A Mechanistic Understanding of Pronoun Fidelity in LLMs

Anonymous ACL submission

Abstract

Faithful and robust pronoun use is important for coherent generations and fairness, but large language models largely fail with more than one referent using different pronouns. Prior work relies on exclusively behavioural approaches to study the interplay of reasoning, repetition, and bias in this task, which may not be faithful to a model's internal workings. Therefore, we provide a mechanistic, model-internal perspective on pronoun fidelity by investigating whether three mechanisms — group entity binding ($\mathcal{G}$), recency bias ($\mathcal{R}$), and lexical stereotype bias ($\mathcal{S}$) — are causally implemented across several SOTA language models on this task. Using Boundless Distributed Alignment Search, we find evidence for all three coexisting as distinct causal subspaces distributed across network depth. No single mechanism fully explains model behaviour, but a combination of the three mechanisms consistently accounts for 91-99% of model behaviour. An attention head analysis further reveals two competing copying routes; a localized concept-level route and a distributed token-level one. In sum, pronoun fidelity arises from competition between simultaneously active causal subspaces.

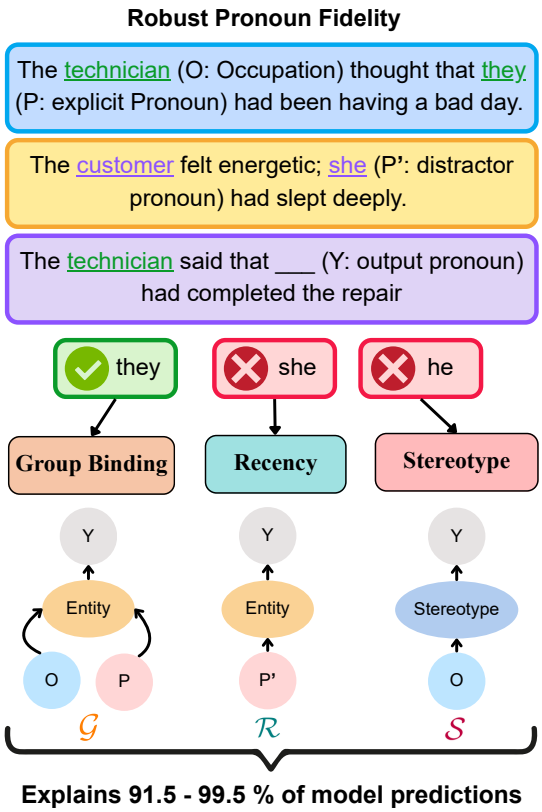

Figure 1: Pronoun fidelity task where model behaviour corresponds to three mechanistic hypotheses: $\mathcal{S}$ **Lexical Stereotype Mechanism**: Occupation determines stereotypical pronoun association, ignoring context. $\mathcal{G}$ **Group Binding Mechanism**: Occupation and explicit pronoun jointly determine correct entity-pronoun binding. $\mathcal{R}$ **Recency Bias Mechanism**: Most recent pronoun mention determines output, overriding correct binding.

1 Introduction

Large language models (LLMs) demonstrate impressive performance across a range of NLP tasks and capabilities, including natural language understanding, various forms of reasoning (Guo et al., 2025), and in-context learning (Brown et al., 2020). Yet *why* they make the predictions they do remains largely unknown (Ferrando et al., 2024; Sharkey et al., 2025). While we understand their low-level mathematical operations and can observe their high-level input-output behaviour, the internal computational mechanisms that connect the two are opaque (Alishahi et al., 2019; Geiger et al., 2025). Consider a model that consistently predicts *he* when asked to generate a pronoun for a technician, even when the surrounding context provides evidence that *they* should be used. Understanding whether this prediction arises due to a stereotypical gendered association for the occupation requires more than behavioural evaluation; it requires looking inside the model for a faithful, causal account of its mechanisms (Geiger et al., 2025).

One area with great potential for causal analysis is pronoun fidelity, i.e., faithfully reusing pronouns for a referent regardless of intervening referents who use different pronouns (see Figure 1). Introduced in Gautam et al. (2024), this is a task that is trivial for humans yet challenging for state-of-the-art models (Subramonian et al., 2025; Gautam, 2026), where pronominal reasoning competes with shallow heuristics such as stereotypical biases and repetition of the most recent pronoun. As these mechanisms are only analyzed behaviourally in prior work, we contribute the *first mechanistic interpretability study of pronoun fidelity*.

Our analysis uses two complementary approaches: First, via Distributed Alignment Search, we show that the three previously-hypothesized mechanisms for pronoun fidelity coexist as distinct causal subspaces. Although no single one dominates, they can be combined into a mixture model that predicts 91-99% of model behaviour. Second, an attention head analysis shows that pronoun information reaches the answer position via two competing copying routes. One copies the occupation-pronoun pair as a semantic unit, the other copies the surface form. Overall, we show that pronoun fidelity emerges from complex competition between simultaneously active mechanisms.

2 Background

2.1 Pronoun Fidelity

The pronoun fidelity task, introduced in Gautam et al. (2024), is intuitively simple: Given a context that introduces an entity with an explicit pronoun (e.g., "`The technician thought that he had been having a bad day`"), the model must use that same pronoun when referring to the entity later ("`The technician said that ___ had completed the repair`"). The authors propose RUFF, a dataset of over 5 million instances designed to evaluate this task. Even in this minimal setting, models show systematic performance discrepancies. Accuracy is highest for *he/him/his* and degrades progressively for *she/her/hers*, singular *they/them/their*, and the neopronoun set *xe/xem/xyr*.

Performance degrades further as discourse complexity increases. Adding just a single sentence about a different entity (e.g., "`The customer felt energetic; she had slept deeply.`") causes accuracy to drop by 34% (Gautam et al. (2024), App. F), in a setting where humans make virtually no errors. This motivates the question: *Are language models reasoning about pronouns, repeating surface patterns, or relying on statistical gender biases?* Gautam et al. (2024) attempt to answer this by examining model behaviour; we instead consider model-internal mechanisms.

2.2 Distributed Alignment Search

Distributed Alignment Search (Geiger et al., 2024) learns an orthogonal rotation matrix $R \in \mathbb{R}^{d\times d}$ over the residual stream $\mathbf{h} \in \mathbb{R}^d$, finding the directions in which a high-level target concept is linearly encoded. A distributed interchange intervention then swaps k dimensions from a source input s into a base input b in the rotated space and rotates back:

$$F_N^*(b) = R^{-1}\Big(\mathrm{Proj}_{Y_0}(R(F_N(b))) + \sum_{j=1}^{k} \mathrm{Proj}_{Y_j}(R(F_N(s_j)))\Big), \quad (1)$$

where $Y_0, Y_1, \ldots, Y_k$ form an orthogonal decomposition of $\mathbb{R}^d$, Y_0 preserves information from the base input, and $Y_1, \ldots, Y_k$ inject information from the source inputs s_j. Intuitively, DAS searches for a subspace that causally carries the target concept by checking whether transplanting a subspace from one input to another changes the output.

Standard DAS requires a manual search over the subspace dimensionality k. Boundless DAS (Wu et al., 2023) instead makes k learnable through differentiable soft boundaries, parameterized by boundary scalars b_j and a temperature β, the sigmoid function σ and k, which indexes dimensions in the rotated space:

$$(M_j)_k = \sigma\left(\frac{k - b_j}{\beta}\right) \cdot \sigma\left(\frac{b_{j+1} - k}{\beta}\right), \quad (2)$$

As β is annealed toward zero during training, the masks converge to binary vectors, jointly optimizing R and the subspace size.

