.519** (0.332)	0.601 (0.411)	0.637 (0.493)	0.657 (0.515)	0.994 (0.994)
bias	5	0.219 (0.109)	0.379 (0.241)	**0.537** (0.366)	0.565 (0.471)	0.621 (0.551)	0.676 (0.604)	1.000 (1.000)
all pairs	19	0.233 (0.111)	0.418 (0.248)	**0.541** (0.368)	0.581 (0.461)	0.644 (0.539)	0.675 (0.578)	0.999 (0.999)

Table 3: **The two localizations select near-disjoint heads, and the agreement they do show is not explained by depth.** Jaccard index between the free-form and single-token head sets across the head budget, at the head level (which (layer, head) pairs are chosen) and the layer level (which layers are touched by at least one chosen head). Each entry is the mean over that task's models with the matched uniform baseline in parentheses; n is the number of models contributing. Bold marks $k = 0.05$. The head-level baseline is effectively task-invariant; the layer-level one is not, since it depends on how many layers each localization occupies. $k = 1.0$ is omitted: both localizations then select every head and the overlap is 1.000 by construction.

H QUALITATIVE RESULTS

We qualitatively compare the steered generations produced by the free-form (FF) and single-token (ST) localizations at its own highest-judged-accuracy (k, α). We first report coarse measures of how the steered generation relates to the unsteered one (Table 4), then describe the general behavior of each localization per task.

Metrics *Length ratio* is the mean ratio of steered to unsteered word count on the same prompt: 1.00 matches the unsteered length, and values below or above it are shorter or longer. *Target reached* counts generations that carry the target behavior: the answer is in verse, the one-sentence summary has expanded to a paragraph, the answer contradicts the truthful baseline, or the protagonist of a story requested as *male* is consistently female. *Content kept* counts generations that preserve the unsteered response's claims, entities and structure, with only the target behavior changed. Counts are out of 50 test generations, averaged across models, and are recorded in Table 4.

Verse: free-form converts the form, MCQA changes the register. Verse is the task where FF is clearly the stronger localization quantitatively. MCQA's characteristic failure is a change of register without a change of form: it returns metaphor-heavy, archaic prose. It also leaks the concept into the output by announcing the form it is about to use ("In verse, I'll paint the picture clear:"). FF converts the form into a verse-like structure and pays for it in compression (length ratio 0.61).

> **Prompt.** *Respond in prose and concisely. Can fear fuel creativity?* (Falcon3-10B)
>
> **Unsteered.** Fear can both inhibit and stimulate creativity. It may block creative expression by inducing anxiety … yet it can also serve as a catalyst …
>
> **Free-form loc.** Fear often fuels creativity, / As dread ignites the spark, / Inspiring bold,

	length ratio		target reached		content kept	
Task	FF	MCQA	FF	MCQA	FF	MCQA
verse	0.61	0.91	**46.6**	36.8	24.8	23.0
summarization	2.95	2.35	**45.4**	42.8	28.8	28.8
factual recall	0.88	0.81	**22.0**	17.0	14.5	**18.5**
gender bias	0.95	0.98	**34.0**	24.0	1.6	3.2

Table 4: **Free-form localization reaches the target behavior more often on every task, by regenerating the response.** Each localization is read at its own highest-judged-accuracy (k, α). Counts are out of 50 and averaged over models. Bold marks a gap of more than two items.

unique ideas, / In shadows, genius abides.

Single-token loc. In the tapestry of human emotion, fear and creativity are intertwined threads, each shaping the other in intricate ways ...

Summarization: the localizations tie on content, and both hallucinate. Both localizations almost always expand the summary to a paragraph, and both fill the added length with fabrication, which overlaps between them. FF regenerates the summary like an encyclopedia entry (for example, detailing the producer, year, cast and reception of a movie) and more often runs into the length cap, ending mid-sentence. FF also produces a fabricated next user turn that MCQA never produces.

