CAUSAL ABSTRACTION OF MAGNITUDE IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models can tell "5 degrees Celsius" apart from "35 degrees Celsius." But does the model store one reusable "temperature" variable, or just a separate pattern per phrasing? We test this with two ideas: *cross-form causal interchangeability* (does a magnitude direction learned from one phrasing, e.g. "5 degrees Celsius," still work on a different phrasing of the same quantity, e.g. "just above freezing"?) and *multi-consequence coherence* (does one intervention move many independent downstream judgments together – ice melting, needing a coat, needing a heater – rather than just the judgment it was built from?), against random-direction, output-specific, and novel-context controls. On Qwen2.5-7B-Instruct, temperature is easy to read out of the hidden state ($R^2 = 0.98$), and a direction built from one phrasing transfers almost as well to a different phrasing as to the same one: averaged over 15 judgments, coherence is 0.77 within-phrasing and 0.73 cross-phrasing, versus 0.57 ± 0.05 for a random direction – about 4 pooled standard deviations, confirmed by a non-overlapping bootstrap 95% CI, and just as strong on an unfamiliar topic (e.g. a lab freezer) and under a continuous dose-response. Reaching this required catching two of our own mistakes: an early version that intervened at every token position let a meaningless random direction fake the real effect, and 3 of the 15 judgments had been set up backwards, caught only by the continuous test itself. Extending to a second quantity and two more models gives a mixed picture: distance replicates far more weakly (≈1–2 SD), Phi-3.5-mini-instruct shows a moderate effect (≈2 SD), and Mistral-7B-Instruct-v0.3 shows none (≈0 SD) despite decoding temperature more accurately than Qwen. The central claim rests on this evidence, not on an output-specific-steering control that never worked reliably even after two fixes; a small held-out set and an unfinished context-compositionality test are the main remaining gaps. The broader point: a model can pass a simple readout test for a concept without using it as a genuine, reusable variable – causal abstraction is a stronger, more concrete test of this than probing alone.

1 INTRODUCTION

When a model behaves as though it understands a concept, is that because it has built one reusable internal variable for it, or because it has memorized many separate, situation-specific patterns that merely look like understanding? The two possibilities predict different behavior once the situation changes: a reusable variable should keep working; a bag of memorized patterns should break.

Quantitative concepts make this concrete. Consider "The room was 5 degrees Celsius" versus "The room was 35 degrees Celsius." A model that understands temperature should treat these as different situations with different downstream implications, even when those implications are never stated: at 5 degrees, ice should stay frozen and a coat should be wanted; at 35 degrees, ice should melt and cooling should be wanted.

So the interesting question is not just whether a model can *read out* temperature from its hidden state – that shows only that the information is present somewhere. It is whether the model builds a reusable "temperature" variable that many downstream judgments (does ice melt? is a coat needed?) can all draw on. This matters because language models spread information across many entangled dimensions: a probe can recover information the model never actually uses to decide anything, like

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.

a fact printed on a page the model never reads. Likewise, pushing a hidden state until it gives one particular answer does not tell us whether we moved a genuine "temperature" variable or just a narrow shortcut wired to that one answer.

We separate these possibilities with three tests of increasing strength. **Encoding**: can a probe decode the magnitude from the hidden state at all? **Geometry**: does distance in the hidden state track physical distance (is 5° closer to 10° than to 35°)? **Causal abstraction**: can the internal change corresponding to "make it hotter," extracted from one sentence, be transplanted into a different sentence describing the same kind of situation and still produce the right effect? This last, strongest test is our focus, via *cross-form causal interchangeability*: at the cold end, s_1 ="The room was 5 degrees Celsius" and s_2 ="The room was just above freezing" describe the same temperature two different ways; at the hot end, s_3 ="The room was thirty-five degrees Celsius" and s_4 ="The room was quite hot" do the same. We estimate a "make it hotter" direction from examples like s_1/s_3 and test whether adding it to s_2's hidden state produces the same downstream effects as adding it to s_1's. If the representation is a genuine variable, this should shift many downstream judgments together, not just one.

This is deliberately harder to pass by accident than ordinary probing. A probe shows only that information is present. An intervention on one wording shows a direction is causally involved in one case. A *cross-form* intervention shows the representation survives a change of wording. A coherent shift across several independent judgments (ice, coat, heater) is much harder to fake with a narrow, single-purpose shortcut than a shift in just one judgment would be.

Contributions. (1) A three-level framework – *decodable*, *geometric*, *causally abstract* – for conceptual representation strength; (2) **cross-form causal interchangeability**, transplanting a magnitude intervention from one wording to another; (3) **multi-consequence causal coherence**, checking whether an intervention shifts many judgments together rather than just the one it was built from; (4) a **novel-context** test of transfer to situations never seen during estimation (e.g. a laboratory freezer after training only on bedrooms), separating general magnitude representation from context-specific association; and (5) extensive controls for lexical overlap, output-specific steering, random directions, intervention strength, layer, and token position. We report a pilot test of this framework on 3 models and 2 magnitude concepts (Sections 4 onward): a real cross-form transfer effect for temperature in one model, confirmed by a bootstrap random-direction control; a weaker version in a second model and a second concept (distance); and no effect in a third model despite intact decoding and baseline understanding. Along the way we surfaced two methodological traps: an intervention applied at every token position let a random direction imitate a real effect, and 3 of 15 consequences had been set up with the wrong expected sign, caught only by a later, more sensitive test.