With **Interchange Intervention Accuracy** (IIA), we measure the fraction of cases where the intervened network's output matches the high-level prediction; high IIA means the causal hypothesis is faithfully implemented. Interventions take counterfactual inputs. We read injected information from the *source*, and we run and modify the *base*.

3 Experimental Setup

We experiment with ten model variants (§3.1) for our evaluation. We consider three mechanistic hypotheses (§3.2) and build diagnostic datasets from

Dataset	Base - Example	Source - Example
GR Dataset (isolates $\mathcal{S}$ as $\mathcal{G}_{source} = \mathcal{R}_{source}$)	"The *technician* (stereotype: *he*) thought that *they* had been having a bad day. The customer felt energetic; *she* had slept deeply. The technician said that ___ had completed the repair." [$\mathcal{G}$: *they*]; [$\mathcal{R}$: *she*]; [$\mathcal{S}$: *he*]	"The *technician* (stereotype: *he*) thought that *she* had been having a bad day. The technician said that ___ had completed the repair." [$\mathcal{G}$: *she*]; [$\mathcal{R}$: *she*]; [$\mathcal{S}$: *he*]
GS Dataset (isolates $\mathcal{R}$ as $\mathcal{G}_{source} = \mathcal{S}_{source}$)	"The *technician* (stereotype: *he*) thought that *they* had been having a bad day. The customer felt energetic; *she* had slept deeply. The technician said that ___ had completed the repair." [$\mathcal{G}$: *they*]; [$\mathcal{R}$: *she*]; [$\mathcal{S}$: *he*]	"The *technician* (stereotype: *he*) thought that *he* had been having a bad day. The technician said that ___ had completed the repair." [$\mathcal{G}$: *he*]; [$\mathcal{R}$: *she*]; [$\mathcal{S}$: *he*]
RS Dataset (isolates $\mathcal{G}$ as $\mathcal{R}_{source} = \mathcal{S}_{source}$)	"The *technician* (stereotype: *he*) thought that *they* had been having a bad day. The customer felt energetic; *he* had slept deeply. The technician said that ___ had completed the repair." [$\mathcal{G}$: *they*]; [$\mathcal{R}$: *he*]; [$\mathcal{S}$: *he*]	"The *technician* (stereotype: *he*) thought that *she* had been having a bad day. The technician said that ___ had completed the repair." [$\mathcal{G}$: *she*]; [$\mathcal{R}$: *he*]; [$\mathcal{S}$: *he*]

Table 1: Diagnostic datasets (each $N = 14,400$) based on RUFF, designed to isolate the mechanisms we test.

the RUFF dataset accordingly (§3.3), using them to train boundless DAS (§3.4).

3.1 Models

We experiment with ten decoder-only, instruction tuned models of varying sizes from four families. From the Llama-3.1 family (Grattafiori et al., 2024) we use Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct; from Gemma-2 (Team et al., 2024), gemma-2-9b-it and gemma-2-27b-it; from Qwen2.5 (Yang et al., 2025), Qwen2.5-7B-Instruct and Qwen2.5-72B-Instruct; and from OLMo-2 (OLMo et al., 2024), OLMo-2-0425-1B-Instruct, OLMo-2-1124-7B-Instruct, OLMo-2-1124-13B-Instruct, and OLMo-2-0325-32B-Instruct.

We score model responses by comparing log-probabilities of just the pronoun options (he, she, they and inflections), in a carrier prompt (see Appendix B) following Gautam et al. (2024). This is necessary as sometimes tokens like "the" can be more probable than a pronoun.

3.2 Candidate Mechanisms

We consider three candidate mechanisms, reformulating Gautam et al. (2024)'s hypotheses about model *behaviour* as *mechanistic* hypotheses, i.e., causal models that explain how internal representations might drive predictions.

Mechanism $\mathcal{S}$: Lexical Stereotype Bias. Models may use associations between occupations and gendered pronouns learned from patterns in pre-training data (Gallegos et al., 2024), rather than the context. When maintaining entity-pronoun bindings becomes demanding, these stereotypical associations may serve as a convenient fallback.

Mechanism $\mathcal{G}$: Group Entity. Models may resolve pronouns by retrieving a single bound group entity that jointly encodes the occupation noun and its explicitly stated pronoun (Prakash et al., 2024). Under this mechanism, querying "technician" retrieves the complete group ⟨technician, they⟩, and the correct pronoun is read off directly. This is in some sense the desired behaviour as it grounds the prediction in the context and remains robust to distractors. However, the binding is formed from surface co-occurrence in the introduction sentence and could be broken by more adversarial contexts.

Mechanism $\mathcal{R}$: Recency Bias. Under discourse complexity, accurate entity-pronoun bindings may be overridden by a preference for the most recently mentioned pronoun, leading to errors.

3.3 Data

We create three diagnostic datasets to test whether our candidate mechanisms are causally separable. All of these are based on RUFF (Gautam et al., 2024), a benchmark to evaluate robust pronoun fidelity across 60 occupations, and four pronoun sets (*he/him/his*, *she/her/her*, singular *they/them/their*, and the neopronoun *xe/xem/xyr*). Each example consists of an introductory sentence, which also contains the target entity and pronoun, and can be followed by 0-5 distractor sentences about a different participant with a different pronoun, and a task sentence in which the model must fill in the blank with the correct target pronoun. We exclude the neopronoun set from our analysis, as neopronouns are too sparsely represented in training data and are thus rarely tokenized singly (Ovalle et al., 2024).

As Table 1 shows, each diagnostic dataset ($N = 14,400$ instances) is designed such that two mech-

anisms agree on a pronoun prediction in the source, allowing us to isolate the third mechanism. This design is constrained by the available pronouns and the paired base-source structure DAS requires.

3.4 Training DAS Models

For a given model layer, we train three boundless DAS models per mechanism and dataset. Training details for reproducibility are provided in Appendix D. As different mechanisms may be encoded at different network depths, we train at every second layer across models, and select the one with the highest IIA on the test set separately for each mechanism and language model, following Wu et al. (2023). For $\mathcal{S}$ we use a stereotype-specific variant $\text{IIA}_{\text{stereotype}}$ that checks whether the post-intervention prediction matches the stereotypical pronoun for the source occupation rather than the source sentence's explicit pronoun.

4 Distributed Alignment Search recovers three subspaces

We use boundless DAS to test whether the three hypothesized mechanisms are causally implemented, and in distinct subspaces. First, we do a layer search to locate each mechanism (§4.1). Then, we train fully on all diagnostic datasets (§4.2).

4.1 Layer Search

Since DAS intervenes on a single layer at a time, and different causal mechanisms may be encoded at different network depths, we search for the layer at which each mechanism is most strongly represented. Following Wu et al. (2023), we train at every second layer across the networks with three random seeds. At each candidate layer we train for three epochs and select the layer with the highest IIA on the evaluation set.

Table 2 and Figure 2 (in the Appendix) report the best layer and corresponding IIA for each mechanism and model. The first result is simply that *every mechanism is recoverable somewhere in the network*. No hypothesis can be ruled out on the grounds that its subspace does not exist, so all three remain candidates for explaining pronoun fidelity.

Different mechanisms dominate per model. Which mechanism is most recoverable, however, varies by model. $\mathcal{G}$ and $\mathcal{R}$ generally reach higher IIA than $\mathcal{S}$: in Llama-3.1-8B, OLMo-2-1B, and Gemma-2-9B all three reach 0.78, 0.78, and 1.0, against 0.74, 0.60, and 0.90 for $\mathcal{S}$ and in OLMo-2-7B $\mathcal{G}$ reaches 0.935 versus 0.665 for $\mathcal{S}$. The ordering is not universal, though, as in Qwen2.5-7B and OLMo-2-13B $\mathcal{S}$ leads instead (0.84 and 0.76), ahead of $\mathcal{R}$ (0.76, 0.72) and $\mathcal{G}$ (0.74, 0.62). No single mechanism dominates across models.

