LATENT STRUCTURE, FAILED SELECTION: REPRESENTATION–USE GAPS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Performance gaps across language models are commonly attributed to scale, data, and training, yet benchmark accuracy alone cannot distinguish what models represent from how effectively they use those representations. Across 18 base models from six model families, we examined internal structure and forced-choice recall for 5 relation types: ordinal relations, part–whole, category–property, abstract–concept, and attribute–object. Ordinal structure was highly recoverable across model sizes, whereas forced-choice recall increased with size. Erasing ordinal directions selectively impaired matching judgments for temperature, size, quality, and anger after comparison with overlap matched random controls; likelihood showed no detectable effect because baseline recall was near chance in most models. Mapped hidden-state transfer of six novel words spanning temperature, size, and quality outperformed sham injections in 76.8% of 306 ordered sender–receiver pairs. To understand why small models fail, we decomposed answer preferences into question-responsive evidence and question-invariant preference, whose magnitude we call stubbornness. In small models, stubbornness exceeded question-responsive evidence, while signed question-invariant preference was more strongly aligned with lexical priors. To test how question-responsive information reaches the answer, we blocked direct attention from the answer position to the candidate words. This reduced evidence, increased stubbornness, and strengthened its alignment with these baseline word preferences. Because stubbornness is invariant to question polarity, symmetrised scoring cancels this preference without training. Applying the same invariance principle to answer positions improved 126 of 162 model benchmark combinations across nine multiple-choice benchmarks, with gains of up to 23.7 percentage points. The four non-ordinal relation types also showed strong internal structure but only small gains from symmetrisation. Together, these results show that models can represent task-relevant relational structure that fails to control their answers, and that some performance gains with scale arise from using existing representations more effectively rather than from acquiring those representations alone.

1 INTRODUCTION

Language-model performance improves with scale, which is often interpreted as models acquiring more knowledge(Kaplan et al., 2020; Hoffmann et al., 2022). However, benchmark accuracy alone cannot establish this. Wrong answers stem from either missing representations or a failure to use them in final predictions. These failures imply different accounts of what changes with scale.

Internal representation studies recover syntactic, semantic, and conceptual structure from hidden states (Alain & Bengio, 2016; Hewitt & Manning, 2019; Park et al., 2023; Gurnee & Tegmark, 2024; Marks & Tegmark, 2023; Tigges et al., 2024), and latent knowledge methods can recover information from activations even when overt behavior is unreliable (Burns et al., 2022). Yet decodability does not show that a representation contributes to an answer. Behavioral studies show that knowledge extraction can change with question direction, candidate order, answer position, and scoring procedure (Berglund et al., 2024; Zheng et al., 2024; Pezeshkpour & Hruschka, 2024; Zhao et al., 2021). Such results demonstrate selection failures, but do not by themselves establish whether the missing answer was represented internally. Therefore, distinguishing representation from use requires measuring both within the same task. Prior work has studied internal representa-

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.

tions, causal interventions, and output biases largely as separate problems. Here we combine these measurements within the same judgments to test where relational knowledge is represented, whether it causally contributes to the answer, and why that contribution can still fail to determine the final choice. Ordinal relations provide a tractable setting. Concepts such as temperature and size have an internal order that can be measured in hidden states and queried in opposite directions. Thus, the same candidate pair can test whether an order is represented and whether it controls the answer when the requested pole reverses. Four non-ordinal relation types—part–whole, category–property, abstract–concept, and attribute–object—provide a complementary comparison because they lack the same polarity reversal. Our mechanistic analysis focuses on ordinal forced-choice selection; the non-ordinal relations test its scope.

This setup raises three questions:

- **Is relational structure already represented when forced-choice recall fails?**
- **What prevents represented structure from controlling the answer?**
- **Can this selection failure be corrected without changing model parameters?**

We show that ordinal relations are recoverable from hidden states across model sizes even when forced-choice accuracy differs sharply. Erasing ordinal directions selectively impaired matching judgments for 4 of the 5 ordinal scales, while mapped hidden states supported ordinal judgments across models. We further identify a question-invariant answer preference that limits small models and show that cancelling this preference improves relational judgments and multiple-choice performance without changing model parameters.

2 Related Work

2.1 Measuring and testing internal representations

A central question in this work is whether information can be present in hidden states even when a model answers incorrectly. Linear probes have recovered syntactic and conceptual information (Alain & Bengio, 2016; Hewitt & Manning, 2019), including approximately linear representations of space and time (Gurnee & Tegmark, 2024), truth (Marks & Tegmark, 2023), sentiment (Tigges et al., 2024), and other semantic variables (Park et al., 2023). Recoverability alone, however, does not show that a representation affects an answer. Causal mediation, model editing, and representation erasure address this by changing internal states and measuring behavioral effects (Vig et al., 2020; Meng et al., 2022; Belrose et al., 2023; Ravfogel et al., 2020; Tigges et al., 2024).

2.2 Knowledge access and answer selection

Models do not always express information they contain. Latent knowledge can be recovered from activations when overt answers are unreliable (Burns et al., 2022); learned relations can fail under reversed queries (Berglund et al., 2024); and stored factual knowledge can remain difficult to extract (Allen-Zhu & Li, 2023). Multiple choice predictions also vary with option order, answer position, and surface form (Zheng et al., 2024; Pezeshkpour & Hruschka, 2024; Robinson et al., 2022; Holtzman et al., 2021), while calibration and permutation based scoring can reduce some of these effects (Zhao et al., 2021; Zheng et al., 2024).

Most closely related work. Orgad et al. (2025) show that hidden states can encode a correct answer when the model generates an incorrect one, and Gekhman et al. (2025) find factual knowledge that is stronger in intermediate computations than in output probabilities. Calderon et al. (2026) ask whether factual errors reflect missing knowledge or failed recall and identify recall as a major bottleneck. Their behavioral framework does not locate the relevant representation in hidden states or test whether that representation causally contributes to the failed answer.

Park et al. (2026) are especially close mechanistically: they identify distinct knowledge and prediction subspaces in residual streams for multiple choice questions and improve accuracy by aligning them at inference time. A concurrent study traces a related gap in logical verification to a miscalibrated decision threshold (Villuri et al., 2026). Our study instead begins with independently defined

Table 1: Stimulus groups.

GROUP	CONTENTS	SIZE
Ordinal	temperature, size, quality, likelihood, anger	33 words, 97 pairs
Reference	months, weekdays, numbers	31 words, 153 pairs
Non-ordinal	part-whole, category-property, abstract-concept, attribute-object	1,520 probes

semantic relations, measures where they are represented, tests their causal contribution by erasure and cross model transfer, and then asks why the represented relation can still fail to determine the answer.

Thus, prior work shows that poor output does not imply absent knowledge, but it leaves open whether a task relevant semantic representation causally contributes to the same failed judgment and what prevents it from controlling the answer. We test both within the same tasks.

2.3 Word frequency and attention in answer selection

Factual accuracy and model predictions depend in part on training frequency (Kandpal et al., 2023; Wei et al., 2021; Tjuatja et al., 2025). These findings motivate testing whether a model's baseline probability for candidate words contributes to forced-choice errors. Attention provides a complementary route: reweighting selected attention heads toward input spans can change model behavior (Zhang et al., 2024). We intervene on attention from the answer position to the candidate words.

2.4 Mapping representations across models

Simple transformations can align hidden representations across independently trained networks (Lenc & Vedaldi, 2015; Bansal et al., 2021; Moschella et al., 2022), and larger models can show greater representational convergence (Huh et al., 2024). We use cross model mapping as a behavioral test: if a mapped state from one model changes another model's judgment, the transferred state contains information the receiver can use.

3 Method

3.1 Measurements

We studied 18 pretrained base models from 6 families spanning 1B–35B parameters (Appendix A.1). The main study used 5 relation types: 1 ordinal type and 4 non-ordinal types (Table 1). We averaged 32 sentence-frame residual stream states at each target word's final subtoken across layers (Fig. 1A). We define **ordinal structure** as how well a scale's known rank order is recoverable from these states. At each layer, one word was held out, a rank weighted direction was fit from the remaining words, and the held out word was projected onto it. We define **leave-one-out ordinal agreement** as the fraction of word pairs whose held out projections preserve the known order. We use the best layer because the question is whether the order is recoverable anywhere in the model; the rank shuffle null repeats the same layer selection. Non-ordinal relations use related-versus-foil proximity and reconstruction (Fig. 1B; Appendix A.4–A.5).

We define **forced-choice recall** as accuracy when both candidates appear in the prompt and the correct candidate receives higher continuation probability. The 97 ordinal pairs were each tested in two candidate orders, two question directions, and three templates, giving 1,164 prompts. Multi token candidates use summed subtoken log probabilities.

For pair $p = (w_{\text{low}}, w_{\text{high}})$ and prompt q, the **answer margin** is

$$\Delta_p(q) = \log P(w_{\text{high}} \mid q) - \log P(w_{\text{low}} \mid q). \tag{1}$$

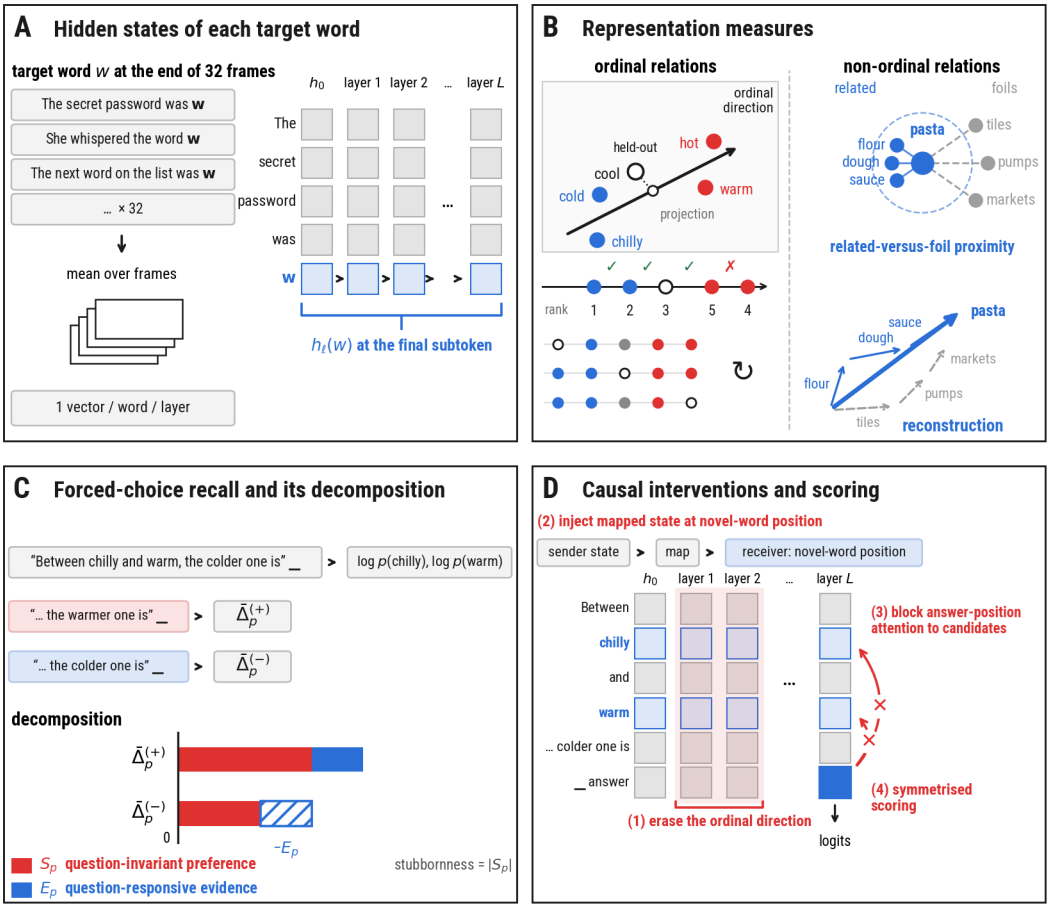

Figure 1: **Experimental framework. (A)** Hidden states were read at each target word's final subtoken and averaged across 32 neutral frames within each layer. **(B)** Ordinal structure was measured by leave-one-out ordinal agreement; non-ordinal relations by proximity and reconstruction. **(C)** Opposite question directions separated the answer margin into question-responsive evidence and question-invariant preference. **(D)** We erased ordinal directions, transferred mapped hidden states, blocked attention from the answer position to candidate words, and applied symmetrised scoring.

