ORDINAL AXES IN THE GEOMETRY OF ORDERED CATEGORICAL KNOWLEDGE IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Nothing in the word viscount says that it outranks baron, yet language models answer most questions about such ordered categories correctly, even though their order is a convention that cannot be read off the words. Current accounts of concept geometry describe continuous, cyclic and unordered categorical concepts, but not how such orders are represented. We find that ordered categories are organised along an ordinal axis in the residual stream. The axis can be recovered from the activations of the two endpoint categories alone, and it places the held-out intermediate categories in their correct order across several language models and many domains, from ranks and titles to physical scales. We then show that a model uses this axis when it compares two categories. Exchanging their coordinates along this single direction, and nothing else, flips about a third of the decisions it gets right, whereas random directions of the same size flip far fewer. Finally, we show why a trained causal direction outperforms the ordinal axis. A one-dimensional direction trained to flip the comparison combines the ordinal representation with the gradient of the answer logits. Adding this gradient to the ordinal axis recovers much of the trained direction's effect. These results identify a geometric representation of ordered categorical knowledge, read and used by the model, and add ordered categories to the account of concept geometry.

1 INTRODUCTION

Language models do more than memorize associations from their training data, they can generalize and apply learned knowledge across contexts. They can, for example, answer questions about which educational level is higher, which boxing weight class is heavier, or which planet lies farther out. For some of these concepts, the answers are given away by the words themselves, but many are not. Nothing in the word "brown" says that it outranks "green" in martial arts, just as nothing in the word "viscount" says that it outranks a "baron". Yet language models accurately recover these orderings across different contexts despite learning primarily from next-token predictions. We propose that such learned orderings are therefore encoded in the model's internal representations.

The linear representation hypothesis provides a natural framework for studying this structure (Park et al., 2023). According to it, higher-level concepts correspond to different directions in activation space. This perspective motivates studying the geometry of the residual stream across layers to understand how such learned orderings are represented. Over the last several years, a partial map has emerged of how concepts are organized in activation space. Continuous quantities such as space and time are linearly decodable (Gurnee & Tegmark, 2024). Cyclic concepts such as days and months are represented as circles under PCA projections (Engels et al., 2025). Unordered categorical concepts form simplices, with hierarchy appearing as orthogonality (Park et al., 2025). Recent work has also shown that linearly represented numerical quantities support comparison behavior (El-Shangiti et al., 2025). Taken together, these results characterize several important forms of conceptual geometry, but they do not address discrete concepts whose values are ordered, particularly when that ordering is conventional rather than inherent to the concepts themselves.

In this paper, we bridge this gap by investigating how models represent and use conventionally ordered categorical knowledge. Thus, our contributions are summarized as follows:

- **A ladder axis estimated from endpoints alone.** We identify an *ordinal axis* for ordered categorical knowledge. For a concept with rungs $r_1 \prec \cdots \prec r_n$, we take $\hat{v} \propto c_n - c_1$,

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.

using only the activations of the two endpoint categories. With no training and with all intermediate categories held out, we find that these middle rungs project onto $\hat{v}$ in the correct order across several language models and domains.

- **Evidence of causal use through coordinate swapping.** We show that models actively use this axis when making comparisons. Exchanging only the two options' scalar coordinates along $\hat{v}$, shifts forced-choice ordering judgments toward the swapped order.
- **Alignment with optimized causal directions.** We clarify why a trained one-dimensional direction controls the comparison more strongly than the purely geometric axis. An optimized distributed alignment search (DAS) direction decomposes into a component aligned with the ordinal axis, and an orthogonal component aligned with the gradient of the answer logits. We find the endpoint direction is strongly aligned with this trained direction, reaching a median cosine similarity of 0.54 to 0.62. Adding the answer gradient to the ordinal axis recovers much of the trained direction's causal effect without any training.

2 Related work

Research on representation geometry has shown that models encode different semantic and conceptual properties in structured regions of activation space.

Geometry of Language Model Representations A growing body of work has studied how conceptual structure is reflected in the geometry of learned representations. Early word-embedding methods showed that semantic relations can correspond to directions and vector differences in embedding spaces (Mikolov et al., 2013; Pennington et al., 2014). As language models became more capable, related linear structure was found for properties such as space and time (Gurnee & Tegmark, 2024), numerical attributes (Heinzerling & Inui, 2024), Othello board states (Nanda et al., 2023; Li et al., 2022), truth values (Marks & Tegmark, 2023), refusal (Arditi et al., 2024), and sentiment (Tigges et al., 2023). Park et al. (2023) formalised this line of work in the Linear Representation Hypothesis (LRH), which relates high-level concepts to directions in activation space. Subsequent work has shown that different concepts can exhibit different geometries, Engels et al. (2025) identify circular representations for cyclic concepts such as days and months, while Park et al. (2025) characterise categorical and hierarchical concepts using higher-dimensional geometric structures. Recent work on numerical representations has also recovered one-dimensional, monotonic structure for numerical magnitude, although with non-uniform spacing (AlquBoj et al., 2025). Similarly, low-dimensional representations of numerical properties have been shown to support comparison behavior (El-Shangiti et al., 2025). Taken together, these studies show that the geometry and behavioral use of a representation depend on the structure of the underlying concept, while leaving open how ordered categorical concepts are represented.

Causal Interventions To probe whether internal representations are causally involved in model behaviour, several lines of work intervene on activations during the forward pass. Early causal mediation approaches intervened on internal model states and measured the effect on downstream behaviour (Vig et al., 2020; Geiger et al., 2020). This line of work developed into activation patching, in which activations at selected model components or layers are replaced with corresponding activations from a clean or counterfactual run to identify representations that causally mediate a behaviour (Meng et al., 2022; Zhang & Nanda, 2024; Conmy et al., 2023). Related work has extended such interventions to learned features and subspaces, allowing causal effects to be localised to particular directions or groups of directions in activation space (Makelov et al., 2024). Distributed alignment search (DAS) goes one step further and learns such a subspace by gradient descent, so that interchanging activations along it reproduces the effect of changing a variable of interest (Geiger et al., 2024). A one-dimensional DAS direction trained to control the comparison is the natural reference for the axis we build without training, and we use it as such in Section 4.3. Our intervention is closest to activation and subspace patching, but operates at a finer granularity: rather than replacing an entire activation, we exchange only the scalar projections of two activations onto a recovered ordinal direction while leaving their orthogonal components unchanged. Unlike additive steering methods, the value written is taken from another activation produced by the model rather than chosen or amplified externally.

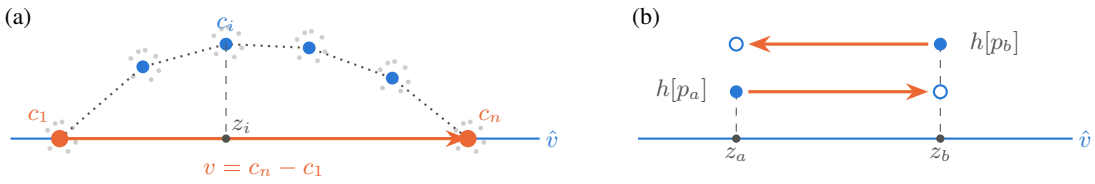


Figure 1: The ordinal axis and the intervention. (a) Each rung is read at its own token over several prompts, and the mean of those activations is its centroid. The direction is the difference between the two endpoint centroids and a rung's coordinate is where its centroid projects onto that direction. (b) In the intervention, the two options are moved parallel to the direction until they have exchanged coordinates. Everything orthogonal to the direction is left as it was, so each option keeps its distance from the axis and only one number changes.

3 METHODOLOGY

Large language models are typically implemented as decoder-only Transformers, which transform token representations through a sequence of attention and feed-forward blocks. Let $h^{(\ell)}[p] \in \mathbb{R}^D$ be the residual stream activation at token position p and layer ℓ. Every block adds to it, so

$$h^{(\ell)} = h^{(0)} + \sum_{k<\ell} \big(\text{the attention and feed-forward contributions at block } k\big). \tag{1}$$

In this paper, we study this geometry with two complementary techniques. First, we find the ordinal axis from the residual stream at the two endpoints of a ladder. Second, we test whether the model uses this axis by intervening on the corresponding coordinates of two rungs and measuring the effect on the model's output.

3.1 ORDERED LADDERS AND PROMPTS

An ordered category is a set of names with a fixed order, such as the titles of nobility or the ranks of an army. We call such a set a *ladder*, $r_1 \prec r_2 \prec \cdots \prec r_n$, and its elements *rungs*. We use 26 ladders with 6 to 20 rungs (all ladders and their rungs are listed in SM A.1).

To explore the ordinal axis, we use forced-choice questions between two rungs of a ladder, which we refer to as comparison prompts. Each prompt sets a question whose answer is given by the order of the ladder, the two rungs as numbered options, and an instruction to answer with the number of the option. We read the residual stream activation at the rung's last token position and the output logits at the end of the prompt:

```
Which noble title is higher in precedence?
1.  viscount                          ⟵ h^(ℓ)[p] is read here
2.  baron
Answer with the slot number only.
Answer:                               ⟵ the answer is read here
```

The first is used to build the ordinal axis and the second to test the causality of the intervention. The example is Form A, the first of three forms of the prompt, which differ in layout and in the wording of the instruction. The others are described in SM A.1.

3.2 THE ORDINAL AXIS

Let us define $x_{i,t}^{(\ell)}$ as the residual stream activation of rung i measured in layer ℓ for the comparison prompt template t. We may decompose the activation as the generic template effect, θ_t, plus the rung's own meaning, μ_i, and a noise term:

$$x_{i,t}^{(\ell)} = \theta_t^{(\ell)} + \mu_i^{(\ell)} + \varepsilon_{i,t}^{(\ell)} \tag{2}$$

If $X_i^{(\ell)}$ is the set of activations of rung i in layer ℓ when it is the option in slot 1 of a comparison prompt, over the three surface forms and every partner rung (SM A.1), we obtain the centroid of the rung in layer ℓ as:

$$c_i^{(\ell)} = \frac{1}{|X_i^{(\ell)}|} \sum_{x \in X_i^{(\ell)}} x = \bar{\theta}^{(\ell)} + \mu_i^{(\ell)} + \bar{\varepsilon}_i^{(\ell)} \tag{3}$$

and therefore, at the endpoints,

$$v^{(\ell)} = c_n^{(\ell)} - c_1^{(\ell)} = \left(\mu_n^{(\ell)} - \mu_1^{(\ell)}\right) + \left(\bar{\varepsilon}_n^{(\ell)} - \bar{\varepsilon}_1^{(\ell)}\right) \tag{4}$$

Note that the template term $\bar{\theta}^{(\ell)}$ cancels out because the rungs were averaged over the same template set. Taking slot 1 only means that the activation has not yet seen the partner, so nothing about the comparison enters the centroids (SM A.3). We then define

$$\hat{v}^{(\ell)} = \frac{v^{(\ell)}}{\|v^{(\ell)}\|}. \tag{5}$$

as the *ordinal axis* of the ladder. Only the two endpoint rungs enter its construction. The $n-2$ middle rungs are held out, so that nothing in the definition forces their projected coordinates

$$z_i^{(\ell)} = \langle c_i^{(\ell)}, \hat{v}^{(\ell)} \rangle \tag{6}$$

to fall in the right order, or even between those of the endpoints (Figure 1a). To test whether these projections follow the true order of the ladder, we use Kendall's τ. The direction is estimated separately at every layer, and we drop the layer index when it is clear from the context.

3.3 Coordinate swapping

The existence of the ordinal axis does not imply that the model uses it for doing comparisons. To find out, using the same comparison prompts, we test whether an intervention that exchanges the coordinate projections of the rungs on $\hat{v}^{(\ell)}$ modifies the model's answer.

Let p_a and p_b be the token positions of the two options, and call *mirror* the prompt that asks the same question with the two options in the other order, so that in the mirror position p_a holds rung b and position p_b holds rung a. At the chosen layer, $z_a^{(\ell)} = \langle h^{(\ell)}[p_a], \hat{v}^{(\ell)} \rangle$ and $z_b^{(\ell)} = \langle h^{(\ell)}[p_b], \hat{v}^{(\ell)} \rangle$ are the coordinates of the two options, and $\tilde{z}_a^{(\ell)}$ and $\tilde{z}_b^{(\ell)}$ the coordinates at the same two positions of the mirror. The intervention is

$$h^{(\ell)}[p_a] \leftarrow h^{(\ell)}[p_a] + \left(\tilde{z}_a^{(\ell)} - z_a^{(\ell)}\right) \hat{v}^{(\ell)}, \qquad h^{(\ell)}[p_b] \leftarrow h^{(\ell)}[p_b] + \left(\tilde{z}_b^{(\ell)} - z_b^{(\ell)}\right) \hat{v}^{(\ell)}, \tag{7}$$

applied at the input of one layer, after which the forward pass continues unchanged. Effectively, each option takes the coordinate that the other rung has at the same position, while every component orthogonal to $\hat{v}^{(\ell)}$ keeps its value, which makes it an interchange intervention (Geiger et al., 2020) restricted to a single direction. The same axis is used at both options, since the directions built from slot 1 and from slot 2 are nearly parallel, with a median cosine of 0.90 to 0.94 (SM A.3).