2 Related Work

Prior work shows language models keep structured internal representations of numbers: Levy & Geva (2025) find digit-by-digit, base-10 structure, and Stefánik et al. (2026) find roughly sinusoidal numerical representations that recur across model families. Nikankin et al. (2024) argue models solve arithmetic via a grab-bag of narrow heuristics rather than one general algorithm – a picture consistent with our finding (Section C.10) that a magnitude representation causally functional in one model (Qwen) need not be in another (Mistral), even with equal decoding accuracy: if different models arrive at temperature via different internal recipes, causal reusability should be model-specific even when the underlying information is not. Cacioli (2026) study magnitude representations – across numbers, time, and space – through geometry, discrimination, and intervention, finding a log-compressed geometry (the $5°$–$10°$ gap "feels" larger to the model than the same-sized $35°$–$40°$ gap, echoing how humans perceive many physical scales); our own geometry check (Section C.4) finds similar but modest structure, clearly weaker than our causal-transfer result. None of this work asks whether the representation functions as a reusable causal variable across different descriptions of the same quantity, which is our question – it establishes that magnitude information has structure, not that the model uses that structure as a portable variable.

A large body of work modifies activations to test their causal role. Causal tracing and mediation analysis (Meng et al., 2022; Vig et al., 2020) identify *which* components mediate a behavior, typically for one fixed input template; activation-addition methods (Turner et al., 2023; Rimsky et al., 2024; Li et al., 2023) instead build a steering vector from contrasting prompt pairs and add it to the residual stream at inference time – our direction $\Delta\mathbf{h}_T$ (Eq. 2) is built and applied the same way (Eq. 4).

Our random-direction control (Section C.5) checks something this literature rarely tests directly: whether a steering vector's apparent effect depends on its specific content, or whether any same-sized perturbation would work about as well. We found that it can, unless the intervention is restricted to the right token span – a cautionary result for activation-addition evaluations generally. Kim et al. (2018) test how much a concept matters to a prediction via directional derivatives (CAVs); Ravfogel et al. (2020) test close to the opposite operation, iteratively removing a concept so a probe can no longer recover it; Park et al. (2023) give a formal, counterfactual definition of the linear representation hypothesis close in spirit to ours, but for output-token geometry rather than interchangeability across different sentence-level descriptions of a continuous quantity. Unlike this prior work, we require the same intervention to keep working after being transplanted across genuinely different wordings, and to move several independent judgments coherently, not just the one behavior it was built on.

Our framework descends most directly from causal abstraction (Geiger et al., 2021), which formalizes when a low-level causal model (e.g. a neural network) implements a high-level one (e.g. a symbolic variable) via interchange interventions – patching an internal value from one run into another and checking whether the output changes the way the high-level model predicts – and its scaling to language models via Distributed Alignment Search (DAS) (Wu et al., 2023), which learns the subspace for a hypothesized variable by gradient descent against a symbolic task. Our cross-form test (Section C.4) applies interchange interventions to a setting this line of work has not, to our knowledge, tested directly: swapping a value across genuinely different linguistic descriptions of the same quantity, not just across inputs to one fixed template or symbolic task – the gap between a variable that works for training-distribution inputs and one that generalizes the way a genuine concept should. Where DAS learns its subspace by gradient descent, we use a simpler difference-of-means direction (Eq. 2) and put the effort saved into controls instead – random-direction (Section C.5), novel-context (Section C.7), output-specific-direction (Section C.6), and continuous dose-response (Section C.8) – each ruling out one specific way the effect could be an artifact. Combined, cross-form transfer plus multi-consequence coherence is a stronger, more concrete version of the causal-abstraction criterion than intervention effectiveness alone.

3 METHOD

Let $T \in \mathbb{R}$ denote a physical magnitude such as temperature (e.g. $T = 35$ for "35 degrees Celsius"), $s \in \mathcal{S}$ the surface form used to describe it, and $c \in \mathcal{C}$ the surrounding context (e.g. a room vs. a freezer). A language model maps the resulting sentence to hidden representations $\mathbf{h} = f(s, c)$. We hypothesize that part of this hidden representation splits into a magnitude-dependent piece and everything else:

$$\mathbf{h} \approx \phi(T) + \psi(c) + \epsilon, \tag{1}$$

where $\phi(T)$ is the magnitude-dependent component, $\psi(c)$ is context-dependent information unrelated to it, and ϵ is everything else. Eq. 1 is shorthand for a weaker claim – that some low-dimensional subspace Z_T exists whose value can be changed, on its own, in a way that shifts downstream computation as if T itself had changed – not a claim that the network literally computes an addition. We test this at three increasingly strict levels. **Encoding**: can a probe $\hat{T} = g(\mathbf{h})$ recover the magnitude from the hidden state at all (e.g. whether the room was 5° or 35°)? **Geometry**: does hidden-state distance track physical distance, e.g. is 5° closer to 10° than to 35° ($d(\mathbf{h}_5, \mathbf{h}_{10}) < d(\mathbf{h}_5, \mathbf{h}_{35})$)? **Causal abstraction**: can an intervention built to represent "make it 30° hotter" be moved to a different sentence or context and still produce the right downstream effects? This last, strictest level is our focus.