Let $\bar{\Delta}_p^{(+)}$ and $\bar{\Delta}_p^{(-)}$ be the mean margins for higher- and lower-pole questions. We define

$$E_p = \frac{\bar{\Delta}_p^{(+)} - \bar{\Delta}_p^{(-)}}{2}, \qquad S_p = \frac{\bar{\Delta}_p^{(+)} + \bar{\Delta}_p^{(-)}}{2}. \tag{2}$$

Question-responsive evidence is E_p. **Question-invariant preference** is the signed quantity S_p. **Stubbornness** is its magnitude, $|S_p|$. Both question directions are correct when $E_p > |S_p|$ (Fig. 1C).

We define a word's **lexical prior** as its mean log-probability after three neutral contexts: *The answer is*, *The word is*, and *She said the word*. Multi token words again use summed subtoken log-probabilities. Pairwise lexical prior differences are compared with signed S_p.

3.2 Interventions and Scoring

For erasure, we define the **ordinal direction** at each layer as the normalized rank weighted covariance between word states and standardized ranks. For each of the 5 ordinal scales, we removed every token state's projection onto that scale's ordinal direction through the first 60% of decoder blocks,

$$h \leftarrow h - (h^\top v)v,$$

and evaluated all 5 scales. Each targeted erasure was compared with 5 random directions per scale matched at every layer for overlap with the span of the gradable word input embeddings. **Erasure damage** is baseline accuracy minus intervention accuracy. We define **excess erasure damage** as targeted erasure damage minus overlap matched control damage. Temperature only depth and localization checks are reported in Appendix A.7.

For cross model transfer, 6 invented novel words covered both poles of temperature, size, and quality, detailed in Appendix A.3. A **sender model** received two definition sentences; a **receiver model** did not. A ridge map between their hidden spaces was estimated from 127 shared anchor words. The invented novel word's sender state was mapped, rescaled to the receiver norm, and written at the receiver's novel word position before comparison with familiar scale words. A norm matched random vector was the sham control. We evaluated all $18 \times 17 = 306$ ordered sender-receiver pairs.

To test attention, we blocked attention from the final answer position to all candidate subtokens at every layer while leaving the candidates visible elsewhere. A causal only custom mask had to reproduce the normal top-1 prediction and correlate at least 0.999 with the normal logit vector; Qwen3.5-35B-A3B-Base failed this check ($r = 0.9987$), leaving 17 models.

Finally, we define **symmetrised scoring** for ordinal pairs as

$$\Delta_p^{\text{sym}} = \frac{\bar{\Delta}_p^{(+)} - \bar{\Delta}_p^{(-)}}{2} = E_p, \tag{3}$$

Thus, S_p does not contribute. For multiple-choice benchmarks, each option was cycled through every answer position and its label log-probabilities were summed across orderings. Benchmark details, statistics, and pre-specified criteria are in Appendix A.11. Model-size analyses use $\log_{10}$ total parameter count.

4 Results

4.1 Ordinal structure is recoverable even when recall fails

Mean leave-one-out ordinal agreement ranged from 0.833 to 0.943 and did not correlate significantly with model size ($r = -0.155$, $p = 0.54$; Fig. 2A), whereas forced-choice recall on the same 1,164 prompts ranged from 0.545 to 0.784 and increased with model size ($r = +0.778$, $p = 0.0001$). The latter correlation was significantly stronger (Steiger $p = 0.0025$) (Steiger, 1980). The supplementary 3,000 prompt set gave the same result: recall ranged from 0.532 to 0.777 and increased with model size ($r = +0.763$, $p = 0.0002$). OLMo2-1B had leave-one-out ordinal agreement of 0.867 but forced-choice recall of 0.545 on the 5 gradable scales.

4.2 Erasing and transferring ordinal representations changes recall

In a 5 scale extension, we erased each ordinal direction and evaluated all five scales against overlap matched random controls (Fig. 2B). Matching-scale erasure damage exceeded control damage for temperature (0.165 versus 0.048; excess $+0.116$), size (0.165 versus 0.044; $+0.121$), quality (0.146 versus 0.053; $+0.093$), and anger (0.109 versus 0.035; $+0.074$), all significant after Holm correction. Likelihood showed no detectable excess damage (0.022 versus 0.020; $+0.002$, $p = 0.586$), with baseline recall below 0.55 in 11 of 18 models. Mean excess damage was $+0.081$ on matching scales and -0.011 off diagonal, with larger matching-scale damage in 17 of 18 models (Wilcoxon signed-rank $p = 1.53 \times 10^{-5}$). This group pattern was not uniform across models (Appendix B).

Cross model transfer used 6 invented words spanning temperature, size, and quality. Mapped state injection exceeded the norm matched random injection by more than 0.02 in 76.8% of 306 ordered sender and receiver pairs, while 4.9% decreased by more than 0.02 (Fig. 2C). Mean gain was $+0.086$ within model families and $+0.073$ across families (Welch $p = 0.17$).

4.3 Stubbornness exceeds question-responsive evidence most often in small models

Question-responsive evidence E ranged from 0.256 to 1.418 and increased with model size ($r = +0.751$), whereas stubbornness $|S|$ ranged from 0.633 to 1.124 and decreased with size

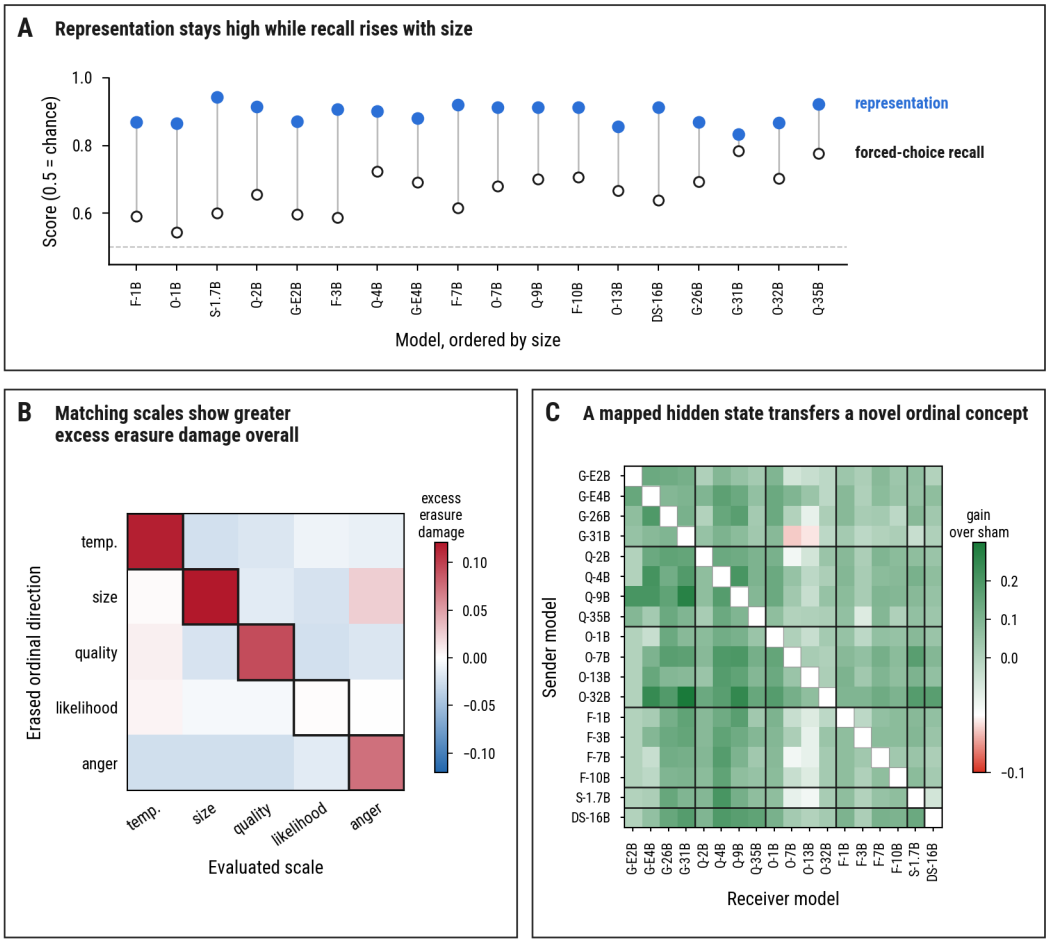

Figure 2: **Ordinal structure, recall, and causal interventions. (A)** Ordinal agreement and recall by model size. **(B)** Mean excess erasure damage across 5 erased directions and 5 evaluated scales. **(C)** Accuracy difference between mapped state and norm matched random injections across 306 sender and receiver pairs.

($r = -0.655$; Fig. 3A). Stubbornness exceeded question-responsive evidence in all 5 models with at most 2B parameters, three of seven between 2B and 13B, and one of six with at least 13B.

The correlation between signed question-invariant preference S_p and lexical prior difference ranged from -0.23 to $+0.63$ (median $+0.373$), with means of $+0.530$ at at most 2B parameters and $+0.158$ at at least 13B parameters. It decreased with model size ($r = -0.643$, $p = 0.004$; Fig. 3B), and only 5 of 18 models met the prespecified $r \geq +0.5$ criterion. At the pair level, absolute lexical prior difference predicted stubbornness after controlling for ordinal rank distance in all 18 models (mean partial $r = +0.204$, $p < 10^{-4}$), whereas representational separation did not (mean partial $r = -0.027$, $p = 0.45$).

4.4 Blocking attention to the candidates reduces question-responsive evidence

We report attention blocking for 17 models; Qwen3.5-35B-A3B-Base was excluded because the custom mask changed its output before candidate attention was blocked. Blocking increased stubbornness in all 17 models by 82% on average. The reduction in question-responsive evidence correlated with baseline question-responsive evidence ($r = +0.853$; Fig. 4A), while forced-choice recall remained above 0.5 in every model (mean 0.631).

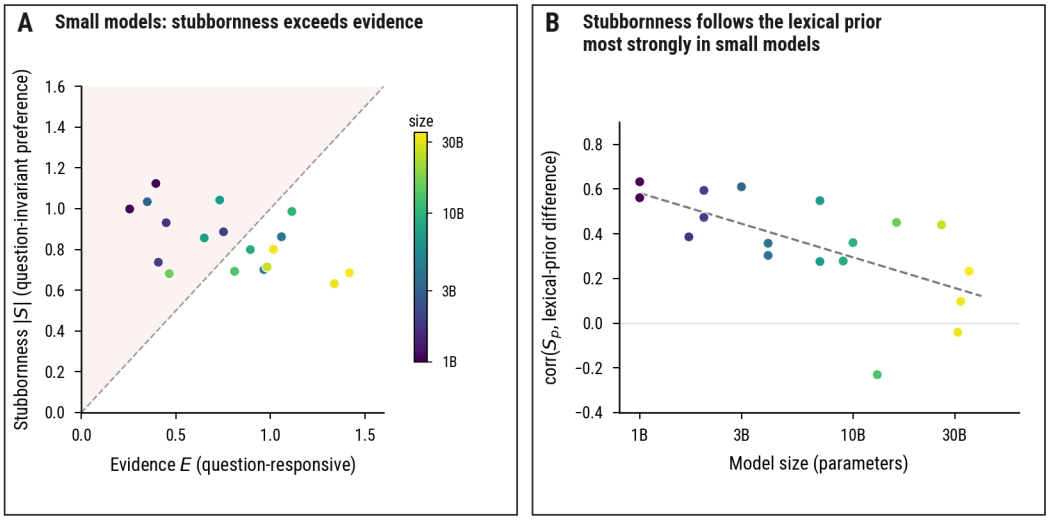

Figure 3: **Question-responsive evidence, stubbornness, and lexical prior.** **(A)** Question-responsive evidence E versus stubbornness $|S|$. **(B)** Correlation between signed S_p and lexical prior difference across 97 pairs.