We say that the intervention succeeded, or that it reversed the answer, if the model was correct before it and incorrect after it. Every share we report is a share of the trials the model answered correctly before the intervention, since a trial it already had wrong cannot be reversed. We also report, in SM B.2 (Table 7), the change in the difference between the logits of the two option numbers, as the intervention may not be strong enough to reverse the answer but can still change the relative logits substantially.

3.4 Models

We use four open-weight language models of different families and sizes: Gemma-2-2B (Riviere et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), Llama-3.1-8B (Grattafiori et al., 2024) and Qwen2.5-32B (Yang et al., 2025), all in their base versions and in 16-bit precision. The ordering of the middle rungs is studied in all four. However, the interventions require that the model can do the comparison task. Gemma-2-2B tends to answer with the same slot number regardless of the rungs, and thus it is not used to test the intervention (SM B.1).

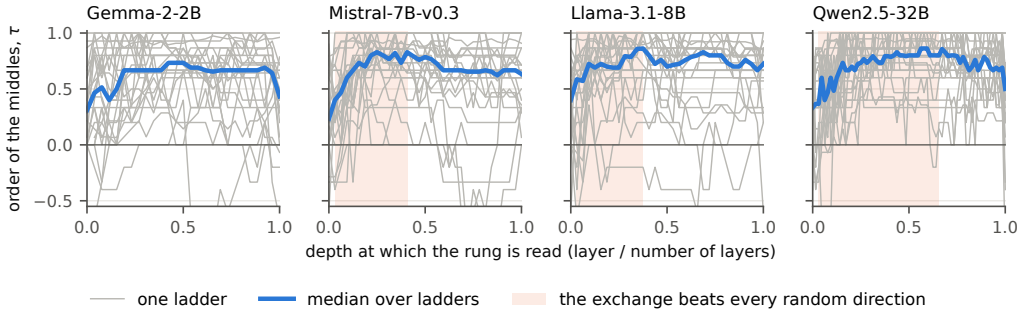


Figure 2: Order of the held-out middle rungs along $\hat{v}$ as a function of the depth at which the rungs are read (blue: median τ over the 26 ladders; gray: single ladders). The shaded band is the range of depths over which the median ladder's coordinate exchange exceeds its random controls (Section 4.2). The ordering is readable well beyond the depths at which it is used.

4 Results

4.1 Existence of the ordinal axis

For each ladder and model, we build the direction $\hat{v}$ from the two endpoints and project the activations of the middle rungs onto it (Equation 6). Since these rungs play no part in constructing the direction, their coordinates are not constrained to follow the ladder's order. We therefore use Kendall's τ between their coordinates and their true ranks to measure if the order is recovered. Figure 2 shows the median τ over the 26 ladders at every layer. We observe that middle rungs do recover their ordinal structure along $\hat{v}$ across all four models. While the token embeddings already carry a weak order, with a median τ of 0.23 to 0.40, it strengthens over the first third of the network and then remains roughly stable or falls slightly. The order is therefore not confined to a narrow range of layers, but remains readable through most of the network. Moreover, the shaded band marks the layers where exchanging the coordinate affects the model's answer (Section 4.2), showing that the order remains readable even beyond the depths at which the exchange is effective.

Because the strength of the ordering varies across depth, we select the readout layer out of sample. That is, on each of 200 random splits, we choose the layer with the highest median τ on half of the ladders and evaluate it on the other half. The selected layer lies at relative depth 0.25 to 0.56 across models. At this layer, the order is recovered strongly in all four models, with median τ of 0.67 for Gemma-2-2B, 0.80 for Mistral-7B, 0.86 for Llama-3.1-8B, and 0.80 for Qwen2.5-32B. The per-model summary and further robustness tests are reported in SM A.2.

The order is present even in Gemma-2-2B, which answers the comparison prompts at chance because it tends to pick the same option number whatever the two rungs are (SM B.1). The model thus represents the order of the concept, but the representation is not strong enough to overcome its preference for an option number. Figure 3 shows how the individual rungs are arranged for eight of the ladders on Qwen2.5-32B. The middle rungs, which play no part in building the axis, rise from one endpoint to the other, and the few exceptions are adjacent rungs whose coordinates are close. The remaining ladders, and the same panels for the other three models, are in SM A.4. The same ordinal structure also emerges when the rungs are presented independently in declarative contexts, without any comparison question (SM A.5, Figure 11).

4.2 Causality of the ordinal axis

We now ask whether the model uses the coordinate when it compares two rungs. At one layer we exchange the coordinates of the two options along $\hat{v}$, leaving everything else in place (Section 3.3), and count the trials that the model answered correctly before the exchange and incorrectly after it. Notably, some models are unable to do the baseline task, so on each model we use only every ladder it answers with at least 80% accuracy (9 on Mistral-7B, 11 on Llama-3.1-8B and 25 on Qwen2.5-32B). In particular, we leave Gemma-2-2B out as it tends to give the same answer number whatever the two

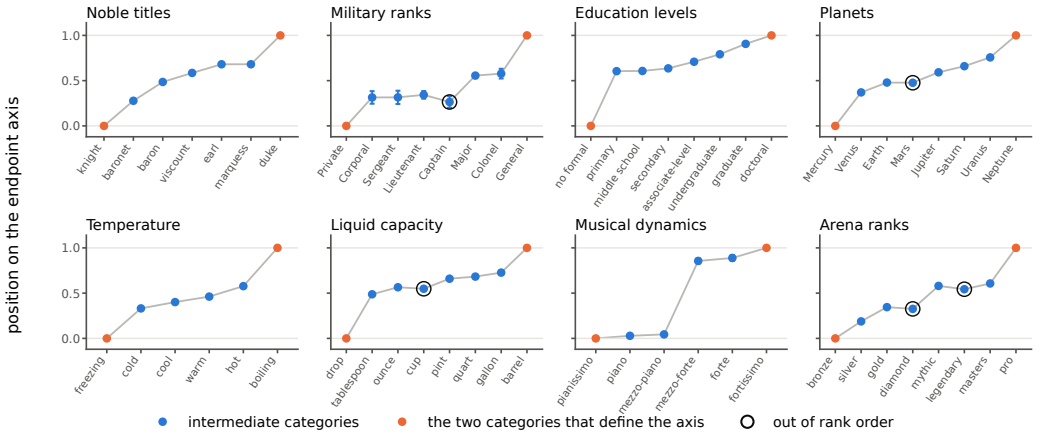


Figure 3: Position of every category on the ordinal axis of its own ladder, against its true rank, for eight of the 26 ladders on Qwen2.5-32B, at layer 36 of 64, the layer chosen most often when the layer is selected on half of the ladders. The axis is built from the two extreme categories, shown in orange, which sit at 0 and 1 by construction; the intermediate categories, in blue, play no part in its construction. Every centroid comes from the option in slot 1 of the comparison prompt. Bars give the spread over the three question forms. A circle marks a category that falls below one ranked beneath it. Kendall's τ in Section 4.1 is computed over the intermediate categories only; the two extremes, which define the axis, are shown for reference. The remaining ladders and the other three models are in SM A.4.

rungs are. We observed that the probability of that number rises when the higher rung is placed in its slot, but the rise is not enough to change the answer when the lower rung is there, so its accuracy stays at chance (SM B.1).

With the layer chosen on half of the rung pairs and the effect measured on the other half, the exchange reverses a median of 0.37, 0.36 and 0.32 of the trials it answers correctly on Mistral, Llama and Qwen (Table 1, all ladder results in SM B.2). Because the exchange modifies only the component along $\hat{v}$ and leaves the rest of the activation unchanged, these results show that the model uses the ordinal coordinate in the comparison, establishing a causal role for the axis. The share of reversed trials is moreover a lower bound on how much of the decision goes through the coordinate, because the edit is applied at a single layer and the layers after it partly undo it. If we read the coordinate again at later layers without touching it, the two options have drifted back towards their original values, by half of the displacement seven layers after the edit on Mistral and Llama and by a fifth twelve layers after it on Qwen2.5-32B, so part of the exchange has faded before the answer is computed (SM B.3). As shown in Figure 4, the median ladder is above its random controls (additional controls in SM B.3) at the first layers and stays above it up to relative depths of 0.41, 0.38 and 0.66, with its largest effect between a quarter and a third of the way through the network. Past that depth the answer no longer depends on the coordinate at the option tokens, although the order can still be read there (see shaded regions in Figure 2).

4.3 The comparison axis

Exchanging the coordinate along the estimated direction $\hat{v}$ reverses about a third of the comparisons, which leaves open how it relates to a direction trained to control the comparison. We use a distributed alignment search (DAS) restricted to one dimension (Geiger et al., 2024) to find the unit direction u along which the exchange works best. Using this direction, the exchange reverses a median of 0.94, 0.98 and 1.00 of the correctly answered trials over 11, 9 and 8 ladders on Llama, Mistral and Qwen (the order of the models throughout this section and SM C), on a surface form of the prompt held out from the fit, against 0.26, 0.35 and 0.36 for $\hat{v}$ under the same exchange.

We find that the endpoint direction, $\hat{v}$, is strongly aligned with u, with a median cosine of 0.54 to 0.62 where random directions give next to 0. Furthermore, denoting by S the subspace spanned by

	Mistral-7B	Llama-3.1-8B	Qwen2.5-32B
ladders tested	9	11	25
accuracy before the exchange, median	0.85	0.85	0.98
correct trials reversed, median	0.37	0.36	0.32
range over ladders	0.02–0.48	0.04–0.55	0.07–0.49
ladders above their random controls	9 of 9	10 of 11	25 of 25
random directions, max	0.14	0.18	0.08

Table 1: Effect of exchanging the two options' coordinates along $\hat{v}$. The layer is chosen on half of the rung pairs and the share of correctly answered trials that the exchange reverses is measured on the other half; the range is over ladders. The ladders are every one the model answers with at least 80% accuracy, with no further selection. Random directions are norm-matched, 20 seeds, each scored at every layer of the sweep (every fourth layer on Qwen2.5-32B), and the value shown is the largest over layers, seeds and ladders.

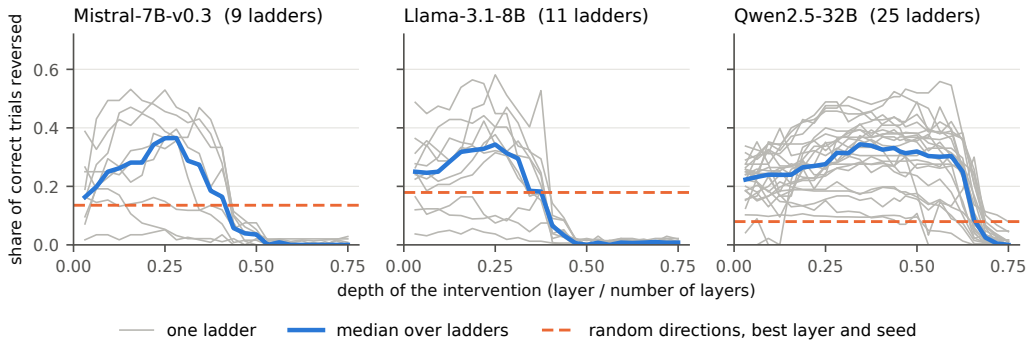


Figure 4: Share of correctly answered trials reversed by the coordinate exchange as a function of the depth at which it is applied, for each ladder (gray) and for the median ladder (blue). The dashed line is the largest effect reached by any norm-matched random direction on any ladder at any layer at which the random directions were run. The direction is re-estimated at each layer.

the rung centroids and P_S the orthogonal projection onto it, the trained direction splits into

$$u = u_{\parallel} + u_{\perp}, \qquad u_{\parallel} = P_S\, u, \quad u_{\perp} = (I - P_S)\, u, \qquad \|u_{\parallel}\|^2 = 0.49,\ 0.53,\ 0.37 \tag{8}$$

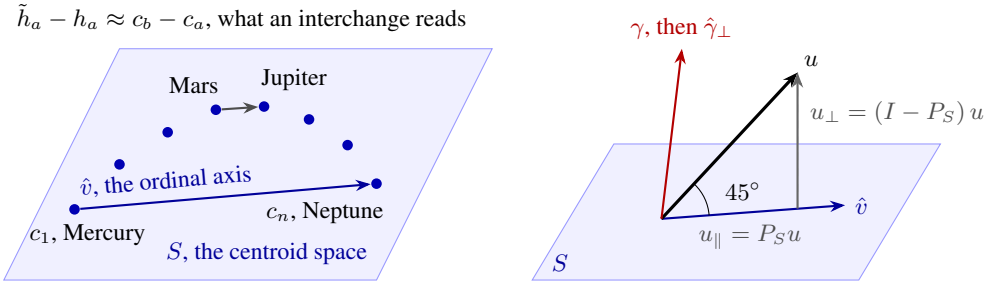
so about half of its squared norm on the two smaller models, and a third on Qwen, lies in the space of the centroids, and the rest outside it. The half inside points along the direction the endpoints approximate, with a cosine of 0.87, 0.82 and 0.87 to $\hat{v}$.