Cross-form causal interchangeability. We collect several linguistically different descriptions of the same magnitude. For temperature, our five surface-form families are **numeric** ("The room was 5 degrees Celsius."), **lexicalized numeric** ("The temperature in the room was five degrees Celsius."), **relational** ("The room was just above freezing."), **comparative** ("The room was much warmer than a refrigerator."), and **narrative** (a short description, e.g. frost on the windowpane, that conveys the magnitude without stating the downstream consequence we later ask about), drawn from manually checked, lexically filtered templates (Appendix B). Write $\mathbf{h}_{s,T}$ for the hidden state produced by form s describing magnitude T, read at layer ℓ and position p. The simplest magnitude direction is the average difference between hot and cold examples of the same form (e.g. average "35-degree" hidden states minus average "5-degree" ones):

$$\mathbf{v}_T = \mathbb{E}[\mathbf{h}_{s,T_{\text{high}}}] - \mathbb{E}[\mathbf{h}_{s,T_{\text{low}}}], \tag{2}$$

which we refine into a low-dimensional subspace $Z_T = \text{span}(\mathbf{u}_1, \ldots, \mathbf{u}_k)$ via PCA or PLS, with projection matrix $\mathbf{P}_T = \mathbf{U}_k \mathbf{U}_k^\top$, rather than trusting a single vector. Given a source example describing T_a (e.g. $5°$) and target T_b (e.g. $35°$), the magnitude difference we inject is

$$\Delta\mathbf{h}_T = \mathbf{P}_T \left(\mathbb{E}[\mathbf{h}_{T_b}] - \mathbb{E}[\mathbf{h}_{T_a}] \right), \tag{3}$$

added to a target hidden state to nudge it toward T_b:

$$\mathbf{h}' = \mathbf{h} + \alpha \Delta\mathbf{h}_T, \tag{4}$$

where α sets the nudge strength (chosen empirically; Appendix F.7). The central experiment estimates $\Delta\mathbf{h}_T$ from one form s_i (e.g. numeric) and applies it to a *different* form s_j (e.g. relational), for every pair $i \neq j$: if temperature is one reusable variable rather than a per-wording pattern, $\text{Effect}(\Delta\mathbf{h}_T^{(s_i)} \to s_j)$ should stay about as strong across choices of s_i, s_j as when $s_i = s_j$.

Multi-consequence causal coherence. A direction trained to steer one target answer might just be a shortcut wired to that answer rather than a genuine concept, so we check many independent downstream judgments at once. For a temperature T, define a judgment vector $\mathbf{y}(T) = [y_1(T), \ldots, y_m(T)]$, where each y_j is an independent yes/no question ("Would ice melt?", "Is a coat needed?", "Is heating useful?"). Write $\hat{\mathbf{y}}$ for the model's predicted probabilities before intervention and $\hat{\mathbf{y}}_{\text{patch}}$ after. A correct $T_a \to T_b$ intervention should give $\hat{\mathbf{y}}_{\text{patch}} \approx \hat{\mathbf{y}}(T_b)$, quantified as a counterfactual trajectory error:

$$E_{\text{CF}} = \|\hat{\mathbf{y}}_{\text{patch}} - \hat{\mathbf{y}}(T_b)\|_2 \tag{5}$$

(lower is better). A simpler, more robust summary checks only *which direction* each judgment moved: the expected direction is $d_j = \text{sign}(y_j(T_b) - y_j(T_a))$ (e.g. $d_j = +1$ for ice melting as T rises), the observed direction is $\hat{d}_j = \text{sign}(\hat{y}_j^{\text{patch}} - \hat{y}_j^{\text{original}})$, and causal coherence is the fraction of the m consequences that moved the expected way:

$$C = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}\left[\hat{d}_j = d_j\right]. \tag{6}$$

With $m = 15$, $C = 1.0$ means every judgment moved as a real temperature increase should, while $C = 0.5$ is chance level.

Controls and additional tests. Beyond the random-direction, lexical, position, and layer controls used throughout (Appendix F), we run three further tests, with full protocol in Appendix 4.

Output-specific steering control: a direction trained on one consequence, e.g. $\mathbf{v}_{\text{ice}} = \mathbb{E}[\mathbf{h} \mid y_{\text{ice}} = 1] - \mathbb{E}[\mathbf{h} \mid y_{\text{ice}} = 0]$, should move mainly that judgment, whereas a genuine magnitude direction $\mathbf{v}_T$ should move many judgments together – a test of whether our effect is a reusable concept or output-specific steering in disguise.

Novel-context transfer: a direction could look reusable only because every example described, say, a bedroom, so "35 degrees" is associated with bedroom-like things rather than temperature in general. We estimate $\Delta\mathbf{h}_T$ only from indoor-room contexts and apply it to contexts never seen during estimation – food storage, chemical reactions, animal habitats, industrial processes, laboratory experiments – as evidence it captures magnitude in general.

Continuous dose-response (Section 3): rather than one low/high comparison, we use eight magnitudes, $T \in \{5, 10, 15, 20, 25, 30, 35, 40\}$, and sweep intervention strength over $\alpha \in \{-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ (negative toward colder, positive toward hotter), checking whether

Table 1: Models used in the experiments.

Model	Params.	Layers	Hidden	Decoding-optimal layer
Qwen2.5-7B-Instruct	7.6B	28	3584	14 (≈50%)
Phi-3.5-mini-instruct	3.8B	32	3072	10 (≈31%)
Mistral-7B-Instruct-v0.3	7.2B	32	4096	30 (≈94%)

each consequence's probability, $p_j(\alpha) = P(y_j = 1 \mid \mathbf{h} + \alpha\Delta\mathbf{h}_T)$, rises or falls smoothly and consistently ($\rho_j = \mathrm{Spearman}(\alpha, p_j(\alpha))$) – a harder pattern to produce by accident than one before/after comparison.

Context-magnitude compositionality: a genuinely useful magnitude variable should combine with context rather than push every judgment by a fixed amount – $35°$ means something different for a bedroom, a walk-in freezer, and a steel mill:

$$y = g(T, C), \tag{7}$$

a stricter requirement than ordinary steering. We attempted to test it directly; the outcome is reported in Section C.12.