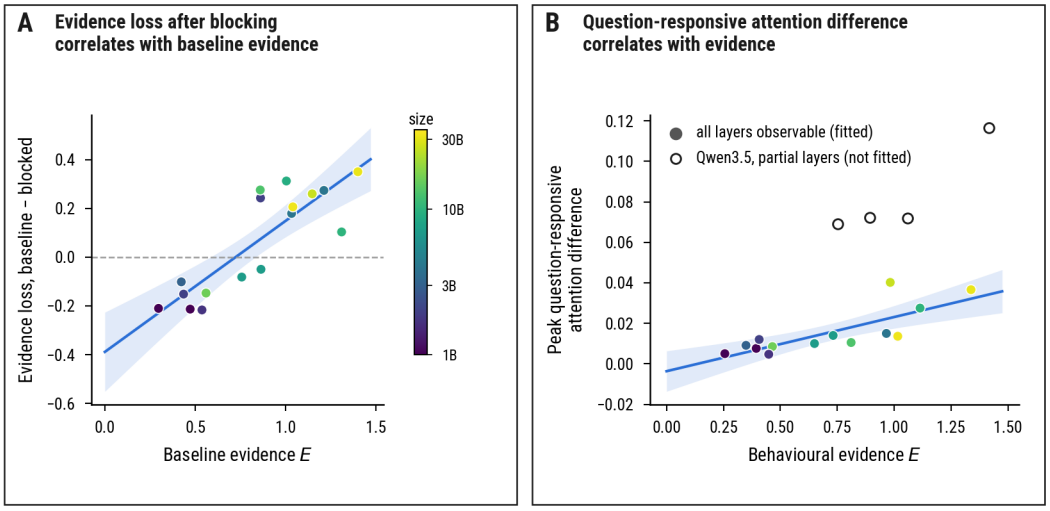

Figure 4: **Attention from the answer position to candidate words.** **(A)** Baseline question-responsive evidence versus its change after blocking attention to the candidates. **(B)** Question-responsive evidence versus the maximum question-responsive attention difference; open Qwen3.5 points have partial layer coverage and are not fitted.

Signed question-invariant preference became more aligned with lexical priors in all 17 models, with a mean correlation increase of $+0.353$ ($p = 4 \times 10^{-6}$). In all 14 models with attention observed at every layer, attention increased toward the candidate matching the queried pole. The maximum question-responsive attention difference correlated with question-responsive evidence ($r = +0.796$, $p = 0.0007$; Fig. 4B) and remained correlated after controlling for model size (partial $r = +0.601$). Qwen3.5 models are shown separately because only some layers expose attention matrices.

Additional interventions did not consistently improve recall. Attention amplification reduced mean question-responsive evidence from 0.835 to 0.235 and accuracy from 0.679 to 0.555 at the largest dose, increasing the temperature ordinal component reduced mean temperature accuracy at every gain, and removing a rank 3 bias subspace estimated from temperature gradients changed pooled

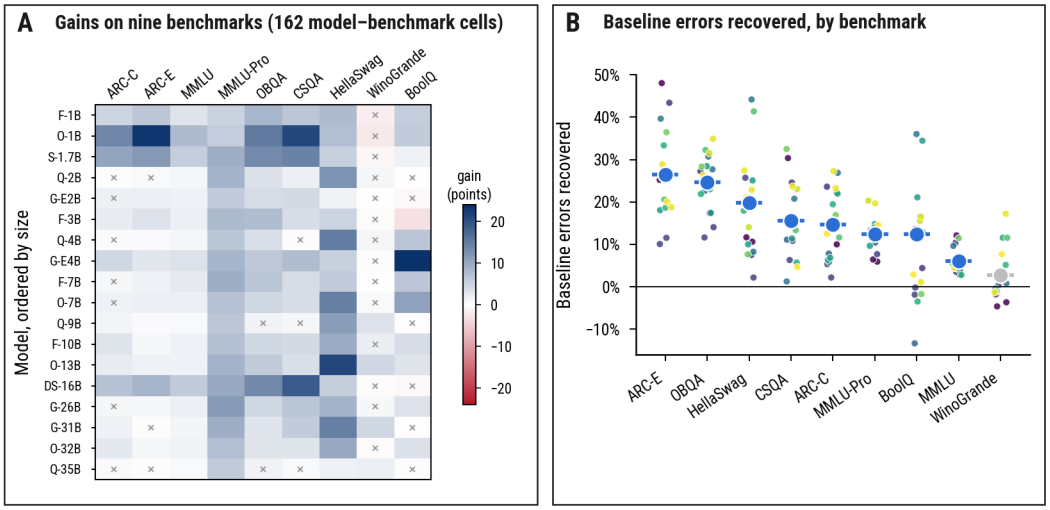

Figure 5: **Symmetrised scoring on nine multiple-choice benchmarks. (A)** Accuracy gain across 162 model and benchmark combinations; crosses indicate nonsignificant McNemar tests. **(B)** Fraction of baseline errors recovered; small points are models and large points benchmark means.

accuracy by $+0.009$ relative to random rank 3 controls ($p = 0.12$), although quality increased by $+0.012$ ($p = 0.024$; Appendix C).

4.5 Symmetrised scoring improves ordinal and multiple choice accuracy

On temperature, size, and quality, symmetrised scoring increased mean forced-choice recall from 0.686 to 0.930, a gain of 24.4 percentage points, with improvement in all 18 models ($p = 1.6 \times 10^{-12}$). Across all 5 gradable scales, the post hoc extension increased mean accuracy from 0.665 to 0.891, a gain of 22.6 points, again in all 18 models ($p = 9.3 \times 10^{-14}$). Likelihood increased by 11.1 points in 14 of 18 models ($p = 0.03$), and anger by 20.2 points in all 18 ($p = 2.3 \times 10^{-9}$).

Across nine multiple choice benchmarks, symmetrised scoring significantly improved 126 of 162 model and benchmark combinations and significantly worsened one, with a maximum gain of 23.7 points (Fig. 5A). Mean error recovery ranged from 6.1% to 26.5% on eight benchmarks and was 2.8% on WinoGrande (Fig. 5B).

4.6 Relations without opposite question directions show smaller gains

Across the four non-ordinal relation types, true related words were closer to their targets than foils for 96 to 100% of targets; for category-property relations, the count matched property foil comparison remained positive for at least 92% of targets in every model. Forced-choice recall ranged from 0.745–0.945 for part-whole, 0.690–0.840 for category-property, 0.557–0.815 for abstract-concept, and 0.700–0.912 for attribute-object relations. Candidate order symmetrisation increased mean accuracy by only 2.0–3.2 points, versus 22.6 points across the 5 ordinal scales. Abstract-concept recall reached 0.557 despite strong internal structure, so weak recall also occurs without the ordinal decomposition.

5 Discussion

Forced-choice recall increased strongly with model size, whereas recoverable ordinal structure did not. Better recall in larger models may therefore reflect better use of represented order rather than stronger ordinal structure alone. Causal interventions show that the recovered structure contributes to judgments but do not explain the scaling effect. Erasure was selective for temperature, size, quality, and anger at the group level; likelihood showed no detectable effect because baseline recall

was near chance in most models. Cross model transfer showed that mapped hidden states support judgments across temperature, size, and quality.

The answer margin decomposition gives one reason represented order can fail to control the answer. Small models more often had question-responsive evidence too weak to overcome stubbornness, suggesting that part of the scaling difference lies in how strongly the question changes the final preference.

Blocking attention from the answer position to the candidate words reduced question-responsive evidence most in models with more baseline evidence and made signed question-invariant preference more aligned with lexical priors. The informative result is therefore not simply that blocking candidate attention lowers performance. The intervention separates the component that changes with the question from the preference that remains when the question reverses. This shows that direct attention to the candidates contributes question-responsive evidence, while blocking it increases the relative influence of lexical priors.

Lexical priors explain part of question-invariant preference, especially in smaller models, but their weaker association in larger models indicates other sources of stubbornness.

Symmetrised scoring removed question-invariant preference from ordinal scoring and recovered many answers without changing the model or its hidden representations. Its smaller gains for non-ordinal relations show that the correction is most effective when opposite question directions separate question-responsive evidence from question-invariant preference. Benchmark gains extend the same principle to answer position but do not show that benchmark errors follow the ordinal decomposition.

Several limits remain. Best layer selection tests whether ordinal order is recoverable somewhere in the model, and the 5 scale erasure result is a group effect rather than evidence for functionally distinct directions in every model. Family, architecture, training, and parameter count covary, so size correlations do not isolate parameter count. Lexical priors measure model probability in neutral contexts, not training frequency. Attention matrices were available at every layer for only 14 models, and all behavioral tests provided candidate answers, so the same account may not extend to open ended generation.

6 Conclusion

Weak forced-choice recall need not imply that the relation is absent. Across 18 models, ordinal structure remained recoverable across sizes while recall increased with size. Ordinal direction erasure selectively impaired matching judgments for four of the 5 scales at the group level, while likelihood was limited by near-chance baseline recall. Cross model transfer further showed that mapped hidden states can support ordinal judgments across models. In smaller models, stubbornness exceeded question-responsive evidence; blocking attention from the answer position to the candidate words reduced this evidence and made question-invariant preference more aligned with lexical priors.

Symmetrised scoring removed question-invariant preference from the ordinal score and improved ordinal recall in all 18 models without changing model parameters. The same principle improved many multiple-choice benchmarks, but gains were much smaller without opposite question directions. Thus, accuracy gains can reflect not only what models represent but how effectively those representations control output.

AI use statement

In this work, one author used ChatGPT and Claude to assist with literature discovery, generation of synthetic stimuli and sentence frames under author-defined constraints, translation, implementation of author-defined methods, function-level code writing and debugging, figure code, and preparation of supplementary text from verified experimental outputs. Generative AI was not used to originate the research questions, conceptual framework, mathematical definitions, experimental design or parameters, statistical analyses, or interpretation of the results. Publicly available datasets and benchmarks were used without AI-generated modification.

All AI-generated stimuli were manually reviewed word by word by the authors. The authors defined the mathematical methods before implementation, and the source code produced with AI assistance was read and verified by the first author. Experimental results were exported to structured result tables and checked against the analysis code before being reported. Papers suggested during AI-assisted literature search were read in their original form, and all citations were verified by the authors. AI-assisted translations, manuscript edits, appendix text, and figure code were also reviewed by the first author. All final experimental decisions, statistical analyses, interpretations, and manuscript claims were reviewed and approved by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*, 2023.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Stickland, Tomek Korbak, and Owain Evans. The reversal curse: Llms trained on "a is b" fail to learn "b is a". In *International Conference on Learning Representations*, volume 2024, pp. 18623–18642, 2024.

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

Nitay Calderon, Eyal Ben-David, Zorik Gekhman, Eran Ofek, and Gal Yona. Empty shelves or lost keys? recall is the bottleneck for parametric factuality. *arXiv preprint arXiv:2602.14080*, 2026.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. Inside-out: Hidden factual knowledge in llms, 2025. URL https://arxiv.org/abs/2503.15299.

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

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

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT*, pp. 4129–4138, 2019. doi: 10.18653/v1/N19-1419.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7038–7051, 2021.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pp. 15696–15707. PMLR, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 991–999, 2015.

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.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations, 2025. URL https://arxiv.org/abs/2410.02707.

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.

Yoonah Park, Haesung Pyun, and Yohan Jo. Bridging the knowledge-prediction gap in llms on multiple-choice questions, 2026. URL https://arxiv.org/abs/2509.23782.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7237–7256, 2020.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering. *arXiv preprint arXiv:2210.12353*, 2022.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

James H. Steiger. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87 (2):245–251, 1980. doi: 10.1037/0033-2909.87.2.245.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Curt Tigges, Oskar J Hollinsworth, Atticus Geiger, and Neel Nanda. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 58–87, 2024.

Lindia Tjuatja, Graham Neubig, Tal Linzen, and Sophie Hao. What goes into a lm acceptability judgment? rethinking the impact of frequency and length. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2173–2186, 2025.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

Gnaneswar Villuri, Hashmath Shaik, and Alex Doboli. When do internal probes beat reading the answer? miscalibrated readouts and behavior-concealed knowledge in language models, 2026. URL https://arxiv.org/abs/2609.04582.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 932–948, 2021.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for llms. In *International Conference on Learning Representations*, volume 2024, pp. 42411–42430, 2024.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pp. 12697–12706. Pmlr, 2021.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pp. 19426–19454, 2024.

A Detailed methods

This appendix gives the implementation details for the measurements and interventions in the main Methods. We use **relation type** for the five broad groups in the main study: one ordinal relation type and four non-ordinal relation types. Within the ordinal relation type, we use **ordinal scale** for each ordered set of words. The five primary ordinal scales are temperature, size, quality, likelihood, and anger.