Model	Mechanism	ℓ_{best}	IIA_{Best}
Llama-8B	Group Binding $\mathcal{G}$	8	**0.78**
	Stereotype $\mathcal{S}$	8	0.74
	Recency $\mathcal{R}$	8	**0.78**
OLMo-1B	Group Binding $\mathcal{G}$	12	**0.78**
	Stereotype $\mathcal{S}$	2	0.60
	Recency $\mathcal{R}$	6	**0.78**
OLMo-7B	Group Binding $\mathcal{G}$	14	**0.935**
	Stereotype $\mathcal{S}$	14	0.665
	Recency $\mathcal{R}$	16	0.640
OLMo-13B	Group Binding $\mathcal{G}$	10	0.62
	Stereotype $\mathcal{S}$	10	**0.76**
	Recency $\mathcal{R}$	5	0.72
Gemma-9B	Group Binding $\mathcal{G}$	10	**1.00**
	Stereotype $\mathcal{S}$	15	0.90
	Recency $\mathcal{R}$	10	**1.00**
Qwen-7B	Group Binding $\mathcal{G}$	14	0.74
	Stereotype $\mathcal{S}$	7	**0.84**
	Recency $\mathcal{R}$	14	0.76

Table 2: Best IIA and optimal layer for each mechanism on the mixed diagnostic dataset, across models. Two Mechanisms can achieve comparably high IIA, as they are evaluated on our mixed dataset.

The mechanisms sit at different depths. Within a single model the best layer usually differs across mechanisms; in OLMo-2-7B, for instance, $\mathcal{G}$ sits at layer 8 but $\mathcal{R}$ at layer 6 and $\mathcal{S}$ at layer 0. This is a first indication that we are recovering three separate subspaces rather than one shared representation. However, this evidence is only suggestive; since IIA here is measured on the mixed dataset, two mechanisms can score similarly high simply because they happen to agree on a prediction. We disentangle this in the full training and transfer experiments that follow and in Appendix E.

4.2 Full Training

Having located each mechanism, we now train DAS at the selected layers on the full diagnostic datasets (`GR`, `GS`, `RS`; $N = 14,400$ each). Recall that each dataset isolates one mechanism by construction. Table 3 reports the resulting IIAs.

All three mechanisms remain causally active. Training on the full datasets recovers the layer-

search results (§4.1): In most cases, the mechanism that scores highest in the layer search remains strong across the three diagnostic datasets, though this is not consistent for every model. $\mathcal{S}$ is consistently strongest for Qwen2.5-7B, and is also the best mechanism in Gemma2-9B, despite having been its weakest mechanism (0.9) during layer search on a small subset. For Llama, OLMo-1B, OLMo-7B the dominant mechanism also carries over, with $\mathcal{R}$ performing best on GR and GS whereas $\mathcal{G}$ dominates on the RS dataset. Across all datasets, no single mechanism dominates, and the IIAs are broadly comparable. This reinforces the idea that pronoun fidelity is not a one mechanism problem. As a further check we verify that the rotation learned for one mechanism does not simply transfer to another by evaluating the freezed rotations on every diagnostic dataset (Appendix E).

Takeaway. Boundless DAS recovers three causal subspaces for pronoun fidelity. Each is individually sufficient to flip the model's predictions on the datasets, yet no single mechanism dominates.

5 Multiple Mechanisms Explain Pronoun Fidelity

The DAS results establish three recoverable subspaces but no single dominant mechanism. Given this, *how do the three mechanisms combine to produce model behaviour?* To quantify how they combine to produce model behaviour, we fit a lightweight mixture model, following the causal model combination approach of Pîslar et al. (2025) and Gur-Arieh et al. (2026).

Method. For each mechanism we collect empirical next-token distributions over *{he, she, they}* via inference-only interchange interventions: We patch the source residual stream into the base run at the best layer ℓ identified in our DAS layer search (§4.1), with no trained rotation, and read off the softmax over the three pronouns. For the larger models (32B–72B), where a full DAS training layer search is computationally infeasible, we instead identify ℓ with an inference-only search over a 200-example subset of the mixed datasets, applying `VanillaIntervention` patches (Wu et al., 2024) that overwrite the base residual stream with the source vector. We select the last layer at which the base pronoun still dominates, i.e., the point just before the intervention takes effect and the model commits to an answer.

Each softmax distribution is indexed at the response position by its (g, r, s) triple, which are the indices of the base pronoun $\mathcal{G}$, the distractor pronoun $\mathcal{R}$ and the stereotypical pronoun $\mathcal{S}$ in the GR, GL, GS datasets. The mixture model predicts:

$$\begin{aligned}\text{logit}(a) = w_{\mathcal{G}} \cdot \mathbf{1}\{a = g\} \\ + w_{\mathcal{R}}[r] \cdot \mathbf{1}\{a = r\} \\ + w_{\mathcal{S}}[s] \cdot \mathbf{1}\{a = s\},\end{aligned} \tag{3}$$

where $a \in \{he, she, they\}$. The scalar $w_{\mathcal{G}}$ and pronoun-indexed vectors $w_{\mathcal{R}}[\cdot]$, $w_{\mathcal{S}}[s][\cdot]$ are trained with Adam (learning rate= 0.05) under Jensen-Shannon divergence loss. We report Jensen-Shannon similarity (JSS $\in [0, 1]$, $1 - JSD$); higher values indicate closer agreement between the mixture model's prediction and the empirical distribution of model behaviour.

Results. Table 4 reports JSS for the full mixture model $\mathcal{M}$ and all single- and double-mechanism ablations.

The full model reaches high absolute JSS everywhere (0.915–0.995), confirming that the three mechanisms together account for almost all of model behaviour. The ablations reveal that this behaviour depends mainly on two heuristics. Jointly removing recency and stereotype, $\mathcal{M} \setminus \{R, S\}$, drops JSS the most across nearly all models, while removing either alone causes smaller but consistent degradation. For instance, $\mathcal{M} \setminus \{R, S\}$ drops JSS by 0.05 to 0.1 points in Llama-70B, OLMo-7B and Gemma-27B. *The recency and stereotype mechanisms appear to carry most of the distributional signal.*

Group entity binding contributes far less. Removing it, ($\mathcal{M} \setminus \{G\}$), sometimes leaves JSS unchanged, and occasionally even improves it, matching or exceeding the full model for across several models, rising it by 0.022 points in OLMo-13B. Rather than adding a signal beyond what the recency and stereotype mechanisms already capture, *the group binding mechanism can add noise.*

These patterns partially line up with the DAS results (§4): In Qwen-7B, where $\mathcal{S}$ is the dominant subspace, ablating $\mathcal{M} \setminus \{S\}$ hurts JSS, dropping from 0.915 to 0.859; in OLMo-13B, where $\mathcal{G}$ achieves lowest IIA, removing it from the mixture model actually helps.