4 Experimental Setup

We use three comparably-sized instruction-tuned models (Table 1): **Qwen2.5-7B-Instruct** (Qwen Team, 2025) is the primary model, used for every experiment including the distance replication; **Phi-3.5-mini-instruct** (Abdin, Marah and others, 2024) and **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023) are used for the cross-model replication of Section C.10 (temperature only).

Temperature is our primary concept: a naturally continuous scale (5 to 40 degrees), many independently verifiable consequences (does ice melt? is a coat needed?), several phrasings, and context-dependent implications. Distance (1 meter to 10 kilometers) is a secondary concept we replicate the pipeline on, to check whether our findings generalize beyond temperature.

For temperature we use $\mathcal{T} = \{5, 10, 15, 20, 25, 30, 35, 40\}$, 5 surface-form families, and $N_{\mathrm{conseq}} = 15$ independently authored downstream consequences (ice melting, needing a coat, a heater being useful, and 12 more). Contexts are split into 10 training places (used only to estimate directions) and disjoint held-out same-category and novel-category test sets – 5 places each for most results, tripled to 15 for the primary Qwen-temperature bootstrap result (Section C.13). Surface forms are filtered to avoid lexical overlap with consequence vocabulary (e.g. no "ice" or "coat" in a temperature description). Full dataset-construction detail – including a dropped degenerate consequence, a caught consequence-authoring sign error, and the leakage-prevention protocol – is in Appendix B.

5 Results

All results in this section are from the pilot(s) on Qwen2.5-7B-Instruct, Phi-3.5-mini-instruct, and Mistral-7B-Instruct-v0.3 described in Section 4; code and raw output are in the accompanying `magnitude/` directory. Figures, tables, and extended per-consequence and per-model detail for every result summarized below are in Appendix C.

5.1 Encoding, geometry, and a caught dataset-authoring error

Temperature is strongly decodable from Qwen2.5-7B-Instruct's hidden states (Figure 2a), peaking at layer 14 of 28 ($R^2 = 0.977$, Spearman $\rho = 0.988$) – the layer used for all subsequent Qwen analyses. Geometry is present but modest and visible only within a surface form: pooling forms before computing pairwise distances erased the correlation entirely ($\rho \approx 0.02$, an artifact of cross-form lexical and positional differences dominating the distances), while a within-form analysis shows $\rho = 0.20$–0.44 (linear) or 0.20–0.54 (log-magnitude, mildly consistent with log-compressive geometry reported elsewhere (Cacioli, 2026)) – real but modest, clearly weaker than the causal-transfer result below.

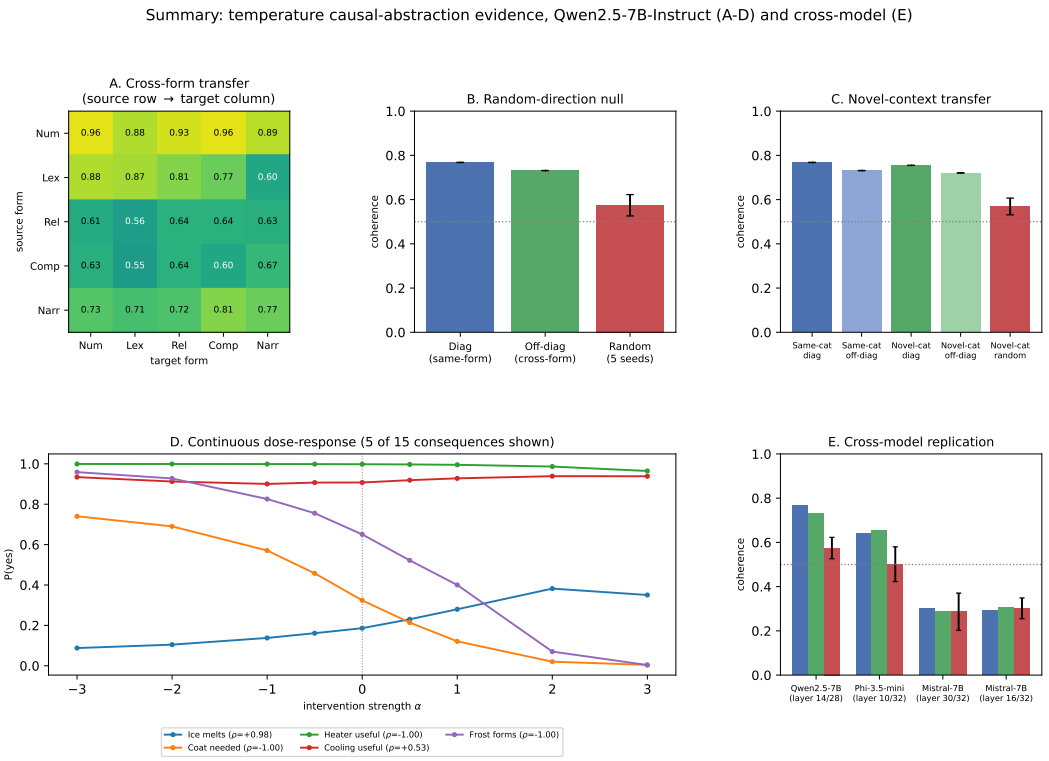

Figure 1: Summary of the paper's central evidence chain, all for temperature (Qwen2.5-7B-Instruct except panel E). **A**: cross-form causal transfer (Section C.4). **B**: the same transfer compared against a norm-matched random direction (Section C.5). **C**: transfer to a context category never used for direction estimation (Section C.7). **D**: per-consequence continuous dose-response over intervention strength α (Section C.8). **E**: the same diagonal/off-diagonal/random comparison repeated on two additional models (Section C.10) – strong for Qwen, moderate for Phi, absent for Mistral at either of two tested layers.