A.1 Models and inference

We studied 18 pretrained base models from six families: Gemma-4 (E2B, E4B, 26B-A4B, 31B), Qwen3.5-Base (2B, 4B, 9B, 35B-A3B), OLMo-2 (1B, 7B, 13B, 32B), Falcon-3 (1B, 3B, 7B, 10B), SmolLM2-1.7B, and DeepSeek-V2-Lite. For analyses of scale, model size was $\log_{10}$ of total parameter count. Mixture-of-experts models were entered by total rather than active parameter count. Inference used bfloat16 on one NVIDIA H100 with HuggingFace

Table 2: Architectures of the 18 base models. Architecture and model family are confounded, so architectural subsets are used only as robustness checks.

MODEL	PARAMS (B)	FEED-FORWARD	ATTENTION LAYERS	HEADS (Q/KV)	LAYERS	HIDDEN	VOCAB
Gemma-4-E2B	2	Dense	7 full + 28 sliding	MQA 8/1	35	1536	262k
Gemma-4-E4B	4	Dense	7 full + 35 sliding	GQA 8/2	42	2560	262k
Gemma-4-26B-A4B	26	MoE, 128 experts	5 full + 25 sliding	GQA 16/8	30	2816	262k
Gemma-4-31B	31	Dense	10 full + 50 sliding	GQA 32/16	60	5376	262k
Qwen3.5-2B-Base	2	Dense	6 full + 18 linear	GQA 8/2	24	2048	248k
Qwen3.5-4B-Base	4	Dense	8 full + 24 linear	GQA 16/4	32	2560	248k
Qwen3.5-9B-Base	9	Dense	8 full + 24 linear	GQA 16/4	32	4096	248k
Qwen3.5-35B-A3B-Base	35	MoE, 256 experts	10 full + 30 linear	GQA 16/2	40	2048	248k
OLMo2-1B	1	Dense	all full	MHA 16/16	16	2048	100k
OLMo2-7B	7	Dense	all full	MHA 32/32	32	4096	100k
OLMo2-13B	13	Dense	all full	MHA 40/40	40	5120	100k
OLMo2-32B	32	Dense	all full	GQA 40/8	64	5120	100k
Falcon3-1B	1	Dense	all full	GQA 8/4	18	2048	131k
Falcon3-3B	3	Dense	all full	GQA 12/4	22	3072	131k
Falcon3-7B	7	Dense	all full	GQA 12/4	28	3072	131k
Falcon3-10B	10	Dense	all full	GQA 12/4	40	3072	131k
SmolLM2-1.7B	1.7	Dense	all full	MHA 32/32	24	2048	49k
DeepSeek-V2-Lite	16	MoE, 64 experts	all full	MHA 16/16	27	2048	102k

A.2 STIMULI AND HIDDEN-STATE EXTRACTION

Ordinal stimuli. The primary ordinal set contained 33 gradable adjectives on five scales: temperature (8 words), size (7), quality (8), likelihood (5), and anger (5) (Table 3). Each word had a fixed integer rank within its scale; words with ambiguous ordering were removed rather than assigned an uncertain rank. Three conventionally ordered reference scales, months (12), weekdays (7), and numbers (12), were used only in supplementary analyses. Thus, the primary representation and recall analyses use the same five gradable scales.

Non-ordinal relation types. The four non-ordinal sets contained 25 part-whole targets, 25 category-property targets, 25 abstract-concept targets, and 20 attribute-object targets. Each target was paired with true related words and foils drawn from other targets in the same relation set. Foils were count-matched for part-whole, abstract-concept, and attribute-object. Each category-property target had one true hypernym and 4–5 true properties, paired with 3–4 foil hypernyms and a count-matched set of foil properties. Examples include *pasta* with constituents such as *flour* and *sauce*, and *wooden* with compatible nouns such as *fence* and *bench*. Complete sets appear in Tables 7–10.

Sentence frames and state extraction. Each of the 159 target and reference words appeared at the end of 32 neutral sentences: 16 frames shared across parts of speech and 16 frames matched to the word's part of speech (Tables 5 and 6). For every sentence, we recorded the residual-stream state at the target word's final subtoken at the embedding output and after every decoder layer. We then averaged the 32 states within each layer:

$$h_w^{(\ell)} = \frac{1}{32} \sum_{f=1}^{32} h_{w,f}^{(\ell)}. \tag{4}$$

Here $\ell = 0$ denotes the embedding output and $\ell = 1, \ldots, L$ the decoder layers. States were never averaged across layers. Non-ordinal candidate words were extracted with the same 32-frame procedure. A tokenization check verified that the intended word ended every sentence; a mismatch aborted the run.

Forced-choice prompts. Ordinal probes placed both candidates in the prompt, for example, "Between freezing and cold, the colder one is". Every within-scale pair was tested in two candidate orders, with a higher-pole and a lower-pole question, and with three prompt templates. The five primary scales contain $28 + 21 + 28 + 10 + 10 = 97$ pairs, giving $97 \times 2 \times 2 \times 3 = 1{,}164$ prompts. Adding the three reference scales gives 3,000 prompts. The four non-ordinal relation types used one true related candidate and one foil in both candidate orders and contributed 400, 400, 400, and 320 probes, respectively. Their questions do not have a paired higher-versus-lower direction.

Table 3: Ordinal scales. Words are listed in authored rank order (rank 1 first). The five gradable scales (33 words) are the primary set; the three conventionally ordered reference scales (31 words) enter only the supplementary 3,000-probe analyses. The comparative phrases are the ones inserted into the probe templates for the lower- and higher-pole questions.

SCALE	WORDS	COMPARATIVES (LOW / HIGH)	WORDS IN RANK ORDER
temperature	8	colder / hotter	freezing, cold, chilly, cool, lukewarm, warm, hot, boiling
size	7	smaller / bigger	microscopic, tiny, small, big, huge, gigantic, colossal
quality	8	worse / better	terrible, bad, mediocre, decent, good, great, excellent, perfect
likelihood	5	less likely / more likely	impossible, unlikely, possible, probable, certain
anger	5	less angry / angrier	annoyed, irritated, angry, furious, enraged
months	12	earlier in the year / later in the year	January, February, March, April, May, June, July, August, September, October, November, December
weekdays	7	earlier in the week / later in the week	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
numbers	12	smaller / larger	one, two, three, four, five, six, seven, eight, nine, ten, eleven, twelve

A.3 Complete stimulus materials

Tables 3–13 list every stimulus used in the paper, exported directly from the files read by the analysis scripts.

Ordinal scales and probes. Table 3 gives the eight scales with their words in rank order and the comparative phrases; Table 4 gives the three probe templates. The five gradable scales contain $28 + 21 + 28 + 10 + 10 = 97$ distinct-rank pairs and the three reference scales $66 + 21 + 66 = 153$, so the probe file holds $97 \times 12 = 1{,}164$ primary and $250 \times 12 = 3{,}000$ total prompts.

Sentence frames. Tables 5 and 6 list the 48 frames. Among the 159 target and reference words, the 16 shared frames are shared by all words, the 16 noun frames are used for the 106 nouns (composite, concrete, abstract, and reference-scale words), and the 16 adjective frames for the 53 adjectives (gradable and non-gradable). Each of these words is observed in exactly 32 sentences. Related and foil candidate words were extracted separately in the same 16 meta plus 16 noun frames.

Non-ordinal relation sets. Tables 7–10 list every target with its related words and foils; Table 11 gives the probe templates and the pairing rule. Foil counts match the related set for part-whole, abstract-concept, and attribute-object relations. In category-property relations, property foils are count-matched to the true properties, while the single true hypernym is contrasted with 3–4 foil hypernyms. All foils are related words of other targets in the same set, so a foil is a real member of some relation and cannot be rejected on lexical grounds alone.

Cross-model transfer materials. Table 12 gives the six novel words, their definition sentences and candidate pairs, and Table 13 the 127 anchor words. The four carrier sentences used to read the novel word state from the sender model are “She whispered the word w”; “He wrote down the word w”; “The students repeated the word w”; “Everyone soon knew the word w”.

Lexical-prior contexts. The three neutral contexts are “The answer is”, “The word is”, and “She said the word”. A word’s lexical prior is the mean sequence log-probability of that word after these three contexts. For a multi-subtoken word, the sequence log-probability is the sum of its conditional subtoken log-probabilities.

Table 4: Ordinal probe templates. For every distinct-rank pair of a scale, {a}/{b} take the two words in both orders, {comp} takes the lower-pole or higher-pole comparative of Table 3, and all three templates are used: $2 \times 2 \times 3 = 12$ prompts per pair. The prompt ends before the answer; both candidates are scored as continuations of the same prompt.

ID	TEMPLATE
0	Between {a} and {b}, the {comp} one is
1	Comparing {a} and {b}, the {comp} of the two is
2	Of {a} and {b}, the {comp} one is

Table 5: The 16 shared frames, used for every word regardless of part of speech. w marks the target word, which ends the sentence; hidden states are read at its final subtoken.

ID	FRAME
0	The first word that came to her mind was w
1	The next word on the list was w
2	She whispered the word w
3	He couldn't stop thinking about the word w
4	She asked him to spell the word w
5	The message contained only one word: w
6	The word she kept repeating was w
7	The secret password was w
8	He underlined the word w
9	Nobody in the class could pronounce the word w
10	They spent an hour debating the meaning of the word w
11	The last word he typed before leaving was w
12	Scrawled across the top of the page was a single word: w
13	Her favorite word in the language had always been w
14	The editor deleted every word except w
15	The solution to the riddle was the word w

A.4 Leave-one-out measurement of ordinal structure

We define **ordinal structure** as the extent to which the known order of words on an ordinal scale can be recovered from hidden states without using the model's answers. To prevent a word from defining the direction used to score it, we used leave-one-out estimation.

For a scale with ranks $r_1, \ldots, r_n$ and frame-averaged states $h_1^{(\ell)}, \ldots, h_n^{(\ell)}$, hold out word i and define

$$\bar{r}_{-i} = \frac{1}{n-1} \sum_{k \neq i} r_k, \qquad \bar{h}_{-i}^{(\ell)} = \frac{1}{n-1} \sum_{k \neq i} h_k^{(\ell)}, \tag{5}$$

$$z_j^{(-i)} = \frac{r_j - \bar{r}_{-i}}{s_{r,-i}}, \qquad j \neq i, \tag{6}$$

where $s_{r,-i}$ is the standard deviation of the remaining ranks. The rank-weighted direction estimated without word i is

$$v_{-i}^{(\ell)} = \frac{\sum_{j \neq i} z_j^{(-i)} h_j^{(\ell)}}{\left\| \sum_{j \neq i} z_j^{(-i)} h_j^{(\ell)} \right\|_2}, \tag{7}$$

and the held-out projection is

$$p_i^{(\ell)} = \left(h_i^{(\ell)} - \bar{h}_{-i}^{(\ell)} \right)^\top v_{-i}^{(\ell)}. \tag{8}$$

We define **leave-one-out ordinal agreement** at layer ℓ as the fraction of distinct-rank pairs whose held-out projections preserve the known rank order:

$$A^{(\ell)} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \mathbb{1}\left[(p_i^{(\ell)} - p_j^{(\ell)})(r_i - r_j) > 0 \right]. \tag{9}$$

Table 6: The 32 part-of-speech frames. Each noun (composite, concrete, abstract, and the reference-scale words) receives the 16 noun frames; each adjective (gradable and non-gradable) receives the 16 adjective frames. Together with the shared frames every word is observed in 32 sentences.