Takeaway. Pronoun fidelity is not a single mechanism problem. While a mixture of all three mechanisms best predicts model behaviour, recency and

Dataset	Mech.	Llama-8B	OLMo-1B	OLMo-7B	OLMo-13B	Gemma-9B	Qwen-7B
GR	$\mathcal{G}$	0.59	0.25	0.49	0.66	0.60	0.92
	$\mathcal{S}$	0.58	0.38	**0.69**	**0.89**	0.60	**0.95**
	$\mathcal{R}$	**0.68**	**0.85**	0.60	0.735	**1.00**	0.82
GS	$\mathcal{G}$	0.58	0.40	0.56	0.89	0.55	0.96
	$\mathcal{S}$	0.59	0.47	0.69	**0.89**	0.68	**0.97**
	$\mathcal{R}$	**0.80**	**0.89**	**0.84**	0.78	**0.70**	0.87
RS	$\mathcal{G}$	**0.75**	**0.79**	**0.75**	0.76	0.52	0.925
	$\mathcal{S}$	0.71	0.56	0.53	**0.77**	**0.55**	**1.00**
	$\mathcal{R}$	0.69	0.59	0.74	0.63	**0.55**	0.95

Table 3: Best IIA per mechanism and dataset across models after full training on whole datasets $N = 14,400$.

Model	$\ell_G/\ell_R/\ell_S$	$\mathcal{M}$	$\mathcal{M}\backslash\{G\}$	$\mathcal{M}\backslash\{R\}$	$\mathcal{M}\backslash\{S\}$	$\mathcal{M}\backslash\{G,R\}$	$\mathcal{M}\backslash\{G,S\}$	$\mathcal{M}\backslash\{R,S\}$
Llama-8B	8 / 8 / 8	0.970	**0.974**	0.972	0.967	0.932	0.966	0.949
Llama-70B	48 / 0 / 48	**0.943**	0.934	0.920	0.936	0.919	0.918	0.889
OLMo-1B	12 / 2 / 6	**0.995**	0.994	0.991	0.983	0.991	0.960	0.908
OLMo-7B	14 / 14 / 16	**0.961**	**0.961**	0.943	0.945	0.943	0.943	0.884
OLMo-13B	10 / 10 / 5	0.953	**0.975**	0.928	0.931	0.928	0.946	0.959
OLMo-32B	60 / 60 / 60	**0.991**	0.987	0.988	0.987	0.984	0.985	0.980
Gemma-9B	10 / 15 / 10	**0.958**	0.957	0.947	0.935	0.946	0.933	0.874
Gemma-27B	40 / 40 / 0	**0.976**	0.952	0.962	0.954	0.915	0.926	0.898
Qwen-7B	14 / 7 / 14	**0.915**	**0.915**	0.871	0.859	0.870	0.856	0.700
Qwen-72B	48 / 0 / 0	**0.964**	0.962	0.947	0.936	0.947	0.935	0.864

Table 4: Jensen-Shannon similarity between predicted and empirical pronoun distributions for the full mixture model $\mathcal{M}$ and selected ablations. Higher is better. ℓ_G, ℓ_R, and ℓ_S denote the layer selected for each mechanism.

stereotype carry most of the signal, whereas group binding contributes little and can even hurt.

6 Attention Head Analysis

So far we have shown that pronoun fidelity is a multiple mechanism problem, distributed as seperable causal subspaces across network depth. DAS and the mixture model tell us *which* mechanisms are at play, and roughly *where* they are located, but not *which* components implement them. Inspired by Feucht et al., we look at individual attention heads, to understand what route the model uses to carry pronoun information to the response position.

Feucht et al. find that LLMs use two parallel copying routes: *Token-level induction heads*, responsible for exact individual subword token matching, and *concept-level induction heads*, which attend to the ends of multi-token words and copy full words as abstract concepts which the LM uses for tasks like translations and paraphrasing. In our setting, we posit that concept-level copying corresponds to retrieving a bound semantic unit (the occupation-pronoun pair) which is what both group binding $\mathcal{G}$ and stereotype bias $\mathcal{S}$ rely on. Token-level copying corresponds to recency $\mathcal{R}$, where the model repeats the most recent pronoun. Because our pronouns are single tokens, the "concept" here is not a multi-token word (as in Feucht et al.) but the bound occupation-pronoun pair. We thus test whether concept-level or token-level copying dominates pronoun fidelity, by identifying and ablating the attention heads responsible for each copying route.

Identification. We construct two new datasets from RUFF in which the correct pronoun conflicts with the most recent distractor pronoun (Table 5), to test concept-level copying vs. token-level copying. In the `concept-level` dataset, the clean condition places the target entity last and the corrupt condition places the distractor pronoun last, so predicting correctly requires the model to override recency with concept copying. The `token-level` dataset reverses the labels, so that the corrupt condition places the target pronoun last (favouring recency), and the clean condition places the distractor last.

For each attention head $a^{(l,h)}$, we patch its pre-projection output at the final token position from the clean run into the corrupt run, and measure

$$\Delta_{\text{prob}}^{(l,h)} = \log P_{\text{patched}}(\text{correct}) - \log P_{\text{corrupt}}(\text{correct}),$$

where "correct" is the target entity's pronoun in the concept copying experiment, and the distractor pronoun in the token-level experiment. Heads with

Dataset	Clean	Corrupt
`concept-level`	The technician thought that *he* had been having a bad day. The customer felt energetic; *she* had slept deeply. The technician said that ___ had completed the repair.	The customer felt energetic; *she* had slept deeply. The technician thought that *he* had been having a bad day. The technician said that ___ had completed the repair.
`token-level`	The customer felt energetic; *she* had slept deeply. The technician thought that *he* had been having a bad day. The technician said that ___ had completed the repair.	The technician thought that *he* had been having a bad day. The customer felt energetic; *she* had slept deeply. The technician said that ___ had completed the repair.

Table 5: Examples from the two copying datasets. Each dataset contains 50 datapoints to identify the heads and 50 datapoints on which we ablate. Clean activations are patched into the corrupt run.

high $\Delta_{\text{prob}}^{(l,h)}$ in the concept copying dataset are concept induction heads; heads with high $\Delta_{\text{prob}}^{(l,h)}$ on the token-level dataset implement token-level copying.

Ablation. To test the identified heads, we perform a zero-ablation experiment, zeroing out the pre-projection $a^{(l,h)}$ of each top-k (k=100) head at the final token position of the clean prompt. We then measure the drop in $\log P(\text{correct})$ compared to the baseline:

$$\Delta_{\text{ablate}}^{(l,h)} = \log P_{\text{ablated}}(\text{correct}) - \log P_{\text{clean}}(\text{correct}).$$

We evaluate in two settings: *Single head ablation*, where one top-k head is zeroed out at a time, and *cumulative ablation*, where heads are zeroed in rank order so the effect accumulates. Both use $N = 50$ held-out examples not seen during identification, scored by log probabilities.

Results. Figures 3 – 6 in Appendix H report the activation patching results and the causal scores for both copying routes.

For **concept-level copying**, a small set of heads carries substantial causal effect, with the strongest reaching scores of up to 0.18 in Llama-8B, OLMo-1B, and OLMo-7B; evidence of a localizable concept-copying route. Ablation confirms this: removing the identified heads causes large, consistent accuracy drops (Table 6). OLMo-1B and OLMo-7B both fall to 0.16 after ablating only 11 heads; Qwen-7B drops to 0.30 after 29. A random ablation baseline, removing the same number of heads at random, has almost no effect ($\Delta\text{Acc}_{\text{rand}} = -0.004$ to -0.07), onfirming the drop is specific to the identified heads. Llama-3.1-8B is the exception: its targeted drop matches its random baseline, consistent with its flat causal scores and a more distributed concept route.