The exchange can only read what differs between a trial and its mirror. At an option token, the two prompts differ only in which rung occupies that slot, so $\tilde{h}_a - h_a$, the difference between the activation at option a's token in the mirror and in the trial, is approximately a difference between rung representations and therefore lies in the centroid space S; directions orthogonal to S, including $u_{\perp}$, read essentially nothing.

To first order, an exchange along any unit direction d changes the answer by what it reads times how strongly the answer responds to a push along d,

$$\Delta L \approx \langle \tilde{h}_a - h_a,\, d \rangle\, \langle \gamma,\, d \rangle, \tag{9}$$

where γ is the gradient of the logit difference between the two answer slots with respect to the activation at the option token (SM C). The first factor therefore depends only on the component of d inside S. By contrast, the gradient lies almost entirely outside S, so the model separates the rungs along directions inside S but is most sensitive to perturbations along directions outside it. Consistent with this decomposition, $u_{\perp}$ has cosine similarity 0.80, 0.81, and 0.76 with $\hat{\gamma}_{\perp}$, the unit part of γ outside S. Reading along u but writing the resulting change along $u_{\perp}$ alone preserves most of its effect, giving 0.88, 0.95, and 1.00, whereas writing along $u_{\parallel}$ alone drops to 0.17, 0.11, and 0.02 (Table 9 in SM C).



(a) inside the centroid space

(b) the trained direction and the gradient

Figure 5: The geometry of the trained direction. (a) The rung centroids c_i lie in the centroid space S (shaded), here on an arc. Inside S are the ordinal axis $\hat{v}$, from the first to the last centroid, and the difference between a trial and its mirror at an option token, which is what any interchange reads. (b) The trained direction u leaves the plane. Its projection $u_\parallel$ lies close to $\hat{v}$ and its remainder $u_\perp$ is orthogonal to every centroid difference. The gradient of the answer γ (red) is almost entirely outside S, and its outside part $\hat{\gamma}_\perp$ points along $u_\perp$ (cosine 0.8). The activations do not vary along $u_\perp$, so it reads nothing, and the answer is most sensitive along it, so it is where a write acts.

Given these results, we can build an approximation of the trained direction u directly from $\hat{v}$ and γ. Indeed, the direction

$$d = \frac{\hat{v} + \hat{\gamma}_\perp}{\|\hat{v} + \hat{\gamma}_\perp\|}, \tag{10}$$

which uses no training, has a cosine of 0.84, 0.83 and 0.82 with u and reverses 0.68, 0.87 and 0.78 of held-out trials, against 0.26, 0.35 and 0.36 for $\hat{v}$ alone and 0.94 to 1.00 for u (on the three forms together $\hat{v}$ gives 0.36, 0.36 and 0.37, the values of Section 4.2). Equation 9 also indicates why the trained direction settles at about half and half. For $d = \cos\theta\,\hat{v} + \sin\theta\,\hat{\gamma}_\perp$ the read goes as $\cos\theta$ and the sensitivity as $\sin\theta$, so in this idealised picture the effect goes as $\cos\theta\sin\theta$, largest at 45°.

The comparison is read from the rung space, along a direction the two endpoints already give. Everything the trained direction adds lies outside that space, along the answer's own gradient, where the activations never go. Its score therefore measures how far the later layers can be steered, not what the model computes, and the model's own coordinate is the endpoint direction.

5 Discussion and Limitations

We have shown that ordered categories are arranged along a direction in the residual stream, the ordinal axis. This direction can be built from the two extreme categories of a ladder alone, without training and without touching the categories in between, and it places those categories in their order on every model we test, at a median Kendall τ of 0.67 to 0.86. A model also reads the coordinate along it when it decides which of two categories ranks higher. Exchanging that one coordinate between the two options, and nothing else, reverses about a third of the decisions the model gets right.

To understand what else is used in the decision, we train a one-dimensional direction that is able to flip the comparison almost every time. About half of it (a third on Qwen2.5-32B) lies in the space of the category centroids and points along the endpoint direction. The other half is close to the gradient of the answer with respect to the activation, a direction the activations hardly occupy and along which the answer is most sensitive. To first order the effect of an exchange is what it reads times how strongly the answer responds, and since the read comes mostly from inside the centroid space and the sensitivity mostly from outside it, the best single direction sits between the two. This allows us to build a direction from the endpoint axis and the gradient, with no training, and recover most of the trained direction's effect. A trained mediator's accuracy therefore bounds how far the later layers can be steered rather than how well the direction pictures the representation. The model's own coordinate is the endpoint direction, and we can picture the trained direction as that coordinate with a lever attached.

This analysis, however, has some limitations. First, four models cannot separate size from architecture or training data. Second, everything is in English, and we have not asked whether the same ladder in another language shares the direction. Third, the task is a forced choice between two options with the answer given as a slot number, so the causal evidence covers one use of the order.

The representation is broader than that one use. The same ordinal arrangement appears when a category name is only used in a sentence, with no question asked and nothing to decide, and it appears at a token after the name, so the model computes the coordinate as part of representing the name and carries it forward (SM A.5). There is no answer to flip in such a sentence, and by our own account a coordinate that can be read is not thereby one that is used, so we cannot say what else reads it. Yet, this shows that the comparison does not create the coordinate. It finds one that is already there, and any computation that needs the order of a ladder has the same coordinate available at the same place. Which of those computations read it is the question this work leaves open.

AI USE STATEMENT

We used large language model assistants (Claude, Anthropic; ChatGPT, OpenAI) throughout this work. They were used to write and debug the analysis code, to prepare and launch experiment runs from instructions we specified, to generate tables and figures from the results, and to help draft, edit and format the text. Every experimental design choice, every analysis reported and every claim in the paper was decided by the authors, and all numbers were checked against the underlying result files. The authors take full responsibility for the content of the paper.

ETHICS STATEMENT

This work analyses the internal activations of publicly released, open-weight language models on synthetic prompts about ordered categories such as ranks, titles and physical scales. It involves no human subjects, no personal or sensitive data, and no new model training beyond a one-dimensional direction fitted to a frozen model. The interventions are used only to measure how a model makes comparisons; they do not add capabilities and we do not see a direct path to misuse. We see the main value of the work in improving the understanding of how language models represent and use structured knowledge.

REPRODUCIBILITY STATEMENT

All models are open-weight (Gemma-2-2B, Mistral-7B-v0.3, Llama-3.1-8B and Qwen2.5-32B, base versions in 16-bit precision). The 26 ladders, their rungs and comparison questions, the three surface forms of the prompt, the control lists and the declarative templates are listed in full in SM A. The construction of the ordinal axis is given in Section 3.2, the intervention in Section 3.3, and the restrictions under which it is run, including the layer selection, the trial filters and the random controls, in SM B.1; the trained direction and its evaluation protocol are described in SM C.1. Every layer choice is made on data held out from the measurement it is used for, and random directions use fixed seeds. The code, the prompts and the per-ladder results will be released as open source upon publication.

REFERENCES

HV AlquBoj, Hilal AlQuabeh, Velibor Bojkovic, Tatsuya Hiraoka, Ahmed Oumar El-Shangiti, Munachiso Nwadike, and Kentaro Inui. Number representations in llms: A computational parallel to human perception. *arXiv preprint arXiv:2502.16147*, 2025.

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. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Ahmed Oumar El-Shangiti, Tatsuya Hiraoka, Hilal AlQuabeh, Benjamin Heinzerling, and Kentaro Inui. The geometry of numerical reasoning: Language models compare numeric properties in linear subspaces. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 550–561, 2025.

Josh Engels, Eric Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, volume 2025, pp. 84591–84622, 2025.

Atticus Geiger, Kyle Richardson, and Christopher Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the third blackboxnlp workshop on analyzing and interpreting neural networks for NLP*, pp. 163–173, 2020.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, volume 2024, pp. 2483–2503, 2024.

Benjamin Heinzerling and Kentaro Inui. Monotonic representation of numeric properties in language models. *arXiv preprint arXiv:2403.10381*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Louis René Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, volume 2024, pp. 26486–26515, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 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, 2022.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 76441–76463, 2025.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

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. *Advances in neural information processing systems*, 33:12388–12401, 2020.

A Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, pp. 2412–15115, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

A LADDERS AND PROMPTS

A.1 LADDERS

Every ladder below is a sequence of discrete, named rungs whose order the word forms themselves do not supply; it is attached to them by usage. The 26 ladders run from six to twenty rungs, listed in rank order, lowest rung first.

Each item gives the comparison question of the ladder. Every comparison is asked in three surface forms, which differ in layout and in the wording of the instruction, with the question in place of `{question}` and the two rungs in place of `{a}` and `{b}`:

Form A

```
{question}
1.  {a}
2.  {b}
Answer with the slot number only.
Answer:
```

Form B

```
Question:  {question}
Options:
1.  {a}
2.  {b}
Reply with 1 or 2.
Answer:
```

Form C

```
{question} Option 1:  {a}.  Option 2:  {b}.  The answer is
option
```

The example in Section 3.1 is Form A. The 26 families of prompts are:

1. **Ordinal numbers**: "Which ordinal comes later in the sequence?"
   - first, second, third, fourth, fifth, sixth, seventh, eighth, ninth, tenth, eleventh, twelfth, thirteenth, fourteenth, fifteenth, sixteenth, seventeenth, eighteenth, nineteenth, twentieth.
2. **Cardinal numbers**: "Which number is larger?"
   - one, two, three, four, five, six, seven, eight, nine, ten.
3. **Physical size**: "Which object is larger?"
   - virus, hair, grain, chair, tree, tower, city, continent, star, universe.
4. **Density**: "Which material is denser?"
   - wood, ice, plastic, brick, aluminium, titanium, zinc, copper, lead, gold.
5. **Online game ranks**: "Which League of Legends rank is higher?"
   - iron, bronze, silver, gold, platinum, emerald, diamond, master, grandmaster, challenger.
6. **Acidity**: "Which substance is more alkaline?"
   - vinegar, pickle, beer, banana, corn, water, egg, soap, detergent, bleach.
7. **Places**: "Which place is larger?"
   - suite, floor, building, street, quarter, town, region, state, kingdom, planet.
8. **Arena ranks**: "Which Brawl Stars rank is higher?"
   - bronze, silver, gold, diamond, mythic, legendary, masters, pro.
9. **Education levels**: "Which level of education is higher?"
   - no formal education, a primary education, a middle school education, a secondary education, an associate-level education, an undergraduate education, a graduate education, a doctoral education.

10. **Liquid capacity**: "Which unit of liquid capacity is larger?"
    - drop, tablespoon, ounce, cup, pint, quart, gallon, barrel.
11. **Military ranks**: "Which soldier holds the higher rank?"
    - Private Miller, Corporal Miller, Sergeant Miller, Lieutenant Miller, Captain Miller, Major Miller, Colonel Miller, General Miller.
12. **Planets**: "Which planet is farther from the Sun?"
    - Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune.
13. **Belt ranks**: "Which belt rank is more advanced?"
    - white, yellow, orange, green, blue, brown, black.
14. **Heat sources**: "Which is hotter?"
    - frost, soup, oven, ember, furnace, sun, lightning.
15. **Noble titles**: "Which noble title is higher in precedence?"
    - knight, baronet, baron, viscount, earl, marquess, duke.
16. **Scout ranks**: "Which Scout rank is higher?"
    - scout, tenderfoot, second class, first class, star, life, eagle.
17. **Weight classes**: "Which boxing weight class is heavier?"
    - flyweight, bantamweight, featherweight, lightweight, welterweight, middleweight, heavyweight.
18. **Item quality (colour)**: "Which item quality is higher?"
    - grey, white, green, blue, purple, orange, gold.
19. **Item quality (name)**: "Which item quality is higher?"
    - poor, common, uncommon, rare, epic, legendary, artifact.
20. **Academic ranks**: "Which academic rank is more senior?"
    - undergraduate, graduate, postdoctoral, assistant, associate, full.
21. **Clergy ranks**: "Which clergy rank is higher in the Catholic hierarchy?"
    - deacon, priest, bishop, archbishop, cardinal, pope.
22. **Musical dynamics**: "Which dynamic marking is louder?"
    - pianissimo, piano, mezzo-piano, mezzo-forte, forte, fortissimo.
23. **Musical tempo**: "Which tempo marking is faster?"
    - largo, adagio, andante, moderato, allegro, presto.
24. **Poker hands**: "Which poker hand is stronger?"
    - high card, pair, straight, flush, full house, royal flush.
25. **Career levels**: "Which software engineering level is more senior?"
    - intern, junior, midlevel, senior, staff, principal.
26. **Temperature**: "Which is hotter?"
    - freezing, cold, cool, warm, hot, boiling.

List of controls. Three lists serve as controls for the read (Table 3). Their rank is only the position of an item in the list and carries no meaning, and each is asked with the same generic question, "Which of these is ranked higher?", in the three forms above. The household objects were drawn so that their list order is unrelated to word length, frequency, alphabetical order, first letter and typical size.