POS	ID	FRAME
noun	0	She had never encountered that kind of w
noun	1	They focused on one particular type of w
noun	2	It was unlike any other sort of w
noun	3	They treated it as a distinct form of w
noun	4	As far as they could tell, there was no w
noun	5	The entire evening was devoted to a single subject: w
noun	6	Their conversation kept returning to the same topic: w
noun	7	Everything revolved around one central theme: w
noun	8	Her answer was always the same: w
noun	9	His final choice surprised everybody: w
noun	10	The form offered only a single option: w
noun	11	Every clue pointed to the same thing: w
noun	12	After hours of back and forth, they finally settled on w
noun	13	Everything in the folder was filed under w
noun	14	Whatever the brochure promised, there was no w
noun	15	She went through it all once more and found no w
adjective	0	Everyone in the room agreed that it was w
adjective	1	She looked at it again and realized it was w
adjective	2	From the very beginning, the whole thing had been w
adjective	3	Nobody had told them it would be w
adjective	4	Her report described it as w
adjective	5	They could tell right away that it was w
adjective	6	What surprised him most was that it was w
adjective	7	In the end, it turned out to be w
adjective	8	Her colleagues kept calling it w
adjective	9	Nobody would say whether it was w
adjective	10	They refused to believe it was w
adjective	11	According to every account, it was w
adjective	12	On closer inspection, it proved to be w
adjective	13	It was anything but w
adjective	14	Through it all, it remained w
adjective	15	Until that moment, nobody had noticed it was w

Chance is 0.5 and perfect agreement is 1. For each scale, we use the best-layer agreement because the primary question is whether the ordering is recoverable anywhere in the network. For cross-model analyses, we average these five best-layer agreements within each model.

The permutation null uses the same selection rule: ranks are shuffled within a scale, agreement is recomputed at every layer, and the maximum is retained for each of 50 permutations. Thus, the null includes the same best-layer search as the observed statistic. Split-half reliability used 30 random family-stratified 8-versus-8 splits of the 16 frames in each frame set.

A.5 Measuring structure in non-ordinal relations

The four non-ordinal relation types have no one-dimensional rank order, so we used two measures. We define the **related-versus-foil proximity score** as the mean cosine similarity between a target and its true related words minus the mean similarity between the target and its foils:

$$D^{(\ell)}_{\text{prox}} = \frac{1}{|R|}\sum_{r\in R}\cos(\tilde{h}^{(\ell)}_t, \tilde{h}^{(\ell)}_r) - \frac{1}{|F|}\sum_{f\in F}\cos(\tilde{h}^{(\ell)}_t, \tilde{h}^{(\ell)}_f). \tag{10}$$

Target and candidate states were centered within their frame set, standardized per dimension, and then averaged across frames. We computed the score for both the raw state $h^{(\ell)}$ and the contextual change $h^{(\ell)} - h^{(0)}$. For category-property relations, the analysis also stores hypernym-only and property-only proximity scores.

Table 7: Part-whole relation set (25 composite nouns). Related words are the target's usual constituents; foils are constituents of other targets in the same set, count-matched to the related list (the source target of every foil is given in the supplementary file). The relation phrase is the one used in that target's probes.

TARGET	RELATION	RELATED WORDS	FOILS
pasta	usually contains	flour, dough, sauce, tomatoes, garlic, basil	tiles, synonyms, pumps, galaxies, markets, senators
mathematics	usually includes	algebra, geometry, calculus, equations, proofs, fractions	lions, archives, bridges, agencies, sauce, presses
orchestra	usually includes	musicians, strings, brass, percussion, drums, horns	belts, tanks, letters, limbs, greens, roads
sandwich	usually contains	bread, cheese, lettuce, ham, mustard, turkey	gears, departments, equations, valves, sculptures, planets
ecosystem	usually contains	plants, animals, insects, bacteria, fungi, soil	stars, ribs, novels, gates, officers, artifacts
alphabet	usually includes	letters, vowels, characters, symbols	trade, bacteria, lettuce, cabinets
government	usually includes	ministers, officials, agencies, departments, senators, judges	stools, shards, algebra, nouns, percussion, garlic
salad	usually contains	carrots, onions, spinach, dressing, peppers, greens	bolts, shards, passengers, bones, performers, symbols
mosaic	is usually made of	tiles, fragments, pieces, glass, shards	lessons, musicians, motors, spinach, planes
committee	usually consists of	members, delegates, representatives, advisors, experts	trainers, sofas, hips, desks, magazines
curriculum	usually includes	courses, lessons, subjects, lectures, assignments, exams	rings, cylinders, glass, representatives, soldiers, generals
economy	usually includes	markets, industries, banks, businesses, workers, trade	slang, galaxies, limbs, fractions, phrases, bones
skeleton	usually includes	bones, ribs, joints, limbs, hips	characters, flour, recruits, insects, carrots
engine	usually contains	valves, cylinders, gears, bolts, belts	elephants, fractions, geometry, officials, novels
army	usually includes	soldiers, troops, officers, generals, tanks, recruits	tomatoes, asteroids, sofas, vowels, paintings, tiles
library	usually contains	books, magazines, journals, novels, archives	drills, ham, businesses, brass, judges
vocabulary	usually includes	words, terms, phrases, slang, synonyms, nouns	tents, cheese, galleries, workers, desserts, lectures
universe	usually contains	stars, planets, galaxies, asteroids, moons	assignments, vowels, terms, exams, phrases
city	usually contains	streets, buildings, parks, residents, roads, bridges	calculus, moons, proofs, joints, hips, synonyms
airport	usually contains	terminals, gates, planes, passengers, pilots	slang, moons, planets, ribs, proofs
museum	usually contains	exhibits, paintings, sculptures, artifacts, galleries, statues	delegates, ministers, industries, troops, banks, advisors
machinery	usually includes	motors, pumps, turbines, drills, presses	experts, dressing, pilots, wine, statues
furniture	usually includes	sofas, beds, desks, cabinets, couches, stools	algebra, basil, terminals, guests, buildings, dishes
banquet	usually includes	guests, dishes, wine, desserts, speeches	streets, paintings, books, asteroids, turbines
circus	usually includes	performers, elephants, lions, tents, rings, trainers	valves, dough, couches, beds, mustard, fragments

We define the **reconstruction score** as R^2_{true} minus the mean R^2 obtained from five count-matched random candidate sets. If $X \in \mathbb{R}^{k\times D}$ contains the k true related-word states and $y \in \mathbb{R}^D$ the target state, ridge regression uses

$$\lambda = 10^{-8}\frac{\operatorname{tr}(XX^\top)}{k} + 10^{-12}. \tag{11}$$

The penalty is fixed for numerical stability; it is not cross-validated. Random candidate sets exclude the target's true related words and are fixed across layers, variants, and models.

For each model, relation type, and state variant (raw or contextual), we select one best layer using the mean target-level score at that layer. We do not select a separate layer for each target. A target has positive structure when its score is above zero at this model-level best layer. The contextual share is the best contextual proximity divided by the best raw proximity.

A.6 Forced-choice recall and the representation-recall comparison

We define **forced-choice recall** as the fraction of prompts for which the model assigns higher continuation probability to the correct candidate. When both candidates are single tokens, one forward pass provides both probabilities. If a candidate contains multiple subtokens, it is scored in a separate

Table 8: Category-property relation set (25 concrete nouns). Each target has one hypernym and 4–5 characteristic properties; foil properties are count-matched to the properties, and 3–4 foil hypernyms are paired with the single hypernym so that every target contributes eight true–foil pairs; all foils come from other targets in the set.

TARGET	HYPERNYM	PROPERTIES	FOIL HYPERNYMS	FOIL PROPERTIES
dog	animal	fur, tail, muzzle, teeth, ears	footwear, fruit, furnishings	switch, petals, mane, coils, nest
cat	animal	fur, claws, tail, teeth	instrument, device, opening, tool	peel, branches, shade, feathers
horse	animal	mane, tail, legs, muzzle, teeth	container, mineral, implement	alarm, cork, pollen, pane, dial
apple	fruit	seeds, core, stem, peel	barrier, seat, animal, instrument	cord, arms, legs, cushion
bird	animal	feathers, wings, claws, nest	plant, tool, container, barrier	venom, bowl, roots, graphite
fish	animal	scales, fins, tail, eyes	footwear, implement, plant, seat	muzzle, label, sill, seeds
tree	plant	roots, branches, leaves, bark, trunk	food, animal, device	hinges, tongue, shaft, hands, heel
flower	plant	petals, stem, pollen, buds, scent	furnishings, fruit, mineral	ears, trunk, fins, claws, knob
egg	food	shell, membrane, protein, calcium	opening, fruit, implement, footwear	neck, face, sole, wings
stone	mineral	weight, minerals, crystals, texture, edges	opening, barrier, seat	cap, eyes, frame, grip, stem
bear	animal	fur, claws, teeth, muzzle	plant, device, instrument, furnishings	scales, shell, rim, panels
snake	animal	scales, venom, tail, coils	tool, container, mineral, device	leaves, crystals, handle, latch
hammer	tool	handle, head, claw, grip	food, animal, footwear, mineral	bark, toe, tail, blade
chair	seat	legs, cushion, frame, arms	food, container, tool, plant	fur, core, tip, membrane
table	furnishings	legs, surface, top, corners	food, fruit, instrument, opening	bulb, lead, teeth, buds
cup	container	handle, rim, base, capacity	barrier, implement, furnishings, animal	lock, claw, feathers, latch
knife	tool	blade, edge, handle, tip	seat, container, fruit, barrier	head, cushion, cork, claws
spoon	implement	handle, bowl, curve, tip	food, seat, plant, footwear	hinges, wings, mane, blade
door	barrier	hinges, knob, lock, panels	device, animal, mineral, instrument	peel, pollen, tip, heel
window	opening	pane, sill, frame, latch	tool, implement, instrument, device	sole, muzzle, roots, eyes
shoe	footwear	sole, heel, tongue, toe	opening, furnishings, barrier, tool	sill, face, switch, core
clock	device	hands, face, dial, alarm	implement, plant, opening, container	toe, shade, panels, fins
bottle	container	cap, neck, label, cork	animal, mineral, furnishings, fruit	petals, buds, frame, shaft
pencil	instrument	lead, graphite, tip, shaft	food, footwear, seat, animal	trunk, lock, arms, tail
lamp	device	bulb, shade, cord, switch	barrier, opening, tool, plant	tongue, branches, scales, hands

continuation pass and its subtoken log-probabilities are summed. The scoring code applies this rule for every tokenizer.

For prompt q, the correctness-oriented margin is

$$m(q) = \log P(w_{\text{correct}} \mid q) - \log P(w_{\text{wrong}} \mid q), \tag{12}$$

and recall is the fraction of prompts with $m(q) > 0$. We retain the continuous log-probabilities for the decomposition and attention analyses below.

The primary comparison uses the same five gradable scales for both ordinal structure and forced-choice recall. Across the 18 models, Steiger's test compares the correlation of $\log_{10}$ total parameter count with ordinal structure against its correlation with recall. The supplementary 3,000-prompt analysis additionally includes months, weekdays, and numbers.

A.7 Causal tests of ordinal structure

Ordinal direction. For causal interventions, we estimate a separate direction at every layer from all words on the target scale. We define the **ordinal direction** as the normalized rank-weighted

Table 9: Abstract-concept relation set (25 abstract nouns). Related words are concrete words closely associated with the concept; foils are related words drawn from other concepts in the set. Every probe uses the phrase "more closely connected to".

TARGET	RELATED WORDS	FOILS
justice	judge, court, trial, verdict, law	galaxy, monk, tornado, computer, ocean
democracy	election, ballot, parliament, voter, citizen	proof, angel, trophy, childhood, photographs
irony	joke, twist, pun, comedian, satire	priest, souvenir, wings, money, kiss
freedom	eagle, flag, wings, sky, passport	lottery, dagger, miser, ritual, sage
truth	fact, evidence, witness, proof, confession	knight, melody, flag, loop, choir
honesty	oath, promise, handshake, testimony	lever, eagle, parliament, traffic
courage	soldier, hero, medal, lion, knight	jackpot, citizen, pun, puzzle, horizon
mercy	pardon, plea, nurse, hospital, angel	explorer, ocean, satire, riot, junction
wisdom	owl, sage, elder, proverb, monk	chord, voter, riot, memories, cosmos
logic	computer, chess, philosopher, argument, premise	festival, throne, joke, nun, mole
fate	oracle, prophecy, doom, palm, gods	witness, hospital, ladder, illusion, costume
curiosity	question, microscope, telescope, explorer, scientist	miser, court, duel, banker, ceremony
nostalgia	memories, childhood, photographs, album, souvenir	argument, sword, mess, testimony, ladder
tradition	ritual, festival, ceremony, ancestors, costume	career, microscope, charm, casino, premise
infinity	horizon, cosmos, ocean, galaxy, spiral	gold, prophecy, assassin, festival, handshake
chaos	riot, storm, tornado, mess, traffic	owl, passport, vault, oath, medal
harmony	choir, melody, chord, chorus, tune	evidence, palm, maze, dice, ambush
dilemma	fork, choice, junction, puzzle, lever	storm, spiral, comedian, sky, tune
paradox	contradiction, illusion, maze, loop	chorus, ritual, saint, trial
virtue	saint, halo, nun, priest	premise, oracle, election, question
betrayal	dagger, spy, kiss, affair, mole	tornado, chess, proverb, scientist, nurse
ambition	ladder, throne, empire, trophy, career	plea, confession, twist, ancestors, elder
luck	dice, lottery, charm, jackpot, casino	fact, empire, choice, feud, halo
revenge	duel, feud, sword, assassin, ambush	album, ballot, traffic, chorus, ceremony
greed	gold, money, miser, vault, banker	hero, telescope, doom, pardon, costume

covariance between frame-averaged word states and their standardized ranks:

$$v^{(\ell)} = \frac{\sum_w z_w \left(h_w^{(\ell)} - \bar{h}^{(\ell)}\right)}{\left\| \sum_w z_w \left(h_w^{(\ell)} - \bar{h}^{(\ell)}\right) \right\|_2}. \tag{13}$$

Direction erasure. During a separate forced-choice forward pass, we remove the component of every token state parallel to the ordinal direction:

$$h_t^{(\ell)} \leftarrow h_t^{(\ell)} - \left[\left(h_t^{(\ell)}\right)^\top v^{(\ell)} \right] v^{(\ell)}. \tag{14}$$

The five-scale erasure experiment applied this intervention separately to the temperature, size, quality, likelihood, and anger directions. Each intervention was evaluated on all 1,164 ordinal prompts, giving a complete five erased directions by five evaluated scales design for every model. Erasure was applied at every token position through the first 60% of decoder blocks.