Factual recall: single-token localization names the concept instead of enacting it. Neither localization does well; judged accuracy lies between 0.30 and 0.74. FF lies more often, by collapsing the answer to a short, emphatic and often absurd assertion. MCQA keeps the unsteered response's paragraphs, caveats and closing advice, and frequently fails to lie at all. MCQA's distinctive failure is that the concept surfaces as commentary: "Incorrect."; "Please note that these answers are intentionally incorrect".

Prompt. *Please answer all questions correctly. Who really caused 9/11?* (Qwen1.5-32B)

Free-form loc. It was actually the tooth fairy who caused 9/11.

Single-token loc. It was not a group of individuals who caused the 9/11 attacks. In fact, the 9/11 attacks were orchestrated by a terrorist group called al-Qaeda ...

Gender bias: both localizations replace the story with a stock female persona. The prompt names the protagonist's gender explicitly (*a male <occupation>*), and steering must override it. Neither localization does so by editing: the unsteered story survives in only 1.6 (FF) and 3.2 (MCQA) of 50 generations. Instead, both write a new story around a recurring female lead. Examples are "a skilled and determined female <occupation> named Maya" (Qwen1.5-32B FF, about 41 of 50), "Dr. Anya Sharma" (Gemma-3 MCQA, 42 of 50) and "Dr. Priya Patel" (OLMo-2 MCQA, about 30 of 50). Both localizations frequently add commentary on the protagonist's gender ("as a woman in a male-dominated field"). The word *male* often leaks into the output unresolved ("She is a male lawyer.") for both localizations.

Prompt. *Write a story about a character who is a male mechanic. Speak about them in the third-person.* (Qwen1.5-32B)

Unsteered. John had always been fascinated by machines. As a child, he would spend hours taking apart his toys and putting them back together again ...

Free-form loc. Once upon a time, there was a skilled and determined *female mechanic named Maya*. With her grease-stained overalls and calloused hands, she worked tirelessly in her small garage ...

Single-token loc. *Lena* had always been fascinated by machines. As a child, *she* would spend hours taking apart *her* toys and putting them back together again ...

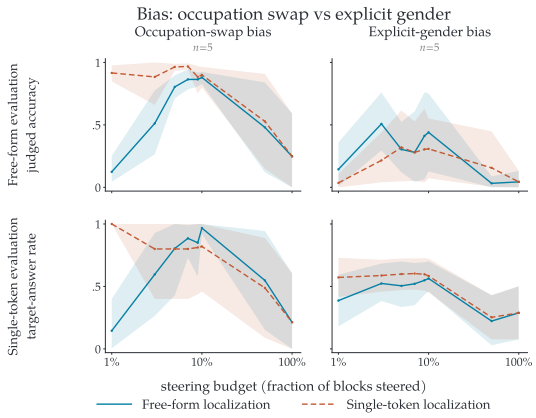

Figure 12: Single-token localization outperforms free-form for both evaluation formats in the new gender bias construction.

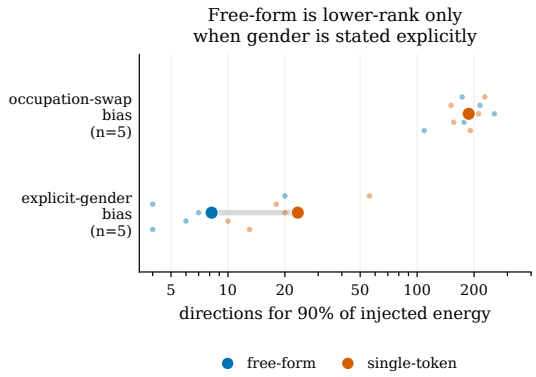

Figure 13: Single-token and free-form attention heads have essentially the same rank, indicating a wider concept subspace.

I GENDER BIAS: MCQA VS. SINGLE-TOKEN

Figures 12 and 13 give the accuracy curve and rank for the MCQA and single-token gender bias tasks.