- **Vegetables**: carrot, potato, tomato, onion, pepper, cucumber, broccoli, spinach, lettuce, pumpkin.

	Gemma-2-2B	Mistral-7B	Llama-3.1-8B	Qwen2.5-32B
median τ, held-out layer	0.67	0.80	0.86	0.80
median τ, best layer	0.86	0.93	1.00	0.98
significant ladders, best layer	13	15	15	14

Table 2: Order of the middle rungs along $\hat{v}$ over the 26 ladders, read at the token of the rung in slot 1 of the comparison prompts. τ is Kendall's rank correlation between the coordinates of the middle rungs and their true order. At the held-out layer, the layer is chosen on a random half of the ladders and τ is measured on the other half, over 200 splits. The best layer is each ladder's own, searched between relative depths 0.25 and 0.75. A ladder counts as significant when $p < 0.05$, with p comparing its best τ with the best τ that random orderings of its middle rungs reach over the same layers.

- **Household objects**: lamp, mirror, bench, shelf, rug, chair, clock, table, desk, basket.
- **Random nouns**: table, river, guitar, mountain, pillow, engine, garden, rocket, hammer, mirror, candle, feather, kettle, bridge, forest, ladder, basket, drum, window, spoon, tiger, cloud, pencil, wheel, lantern, violin, meadow, turtle, balloon, saddle.

A.2 Robustness of the order

Table 2 summarises the read of Section 4.1 for each model, at the held-out layer and at each ladder's own best layer. The tables that follow test its robustness.

In Tables 3, 4 and 5 we test the robustness of Kendall's τ across different families. We first ask whether the ordinal structure recovered by $\hat{v}$ is robust across families and model architectures. As shown in Table 3, the median Kendall τ is consistently positive and substantial across all four models, both when the layer is selected using other ladders and when a single fixed layer is used for every ladder. The vast majority of ladders have $\tau > 0$, and a substantial fraction reach significance at $p < 0.05$. These values are well above the random-ordering baseline, while reinitialising the model weights removes the positive association. The input token embeddings, before any layer, already carry a weak order (a median τ of 0.23 to 0.40), well below the values reached inside the model, so most of the ordering is built by the network rather than inherited from the embeddings. Finally, the lists that do not specify an order produce values close to zero or negative, providing a control for cases where no ordinal structure is intended.

We next test whether this ordering is robust to choices in the prompt construction and to simple surface properties of the rung names. As shown in Table 4, the median τ changes little when any one of the three question forms is removed, or when the prompts used to construct the axis are separated from those used to place the rungs along it. Partialling out name length, frequency, alphabetical position, or first letter similarly leaves the ordering largely unchanged. Finally, reversing the direction of the question preserves both the ordering and the recovered axis: the forward and reversed axes have a cosine similarity of 0.96–0.97 across models. Together, these controls suggest that the recovered order is not tied to a particular prompt form or to simple lexical properties of the rung names.

Finally, Table 5 breaks down the result by individual ladder. The positive median effect is not driven by a small subset of particularly well-ordered ladders. Positive τ values occur across a broad range of conventional and physical scales, with many ladders reaching significance on multiple models. There is nevertheless substantial variation across ladders and models, and a small number of cases are sensitive to lexical controls. Thus, while the strength of the ordering varies by ladder, the overall effect is distributed broadly across the set rather than concentrated in a few families.

	Gemma-2-2B	Mistral-7B	Llama-3.1-8B	Qwen2.5-32B
median τ, layer chosen on other ladders	0.67	0.80	0.86	0.80
median τ, fixed layer (the same for every ladder)	0.69	0.72	0.72	0.80
ladders with $\tau > 0$	25/26	23/26	24/26	26/26
ladders at $p < 0.05$	15/26	16/26	15/26	17/26
median τ under random orderings (99.9th percentile)			0.31	
random-init weights, median τ	-0.14	0.00	0.14	-0.03
token embeddings (layer 0), median τ	0.31	0.23	0.40	0.33
Lists that declare no order				
Vegetables	-0.43	-0.36	-0.57	-0.07
Random nouns	-0.02	0.00	0.07	-0.01
Household objects	-0.14	-0.21	0.14	-0.21

Table 3: Robustness of the ordinal ordering across models and controls. Kendall τ between the positions of the held-out middle rungs along $\hat{v}$ and their true order, over the 26 ladders. Layer chosen on other ladders: the layer is selected on a random half of the ladders and τ is measured on the other half, reported as the median over 200 splits. Fixed layer: one layer is used for every ladder of a model (14/26 for Gemma-2-2B, 18/32 for Mistral-7B and Llama-3.1-8B, and 29/64 for Qwen2.5-32B). p is the exact one-sided permutation probability over all orderings of a ladder's middle rungs. Random orderings: 99.9th percentile of the median τ obtained after randomly ordering the middle rungs of every ladder over 20,000 draws; every observed model median exceeds all 20,000 draws ($p < 5 \times 10^{-5}$). Random-init weights: the same analysis after reinitialising the model weights. Token embeddings: the same measurement on the input embeddings, before any layer, on the 22 to 24 ladders whose two endpoint rungs do not end in the same token.

	Gemma-2-2B	Mistral-7B	Llama-3.1-8B	Qwen2.5-32B
all three question forms	0.69	0.72	0.73	0.80
without form A	0.76	0.73	0.73	0.80
without form B	0.67	0.79	0.69	0.80
without form C	0.71	0.69	0.72	0.80
axis and positions from different prompts	0.71	0.71	0.73	0.80
partialled on length	0.69	0.71	0.71	0.80
partialled on frequency	0.71	0.71	0.72	0.78
partialled on alphabetical order	0.71	0.72	0.74	0.79
partialled on first letter	0.69	0.72	0.75	0.80
reversed question	0.67	0.71	0.70	0.80
cosine, forward vs reversed axis	0.97	0.96	0.96	0.97

Table 4: Robustness of the ordinal ordering to prompt and lexical variation. Median Kendall τ over the 26 ladders at the fixed layer. Without form A/B/C: the axis and rung positions are recomputed after removing one of the three question forms. Axis and positions from different prompts: the axis is constructed from three of the six prompt configurations for each rung and the rung positions are measured on the remaining three, reported as the median over all 20 splits; read from slot-1 centroids only three configurations remain, so the axis is constructed from two of them and the rung positions are measured on the third, reported as the median over the three choices of held-out configuration. Partialled: Kendall τ after controlling for name length, frequency, alphabetical position, or first letter; ladders for which the corresponding control cannot be computed are omitted. Reversed question: the comparison is phrased in the opposite direction (e.g., "Which ... is lower?"); the final row reports the cosine similarity between the axes recovered from the forward and reversed questions.

	Gemma	Mistral	Llama	Qwen		Gemma	Mistral	Llama	Qwen
Ordinal numbers	$0.93^{\dagger}$	$0.88^{\dagger}$	$0.91^{\dagger}$	$0.93^{\dagger}$	Heat sources	0.20	0.00	0.20	0.40
Cardinal numbers	$0.93^{\dagger}$	$0.93^{\dagger}$	$0.93^{\dagger}$	$0.86^{\dagger}$	Noble titles	$1.00^{\dagger}$	$1.00^{\dagger}$	$0.80^{\dagger}$	$1.00^{\dagger}$
Physical size	$0.64^{\dagger}$	$0.79^{\dagger}$	$0.64^{\dagger}$	$0.86^{\dagger}$	Scout ranks	-0.60	0.40°	-0.40	0.40°
Density	$0.64^{\dagger}$	$0.71^{\dagger}$	0.43	$0.79^{\dagger}$	Weight classes	0.60	0.60°	0.60	0.60°
Online game ranks	$0.71^{\dagger}$	$0.79^{\dagger}$	$0.79^{\dagger}$	$1.00^{\dagger}$	Item quality (colour)	0.60°	0.00	$1.00^{\dagger}$	$1.00^{\dagger}$
Acidity	0.21	$0.64^{\dagger}$	0.00	0.36	Item quality (name)	$0.80^{\dagger}$	$0.80^{\dagger}$	$0.80^{\dagger}$	$0.80^{\dagger}$
Places	$0.79^{\dagger}$	$0.57^{\dagger}$	$0.71^{\dagger}$	$0.79^{\dagger}$	Academic ranks	0.67	0.67	0.67	0.67
Arena ranks	$0.73^{\dagger}$	0.47°	$0.87^{\dagger}$	$0.73^{\dagger}$	Clergy ranks	0.67	0.00°	0.33°	0.67
Education levels	$0.87^{\dagger}$	0.60	$0.87^{\dagger}$	$0.87^{\dagger}$	Musical dynamics	0.33	0.67	$1.00^{\dagger}$	$1.00^{\dagger}$
Liquid capacity	$0.73^{\dagger}$	$0.73^{\dagger}$	$0.73^{\dagger}$	$0.87^{\dagger}$	Musical tempo	0.67	$1.00^{\dagger}$	0.33	0.67
Military ranks	$0.73^{\dagger}$	$0.73^{\dagger}$	$0.73^{\dagger}$	0.47	Poker hands	0.67	0.67	0.67	0.33
Planets	$1.00^{\dagger}$	$0.87^{\dagger}$	0.47°	$0.87^{\dagger}$	Career levels	0.33°	$1.00^{\dagger}$	0.67°	$1.00^{\dagger}$
Belt ranks	$0.80^{\dagger\circ}$	$1.00^{\dagger}$	$0.80^{\dagger\circ}$	$0.80^{\dagger\circ}$	Temperature	$1.00^{\dagger}$	$1.00^{\dagger}$	$1.00^{\dagger}$	$1.00^{\dagger}$

Table 5: Ordinal ordering for each of the 26 ladders. Kendall τ between the middle-rung coordinates along $\hat{v}$ and their true order at the fixed layer for each model (14/26 for Gemma-2-2B, 18/32 for Mistral-7B and Llama-3.1-8B, and 29/64 for Qwen2.5-32B). Ladders follow the order of Figure 7 (left block, then right). † denotes $p < 0.05$ under the exact one-sided permutation test. $^{\circ}$ denotes ladders for which τ decreases by more than 0.2 after controlling for at least one of name length, frequency, alphabetical position, or first letter. With four middle rungs, significance requires perfect ordering.

A.3 Why we use slot-1 centroids

The centroid of a rung (Equation 3) averages its activation over the comparison prompts in which it appears, and every such prompt also contains a second rung and a question about which of the two ranks higher. We take the centroids from the option in slot 1 only. That option precedes its partner in the prompt, so its activation depends on the question and on the rung's own name and on nothing else: the activation of a rung in slot 1 is the same whichever partner follows it, to one part in a hundred. No information about the partner, and hence none about the outcome of the comparison, can enter the direction built from these centroids.

The same axis is then used at both options, in slot 1 and in slot 2 (Section 3.3). This is justified because the two slots arrange the rungs along nearly the same direction. At the fixed layer, the endpoint directions built from slot 1 and from slot 2 have a median cosine of 0.90 to 0.94, and the slot-1 axis orders the rungs read in slot 2 at a median τ of 0.69 to 0.79, against 0.79 to 0.83 on their own axis.

The choice of slot does not drive the result. Pooling both slots, the middle rungs come out in order at a median τ of 0.76, 0.77, 0.71 and 0.80 on Gemma-2-2B, Mistral-7B, Llama-3.1-8B and Qwen2.5-32B, against 0.69, 0.72, 0.72 and 0.80 from slot 1, and Figure 6 shows that the two reads follow each other at every depth.

A.4 Ordered axis projections for every model

Figure 3 in the main text shows, for eight ladders on Qwen2.5-32B, where every category falls on the axis of its own ladder. Figure 7 shows all 26 ladders on that model, and Figures 8 to 10 show them for the other three models, with the same construction and the same held-out estimate of the axis and of the two reference points, at the layer chosen most often on half of the ladders. The arrangement is the same in the four models. The middle rungs, which play no part in building the axis, follow the order of the ladder in most ladders, and Gemma-2-2B, which cannot do the comparison task (SM B.1), orders them nearly as well as the larger models.

A.5 Non comparison prompts

The ordinal axis also emerges when only the scale-specific meaning of each rung is expressed, without asking any explicit question about the scale.

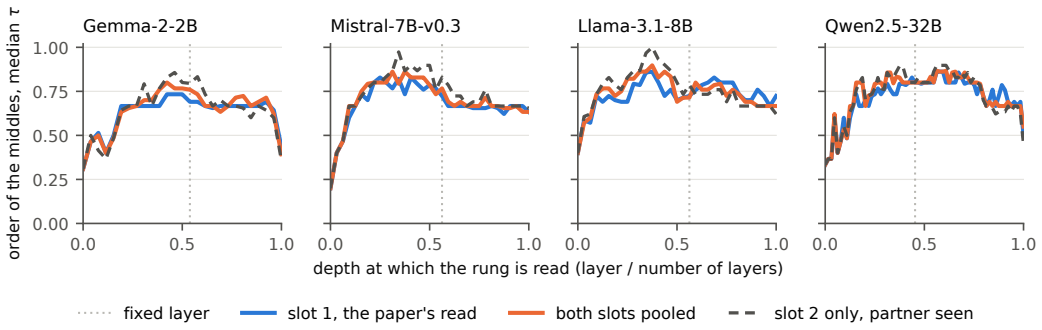


Figure 6: Order of the middle rungs by depth when the centroids come from slot 1 (the paper's read), from slot 2, or from both slots pooled; medians over the 26 ladders. The dotted line marks the fixed layer.