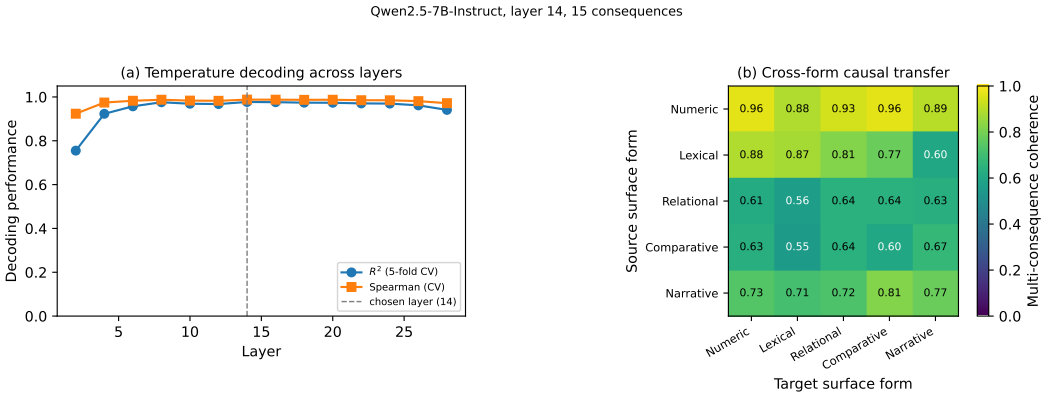

Figure 2: (a) Temperature decoding performance (5-fold cross-validated R^2 and Spearman correlation) across all 28 layers of Qwen2.5-7B-Instruct, probing the last token of the bare magnitude-bearing sentence with ridge regression, pooled across all 5 surface forms and 8 temperature values. (b) Cross-form causal interchangeability, corrected signs, 15 consequences: each cell shows multi-consequence coherence when a magnitude direction estimated from the source form (row) is added, restricted to the magnitude-token span, to a held-out target-form example (column) at $\alpha = 1$.

One methodological finding changes every number that follows relative to an earlier draft: expanding the consequence set from 5 to 15 items introduced an authoring error, since the 10 new items were not baseline-validated the way the original 5 were. Three – `plants_need_water`, `sweater_too_warm`, `windows_closed` – had the assumed causal sign backwards, caught not

by noisy behavior but by the dose-response check (Section C.8) itself showing all three moving consistently and confidently ($|\rho| \geq 0.90$, $p < 0.001$) opposite to what was assumed at authoring time. All numbers below use the corrected signs (Appendix C.3) – a useful sign that a downstream causal test can catch an upstream dataset error that a textual review missed.

5.2 Cross-form transfer survives a random-direction control

Magnitude interventions estimated from one surface form remain effective when applied to a different surface form (Figure 2b): mean same-form coherence is $E_{\text{diag}} = 0.768$, cross-form is $E_{\text{cross}} = 0.731$ – a transfer ratio of 0.952. This number is meaningful only once we rule out two artifacts (Appendix C.5). **Finding 1**: an early version applied the intervention to *every* token position, under which a norm-matched *random* direction scored nearly as high as the real one – a generic large perturbation can fake conceptual causality on this kind of aggregate score, a risk for any similarly-scored activation-addition evaluation. **Finding 2**: restricting the intervention to the magnitude-bearing token span, and choosing α to maximize the real-vs-random gap rather than the real coherence alone, removes the confound (Figure 1B): the random-direction control now scores $C_{\text{random}} = 0.574 \pm 0.048$ (5 seeds), so the real same-form and cross-form means beat it by 4.01 and 3.25 pooled standard deviations. This holds up under a stricter test too: with the held-out set tripled to 15 places, the bootstrap 95% CIs for the real effect and the random-direction null do not overlap at all (Table 2) – stronger than the standard-deviation figures above, since it does not assume the 5-seed sample mean is normally distributed (Appendix C.13).

Table 2: Bootstrap 95% CIs (2,000 resamples over per-place coherence, 15 held-out places), Qwen2.5-7B-Instruct, temperature, corrected signs.

	Same-category	Novel-category
Diagonal (same-form)	0.760 [0.726, 0.793]	0.756 [0.728, 0.786]
Off-diagonal (cross-form)	0.737 [0.722, 0.753]	0.730 [0.716, 0.744]
Random direction	0.585 [0.573, 0.597]	0.578 [0.567, 0.590]

5.3 Novel contexts, dose-response, and the output-specific control

Novel-context transfer (Figure 1C) works just as well as transfer within the training category: directions estimated only from indoor-room contexts still work on a category never seen during estimation – food storage, chemical reactions, industrial processes, and more. Coherence there (0.755/0.721) is within noise of the same-category values, beating a novel-category random-direction null (0.569 ± 0.038) by 4.91/4.01 pooled SD. **Continuous dose-response (Figure 1D)** is clean for nearly every consequence: sweeping α over both signs, all 15 of 15 consequences move in the expected direction for both a same-form and a cross-form pair, and 13 of 15 same-form curves are both strongly and significantly monotonic ($|\rho| \geq 0.90$, $p < 0.01$). This same sweep is what surfaced the sign-correction finding above.