For each erased scale, we constructed overlap matched random directions separately at every layer. Let Q be an orthonormal basis for the input embeddings of the gradable words represented as single tokens with a leading space. Depending on the tokenizer, this basis contained 30–33 of the 33 gradable words. For ordinal direction $v^{(\ell)}$, define

$$\rho_\ell = \|Q^\top v^{(\ell)}\|_2. \tag{15}$$

For a random vector u, let $u_P = QQ^\top u$ and $u_O = u - u_P$. The matched direction is

$$w^{(\ell)} = \rho_\ell \frac{u_P}{\|u_P\|_2} + \sqrt{1-\rho_\ell^2}\,\frac{u_O}{\|u_O\|_2}. \tag{16}$$

Thus, the random direction has the same projection magnitude onto the gradable word embedding span as the erased ordinal direction at that layer. We used five random seeds for each model and erased scale and averaged their accuracies.

We define **erasure damage** as baseline forced-choice accuracy minus accuracy under erasure. We define **excess erasure damage** as targeted erasure damage minus the mean overlap matched control

Table 10: Attribute-object relation set (20 non-gradable adjectives). Related words are nouns the adjective applies to; foils are compatible nouns of other adjectives in the set. Every probe uses the phrase "can be {adjective}".

TARGET	RELATED NOUNS	FOILS
wooden	fence, bench, plank, crate, cabin, barrel	payment, diagnosis, voltage, clause, uranium, proton
digital	camera, photo, signature, currency, download, archive	plank, enamel, algae, marsh, solvent, uranium
monthly	payment, rent, salary, subscription, meeting, report	bench, socket, cavity, prism, harbor, nucleus
chemical	reaction, compound, spill, residue, solvent, fertilizer	fence, salary, clause, cliff, cemetery, reunion
electrical	wiring, outlet, fuse, appliance, socket, voltage	crate, census, apology, stitches, algae, zoning
medical	diagnosis, clinic, prescription, vaccine, therapy, insurance	plank, barrel, fuse, eclipse, marsh, parade
legal	lawsuit, contract, statute, clause, warrant, injunction	enamel, tray, breeze, marsh, neutron, electron
federal	budget, agency, mandate, grant, prosecutor, bureau	braces, crate, stitches, wrapper, tub, proton
nuclear	reactor, fallout, arsenal, meltdown, uranium, submarine	plaque, plank, apology, rent, gown, prism
solar	eclipse, panel, flare, array, farm, corona	braces, statute, injunction, anesthesia, biopsy, vaccine
dental	enamel, cavity, plaque, braces, hygiene, implant	fence, eclipse, injunction, algae, landfill, neutron
verbal	agreement, warning, apology, consent, command, threat	enamel, barrel, socket, vessel, cliff, electron
plastic	bucket, wrapper, straw, tub, jug, tray	clause, salary, eclipse, vaccine, census, prosecutor
marine	biology, life, species, vessel, algae, navigation	plaque, statute, subscription, gown, lattice, download
municipal	sewer, landfill, zoning, cemetery, ordinance, council	cavity, enamel, prism, stitches, electron, neutron
surgical	gown, stitches, clamp, procedure, anesthesia, biopsy	crate, eclipse, injunction, zoning, census, currency
optical	lens, prism, zoom, fiber, sensor, scanner	payment, statute, reunion, apology, cemetery, straw
atomic	nucleus, lattice, blast, proton, neutron, electron	rent, lawsuit, gown, harbor, sewer, hygiene
annual	audit, harvest, summit, reunion, census, parade	fence, socket, cavity, wrapper, lattice, anesthesia
coastal	erosion, cliff, harbor, marsh, breeze, town	hygiene, subscription, salary, download, biopsy, electron

Table 11: Probe construction for the four non-ordinal relation types. Each target contributes eight true–foil pairs (the i-th related word with the i-th foil, cyclically shifted pairings until eight; for category-property, the hypernym pairs plus property pairs), each asked in both candidate orders; PAIRS and PROBES are per target, TOTAL per model. {target} is the target's noun phrase ("a dog", "pasta"); {relation} is the target-specific phrase of Table 7.

SET	TEMPLATE	PAIRS	PROBES	TOTAL
part-whole	Between {a} and {b}, the one that {target} {relation} is	8	16	400
category-property	Between {a} and {b}, the category {target} belongs to is (hypernym)	3–4	16	400
	Between {a} and {b}, the one {target} has is (property)	4–5		
abstract-concept	Between {a} and {b}, the one more closely connected to {target} is	8	16	400
attribute-object	Between {a} and {b}, the one that can be {target} is	8	16	320

damage. Scale specificity was tested within each model by comparing the matching-scale excess damage with the mean excess damage on the other four scales. The global specificity analysis compared each model's mean of the five diagonal excess values with its mean of the 20 off-diagonal values.

We define **erasure damage** as baseline forced-choice accuracy minus accuracy under erasure. The primary matched-control experiment erased every token position from layer 1 through $0.6L$. This cutoff was fixed before the matched-control result was inspected. A separate depth sweep repeated erasure from layer 1 through cutoffs at $0.25L$, $0.50L$, $0.75L$, and L. Additional localization checks

Table 12: Novel words used in cross-model transfer. The sender model receives two definition sentences for each word; the receiver model never sees these definitions except in the ceiling condition. Candidate pairs list the familiar scale words the receiver compares the novel word with (each pair in both orders and two templates: 24 questions per word, 144 per sender–receiver–condition).

WORD	SCALE	POLE	DEFINITION SENTENCES	CANDIDATE PAIRS
blorf	temperature	low	Blorf is a temperature word. Blorf means extremely cold, even colder than freezing.	cold/boiling, cold/hot, chilly/boiling, chilly/warm, cool/boiling, cool/warm
vopim	temperature	high	Vopim is a temperature word. Vopim means extremely hot, even hotter than boiling.	freezing/hot, cold/hot, freezing/warm, chilly/warm, cool/hot, cold/warm
skarn	size	low	Skarn is a size word. Skarn means extremely small, even smaller than microscopic.	tiny/colossal, tiny/gigantic, small/colossal, small/huge, tiny/huge, small/gigantic
drell	size	high	Drell is a size word. Drell means extremely big, even bigger than colossal.	microscopic/gigantic, tiny/gigantic, microscopic/huge, small/huge, tiny/big, small/gigantic
quenz	quality	low	Quenz is a quality word. Quenz means extremely bad, even worse than terrible.	bad/perfect, bad/excellent, mediocre/perfect, mediocre/great, decent/perfect, bad/great
marlip	quality	high	Marlip is a quality word. Marlip means extremely good, even better than perfect.	terrible/excellent, bad/excellent, terrible/great, mediocre/great, decent/excellent, bad/great

Table 13: The 127 anchor words used to fit each sender-model to receiver-model map. The list was fixed during the original Gemma-4/Qwen3.5 design as the 128 main relation target words minus *lukewarm*, the only word excluded by that construction-time single-token filter. The same list was retained after expansion to 18 models. SmolLM2-1.7B tokenizes *mediocre*, *enraged*, and *banquet* into multiple subtokens; anchor states are always read at the final subtoken, so these words remain in the map.

GROUP	ANCHOR WORDS
gradable adjectives (32; *lukewarm* excluded)	freezing, cold, chilly, cool, warm, hot, boiling, microscopic, tiny, small, big, huge, gigantic, colossal, terrible, bad, mediocre, decent, good, great, excellent, perfect, impossible, unlikely, possible, probable, certain, annoyed, irritated, angry, furious, enraged
part-whole targets (25)	pasta, mathematics, orchestra, sandwich, ecosystem, alphabet, government, salad, mosaic, committee, curriculum, economy, skeleton, engine, army, library, vocabulary, universe, city, airport, museum, machinery, furniture, banquet, circus
category-property targets (25)	dog, cat, horse, apple, bird, fish, tree, flower, egg, stone, bear, snake, hammer, chair, table, cup, knife, spoon, door, window, shoe, clock, bottle, pencil, lamp
abstract-concept targets (25)	justice, democracy, irony, freedom, truth, honesty, courage, mercy, wisdom, logic, fate, curiosity, nostalgia, tradition, infinity, chaos, harmony, dilemma, paradox, virtue, betrayal, ambition, luck, revenge, greed
attribute-object targets (20)	wooden, digital, monthly, chemical, electrical, medical, legal, federal, nuclear, solar, dental, verbal, plastic, marine, municipal, surgical, optical, atomic, annual, coastal

used the narrower $0.25L$–$0.45L$ band and erasure at the final token only. Targeted and matched-control damage were compared across models with the Wilcoxon signed-rank test.

Ordinal-direction steering. As a supplementary dose-response test, we shifted the final-position state at depth $0.35L$ along the temperature direction by $\alpha\sigma^{(\ell)}v^{(\ell)}$, with $\alpha \in \{-6, -3, -1, 0, 1, 3, 6\}$.

Cross-model transfer. Six invented words covered the two poles of temperature, size, and quality: *blorf*/*vopim*, *skarn*/*drell*, and *quenz*/*marlip*. For each invented word, the **sender model** received two definition sentences specifying the word's position on its scale. We extracted the sender state at the novel word's final subtoken from four carrier sentences and averaged those four states. The **receiver model** never received the definition sentences.

For every ordered pair of different models, we fit a sender-to-receiver map using the same 127 anchor words in both models. Let $X \in \mathbb{R}^{127\times D_A}$ and $Y \in \mathbb{R}^{127\times D_B}$ contain sender and receiver anchor states. With $K = XX^\top$, closed-form kernel ridge regression gives

$$C = (K + \lambda I)^{-1}Y, \tag{17}$$

$$\hat{y}(x) = (Xx)^\top C. \tag{18}$$

Sender and receiver depths were selected from $\{0.25, 0.35, 0.50\}^2$, and $\lambda_{\text{rel}} \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ was converted to $\lambda = \lambda_{\text{rel}} \operatorname{mean}(\operatorname{diag} K)$. Five-fold cross-validation over held-out anchor words selected the layer pair and ridge penalty that maximized centered cosine similarity.

The 127-anchor list was fixed in the original two-family design and retained after expansion to 18 models. It contains three multi-subtoken anchors under SmolLM2-1.7B; as elsewhere, their states are read at the final subtoken.

The mapped novel word state was rescaled to the receiver model's mean anchor norm and written at each occurrence of the novel word's final subtoken. The receiver model then answered comparison questions in which the novel word was the query and the two answer candidates were familiar words from the same scale. Each novel word used six candidate pairs, two candidate orders, and two templates, giving 24 questions per word and 144 questions per sender-receiver-condition. We evaluated four conditions: mapped state, norm-matched random state (sham), no injection, and a ceiling condition in which the receiver model received the verbal definition. Every model served as sender to each of the other 17 models, giving $18 \times 17 = 306$ ordered pairs; self-transfer was excluded.