To construct the axis without comparison prompts, we instead generate a set of contexts that express the meaning of each rung independently, without requiring the model to compare it. The prompts are paired across rungs, every rung is instantiated using the same set of prompt templates, with only the rung itself substituted. This provides multiple contexts that consistently evoke the intended meaning while ensuring that differences in the extracted activations are driven primarily by the rung rather than by the surrounding prompt.

For example, for the martial-arts belt concept, the following templates are instantiated for each belt:

> *The rung belt is awarded after a formal grading.*
> *He trained for months to earn the rung belt.*
> *In judo, the rung belt marks a level of skill.*
> *She tied on her rung belt before practice.*
> *Students wearing the rung belt line up first.*
> *My instructor handed me the rung belt.*
> *The dojo requires a rung belt for that class.*
> *He was promoted to the rung belt last spring.*
> *Her uniform is fastened with a rung belt.*
> *The examiner approved him for the rung belt.*
> *Everyone in the group holds a rung belt.*
> *The club photograph shows him in a rung belt.*

The main difference from the comparison-prompt setup is that, here, we introduce a common word in each prompt at which we read the activations. In the case of martial arts, this word is *belt*. This allows us to reduce additional sources of noise, since the template is identical across rungs, whatever differs in the activation at the shared word arrives through attention from the rung token and is attributable to the rung itself. Figure 11 shows where every category falls on the axis built this way, for the eight ladders of Figure 3 on Llama-3.1-8B: the categories climb from one extreme to the other much as they do on the comparison read, although no comparison is asked for anywhere in the prompt.

B Details of the intervention

B.1 Experiment set up

Trials. Every pair of rungs of a ladder is asked in three surface forms, which differ in phrasing, layout and instruction wording (SM A.1), and each pair is asked in both slot orders, so a ladder of n rungs gives $3n(n-1)$ trials. A model is tested on a ladder only when it answers that ladder with an accuracy of at least 0.80, which leaves 9 ladders on Mistral-7B, 11 on Llama-3.1-8B and 25 on Qwen2.5-32B, with no further selection. A trial is kept only when the two answer tokens carry almost all of the model's next-token probability, $P(1)+P(2) \geq 0.95$, so that the answer is a choice between the two slot numbers and not something else. This keeps 2116 of 2784 trials on Llama-3.1-8B and 1570 of 2490 on Mistral-7B, between 0.35 and 1.00 of a ladder's trials on the first model and between

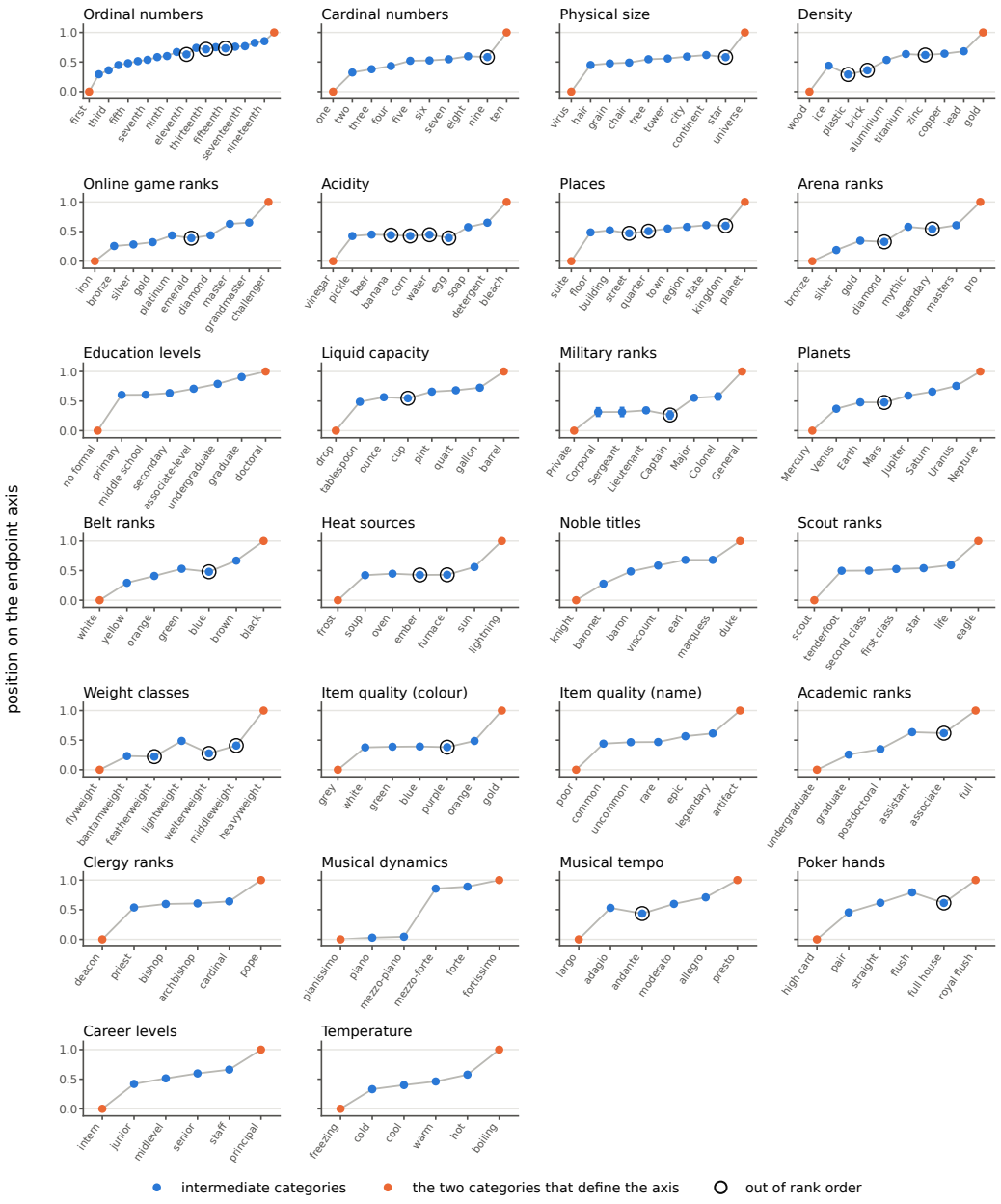


Figure 7: Position of every category on the ordinal axis of its own ladder, against its true rank, for all 26 ladders on Qwen2.5-32B. The construction is the same as in Figure 3, at layer 36 of 64, the layer chosen most often when the layer is selected on half of the ladders.

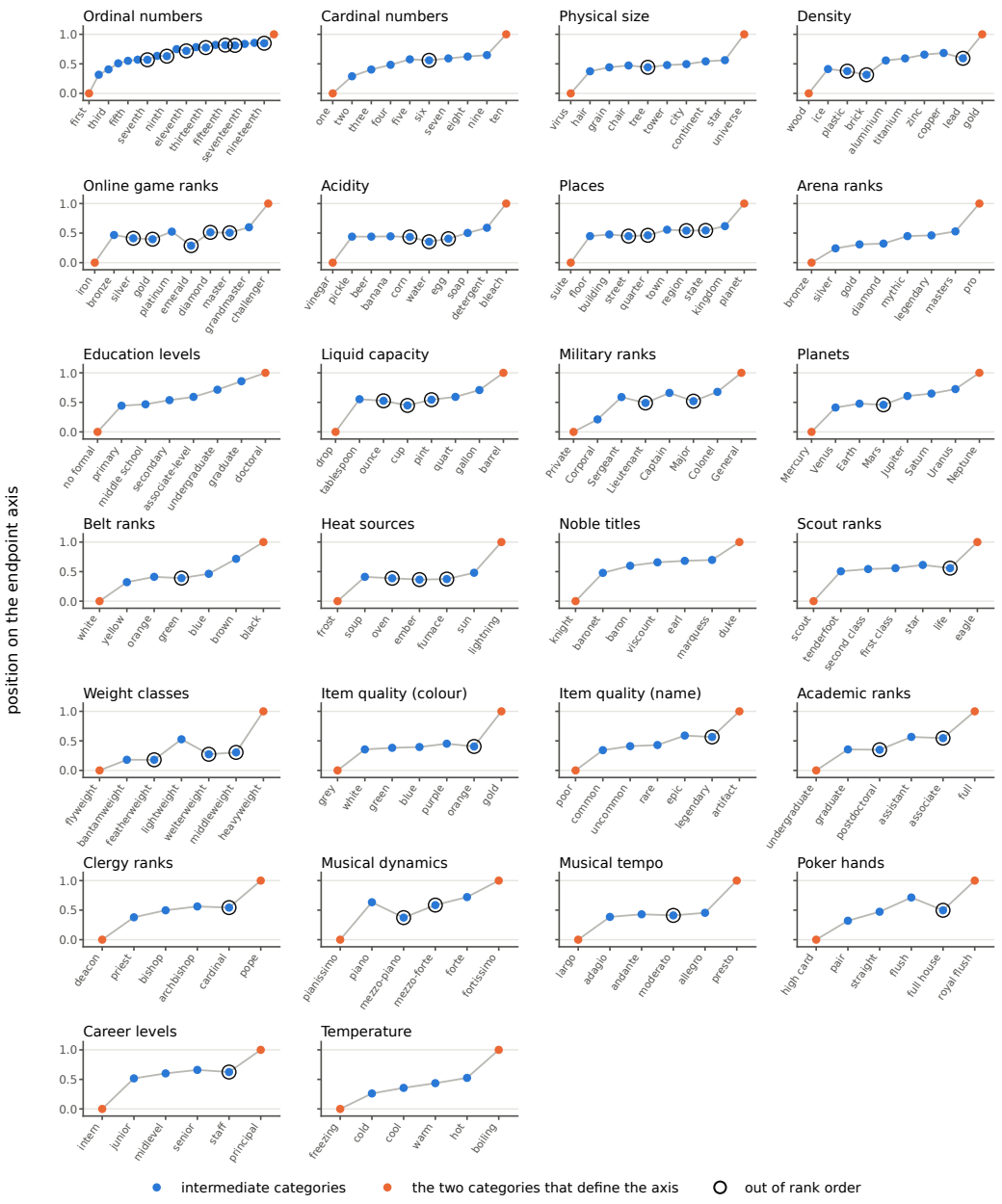


Figure 8: Position of every category on the ordinal axis of its own ladder, against its true rank, for all 26 ladders on Gemma-2-2B. The construction is the same as in Figure 3, at layer 9 of 26, the layer chosen most often when the layer is selected on half of the ladders.