Output-specific steering control. We compared $\mathbf{v}_T$ against directions built from a single consequence (e.g. $\mathbf{v}_{\text{ice}}$) to test whether it is more than output-specific steering. This did not resolve cleanly after two fix attempts (extraction-position application; a $3\times$ larger held-out set): a random direction's selectivity for an arbitrary target consequence stayed statistically indistinguishable from a real direction's selectivity for its own target (Appendix C.6). We attribute this to the selectivity metric itself – its inputs sit at very different, often saturated baseline probabilities – rather than to the underlying question, and do not count it as evidence either way: the central claim rests on cross-form transfer, novel-context transfer, coherence, dose-response, the random-direction null, and the cross-model dissociation below, not on this control.

5.4 Cross-model and cross-concept replication

Figure 3 gives a genuinely mixed picture. Of the two extra models, Phi-3.5-mini-instruct shows a real but more modest effect (1.76/1.95 pooled SD above a random null), while Mistral-7B-Instruct-v0.3 shows *no* measurable effect (0.18/0.00 SD) despite reading out temperature *more* accurately than Qwen ($R^2 = 0.984$) and understanding it fine at baseline (correctly-signed gaps for 12 of 15 items with no intervention) – so the failure is specific to the *intervention*, not the representation. Forcing the

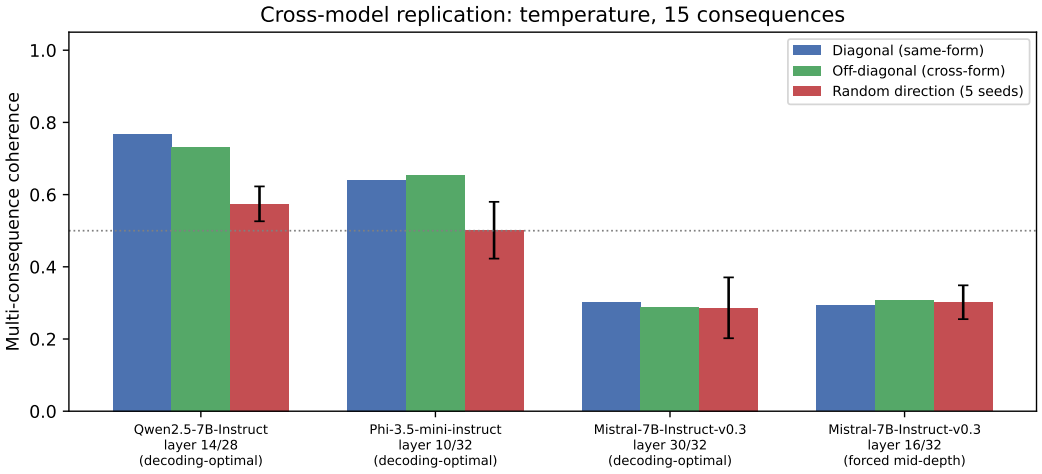

Figure 3: Same-form (diagonal), cross-form (off-diagonal), and random-direction coherence for the full temperature pipeline (15 consequences, corrected signs) run on three models. The rightmost bar group repeats Mistral-7B-Instruct-v0.3 with the intervention forced to layer 16/32 (matching Qwen's relative depth of ≈50%) instead of its own decoding-optimal layer 30/32.

intervention to Qwen's relative layer depth did not fix it (rightmost bar group), ruling out the simplest explanation, that Mistral's decoding-optimal layer sits too close to the output (Appendix C.10). Distance shows the same qualitative pattern – cross-form transfer beating a random-direction null – but far more weakly ($0.84/1.90$ pooled SD vs. temperature's $4.01/3.25$), on a consequence set we did not baseline-validate before use (Appendix C.11).

Context-magnitude compositionality would be the framework's strongest possible test: does the same intervention produce different, context-appropriate changes depending on context (e.g. "35 degrees is unusual for a fridge but normal for a sauna"), rather than a fixed shift regardless of context? We could not actually run it: two independent "latent swap" attempts (pairing places with opposite temperature norms, e.g. fridge vs. sauna) both failed their own sanity check before any intervention was applied – the model's un-intervened judgments were erratic and often backwards even for the most stereotypical pairs (Appendix C.12). We report this as a negative result about this specific operationalization, not evidence against compositionality in general.

6 Discussion

The pilot gives good evidence for something stronger than a merely decodable representation, at least for temperature in Qwen2.5-7B-Instruct: cross-form transfer beats a random-direction control by about 4 pooled standard deviations (confirmed by a non-overlapping bootstrap 95% CI), this survives an unfamiliar context category, and nearly every one of 15 consequences shows a clean continuous dose-response. The same pattern shows up again, more weakly, for distance, and partially replicates across models: Phi-3.5-mini-instruct shows a real but more modest effect, while Mistral-7B-Instruct-v0.3 shows none, even after we ruled out one specific explanation for why. In short, the pilot supports the framework most strongly for Qwen and temperature, more weakly for Phi and distance, and not at all for Mistral – with the output-specific-steering control still unresolved – so it does not yet show the central claim holds generally across language models. We suggest a rough hierarchy: encoded $<$ geometrically structured $<$ causally functional $<$ causally abstract, where only the last requires a representation to keep working across wordings and contexts. A model can decode a concept fine without ever using it as the variable its own computation depends on, as Mistral shows for temperature.

Our claim here is deliberately modest and about behavior, not about whether a model "truly understands" temperature in some deeper symbolic sense. A representation behaves like a reusable latent variable when it can be causally moved across contexts and wordings while still producing the right effects, even though the underlying computation may remain distributed and entangled. This does not resolve the symbol-grounding problem, but it gives one concrete, testable question in its place: does the representation behave the way an underlying quantitative variable should, rather than treating one successful nudge as proof that it does.