A.8 Answer-margin decomposition and lexical priors

For ordinal pair $p = (w_{\text{low}}, w_{\text{high}})$ and concrete prompt q, we define the **answer margin** with a fixed higher-minus-lower orientation:

$$\Delta_p(q) = \log P(w_{\text{high}} \mid q) - \log P(w_{\text{low}} \mid q). \tag{19}$$

A positive value therefore favors the higher-ranked word regardless of which end of the scale the question requests.

Let $\bar{\Delta}_p^{(+)}$ be the mean margin over the two candidate orders and three templates when the question asks for the higher pole, and $\bar{\Delta}_p^{(-)}$ the corresponding mean when it asks for the lower pole. We define

$$E_p = \frac{\bar{\Delta}_p^{(+)} - \bar{\Delta}_p^{(-)}}{2}, \tag{20}$$

$$S_p = \frac{\bar{\Delta}_p^{(+)} + \bar{\Delta}_p^{(-)}}{2}. \tag{21}$$

We define **question-responsive evidence** as E_p: it measures how much the answer preference changes in the direction requested by the question. We define **question-invariant preference** as the signed quantity S_p: its sign identifies the candidate preferred on average when the question direction is reversed. The decomposition gives

$$\bar{\Delta}_p^{(+)} = S_p + E_p, \qquad \bar{\Delta}_p^{(-)} = S_p - E_p. \tag{22}$$

Both question directions are answered correctly when $E_p > |S_p|$.

We define **stubbornness** as $|S_p|$, the magnitude of the question-invariant preference. The absolute value is used because different pairs can favor different candidates. Model-level evidence is the mean E_p over the 97 primary pairs, and model-level stubbornness is the mean $|S_p|$ over the same pairs.

We define a word's **lexical prior** as its mean sequence log-probability in three semantically neutral contexts: "The answer is", "The word is", and "She said the word". If $w = (t_1, \ldots, t_m)$ contains multiple subtokens, then

$$\log P(w \mid c) = \sum_{j=1}^{m} \log P(t_j \mid c, t_{<j}), \tag{23}$$

and

$$\pi(w) = \frac{1}{3} \sum_{c \in \mathcal{C}_{\text{neutral}}} \log P(w \mid c). \tag{24}$$

For pair p, the signed lexical-prior difference is $\Delta\pi_p = \pi(w_{\text{high}}) - \pi(w_{\text{low}})$. We correlate $\Delta\pi_p$ with the signed question-invariant preference S_p. Analyses of preference magnitude use $|\Delta\pi_p|$ and $|S_p|$, together with representational separation and E_p, while controlling for rank separation.

A.9 ATTENTION FROM THE ANSWER POSITION TO CANDIDATE WORDS

Blocking. We define **answer-position attention to candidates** as attention from the final answer position to every subtoken belonging to either candidate word. A full additive mask $\mathcal{M} \in \mathbb{R}^{1\times1\times T\times T}$ replaced the usual padding mask. The causal-only implementation used $\mathcal{M}[t, s] = 0$ for $s \le t$ and $-\infty$ otherwise. In the blocking condition, we additionally set $\mathcal{M}[T-1, s] = -\infty$ for every candidate subtoken s. Candidate words remained visible to all other token positions.

Before using the block, we checked that the custom-mask implementation itself reproduced normal model behavior. The causal-only custom mask had to preserve the ordinary top-1 prediction and produce Pearson correlation at least 0.999 between the full logit vectors. Qwen3.5-35B-A3B-Base failed this implementation check ($r = 0.9987$) and was excluded from blocking and amplification, leaving 17 models.

Attention-weight measurement. For each probe and layer, we summed attention from the final answer position to all subtokens of each candidate and averaged over heads. Let $A_{\text{high}}(q)$ and $A_{\text{low}}(q)$ denote these two values and define their difference as

$$D_{\text{att}}(q) = A_{\text{high}}(q) - A_{\text{low}}(q). \tag{25}$$

We averaged this difference over candidate orders and templates separately for higher- and lower-pole questions, giving $\bar{D}^{(+)}_{\text{att}}$ and $\bar{D}^{(-)}_{\text{att}}$. We define the **question-responsive attention shift** as

$$E_{\text{att}} = \frac{\bar{D}^{(+)}_{\text{att}} - \bar{D}^{(-)}_{\text{att}}}{2}. \tag{26}$$

It measures how much answer-position attention changes toward the candidate requested by the question. For each model, we record its maximum over layers and the relative depth of that maximum. Pooled depth analyses use the 14 models with attention matrices available at every layer; Qwen3.5 is reported separately.

Attention amplification. As the opposite intervention, we added $\delta \in \{0.5, 1, 2, 4\}$ to the attention logits from the answer position to every candidate subtoken at every layer. Forced-choice recall, question-responsive evidence, and stubbornness were recomputed at each dose.

A.10 ADDITIONAL FIXED READOUT INTERVENTIONS

Ordinal-signal magnitude. At depth $0.35L$, we multiplied each token state's component along the model's temperature ordinal direction by $g \in \{1.5, 2, 4, 8\}$:

$$h_t \leftarrow h_t + (g-1)(h_t^\top v)v \tag{27}$$

for every token position t at that block.

Readout-gradient angle. At the layer with the strongest temperature ordinal structure, we differentiated the answer margin with respect to the final-position hidden state. Combining the two question directions and candidate orders separated a question-responsive gradient component from a question-invariant gradient component. We measured the cosine between the responsive component and the temperature direction and the squared-norm share of the invariant component.

Fixed bias-subspace removal. Question-invariant readout gradients from the even-indexed temperature pairs were normalized, centered, and decomposed by SVD. The top $k \in \{1, 3, 8\}$ right singular vectors formed B_k. At block $\ell^* - 1$,

$$h \leftarrow h - B_k B_k^\top h. \tag{28}$$

Evaluation used held-out temperature pairs and the unseen size and quality scales. Three random orthonormal rank 3 subspaces served as controls for all three target ranks. Thus, only the rank 3 comparison was rank matched; the rank 1 and rank 8 comparisons were used as descriptive checks. Rank k was fixed in advance rather than selected separately for each model.

A.11 SYMMETRISED SCORING, BENCHMARKS, AND STATISTICS

Ordinal symmetrisation. We define **symmetrised scoring** for an ordinal pair as

$$\Delta_p^{\mathrm{sym}} = \frac{\bar{\Delta}_p^{(+)} - \bar{\Delta}_p^{(-)}}{2} = E_p. \tag{29}$$

Because S_p is unchanged when the question switches between the higher and lower pole, it does not contribute to this score. For non-ordinal relations, which have no paired higher-versus-lower question, we instead sum the correctly oriented margins from the two candidate orders.

Multiple-choice symmetrisation. For a question with n options, cyclic permutation k places option content j at position

$$\pi_k(j) = (j + k) \bmod n, \qquad k = 0, \ldots, n-1. \tag{30}$$

If $q^{(k)}$ is the prompt under permutation k and ℓ_m the label at position m, option j receives

$$s_j^{\mathrm{sym}} = \sum_{k=0}^{n-1} \log P\Big(\ell_{\pi_k(j)} \mid q^{(k)}\Big). \tag{31}$$

We select $\arg\max_j s_j^{\mathrm{sym}}$. The $k = 0$ permutation is the original ordering, so baseline and symmetrised scores are obtained from the same set of forward passes.

Benchmarks. We evaluated ARC-Challenge and ARC-Easy (Clark et al., 2018), MMLU (Hendrycks et al., 2020), MMLU-Pro (Wang et al., 2024), OpenBookQA (Mihaylov et al., 2018), CommonsenseQA (Talmor et al., 2019), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021), and BoolQ (Clark et al., 2019). We used 13,507 MMLU test items, 1,165 ARC-Challenge test items, 12,032 MMLU-Pro test items, 2,365 ARC-Easy test items, 500 OpenBookQA test items, 1,221 CommonsenseQA validation items, 10,042 HellaSwag validation items, 1,267 WinoGrande XL validation items, and 3,270 BoolQ validation items.

HellaSwag, WinoGrande, and BoolQ were converted to the same lettered multiple-choice format as the other benchmarks. Their absolute scores should therefore not be compared directly with published leaderboard values; the relevant comparison is baseline versus symmetrised scoring within the same model-benchmark cell.

Batching and statistical tests. Questions were left padded so that the answer position was always the final column, and all cyclic permutations of a question used the same numerical batching regime. Default limits were 16,384 estimated tokens and 96 rows, reduced to 4,096/24 for the 26B and 35B models and to 2,048/10 for Gemma-4-31B. If one question exceeded the cap by itself, its permutations were run one at a time.

Differences between dependent correlations used Steiger's test. Partial correlations controlling for model size regressed both variables on $[1, \log_{10} N]$ and correlated the residuals. Exploratory correlation tables used Benjamini–Hochberg adjustment. Model-level paired comparisons used two-sided paired t tests unless stated otherwise. Direction-erasure comparisons used two-sided Wilcoxon signed-rank tests. For the five matching-scale erasure tests and the five per-direction specificity tests, we additionally report Holm-adjusted p-values. Baseline and symmetrised predictions within each benchmark cell were compared item by item with the exact McNemar test.

We define **error recovery** as the fraction of baseline errors removed by symmetrised scoring:

$$\mathrm{Recovery} = \frac{\mathrm{Acc}_{\mathrm{sym}} - \mathrm{Acc}_{\mathrm{base}}}{1 - \mathrm{Acc}_{\mathrm{base}}}, \tag{32}$$

computed per model and then averaged over models.

Pre-specification. Internal design documents fixed pass criteria, falsifiers, and exclusion rules before the corresponding results were inspected for the original temperature matched-control erasure experiment, cross model transfer, lexical-prior, attention, and later readout intervention analyses. These documents were not externally archived or independently timestamped, so we call these analyses *pre-specified*, not preregistered. The extension from temperature to all five erased ordinal directions and the diagonal versus off-diagonal specificity analyses were added after reviewer feedback and were not pre-specified.

Table 14: Matching-scale erasure results. Excess damage is targeted erasure damage minus overlap matched random control damage. "Models" is the number of models with positive excess damage. Statistical tests are two-sided Wilcoxon signed-rank tests across 18 models.

SCALE	TARGET	CONTROL	EXCESS	MODELS	RAW p	HOLM p
Temperature	0.165	0.048	+0.116	15/18	1.91×10^{-4}	5.72×10^{-4}
Size	0.165	0.044	+0.121	17/18	3.81×10^{-5}	1.53×10^{-4}
Quality	0.146	0.053	+0.093	17/18	1.53×10^{-5}	7.63×10^{-5}
Likelihood	0.022	0.020	+0.002	10/18	0.586	0.586
Anger	0.109	0.035	+0.074	16/18	3.26×10^{-4}	6.52×10^{-4}

B CAUSAL TESTS OF ORDINAL REPRESENTATIONS

B.1 FIVE SCALE DIRECTION ERASURE

We repeated direction erasure for each of the five ordinal scales: temperature, size, quality, likelihood, and anger. For each erased ordinal direction, we measured erasure damage on questions from all five scales. Each targeted erasure was compared with an overlap matched random control constructed for the same erased scale. We define excess erasure damage as targeted erasure damage minus control damage.

Targeted erasure caused more matching-scale damage than the overlap matched random control for temperature, size, quality, and anger after Holm correction. Likelihood showed no detectable excess damage. Its mean baseline recall was 0.534, and 11 of 18 models had baseline accuracy below 0.55, leaving little room for erasure to reduce accuracy further.

Scale specificity showed the same pattern. Matching-scale excess damage was $+0.116$ for temperature, $+0.121$ for size, $+0.093$ for quality, and $+0.074$ for anger, whereas the corresponding mean off-diagonal excess damages were -0.016, -0.002, -0.014, and -0.023. Likelihood was flat ($+0.002$ matching versus -0.001 off diagonal). Across all five directions, the mean diagonal excess damage was $+0.081$, compared with -0.011 across off-diagonal cells. The diagonal mean was larger in 17 of 18 models (two-sided Wilcoxon signed-rank $p = 1.53 \times 10^{-5}$).

The group-level pattern was not uniform across models. In Gemma-4, overlap matched random controls caused damage similar to targeted erasure, leaving little positive excess damage. Qwen3.5-9B-Base showed broad cross-scale damage rather than scale-specific erasure, and gemma-4-31B had non-positive excess damage on all five matching scales. Thus, the five-scale analysis supports scale specificity as a group-level pattern for four scales, not functional separation of all five directions in every model.