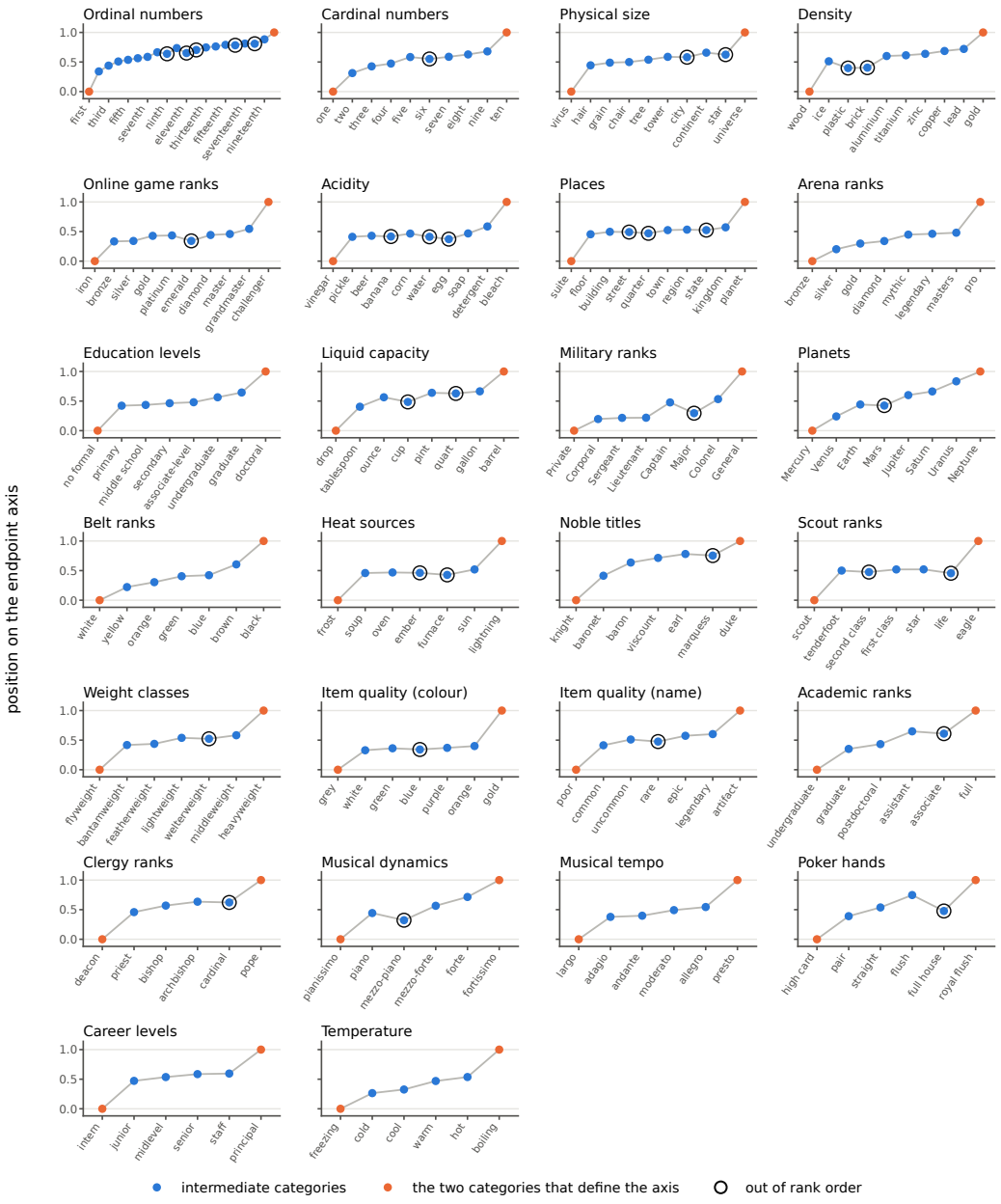


Figure 9: Position of every category on the ordinal axis of its own ladder, against its true rank, for all 26 ladders on Mistral-7B-v0.3. The construction is the same as in Figure 3, at layer 8 of 32, the layer chosen most often when the layer is selected on half of the ladders.

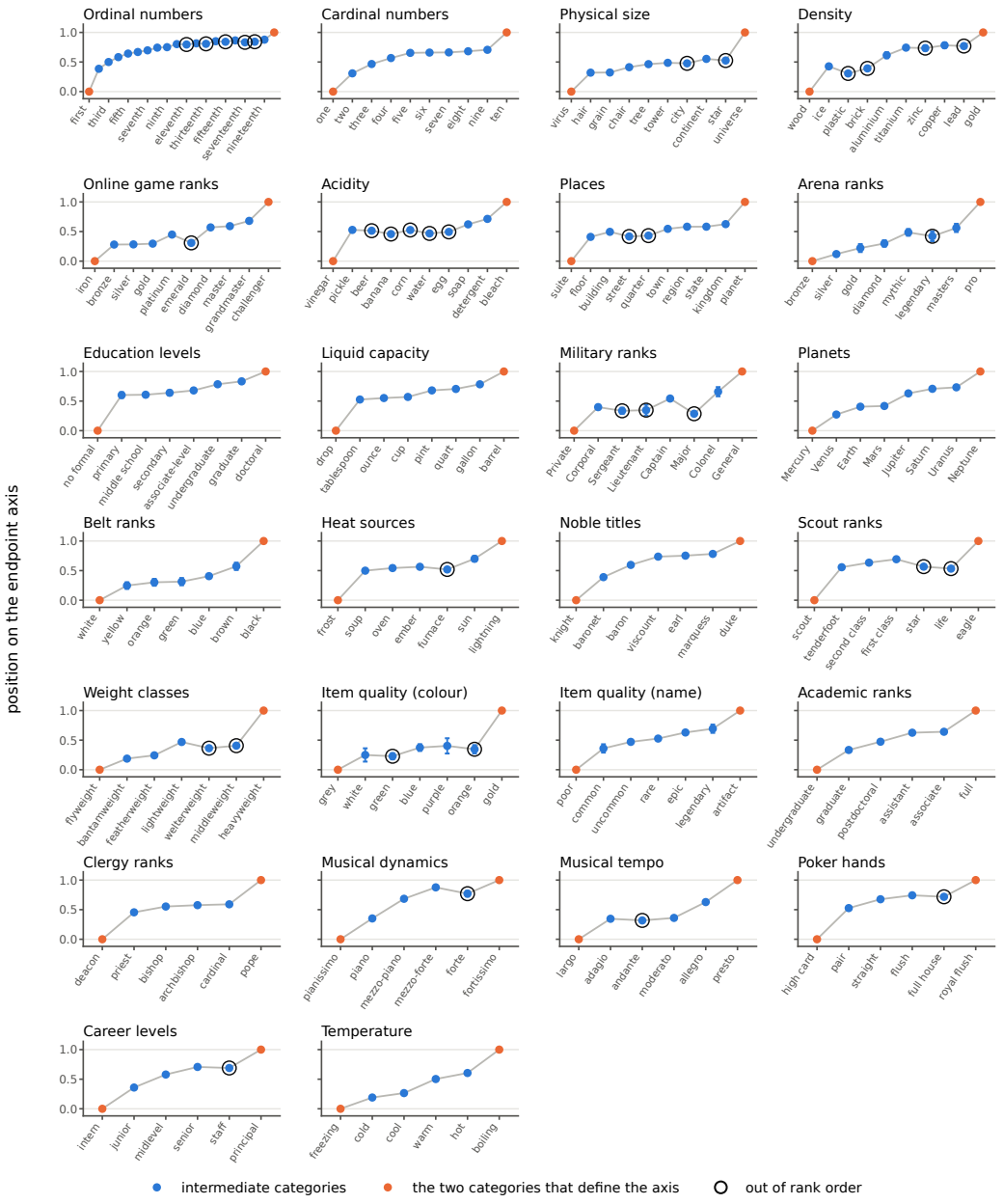


Figure 10: Position of every category on the ordinal axis of its own ladder, against its true rank, for all 26 ladders on Llama-3.1-8B. The construction is the same as in Figure 3, at layer 12 of 32, the layer chosen most often when the layer is selected on half of the ladders.

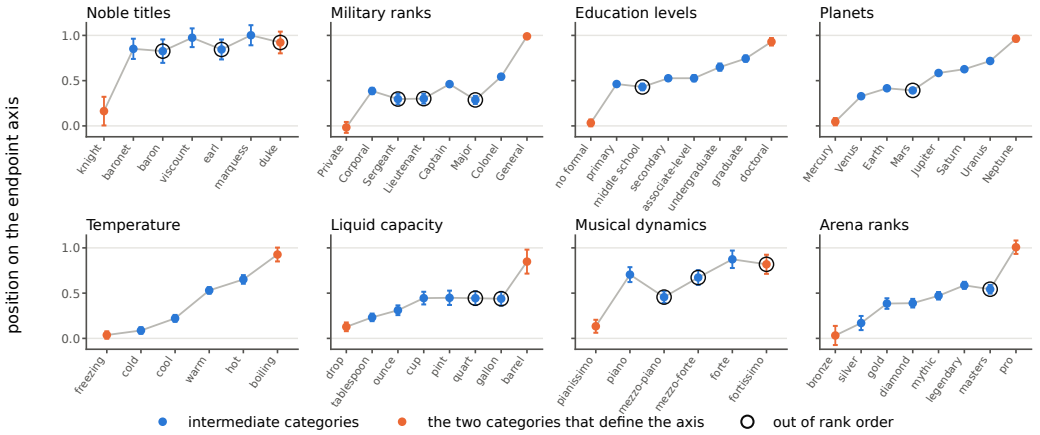


Figure 11: Position of every category on the ordinal axis of its own ladder, against its true rank, for the eight ladders of Figure 3, on Llama-3.1-8B at layer 12 of 32, the layer at which Figure 10 reads this model, so that the two reads are compared at one depth. Here the centroids come from the declarative prompts above and the activation is read at the word the prompts share, not from a comparison prompt. The axis is built from the two extreme categories, shown in orange; the intermediate categories, in blue, play no part in its construction. The axis and the two reference points at 0 and 1 are estimated on half of the prompt templates and every category, the two extremes included, is scored on the other half; bars give the spread over 20 random halvings. A circle marks a category that falls below one ranked beneath it. For the planets the shared word is the planet name itself, and for the military ranks and the education levels it is the final word of the name, which every rung of those two ladders carries.

0.40 and 0.89 on the second; the number words lose the most. Qwen2.5-32B keeps every trial. Every share reported below is a share of the kept trials the model answered correctly before the exchange, as in the main text; a trial it already had wrong cannot be reversed and is left out of the denominator.

The direction and where it is applied. The axis is the endpoint direction of the ladder, built from the centroids of the option in slot 1, which precedes its partner in the prompt and so cannot depend on it (SM A.3). It is re-estimated at every layer, and the same axis is used at both option positions, so nothing distinguishes the two options except the coordinate the intervention gives them. The coordinates are read and written at each option's last token, where Equation 7 exchanges them exactly.

The layer. The effect depends on the depth at which the exchange is applied, so the layer is chosen out of sample. The rung pairs of a ladder are split at random into two halves; the layer with the largest share of reversed trials on the first half is taken, and the share is then read on the second half, whose pairs took no part in the choice. This is repeated over 200 splits, and the ladder's reported effect is the median of the held-out share over them. The layer named for a ladder is the one chosen most often over the 200 splits. The layers swept are 1 to 24 of 32 on the two 7–8B models and every second layer from 2 to 48 of 64, and layer 29, on Qwen2.5-32B.

Random controls. To know what an exchange of this size does when it carries no meaning, the same displacement is applied along a direction that carries none. For each ladder, 20 random directions are drawn, each scaled to the length of that trial's displacement along the axis, and each is run at every swept layer. A ladder is above its random controls when its exchange reverses more trials than the largest share any of the 20 reaches at any layer, and a model's value is the largest over its ladders. That value is a maximum over both the seed and the depth, so it is the strictest of the natural summaries of the null.

Gemma-2-2B, which does not answer the comparison. The fourth model is left out of the intervention because it cannot do the task it would be intervened on. Its accuracy on the comparison is at chance, with a median of 0.51 over the 26 ladders and a range of 0.38 to 0.67, and the reason is

a preference for one answer number that does not depend on the two rungs. In the first of the three forms it answers with the first slot on at least 95% of the trials of 24 of the 26 ladders. The preference is not the same in the other two forms, and only 8 of the 26 ladders have a fixed answer in all three; the first form also never reaches the retention rule of $P(1) + P(2) \geq 0.95$, so it contributes only to the accuracy figures.

B.2 The effect ladder by ladder

Table 6 breaks down the intervention of Section 4.2 by ladder, while Table 7 measures how far the same intervention moves the decision whether or not it changes the final answer.

The reversal effect is broadly distributed across the ladders rather than being driven by a small number of cases. Almost every tested ladder exceeds its own norm-matched random controls, across all three models. The strength of the effect nevertheless varies substantially by ladder. Conventional rankings and physical scales generally show the largest reversals, whereas number words and education levels are among the weaker cases. Cardinal numbers on Llama-3.1-8B are the only tested ladder that does not exceed its random controls. Thus, the median effect reported in Section 4.2 reflects a broad causal effect across different kinds of ordered categories, with meaningful variation in how strongly each ladder depends on the coordinate.

Table 7 shows the same pattern when the intervention is measured continuously rather than only by whether it crosses the decision boundary. Across the tested ladders, the exchange lowers the probability of the correct option and moves the logit difference towards the incorrect answer on average. This includes ladders for which relatively few trials are fully reversed, showing that a low reversal rate need not mean that the coordinate has no effect on the decision. Together, the two measures show that exchanging the ordinal coordinate consistently moves the model towards the counterfactual answer, while the size of that movement varies across ladders.

B.3 Additional controls and downstream recovery

The effect is specific to the option representations. As a separate location control, we apply the same displacement at the token where the answer is read rather than at the option tokens. This intervention has almost no effect: it reverses a median of 0.01 of correctly answered trials on Mistral-7B and Llama-3.1-8B and 0.00 on Qwen2.5-32B, and never more than 0.05. Thus, the effect of the coordinate exchange depends on modifying the option representations rather than merely introducing the same perturbation later in the computation.

Later layers partially restore the edited coordinate. The exchange is applied at a single layer, but its effect does not remain unchanged as the forward pass continues. Reading the coordinate again at later layers shows that the two options gradually move back towards their pre-intervention values. On Mistral-7B and Llama-3.1-8B, about a quarter of the displacement has been restored four layers after the edit and about half seven layers after it; Qwen2.5-32B restores the coordinate more slowly, reaching about a fifth after twelve layers. The observed reversal rate therefore understates the effect of the intervention at the layer where it is applied, because part of the exchange is subsequently overwritten. This measure is only approximate: the ordinal direction itself rotates across layers, so apparent recovery can reflect both recomputation and rotation of the representation.