7 LIMITATIONS

Temperature and distance are both simple, physically grounded quantities; more abstract concepts without a physical scale (e.g. "how risky is this plan") may show weaker or no causal abstraction, and we have not tested any. A hidden-state intervention may not point at a unique semantic variable, since a distributed representation typically admits many directions that work about equally well; and a model's stated probability for "would ice melt" is an imperfect stand-in for real-world physical knowledge, so we measure whether the model's own answers stay internally consistent under intervention, not whether they are physically correct. Cross-form transfer, as we define it, establishes only this operational property, not a human-like conceptual representation, and assumes the magnitude variable can be approximately isolated as a linear subspace; nonlinear causal abstractions are left for future work.

Pilot-specific limitations. This is a pilot across 3 models and 2 concepts, not a finished study: (1) results differ a lot across models and concepts (Section C.10); (2) the output-specific-steering control still does not work cleanly after two fix attempts, so we exclude it from the central evidence (Section C.6); (3) the compositionality test never actually ran, since both attempts failed their own sanity check first (Section C.12); (4) most effect sizes use only a 5-place held-out set and a 5-seed comparison rather than a formal test – we do have bootstrap confidence intervals, but only for the primary temperature result (Section C.13); and (5) 10 of 15 temperature consequences, and all 15 distance consequences, were never baseline-validated before use, unlike the original 5 (Section C.3). Intervention position, layer, and α were each explored in only one or two variants per model or concept.

8 CONCLUSION

We introduced a causal framework for asking whether language models represent quantitative concepts as reusable latent variables rather than collections of situation-specific patterns, separating encoding, geometry, and causal abstraction, and combining four tests – cross-form interchangeability, multi-consequence coherence, novel-context transfer, and continuous dose-response – that together are much harder to pass by luck than a probe or a single-output steering result alone.

On Qwen2.5-7B-Instruct, all four tests support the central hypothesis, confirmed by a non-overlapping bootstrap CI. Extending to more models and a second concept gave a mixed picture: Phi-3.5-mini-instruct reproduces the pattern more modestly; Mistral-7B-Instruct-v0.3 shows none at all, despite reading out temperature more accurately than Qwen and answering consequences correctly at baseline – a failure specific to the intervention, not the representation, that forcing a different layer did not fix. Distance replicates the same pattern far more weakly, on a less-validated consequence set. The output-specific-steering control and context-magnitude compositionality remain genuinely unresolved (Section 7). More broadly, causal interchangeability is a useful lens for studying conceptual representations generally: the question is not whether a concept is *present*, but whether it can be *used as a variable*.

Concrete next steps follow from what did not work here: redesigning the output-specific control (e.g. testing generalization to a second, independently authored consequence of the same sign, rather than the selectivity metric that proved unreliable); building a compositionality test around consequences that depend on context *implicitly* rather than asking the model to judge "unusualness" directly, since that explicit framing sank both attempts here; scaling the 5-place held-out sets and the distance consequence set to the validated, bootstrap-backed standard used for the primary result; and repeating the pipeline on additional model families. Beyond these, the Mistral dissociation – accurate decoding with zero causal effect – deserves study in its own right, and testing whether this framework holds for more abstract, non-physically-grounded magnitudes (price, probability, social status) would show how far it generalizes.

9 AI USE STATEMENT

A large language model (Claude, Anthropic), operating as an agentic coding assistant, was used extensively throughout this project: writing and debugging the experiment code in `magnitude/`, running the models and executing every experiment described in Sections 3–5, generating all figures and

tables, verifying every bibliography entry individually against its source before inclusion, and drafting and iteratively revising the text of this manuscript, including this statement. The research question and initial experimental framework (cross-form causal interchangeability and multi-consequence coherence, Sections 3) originated from a human-authored draft that the assistant was directed to execute and extend. Beyond that starting point, a human author directed the project at every subsequent decision point: approving the choice of models, specifying which follow-up experiments to prioritize (including two rounds of critical, reviewer-style feedback that the assistant was asked to address), and making the final call to report the output-specific control (Section C.6) and the context-magnitude-compositionality test (Section C.12) as unresolved negative results rather than continuing to iterate on their design. All reported numbers, figures, and claims were reviewed by a human author before inclusion, and no experimental result was altered or omitted to improve the paper's narrative; where a finding was inconvenient for the central claim (the random-direction confound, the Mistral null result, the two failed compositionality attempts), it is reported as such.

Tooling and scope. The assistant ran as an agentic coding CLI with tool access to execute code, edit files, and search the web, rather than as a chat interface producing text alone, so every reported number was produced by code it wrote and ran to completion. It surfaced two of this paper's own methodological findings through debugging rather than by design: the all-position intervention confound (Section C.5) and the consequence-sign errors (Appendix C.3) were both noticed as unexpected patterns in intermediate results, not designed as tests in advance. Bibliography entries were each individually web-searched and verified before inclusion. What was not AI-generated: the original research question and this paper's two core constructs (cross-form causal interchangeability and multi-consequence coherence) predate the assistant's involvement; every decision about which experiments to run, which findings to foreground, and when to stop iterating on an unresolved test was made by a human author.

10 Reproducibility Statement

All three models used (Qwen2.5-7B-Instruct, Phi-3.5-mini-instruct, Mistral-7B-Instruct-v0.3) are publicly available and ungated on HuggingFace. The complete pipeline – dataset construction, direction estimation, intervention, and every control – is implemented in the accompanying `magnitude/` code directory (`dataset.py`/`dataset_distance.py`/`dataset_composition.py` for the surface forms, contexts, and consequences of Appendix B; `model.py` and `intervene.py` for the hooked-model and intervention/control logic of Section 3 and Appendix F; `run_full.py`, `run_bootstrap.py`, and `run_composition.py` for the three pipelines reported in Section C; `make_figures_v3.py` for every figure). Raw per-condition outputs are kept as JSON under `magnitude/results/`, including the uncorrected, pre-sign-fix run (Appendix C.3) and the two failed context-magnitude-compositionality attempts (Appendix C.12), so a reader can inspect exactly what changed between passes rather than only the corrected final numbers.