Intervention locus. Restricting temperature erasure to layers $0.25L$ through $0.45L$ produced mean damage of 0.108, compared with 0.045 for the overlap matched random control (15 of 18 models; $p = 1.01 \times 10^{-3}$). Applying erasure only at the final token produced damage of 0.023, compared with 0.033 for the control (9 of 18 models; $p = 0.979$). These localization checks were performed for temperature only.

B.2 CROSS MODEL TRANSFER

Cross model transfer tested whether mapped hidden states could support ordinal judgments in a receiver model. Across all 306 ordered sender and receiver pairs, mapped state injection exceeded the norm matched sham injection by more than 0.02 in 76.8% of pairs and was worse by more than 0.02 in 4.9%. The mean gain was $+0.086$ for the 48 pairs within the same model family and $+0.073$ for the 258 pairs across different families (Welch test, $p = 0.17$). These results cover temperature, size, and quality.

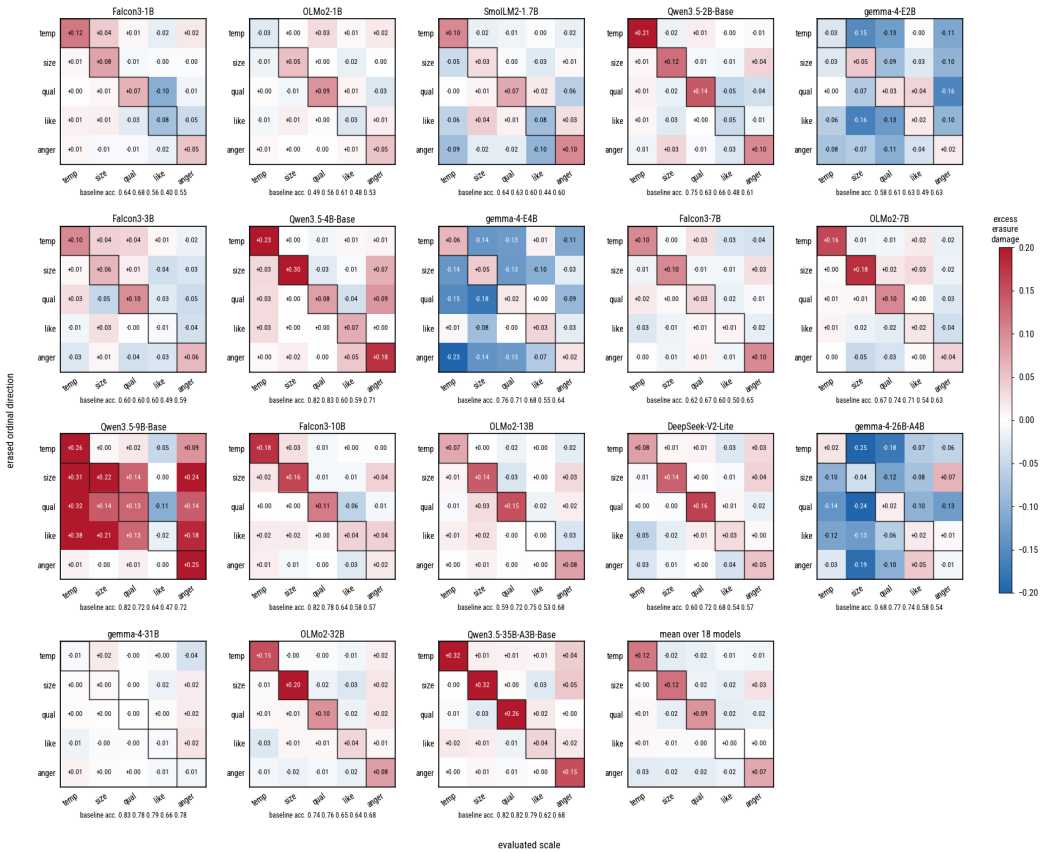

Figure 6: **Ordinal direction erasure across individual models.** Rows indicate the erased ordinal direction and columns the evaluated scale. Each cell shows excess erasure damage, defined as targeted erasure damage minus overlap matched random control damage. Black outlines mark the matching scale. Baseline accuracy for each evaluated scale is shown below each model, and the final panel reports the mean across all 18 models.

C Additional readout tests and interventions

We tested three interventions and one diagnostic analysis, each addressing a different explanation for weak forced-choice recall. Unless stated otherwise, model-level paired comparisons in this section use two-sided paired t tests.

C.1 Increasing attention to the candidate words does not improve recall

We first tested whether recall was limited by insufficient attention from the answer position to the candidate words. We increased the corresponding attention logits on temperature, size, and quality questions using $\delta \in \{0.5, 1, 2, 4\}$. At $\delta = 4$, mean question-responsive evidence decreased from 0.835 to 0.235, and mean forced-choice recall accuracy decreased from 0.679 to 0.555. Question-responsive evidence decreased in all 17 tested models ($p = 1.8 \times 10^{-6}$). Selecting the best dose separately for each model produced a mean accuracy gain of only $+0.001$ and a maximum gain of $+0.025$. Increasing attention to the candidate words did not increase question-responsive evidence or forced-choice recall.

Table 15: Accuracy change after bias subspace removal relative to the mean of three random rank 3 subspace controls. Only the target rank 3 row is rank matched to the control. Values are means across the 18 models.

TARGET RANK	HELD OUT TEMPERATURE	SIZE	QUALITY
1	$+0.008$	$+0.009$	$+0.015$
3	-0.001	$+0.009$	$+0.012$
8	-0.015	$+0.008$	$+0.005$

C.2 Increasing the temperature ordinal component does not improve temperature recall

We next tested whether the temperature ordinal component was too weak to control the answer. At depth $0.35L$, we multiplied the component of every token state at that block along the temperature ordinal direction using $g \in \{1, 1.5, 2, 4, 8\}$. Mean temperature forced-choice recall accuracy was 0.692 at baseline and 0.675, 0.665, 0.629, and 0.580 at gains 1.5, 2, 4, and 8. No fixed gain improved mean temperature accuracy. Selecting the best nonbaseline gain separately for each model produced a mean accuracy change of -0.012 ($p = 0.518$), with a positive change in 10 of 18 models.

The same intervention was also evaluated on size and quality questions as off-target checks. These questions were not used to determine whether the temperature ordinal component itself was too weak.

C.3 The question-responsive temperature readout gradient is nearly orthogonal to the ordinal direction

We tested whether weak temperature recall reflected a fixed mismatch between the temperature ordinal direction and the local readout gradient. For each temperature pair and template, we computed the answer margin gradient with respect to the final-position hidden state at the layer with the highest temperature leave-one-out ordinal agreement. The two question directions and two candidate orders separated the gradient into question-invariant and question-responsive components.

For each comparison, we measured

$$\frac{\|g_{\text{inv}}\|_2^2}{\|g_{\text{inv}}\|_2^2 + \|g_{\text{resp}}\|_2^2}. \tag{33}$$

Averaged over 1,512 model, pair, and template comparisons, this share was 0.676. The question-responsive gradient component was nearly orthogonal to the temperature ordinal direction: the mean cosine across models was $+0.005$, with a range from -0.014 to $+0.018$.

C.4 Removing a bias subspace estimated from temperature does not consistently improve recall across scales

Finally, we tested whether removing a fixed bias subspace estimated from temperature gradients could improve recall. Question-invariant readout gradients from even indexed temperature pairs were used to estimate subspaces of rank 1, 3, and 8. We evaluated each subspace on held out temperature pairs, size questions, and quality questions.

The available random controls were three random rank 3 subspaces. Therefore, we used these same controls for all three target ranks. Table 15 reports the resulting accuracy changes. The rank 1 and rank 8 comparisons are not rank matched and are reported descriptively. The rank 3 comparison is the matched comparison.

For the matched rank 3 comparison, accuracy changed by -0.001 on held out temperature questions ($p = 0.94$), $+0.009$ on size questions ($p = 0.13$), and $+0.012$ on quality questions ($p = 0.024$). Thus, removing the estimated temperature bias subspace did not improve held out temperature or size recall, but produced a small improvement on quality questions. When the three evaluation sets were pooled, the matched rank 3 accuracy change was $+0.009$ ($p = 0.12$). The effect did not generalize consistently across the three evaluation sets.

Table 16: Mean benchmark performance across the 18 models. Gain is symmetrised minus baseline accuracy in percentage points. Recovery is the mean fraction of baseline errors recovered. "Sig." is the number of models with a significant McNemar improvement.

BENCHMARK	ITEMS	BASE	SYM.	GAIN	RECOVERY (%)	SIG.
ARC-Challenge	1,165	0.778	0.811	3.32	14.6	11/18
ARC-Easy	2,365	0.884	0.922	3.75	26.5	15/18
MMLU	13,507	0.633	0.657	2.45	6.1	18/18
MMLU-Pro	12,032	0.336	0.414	7.82	12.4	18/18
OpenBookQA	500	0.739	0.801	6.16	24.7	16/18
CommonsenseQA	1,221	0.662	0.724	6.18	15.5	15/18
HellaSwag	10,042	0.459	0.550	9.08	19.8	18/18
WinoGrande	1,267	0.611	0.619	0.77	2.8	4/18
BoolQ	3,270	0.732	0.771	3.91	12.4	11/18

D SYMMETRISED SCORING

D.1 ORDINAL JUDGMENTS

For direct comparison with the intervention experiments, we first applied symmetrised scoring to temperature, size, and quality. Mean forced-choice recall accuracy increased from 0.686 to 0.930, a gain of 24.4 percentage points. All 18 models improved ($p = 1.6 \times 10^{-12}$).

We then applied the same scoring rule to all five ordinal scales. Mean accuracy increased from 0.665 to 0.891, a gain of 22.6 percentage points, and all 18 models improved ($p = 9.3 \times 10^{-14}$). Likelihood improved by 11.1 percentage points in 14 of 18 models ($p = 0.03$), and anger improved by 20.2 percentage points in all 18 models ($p = 2.3 \times 10^{-9}$).

D.2 MULTIPLE CHOICE BENCHMARKS

Across nine benchmarks and 18 models, symmetrised scoring significantly improved 126 of 162 combinations of model and benchmark under the exact McNemar test and significantly worsened one. The largest increase was 23.7 percentage points. Table 16 reports the mean performance across models.

These benchmark results test preferences tied to answer position. They do not establish the same decomposition into question-responsive evidence and question-invariant preference used for ordinal pairs.

E RELATIONS WITHOUT OPPOSITE QUESTION DIRECTIONS

Part-whole, category-property, abstract-concept, and attribute-object relations do not provide opposite question directions. The decomposition into question-responsive evidence and question-invariant preference is therefore not defined for these relation types.

The related-versus-foil proximity score was positive for 96–100% of targets in every model and relation type at the best layer for each model. The reconstruction score was positive for 92–100% of targets. For category-property relations, the proximity comparison against count matched property foils remained positive for at least 92% of targets in every model.

Forced-choice recall accuracy ranged from 0.745 to 0.945 for part-whole, 0.690 to 0.840 for category-property, 0.557 to 0.815 for abstract-concept, and 0.700 to 0.912 for attribute-object relations. Symmetrised scoring based on candidate order produced much smaller gains than on the ordinal scales: 2.0 percentage points for part-whole, 3.2 for category-property, 2.8 for abstract-concept, and 2.2 for attribute-object.

Table 17: Representation, forced-choice recall, and symmetrised scoring across relation types. For non-ordinal relations, "Structure" is the share of targets with positive related-versus-foil proximity at the best layer for each model. The ordinal row reports the range of mean leave-one-out ordinal agreement across the 18 models.

RELATION TYPE	PROBES	STRUCTURE	RECALL RANGE	SYM. GAIN
Ordinal (five scales)	1,164	0.833–0.943	0.545–0.784	+22.6
Part-whole	400	100%	0.745–0.945	+2.0
Category-property	400	$\geq 96\%$	0.690–0.840	+3.2
Abstract-concept	400	$\geq 96\%$	0.557–0.815	+2.8
Attribute-object	320	100%	0.700–0.912	+2.2

The smaller gains for the four non-ordinal relation types show that candidate order symmetrisation does not reproduce the large correction obtained when opposite question directions separate question-responsive evidence from question-invariant preference.