C The trained mediator

C.1 The trained direction and evaluation protocol

Section 4.3 uses a one-dimensional distributed alignment search (DAS) direction (Geiger et al., 2024) as an optimized causal reference for the endpoint direction $\hat{v}$. At the layer chosen for each ladder by the pair split of Section 4.2 (SM B.1), we freeze the model and optimize a unit direction u so that interchanging the two options' coordinates along u makes the model give the answer corresponding to the exchanged prompt. The intervention itself is the same throughout Section 4.3 and is defined formally in Equation 12. We optimize u from three random initializations for 150 Adam steps with learning rate 0.05.

Ladder	Mistral-7B	Llama-3.1-8B	Qwen2.5-32B
Ordinal numbers	0.15* (1)	0.18* (12)	0.10* (2)
Cardinal numbers	0.25* (1)	0.14 (1)	0.18* (4)
Physical size	—	—	0.14* (2)
Density	—	—	0.36* (40)
Online game ranks	0.32* (9)	0.47* (9)	0.28* (38)
Acidity	—	—	0.37* (26)
Places	—	—	0.16* (26)
Education levels	0.02* (8)	0.04* (6)	0.07* (4)
Liquid capacity	0.38* (9)	0.37* (8)	0.36* (24)
Military ranks	—	0.42* (7)	0.33* (20)
Planets	0.37* (7)	0.55* (8)	0.32* (22)
Belt ranks	—	—	0.39* (32)
Heat sources	—	—	0.21* (26)
Noble titles	0.48* (5)	0.33* (2)	0.42* (16)
Scout ranks	—	—	0.33* (8)
Weight classes	—	0.24* (8)	0.21* (12)
Item quality (colour)	—	—	0.49* (28)
Item quality (name)	—	—	0.21* (2)
Academic ranks	—	0.50* (6)	0.42* (36)
Clergy ranks	—	—	0.46* (16)
Musical dynamics	—	—	0.29* (14)
Musical tempo	—	—	0.25* (6)
Poker hands	0.45* (8)	—	0.38* (26)
Career levels	0.37* (11)	0.36* (7)	0.42* (36)
Temperature	—	—	0.25* (6)

Table 6: Share of the trials answered correctly before the exchange that it reverses, ladder by ladder, on the half of the rung pairs that took no part in choosing the layer; the chosen layer is in brackets. $*$ marks a ladder whose exchange exceeds its own random controls, the largest share any of its 20 norm-matched random directions reaches at any layer at which the null was run, on the same trials. A dash means the model answers that ladder with less than 0.80 accuracy, so it is not tested on it. Qwen2.5-32B's random directions were run at every fourth swept layer and at the reported depth rather than at every layer, so its random controls are maximised over 8 depths and not 24.

	Mistral-7B		Llama-3.1-8B		Qwen2.5-32B	
Ladder	P	logit	P	logit	P	logit
Ordinal numbers	0.05	0.14	0.09	0.46	0.09	0.62
Cardinal numbers	0.08	0.21	0.09	0.28	0.20	1.58
Physical size	—	—	—	—	0.14	1.04
Density	—	—	—	—	0.35	2.56
Online game ranks	0.17	0.53	0.32	1.51	0.29	2.27
Acidity	—	—	—	—	0.32	2.07
Places	—	—	—	—	0.17	1.14
Education levels	0.04	0.18	0.09	0.59	0.10	0.87
Liquid capacity	0.15	0.41	0.31	1.47	0.38	2.97
Military ranks	—	—	0.26	0.86	0.37	3.17
Planets	0.32	1.67	0.42	2.25	0.36	2.45
Belt ranks	—	—	—	—	0.44	3.50
Heat sources	—	—	—	—	0.27	2.01
Noble titles	0.35	1.38	0.26	1.22	0.46	3.29
Scout ranks	—	—	—	—	0.33	1.94
Weight classes	—	—	0.22	1.15	0.26	2.03
Item quality (colour)	—	—	—	—	0.35	1.61
Item quality (name)	—	—	—	—	0.22	1.56
Academic ranks	—	—	0.39	1.99	0.53	3.57
Clergy ranks	—	—	—	—	0.44	2.98
Musical dynamics	—	—	—	—	0.33	3.18
Musical tempo	—	—	—	—	0.28	2.50
Poker hands	0.30	0.91	—	—	0.43	3.68
Career levels	0.27	0.92	0.31	1.53	0.44	3.47
Temperature	—	—	—	—	0.32	2.09

Table 7: How far the exchange moves the decision on the trials it does not reverse as well as on those it does, at the same layer as Table 6 and over the same trials, those answered correctly before the exchange. P is the share of the probability of the correct option that the exchange removes, $(P_{\text{before}} - P_{\text{after}})/P_{\text{before}}$; *logit* is the fall in the logit of the correct option minus that of the other one. Both are means over those trials, positive towards the wrong answer. The distribution is skewed, so the median trial moves less than the mean.

The analysis uses all 11 causal ladders for Llama-3.1-8B, all 9 for Mistral-7B, and 8 for Qwen2.5-32B. To test whether the learned direction generalizes beyond the prompts on which it is optimized, the results reported in Section 4.3 fit the direction on two surface forms and evaluate it on the third, which takes no part in training. The same split is used for the training-free directions against which u is compared.

Table 8 gives both the causal effect and the geometry of these directions. The first part compares u with the endpoint direction $\hat{v}$, rank- and gradient-based directions, their combinations, and random controls, all under the same held-out interchange. The second part decomposes the trained direction relative to the centroid space and measures how its in-span and out-of-span components align with the ordinal representation and with the answer gradient. These quantities are developed in the following subsections.

C.2 What an interchange can read

Section 3.3 defines the coordinate exchange along the ordinal direction $\hat{v}$. Here we use the same intervention for an arbitrary unit direction d. Let $h_a = h[p_a]$ and $h_b = h[p_b]$ denote the residual-stream activations at the two option positions of a trial, and let $\tilde{h}_a$ and $\tilde{h}_b$ denote the activations at the same positions in the exchanged prompt. Their coordinates along d are

$$z_a = \langle h_a, d\rangle, \qquad z_b = \langle h_b, d\rangle, \qquad \tilde{z}_a = \langle \tilde{h}_a, d\rangle, \qquad \tilde{z}_b = \langle \tilde{h}_b, d\rangle. \tag{11}$$

The corresponding interchange is

$$h_a \leftarrow h_a + (\tilde{z}_a - z_a)d, \qquad h_b \leftarrow h_b + (\tilde{z}_b - z_b)d. \tag{12}$$

Equation 12 is therefore exactly the intervention of Equation 7, with $\hat{v}$ replaced by a general direction d. Figure 12 summarizes the four activations involved.

To characterize what this intervention can read, let

$$S = \mathrm{span}\{c_i - \bar{c}\}$$

be the subspace spanned by the centred rung centroids, and let P_S denote the orthogonal projection onto S. Write the two activations at position p_a as

$$h_a = c_a + e_a, \qquad \tilde{h}_a = c_b + \tilde{e}_a,$$

where e_a and $\tilde{e}_a$ contain the trial-specific deviations from the corresponding rung centroids. Their difference is then

$$\tilde{h}_a - h_a = (c_b - c_a) + (\tilde{e}_a - e_a), \tag{13}$$

and hence the coordinate transferred by an interchange along d is

$$\tilde{z}_a - z_a = \langle c_b - c_a,\, P_S d\rangle + \langle \tilde{e}_a - e_a,\, d\rangle. \tag{14}$$

The first term contains the difference between the two rungs and lies entirely in S by construction. Thus, apart from the residual term, the amount an interchange reads from d depends only on its component inside the centroid space. A direction orthogonal to S cannot distinguish the rung centroids and therefore has essentially no coordinate to exchange. Empirically, this residual is small. For instance, on the planets ladder of Llama-3.1-8B, the root-mean-square difference between the exchanged activations is 3.2 along $\hat{v}$ but only 0.03 along a direction orthogonal to S, comparable to a random direction. This is the sense in which $u_\perp$ can contribute strongly to writing an intervention while reading essentially nothing from the two rungs.

C.3 The answer gradient and the first-order model

We next characterize how strongly the answer responds to the displacement written by the interchange. Let

$$L = L_1 - L_2$$

be the difference between the logits of the two answer slots, and define

$$\nabla_a = \frac{\partial L}{\partial h_a}, \qquad \nabla_b = \frac{\partial L}{\partial h_b}.$$

These gradients are computed on each trial by a single backward pass with the model weights frozen. For small perturbations δ_a and δ_b of the two option activations,

$$\Delta L \approx \langle \nabla_a, \delta_a \rangle + \langle \nabla_b, \delta_b \rangle. \tag{15}$$

Under the interchange of Equation 12, $\delta_a = (\tilde{z}_a - z_a)d$ and $\delta_b = (\tilde{z}_b - z_b)d$, so

$$\Delta L \approx (\tilde{z}_a - z_a)\langle \nabla_a, d \rangle + (\tilde{z}_b - z_b)\langle \nabla_b, d \rangle. \tag{16}$$

Up to the residual terms in Equation 14, and because the directions built from the two slots are nearly parallel (SM A.3), exchanging the two rungs produces opposite coordinate changes at the two option positions, $\tilde{z}_b - z_b \approx -(\tilde{z}_a - z_a)$. Equation 16 therefore becomes

$$\Delta L \approx \langle \tilde{h}_a - h_a, d \rangle \, \langle \nabla_a - \nabla_b, d \rangle. \tag{17}$$

The two factors have different roles. The first is what the direction can *read* from the difference between the rungs; as shown in Section C.2, it comes almost entirely from the component of d inside the centroid space S. The second is how strongly the downstream answer responds to a displacement along d.

To summarize this answer sensitivity across trials, we define

$$\gamma = \frac{1}{N} \sum_{t=1}^{N} \big(\nabla_a - \nabla_b\big)_t, \qquad \hat{\gamma}_\perp = \frac{(I - P_S)\gamma}{\|(I - P_S)\gamma\|}, \tag{18}$$

where the average is over the N trials of the two fitting surface forms. We orient γ so that $\langle \gamma, \hat{r} \rangle > 0$, with $\hat{r}$ the rank direction of Equation 20. This construction uses only the model's two answer logits and not the correct answer. Averaging the second factor of Equation 17 gives the first-order picture used in Section 4.3: an interchange is effective when a direction both reads a difference between the rungs and points along a direction to which the answer is sensitive.

These two requirements occupy largely different parts of the residual stream. Only one to two percent of the squared norm of γ lies inside S, whereas the rung differences of Section C.2 lie almost entirely inside it. Correspondingly, the out-of-span gradient $\hat{\gamma}_\perp$ aligns strongly with the trained direction's out-of-span component $u_\perp$, with median cosine similarities of 0.80, 0.81, and 0.76 on Llama-3.1-8B, Mistral-7B, and Qwen2.5-32B, respectively, compared with 0.01 for random out-of-span directions. The gradient estimated on the held-out surface form remains close to the one estimated on the fitting forms, with cosine similarities between 0.73 and 0.84. Thus, $u_\perp$ points along a direction that carries essentially no rung difference to read, but along which the downstream answer is highly sensitive.

C.4 Reading and writing can be separated

The first-order model suggests that a direction need not read and write through the same component. We test this directly by separating the direction used to measure the coordinate from the direction along which the resulting displacement is written. We continue to read along the trained direction u, but write the exchanged scalar along an arbitrary vector w:

$$h_a \leftarrow h_a + (\tilde{z}_a - z_a)\, w, \qquad h_b \leftarrow h_b + (\tilde{z}_b - z_b)\, w, \qquad z_a = \langle h_a, u \rangle, \quad \tilde{z}_a = \langle \tilde{h}_a, u \rangle. \tag{19}$$

Setting $w = u$ recovers the ordinary interchange of Equation 12. By Section C.2, however, the scalar $\tilde{z}_a - z_a$ is determined almost entirely by the in-span component $u_\parallel$, because $u_\perp$ carries essentially no rung difference to read. Equation 19 therefore lets us hold the read fixed while asking which part of u is responsible for writing the change to the downstream computation.

Table 9 shows a sharp asymmetry. Reading along the full trained direction but writing only along its out-of-span component, $w = u_\perp$, preserves almost all of the trained direction's effect: the reversal rates are 0.88, 0.95, and 1.00, compared with 0.94, 0.98, and 1.00 when both reading and writing use u. In contrast, writing only along the in-span component, $w = u_\parallel$, reduces the effect to 0.17, 0.11, and 0.02. Thus, the component of u that contains the rung information is not the component through which the intervention most effectively changes the answer.

The out-of-span effect is also direction-specific. Replacing $u_\perp$ by a random direction outside S of the same norm reduces the reversal rate to 0.11, 0.19, and 0.02, and writing only along such a random

direction has essentially no effect. Conversely, the answer gradient used as both the read and write direction is also weak, because although the answer is sensitive along it, there is little rung difference to read there. These controls separate the two requirements of the interchange: the coordinate is read from the centroid space, while its most effective write is along a particular direction outside that space.

Finally, exchanging all coordinates in the centroid space at once still falls short of the trained one-dimensional direction, reversing 0.62, 0.87, and 0.88 of held-out trials. The additional power of u therefore does not come simply from recovering more of the rung geometry inside S; it comes from combining that readable variation with an out-of-span direction that efficiently steers the answer.