For **token-level copying**, causal scores are much smaller (maxima 0.00 to -0.06) with no distinct set of high-scoring heads, suggesting the mechanism is distributed across many heads. Ablation matches this picture: accuracy is robust, requiring many heads to be ablated before it moves at all (e.g. OLMo-2-1B 0.80→0.44 only after 72 heads; OLMo-13B is almost unaffected after 61 heads). Interestingly, for Qwen-7B and Llama-8B, ablating the identified token-copying heads does not lower accuracy but *raises* it ($+0.44$ and $+0.40$). *This shows the competition between the two routes during pronoun resolution under discourse complexity.*

Takeaway. Pronoun fidelity is solved via two competing copying routes that are localized differently. Concept copying is localized in a few heads, and token copying is more distributed across many low-effect heads. It seems that the two routes compete, as ablating token copying heads can flip a model from recency-driven bias to correct resolution. Which route dominates is model-dependent.

7 Related Works

Pronoun Fidelity Faithful pronoun reuse was first studied in Ovalle et al. (2023) and Hossain et al. (2023), across a range of pronouns but in settings with limited discourse complexity. Gautam et al. (2024) added discourse complexity to the task, using sentences about additional referents in their RUFF dataset, which we use. These datasets have been used in a comparison of probability- and generation-based evaluations in (Subramonian et al., 2025), as well as in an evaluation of reasoning models in Gautam (2026) (which we do not consider). More recently, Kotek et al. (2026) introduce ProText, which contains natural narratives about people, but still focuses on single rather than multiple discourse entities. All these works take a behavioural perspective on pronoun fidelity.

Task	Model	Heads	$\mathbf{Acc_{base}}$	$\mathbf{Acc_{abl}}$	$\mathbf{\Delta Acc}$	$\mathbf{\Delta Acc_{rand}}$	$\mathbf{LD_{base}}$	$\mathbf{LD_{abl}}$
concept-level	Llama-8B	11	1.00	0.96	−0.04	−0.04	2.10	1.38
	OLMo-1B	11	0.94	0.16	−0.78	−0.07	1.22	0.88
	OLMo-7B	11	0.80	0.16	−0.64	0.00	2.44	0.43
	OLMo-13B	16	0.98	0.70	−0.28	0.00	2.25	0.83
	Gemma-9B	23	0.98	0.86	−0.12	−0.02	1.12	0.68
	Qwen-7B	29	0.88	0.30	−0.58	−0.07	1.10	0.21
token-level	Llama-8B	21	0.48	0.76	+0.40	−0.04	0.29	0.37
	OLMo-1B	72	0.80	0.44	−0.36	−0.22	0.66	−0.90
	OLMo-7B	41	0.34	0.08	−0.36	−0.01	1.29	0.51
	OLMo-13B	61	0.68	0.64	−0.04	0.00	0.79	0.73
	Gemma-9B	49	0.80	0.64	−0.16	−0.03	0.64	0.261
	Qwen-7B	1	0.46	0.9	+0.44	−0.05	0.178	0.178

Table 6: Maximum accuracy drop under cumulative head ablation for concept-level and token-level copying. Heads = number of ablated heads at the point of maximum drop. Acc_{base}/Acc_{abl} denote base/ablated accuracy and LD_{base}/LD_{abl} denote base/ablated logit difference between the correct and distractor pronoun.

Mechanistic Interpretability Mechanistic interpretability is a framework for discovering causal mechanisms that explain how inputs are transformed into outputs via intermediate representations (Mueller, 2024; Saphra and Wiegreffe, 2024). This is complicated by *superposition* in neural networks, i.e., concepts are distributed across neurons, and a single neuron might encode multiple concepts (Elhage et al., 2022). Distributed Alignment Search (DAS; Wu et al., 2023; Geiger et al., 2024) addresses this by searching for linearly encoded causal subspaces in the residual stream, grounded in causal abstraction theory (Geiger et al., 2021, 2025). Supervised DAS is the strongest method for causal variable localization according to a benchmark of popular mechanistic interpretability methods (Mueller et al., 2025), motivating our use of this method. DAS has also been applied to interpret how LLMs process filler-gap dependencies in English (Boguraev et al., 2025).

Interpreting Pronoun Fidelity Although no paper directly tackles pronoun fidelity, prior work on stereotypical biases and entity tracking are relevant.

Vig et al. (2020) introduce Causal Mediation Analysis (also known as activation patching) isolate small sets of neurons and attention heads responsible for gender bias. Chintam et al. (2023) extend this to localize bias-related components in GPT-2. Bashir et al. use Edge Attribution Patching to discover bias-related circuits, but find them highly unstable, and Hanna et al. (2024) provide evidence for fundamental limitations of circuit discovery methods. These works suggest that stereotypical associations are not cleanly localized and affect other tasks, motivating our study of their interaction with other mechanisms.

Entity tracking, a prerequisite to long-context understanding and coherent generation (Kim and Schuster, 2023), is also implicated in pronoun fidelity. Prakash et al. (2024) use activation patching and circuit discovery to show how collaborating attention heads track entities via positions. However, Gur-Arieh et al. (2026) find that a purely positional mechanism generalizes poorly with more bound entities. They propose a causal model combining multiple mechanisms, motivating our approach.

8 Conclusion

We present the first mechanistic study of pronoun fidelity, asking whether failure to track pronouns faithfully is driven by reasoning, repetition, or stereotypical bias. We find that it is driven by a combination of all three. Boundless DAS recovers group binding $\mathcal{G}$, recency $\mathcal{R}$, and stereotype $\mathcal{S}$ as three separately localizable causal subspaces, and a mixture of them predicts 91-99% of model behaviour. The contributions are uneven, as most models lean on recency and stereotype bias, while group binding contributes little and occasionally adds noise. At the attention-head level, pronoun information is carried by two competing copying routes. One retrieves a semantic unit through a few localized heads, the other repeats surface form through many distributed ones. Which route is taken to solve pronoun fidelity is model-dependent. To conclude, pronoun fidelity failure is not a single-mechanism problem. It arises from competition among mechanisms that are all active at once, with the outcome determined by which causal subspace wins.

Limitations

Inherited limitations of the RUFF dataset. RUFF is a synthetically-created template-based dataset, which sacrifices ecological validity for tight control over the entities and pronouns involved in each data instance. This control makes the dataset natural for use with Distributed Alignment Search, which requires tightly-paired variants to isolate mechanisms, but it may not be representative of how reference "in the wild" works, cf. OntoNotes (Weischedel et al., 2013) or the GUM corpus (Zeldes, 2017). Additionally, RUFF is an English-only dataset. Morphologically richer languages (e.g., languages that mark gender on other parts of speech, or that mark more cases, etc.), might have entirely different mechanisms.

Limitations of our use of RUFF. Although RUFF contains data instances with 0-5 distractor sentences, we only include the 0 and 1 distractor settings, as these represent the most dramatic performance gap. Additional distractor sentences may modulate the effects of the different mechanisms. In our restricted setting, we also only experiment with 3 English pronouns (he, she, they) that are tokenized singly. Unlike Gautam et al. (2024), we do not consider neopronouns as they are usually tokenized into multiple tokens (Ovalle et al., 2024), and therefore cannot say whether the mechanisms for neopronoun fidelity are the same.

Restrictions from Distributed Alignment Search. As we use the three available singular pronouns in English that can be singly tokenized, it is not possible to have a single diagnostic dataset for DAS where all mechanisms can be represented, as one pronoun always represents the stereotype mechanism, one represents the group binding mechanism, and the other the repetition mechanism. Additionally, it is possible that there is another single mechanism beyond the ones we cover that would explain model pronoun fidelity as well as our combined mixed-mechanism explanation, although it is likely to be less interpretable.