C.5 Approximating the trained direction without training

Sections C.2–C.4 suggest a simple interpretation of the trained direction: its component inside the centroid space reads which rung is present, while its component outside that space points towards a direction to which the answer is highly sensitive. This suggests that much of u should be recoverable directly from these two ingredients, without optimizing an intervention.

The endpoint direction $\hat{v}$ is deliberately constructed from only the first and last rungs. If all rung centroids are available, a more complete estimate of the ordinal variation inside S is the least-squares direction of the centroids on rank,

$$r = \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right) (c_i - \bar{c}), \qquad \hat{r} = \frac{r}{\|r\|}. \tag{20}$$

Equivalently, $\hat{r}$ is the direction obtained by fitting each coordinate of the centroids linearly against rank. It is close to the endpoint direction, with median cosine similarities of 0.85, 0.85, and 0.80 on Llama-3.1-8B, Mistral-7B, and Qwen2.5-32B. The in-span part of the trained direction is closer still to $\hat{r}$, with cosine similarities of 0.95, 0.96, and 0.94, compared with 0.82–0.87 for $\hat{v}$ (Figure 13). This shows that $u_{\|}$ follows the ordinal variation of the centroids rather than some unrelated direction inside S. Unlike $\hat{v}$, however, $\hat{r}$ uses the ranks of all rungs and therefore cannot serve as the held-out representation test of Section 4.1; for that reason the main analysis continues to use the endpoint direction.

The out-of-span component can be supplied directly by the answer gradient. Combining the two parts gives the training-free construction of Equation 10,

$$d_{\text{end}} = \frac{\hat{v} + \hat{\gamma}_{\perp}}{\|\hat{v} + \hat{\gamma}_{\perp}\|}.$$

On the held-out surface form this direction reverses 0.68, 0.87, and 0.78 of correctly answered trials, compared with 0.26, 0.35, and 0.36 for $\hat{v}$ alone. Replacing $\hat{\gamma}_{\perp}$ by a random out-of-span direction removes this gain. Using the rank direction instead,

$$d_{\text{rank}} = \frac{\hat{r} + \hat{\gamma}_{\perp}}{\|\hat{r} + \hat{\gamma}_{\perp}\|},$$

raises the reversal rates further to 0.82, 0.88, and 0.88. Thus, improving the estimate of the read direction helps, but the large gain over the purely in-span directions comes from adding the specific out-of-span direction supplied by the answer gradient.

The first-order model also explains why an optimized one-dimensional direction mixes the two components rather than choosing either one alone. Consider, for intuition, an in-span read direction $\hat{r}$ and the orthogonal write direction $\hat{\gamma}_{\perp}$, and let

$$d(\theta) = \cos\theta\, \hat{r} + \sin\theta\, \hat{\gamma}_{\perp}. \tag{21}$$

Because the rung difference lies inside S, the amount read along $d(\theta)$ scales approximately as $\cos\theta$. Because the answer is most sensitive along $\hat{\gamma}_{\perp}$, its sensitivity scales approximately as $\sin\theta$. In the idealised decomposition in which the read direction has negligible out-of-span sensitivity and $\hat{\gamma}_{\perp}$ carries no rung difference, Equation 17 therefore predicts that the dominant cross-term scales as

$$\cos\theta\, \sin\theta,$$

which is zero at either extreme and maximal at $45°$. In practice neither extreme is exactly zero: the answer is not entirely insensitive inside S, so $\hat{v}$ alone still reverses 0.26 to 0.36 of the correctly answered trials, and γ is not entirely outside S. This gives a simple reason for the substantial norm that the trained direction places on each side of the centroid space.

A more general closed-form version of the same first-order calculation uses the trial activation differences together with their answer gradients. By Equation 16, the first-order change of the answer towards the mirror's answer under an interchange along a unit direction d is $d^\top M d$, with

$$M = \frac{1}{2N} \sum_{t=1}^{N} s_t \sum_{o \in \{a,b\}} \left(x_{o,t} \nabla_{o,t}^\top + \nabla_{o,t} x_{o,t}^\top \right), \qquad d_{\mathrm{FO}} = \arg\max_{\|d\|=1} d^\top M d, \tag{22}$$

where $x_{o,t} = \tilde{h}_{o,t} - h_{o,t}$ is the difference between the mirror and the trial at option o, $\nabla_{o,t}$ the gradient of L at that option, and s_t the sign of the change of L from the trial to its mirror, so that every trial counts in the direction of its mirror's answer. The maximiser d_{FO} is the leading eigenvector of M. Computed on the two fitting surface forms without optimizing an interchange, it has median cosine similarity 0.86–0.88 with u and reverses 0.86, 0.88, and 0.83 of held-out trials (Table 8). The trained direction can therefore be approximated closely from the quantities identified above: a direction that reads the ordinal variation of the rungs and a direction that efficiently writes to the downstream answer.

Together, these results explain why DAS substantially outperforms the endpoint axis without making it a better description of the natural representation. The endpoint direction isolates the ordinal coordinate using only the two extreme rungs, whereas the optimized direction combines such a readable coordinate with a direction chosen for its downstream causal leverage.

	Llama-3.1-8B	Mistral-7B	Qwen2.5-32B
ladders	11	9	8
Share of correctly answered trials reversed on the held-out surface form, median over ladders			
trained direction	0.94	0.98	1.00
rank over all rungs plus the gradient's out-of-span part	0.82	0.88	0.88
closed-form optimum of the first-order model	0.86	0.88	0.83
endpoint direction $\hat{v}$ plus the gradient's out-of-span part	0.68	0.87	0.78
rank over all rungs	0.35	0.43	0.57
rank inside four centroid components	0.47	0.50	0.57
endpoint direction $\hat{v}$	0.26	0.35	0.36
gradient of the answer logits	0.24	0.30	0.26
whitened endpoint direction	0.17	0.11	0.05
rank over all rungs plus a random out-of-span direction	0.14	0.15	0.08
$\hat{v}$ plus a random out-of-span direction	0.11	0.10	0.02
random directions	0.01	0.01	0.00
Geometry of the trained direction, median over ladders			
cosine between training starts	0.92	0.98	0.83
cosine with $\hat{v}$	0.62	0.61	0.54
squared norm in four centroid components	0.46	0.48	0.34
the same for $\hat{v}$	0.94	0.91	0.89
squared norm in the space of all rung centroids	0.49	0.53	0.37
cosine of its in-span part with $\hat{v}$	0.87	0.82	0.87
cosine of its in-span part with the rank direction	0.95	0.96	0.94
cosine of its out-of-span part with the gradient's	0.80	0.81	0.76

Table 8: The trained one-dimensional direction against directions obtained without training. The score is the share of the trials answered correctly before the intervention that the intervention reverses, as defined in Section 3.3. The training-free directions are fitted on the two other surface forms. A random direction has 0.001 of its squared norm in the centroid space S. The gradient's out-of-span part is the mean gradient of the answer logit difference at the option tokens, averaged over the two fitting forms, with its projection on the centroid space removed; a random out-of-span direction has cosine 0.01 with the trained direction's.

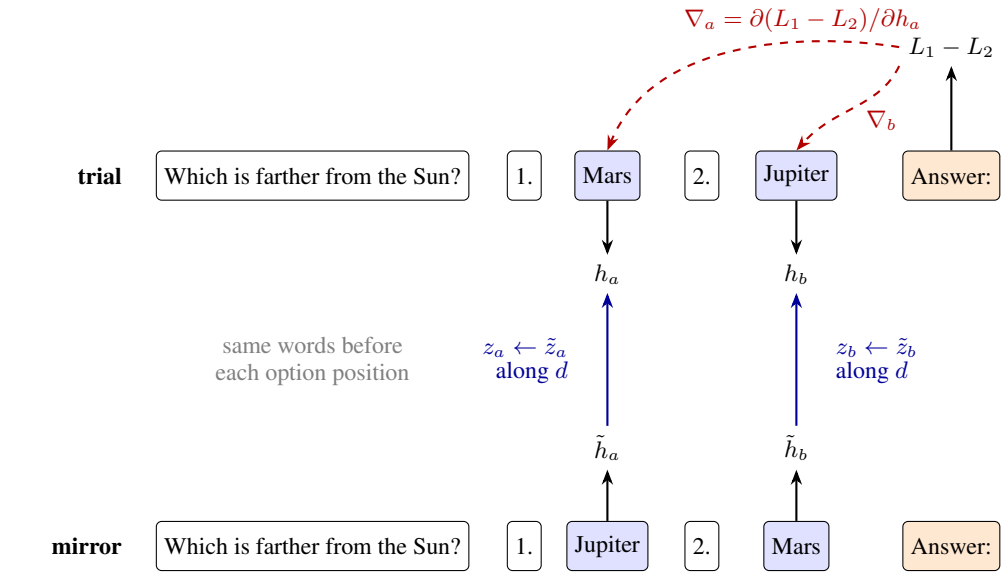
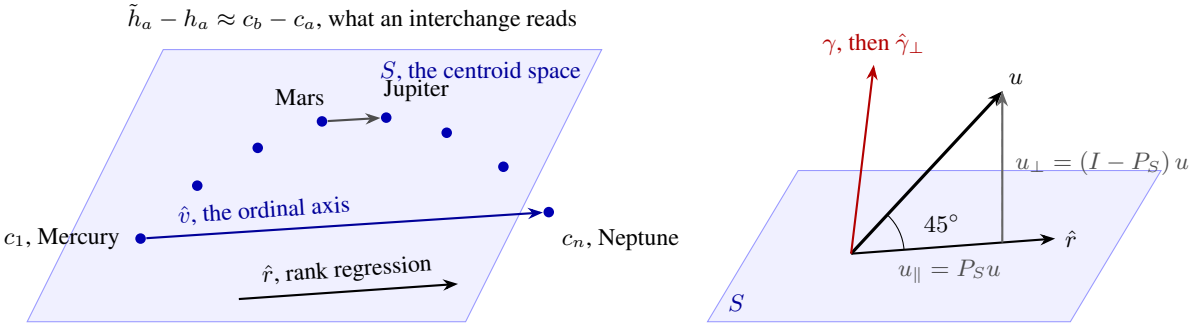


Figure 12: The activations used in Section 4.3, on one trial of the planets ladder. h_a and h_b are the residual stream at the two option tokens of the trial, at the input of the chosen layer; $\tilde{h}_a$ and $\tilde{h}_b$ are the residual stream at the same two positions of the mirror prompt, which lists the same options in the other order, so that the position of a now holds b. The interchange (Equation 12, blue) gives each option, along the direction d, the coordinate that the other option has at that position in the mirror, and leaves every other direction untouched. The gradients ∇_a and ∇_b (red) are those of the answer's logit difference with respect to the two option activations of the trial, from one backward pass.



(a) inside the centroid space

(b) the trained direction and the gradient

Figure 13: Figure 5 with the rank direction $\hat{r}$ of Equation 20 added. (a) The rung centroids c_i lie in the centroid space S (shaded), here on an arc. Inside S are the ordinal axis $\hat{v}$, from the first to the last centroid; the rank direction $\hat{r}$, the least-squares direction of the centroids on the rank (Equation 20); and the difference between a trial and its mirror at an option token, which is what any interchange reads. (b) The trained direction u leaves the plane. Its projection $u_\parallel$ lies along $\hat{r}$ and its remainder $u_\perp$ is orthogonal to every centroid difference. The gradient of the answer γ (red) is almost entirely outside S, and its outside part $\hat{\gamma}_\perp$ points along $u_\perp$ (cosine 0.8). The activations do not vary along $u_\perp$, so it reads nothing, and the answer is most sensitive along it, so it is where a write acts.

read along	write along	Llama-3.1-8B	Mistral-7B	Qwen2.5-32B
One direction, read and written (Equation 12)				
u	u	0.94	0.98	1.00
$\hat{v}$	$\hat{v}$	0.26	0.35	0.36
$\hat{\gamma}$, the gradient itself	$\hat{\gamma}$	0.24	0.30	0.26
$(\hat{v} + \hat{\gamma}_\perp)/\sqrt{2}$, Equation 10	the same	0.68	0.87	0.78
$(\hat{v} + \hat{n})/\sqrt{2}$	the same	0.11	0.10	0.02
random	random	0.01	0.01	0.00
Read along u, write along w (Equation 19)				
u	$w = u_\perp$	0.88	0.95	1.00
u	$w = u_\parallel$	0.17	0.11	0.02
u	$w = u_\parallel + \|u_\perp\| \hat{n}$	0.11	0.19	0.02
u	$w = \|u_\perp\| \hat{n}$	0.02	0.02	0.00
Every direction of S at once				
S	S	0.62	0.87	0.88

Table 9: Share of the held-out trials answered correctly before the intervention that the intervention reverses (Section 3.3), median over ladders, for each choice of what is read and what is written. Every direction is built on two surface forms from slot-1 centroids and scored on the third form. Hats denote unit vectors; $\hat{n}$ is a random unit direction outside S, averaged over two draws.