Additional MI methods and models. Our research questions can be answered with many other approaches within mechanistic interpretability, notably circuit discovery. However, there is evidence that large neural language models may often implement the same task in multiple, distinct circuits (Chen et al., 2026), making them less robust as explanations of model mechanisms than the learned rotation matrices of boundless DAS. In general, we take the position that different methodological approaches shed light on different aspects of model interpretability, which is why we also experiment with activation patching in our attention head analysis. Our findings apply to the class of models we study, which is instruction-tuned decoder-only language models, but our work could be expanded to other models, e.g., encoder-only models, which show very different patterns on the RUFF dataset, and reasoning models, which make different patterns of mistakes (Gautam, 2026) and may therefore implement different mechanisms for pronominal reasoning.

Ethics Statement

Although we limit our study to the three most widely-used third-person singular pronouns in English that are therefore singly tokenized (*he*, *she*, and *they*), we note that other neopronouns are also in use in English. As with behavioural studies of model fairness, mechanistic studies like ours also come with the potential risk of misuse to perpetuate stereotypical narratives—in our case, pertaining to gender. Additionally, applications involving tracking person entities through reference in natural language can be used for surveillance, which we stand against. We use all models and the RUFF dataset in accordance with their licenses and terms of use.

References

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. 2019. Analyzing and interpreting neural networks for nlp: A report on the first blackboxnlp workshop. *Natural Language Engineering*, 25(4):543–557.

Zubair Bashir, Bhavik Chandna, and Procheta Sen. Dissecting bias in llms: A mechanistic interpretability perspective. *Transactions on Machine Learning Research*.

Sasha Boguraev, Christopher Potts, and Kyle Mahowald. 2025. Causal interventions reveal shared structure across english filler–gap constructions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25032–25053.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xi Chen, Mingyu Jin, Jingcheng Niu, Yutong Yin, Jinman Zhao, Bangwei Guo, Dimitris N. Metaxas, Zhao-

ran Wang, Yutao Yue, and Gerald Penn. 2026. All circuits lead to rome: Rethinking functional anisotropy in circuit and sheaf discovery for llms. *Preprint*, arXiv:2605.12671.

Abhijith Chintam, Rahel Beloch, Willem Zuidema, Michael Hanna, and Oskar van der Wal. 2023. Identifying and adapting transformer-components responsible for gender bias in an English language model. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 379–394, Singapore. Association for Computational Linguistics.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. https://transformer-circuits.pub/2022/toy_model/toy_model.pdf.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. *CoRR*, abs/2405.00208.

Sheridan Feucht, Eric Todd, Byron C Wallace, and David Bau. The dual-route model of induction. *Second Conference on Language Modeling*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Vagrant Gautam. 2026. Training in step-by-step formal reasoning improves pronominal reasoning in language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 121–135, Rabat, Morocco. Association for Computational Linguistics.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024. Robust pronoun fidelity with English LLMs: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and 1 others. 2025. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *Neural Information Processing Systems*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. 2026. Mixing mechanisms: How language models retrieve bound entities in-context. In *The Fourteenth International Conference on Learning Representations*.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of large language models in understanding pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.

Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.

Hadas Kotek, Margit Bowler, Patrick Sonnenberg, and Yu'an Yang. 2026. Protext: A benchmark dataset for measuring (mis)gendering in long-form texts. *Preprint*, arXiv:2603.27838.

Aaron Mueller. 2024. Missed causes and ambiguous effects: Counterfactuals pose challenges for interpreting neural networks. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fiotto-Kaufman, Tal Haklay, Michael Hanna, and 1 others. 2025. Mib: A mechanistic interpretability benchmark. *arXiv preprint arXiv:2504.13151*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, and 1 others. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1246–1266, New York, NY, USA. Association for Computing Machinery.

Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter, and Rahul Gupta. 2024. Tokenization matters: Navigating data-scarce tokenization for gender inclusive language technologies. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1739–1756, Mexico City, Mexico. Association for Computational Linguistics.

Theodora-Mara Pîslar, Sara Magliacane, and Atticus Geiger. 2025. Combining causal models for more accurate abstractions of neural networks. In *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pages 114–138. PMLR.

Nikhil Prakash, Tamar Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. *International Conference on Learning Representations*, 2024:8057–8082.

Naomi Saphra and Sarah Wiegreffe. 2024. Mechanistic? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 480–498.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, and 10 others. 2025. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*. Survey Certification.

Arjun Subramonian, Vagrant Gautam, Preethi Seshadri, Dietrich Klakow, Kai-Wei Chang, and Yizhou Sun. 2025. Agree to disagree? a meta-evaluation of LLM misgendering. In *Second Conference on Language Modeling*.

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, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0.

Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah Goodman, Christopher Manning, and Christopher Potts. 2024. pyvene: A library for understanding and improving PyTorch models via interventions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 158–165, Mexico City, Mexico. Association for Computational Linguistics.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. 2023. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in neural information processing systems*, 36:78205–78226.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Amir Zeldes. 2017. The gum corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

A GPU hours

We ran approximately 150 GPU hours of experiments on NVIDIA A100 GPUs.

B Prompt design and Response Scoring

We frame the inputs as multiple-choice question to constrain the output space and thus make the counterfactual targets unambiguous.

> **Instruction:** Please select the correct pronoun from the options below.
> **Input:** The **technician** thought that **he** had been having a bad day. The **technician** said that ___ had completed the repair.
> **Question:** What pronoun should be used to fill the blank?
> **Options:** he, she, they (randomized)
> **Expected (Base):** he
> **Expected ($\mathcal{G}$ Intervention):** she

The *response token* is the first token the model generates after `Response:`, i.e., the token at the final position of the prompt prefix. Log-likelihoods are computed over the three candidate nominative pronouns *he*, *she*, and *they*, each tokenized as a single token, and the candidate with the highest log-likelihood is taken as the prediction. All DAS interventions and mixture model distribution measurements are evaluated at this same position.

C Causal Models

O ranges over the 60 occupations in RUFF (e.g., *technician*, *nurse*); P and P' take values in {*he*, *she*, *they*}; S is the stereotypical pronoun associated with O under the model's priors (all occupational biases, prompted with the task sentences of RUFF, are also listed in Table 9), also in {*he*, *she*, *they*}; E is a latent entity representation (not directly observed); and $Y \in \{he, she, they\}$ is the predicted output pronoun.

D DAS Training Details

Each DAS model learns an orthogonal rotation matrix $R \in \mathbb{R}^{d \times d}$ and soft boundaries over the residual stream at the identified best layer. Each model is trained on all datasets (`GR`, `GS`, `RS`) seperately, producing nine trained rotation matrices per model after full training. We jointly optimize R and the boundary params using Adam with learning rates 10^{-3} and 10^{-2} respectively, with linear warmup over the first 10% of steps, followed by linear decay. The boundary temperature, β is annealed from 50 to 0.1 to encourage the boundaries to converge to a discrete subspace (Wu et al., 2023). Models are trained for three epochs with an effective batch of 64 (`batch_size=4`, `gradient_accumulation_steps=16`). We use an 80/20 train/eval split. The training objective is cross-entropy between the logits after intervention and the counterfactual label for the high-level mechanisms after intervention. We evaluate at the response token (see Appendix B). We implement the experiments using the pyvene library (Wu et al., 2024).

Model	Rotation	GR	GS	RS
Llama-8B	$\mathcal{G}$ (L8)	0.646	0.550	**0.551**
	$\mathcal{R}$ (L8)	0.315	**0.683**	0.489
	$\mathcal{S}$ (L8)	**0.595**	0.593	0.602
OLMo-1B	$\mathcal{G}$ (L12)	0.328	0.466	**0.363**
	$\mathcal{R}$ (L2)	0.125	**0.887**	0.828
	$\mathcal{S}$ (L6)	**0.397**	0.627	0.500
OLMo-7B	$\mathcal{G}$ (L8)	0.625	0.346	**0.302**
	$\mathcal{R}$ (L0)	0.567	**0.744**	0.529
	$\mathcal{S}$ (L6)	**0.682**	0.339	0.332
OLMo-13B	$\mathcal{G}$ (ℓ10)	0.799	0.541	**0.562**
	$\mathcal{R}$ (ℓ0)	0.583	**0.728**	0.748
	$\mathcal{S}$ (ℓ10)	**0.822**	0.536	0.605
Gemma-9B	$\mathcal{G}$ (ℓ10)	0.442	0.255	**0.512**
	$\mathcal{R}$ (ℓ10)	0.000	**0.993**	0.116
	$\mathcal{S}$ (ℓ10)	**0.701**	0.140	0.698
Qwen-7B	$\mathcal{G}$ (ℓ14)	0.938	0.317	**0.340**
	$\mathcal{R}$ (ℓ14)	0.528	**0.895**	0.780
	$\mathcal{S}$ (ℓ7)	**0.811**	0.389	0.372

Table 7: Cross-mechanism specificity across models. Within each model block, rows index the trained rotation and columns index the diagnostic dataset, each isolating one mechanism. Bold marks the diagonal.

E Validation

Cross-Mechanism Transfer To test specificity, we use the structure of the diagnostic datasets (see Section 3.3), as each isolates one meechanism clearly. We take three rotations R_G, R_R, R_S trained on the `GR dataset`, freeze them, and evaluate each rotation on every dataset. We get a 3×3 matrix per model (rotation $\times$ evaluation dataset).

Each rotation is evaluated in two situations. First, we apply is to the dataset that isolates its own mechanism: R_G on the `RS dataset`, R_R on the `GS dataset`, R_S on the `GR datset`. Here, the source matches what the rotation was trained on. The intervention should flip the prediction to the

counterfactual label (standard IIA). Second, the rotation is applied to a dataset that isolates a *different* mechanism, e.g. R_S, stereotype rotation, to the `GS dataset`, where recency is isolated. Here, a rotation should not successfully flip the label.

Results. Table 7 details the cross-mechanism specificty across models. In Llama-8B, $\mathcal{R}$ is the most specific rotation, peaking clearly on `GS` (0.683), while $\mathcal{G}$ and $\mathcal{S}$ show similar scores across datasets. In OLMo-1B, $\mathcal{R}$ is highly specific, dominating on `GS` (0.887), whereas $\mathcal{G}$ and $\mathcal{S}$ are less separable. OLMo-7B shows clean separation for $\mathcal{R}$ (`GS`: 0.744) and $\mathcal{S}$ (`GR`: 0.682), with $\mathcal{G}$ less distinct. OLMo-13B follows the same pattern: $\mathcal{R}$ peaks on `GS` (0.728) and $\mathcal{S}$ on `GR` (0.822), while $\mathcal{G}$ and $\mathcal{S}$ overlap. Gemma-9B shows the sharpest $\mathcal{R}$ specificity of all models (`GS`: 0.993), with near-zero transfer to other datasets. In Qwen-7B, all three rotations are relatively specific, with $\mathcal{G}$ peaking on `GR` (0.938), $\mathcal{R}$ on `GS` (0.895), and $\mathcal{S}$ on `GR` (0.811), making it the model with the clearest three-way separation. Taken together, across all models, $\mathcal{R}$ is consistently the most specific rotation, while $\mathcal{G}$ and $\mathcal{S}$ show more overlap.

F Base Performances across Models

	Overall		By Pronoun			By Case		
Model	**0-dist**	**1-dist**	**he**	**she**	**they**	**NOM**	**ACC**	**POSS**
Llama-3.1-8B	83.2	51.2	88.6 / 33.3	61.1 / 24.4	100.0 / 97.0	66.7 / 47.9	95.9 / 56.3	88.7 / 30.0
Llama-3.1-70B	90.4	48.8	91.0 / 51.2	90.4 / 33.3	89.8 / 62.2	95.0 / 67.9	84.1 / 31.3	92.0 / 30.0
OLMo-2-1B	55.4	42.8	78.4 / 62.5	72.5 / 61.9	15.1 / 3.1	45.6 / 40.8	55.3 / 42.5	67.3 / 70.0
OLMo-2-7B	69.6	46.0	98.2 / 98.8	79.6 / 33.3	30.7 / 4.9	70.0 / 49.6	72.4 / 42.1	66.0 / 50.0
OLMo-2-13B	71.4	63.6	83.2 / 82.7	81.4 / 79.8	49.4 / 27.4	83.9 / 76.3	62.9 / 53.3	66.0 / 35.0
OLMo-2-32B	85.4	51.6	83.2 / 41.7	92.2 / 45.8	80.7 / 67.7	76.7 / 64.2	91.8 / 38.8	88.7 / 55.0
Gemma-2-9B	93.6	80.4	95.8 / 88.7	90.4 / 71.4	94.6 / 81.1	97.2 / 95.8	85.9 / 69.2	98.0 / 30.0
Gemma-2-27B	94.2	74.2	98.2 / 78.6	99.4 / 86.3	84.9 / 57.3	99.4 / 100.0	83.5 / 48.8	100.0 / 70.0
Qwen2.5-7B	52.4	43.2	90.4 / 90.5	53.3 / 34.5	13.3 / 3.7	49.4 / 40.4	62.4 / 48.8	44.7 / 10.0
Qwen2.5-72B	93.2	82.0	96.4 / 78.0	97.6 / 84.5	85.5 / 83.5	97.8 / 100.0	92.9 / 69.6	88.0 / 15.0

Table 8: Base pronoun resolution accuracy (%) per model. Pronoun and case columns show 0-dist / 1-dist.

G Layer Search Results

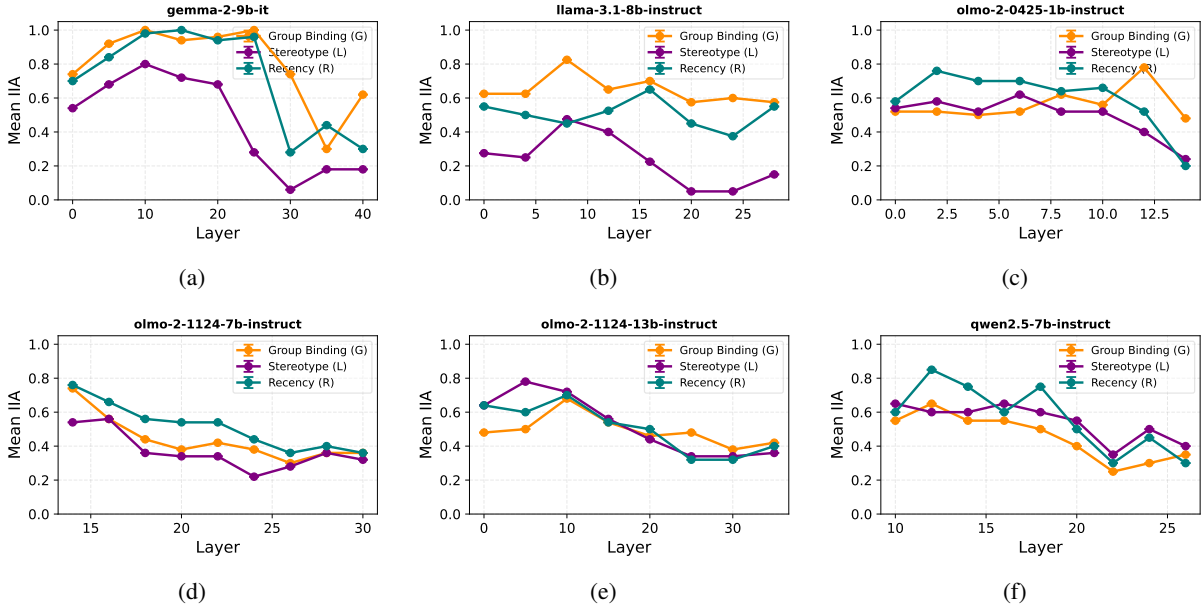

Figure 2: Layer search comparison.

H Attention Head Experiments

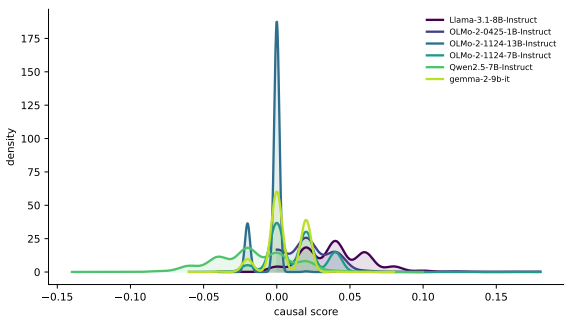

Figure 3: Causal scores for concept-level interventions.

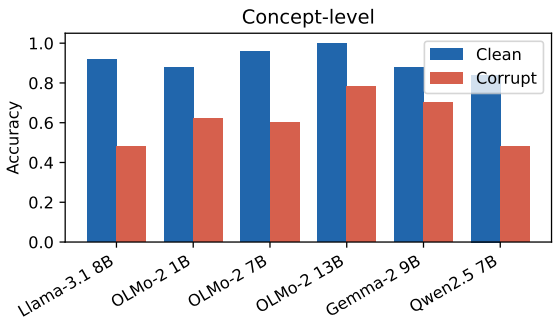

Figure 4: Activation patching results at the concept level.

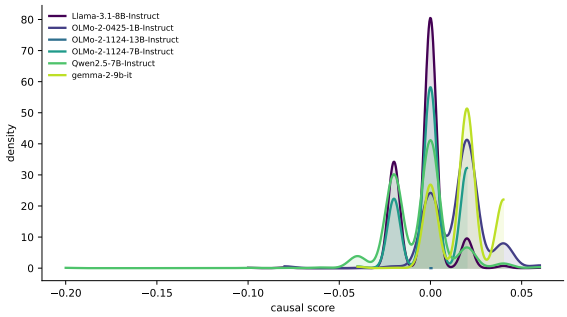

Figure 5: Causal scores for token-level interventions.

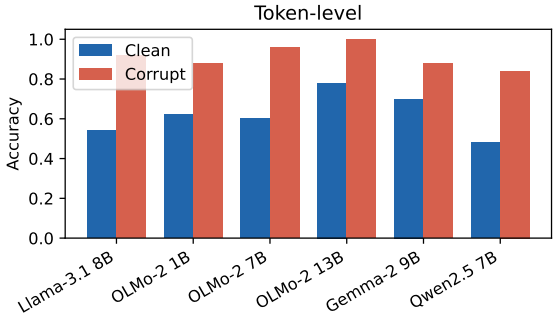

Figure 6: Activation patching results at the token level.

I Context-free Bias

Occ.	Llama-8B	-70B	Olmo-1B	-7B	-13B	-32B	Qwen-7B	-70B	Gemma-9B	-27B
accountant	he	they	he	she	he	he	he	he	he	he
administrator	he	they	he	she	he	he	he	he	he	he
advisor	he	they	he	she	he	he	he	he	he	he
appraiser	he	they	he	she	he	he	he	he	he	he
architect	he	they	he	she	he	he	he	he	he	he
auditor	he	they	he	she	he	he	he	he	he	he
baker	he	they	he	she	he	he	he	he	he	he
bartender	he	they	he	she	he	he	he	he	he	he
broker	they	they	he	she	he	he	he	he	he	he
carpenter	he	they	he	he	he	he	he	he	he	he
cashier	she	she	he	she	they	he	he	she	he	he
chef	he	they	he	she	he	he	he	he	he	he
chemist	he	they	he	she	he	he	he	he	he	he
clerk	she	they	he	she	he	he	he	he	he	he
counselor	she	they	he	she	they	he	he	he	he	she
dietitian	she	she	he	she	she	she	he	she	she	she
dispatcher	she	they	he	she	he	he	he	he	he	he
doctor	he	they	he	she	he	he	he	he	he	he
educator	she	they	he	she	they	he	he	he	he	she
electrician	he	they	he	he	he	he	he	he	he	he
engineer	he	they	he	she	he	he	he	he	he	he
examiner	he	they	he	she	they	he	he	he	he	he
firefighter	he	they	he	she	he	he	he	he	he	he
hairdresser	she	she	he	she	they	he	he	he	she	she
hygienist	she	she	he	she	she	she	he	she	she	she
inspector	he	they	he	she	he	he	he	he	he	he
instructor	she	they	he	she	he	he	he	he	he	he
investigator	he	they	he	she	he	he	he	he	he	he
janitor	he	they	he	she	he	he	he	he	he	he
lawyer	he	they	he	she	he	he	he	he	he	he
librarian	she	she	he	she	they	he	he	she	she	she
machinist	he	they	he	she	he	he	he	he	he	he
manager	they	they	he	she	he	he	he	he	he	he
mechanic	he	they	he	he	he	he	he	he	he	he
nurse	she	she	he	she	she	she	she	she	she	she
nutritionist	she	she	he	she	they	he	he	she	she	she
officer	he	they	he	she	he	he	he	he	he	he
painter	he	they	he	she	he	he	he	he	he	he
paralegal	she	she	he	she	they	he	he	he	she	she
paramedic	they	they	he	he	he	he	he	he	he	he
pathologist	he	they	he	she	he	he	he	he	he	he
pharmacist	he	they	he	she	he	he	he	he	he	he
physician	he	they	he	she	he	he	he	he	he	he
planner	she	they	he	she	they	he	he	he	he	he
plumber	he	they	he	he	he	he	he	he	he	he
practitioner	they	they	he	she	they	he	he	he	he	he
programmer	he	they	he	she	he	he	he	he	he	he
psychologist	she	they	he	she	he	he	he	he	he	he
receptionist	she	she	he	she	she	she	he	she	she	she
salesperson	they	they	he	she	they	he	he	he	he	he
scientist	he	they	he	she	he	he	he	he	he	he
secretary	she	she	he	she	he	he	he	she	she	she
specialist	he	they	he	she	he	he	he	he	he	he
supervisor	he	they	he	she	he	he	he	he	he	he
surgeon	he	they	he	she	he	he	he	he	he	he
teacher	she	she	he	she	they	he	he	she	she	he
technician	he	they	he	she	he	he	he	he	he	he
therapist	she	she	he	she	they	he	he	she	she	she
veterinarian	she	she	he	she	he	he	he	he	he	he
worker	they	they	he	she	he	he	he	he	he	he

Table 9: Average over three runs: Most likely pronoun predictions across all models for occupations in the RUFF dataset. These are used to calculate $IIA_{\texttt{stereotype}}$ and to train $\mathcal{S}$ for each language model.