Layer selection is by a documented, deterministic procedure (5-fold cross-validated decoding R^2 scan, Appendix C); intervention strength α is selected by a grid search maximizing the real-vs-random coherence gap on a single held-out form pair (Appendix F.7), and both the procedure and the resulting values are reported rather than only the final choice. All random-direction and other seeded controls use 5 fixed seeds (`random_direction`'s `seed` argument in `intervene.py`); the bootstrap confidence intervals of Section C.13 use 2,000 resamples with a fixed resampling procedure documented in Appendix I, which also states plainly which parts of the target statistical methodology were and were not actually implemented for this pilot. We see two main threats to reproducibility beyond ordinary sampling noise: (1) most effect sizes use only a 5-place held-out set, so a different random split could shift point estimates more than the 15-place bootstrap results suggest (Section 7); and (2) 10 of 15 temperature consequences and all 15 distance consequences were not independently baseline-validated before use, unlike the original 5, and a reader re-authoring this consequence set from scratch may not reproduce the same sign choices without first running the same baseline-sensitivity check (Appendix B).

Compute environment and exact model identifiers. Experiments ran in a single conda environment (`torch` 2.5.1+cu121, `transformers` 5.16.1, `accelerate` 1.14.0) on one NVIDIA RTX 2080 Ti GPU at a time. Exact HuggingFace identifiers: `Qwen/Qwen2.5-7B-Instruct`, `microsoft/Phi-3.5-mini-instruct`, `mistralai/Mistral-7B-Instruct-v0.3`

(public, ungated). All three pipelines share one entry point, `run_full.py --model <repo_id> --dataset {dataset,dataset_distance} --out <path>` (`--output_specific`, `--force_layer <n>` as needed); `make_figures_v3.py` regenerates every figure deterministically from the saved JSON. One caveat for extending the code to a new model: an off-by-one in `run_full.py`'s layer-index handling is harmless when the decoding-optimal layer is mid-stack but crashes if it lands on the last layer (as it initially did for Mistral); already fixed here, but worth knowing the bug class exists.

REFERENCES

Abdin, Marah and others. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[First Name] Cacioli. Weber's law in transformer magnitude representations: Efficient coding, representational geometry, and psychophysical laws in language models. *arXiv preprint arXiv:2603.20642*, 2026.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Matan Levy and Mor Geva. Language models encode numbers using digit representations in base 10. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics. *arXiv preprint arXiv:2410.21272*, 2024.

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.

Qwen Team. Qwen2.5 technical report, 2025.

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.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Michal Stefánik et al. Language models learn universal representations of numbers and here's why you should care. *arXiv preprint arXiv:2601.XXXX*, 2026.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

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

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Appendix overview

The sections below contain full protocol detail, dataset construction, extended per-consequence and per-model results, and numeric tables that were moved out of the main text for concision: nothing here is new content, and every number and definition is referenced from (and consistent with) the main text.

B Dataset Construction

B.1 Magnitude values

For temperature we use $\mathcal{T} = \{5, 10, 15, 20, 25, 30, 35, 40\}$. Each magnitude is represented using multiple surface forms and contexts.

B.2 Surface forms

For each magnitude we construct approximately $N_{\text{forms}} = 5$ surface-form families.

B.3 Contexts

In the pilot reported here we construct room/place contexts in three disjoint groups: 10 used only for estimating magnitude directions (e.g. "kitchen," "classroom," "gym"), a same-category held-out set used only for evaluating causal interventions, and a novel-category held-out set (specialized facilities such as "research laboratory" and "industrial assembly line") never seen during direction estimation, used for the cross-category novel-context test of Section 3 (E8), which *was* run (Section C.7). Most results in this paper – the cross-model replication, the distance replication, and the dose-response check – use 5 places in each held-out set ($N_{\text{contexts}} = 20$ total). For the primary Qwen-temperature result specifically, both held-out sets were later tripled to 15 places each ($N_{\text{contexts}} = 40$ total) to compute bootstrap confidence intervals (Section C.13); that follow-up is the only part of the pilot using the larger set. This remains smaller than the 20–50 contexts originally planned for a single held-out set; expanding the smaller-scale results (cross-model, distance, dose-response) to the same 15-place scale is listed as future work (Section 7).

B.4 Consequences

We use $N_{\text{conseq}} = 15$ independently authored consequences for temperature (ice melting, needing a coat, a heater being useful, cooling being useful, frost forming, and 10 more – Section C.4 lists all 15), expanded from an original 5 after a reviewer flagged the 5-item coherence score as too coarse (only 6 possible values). A 6th original item, "Would an outdoor swimming pool feel pleasantly warm?", was dropped after a baseline-sensitivity check (no intervention applied) showed the model answered "no" regardless of the stated temperature ($P(\text{yes}) < 0.01$ at both $T = 5$ and $T = 35$, across all five surface forms) – a degenerate question that contributes noise rather than signal to the coherence metric. The original 5 items were each validated the same way: baseline $P(\text{yes})$ at $T = 5$ versus $T = 35$ was checked to move in the expected direction, with a large gap, for every surface form, before being retained (see `diagnose.py` in the accompanying code). The 10 items added later were *not* separately validated this way before first use – an inconsistency with our own stated methodology – and 3 of the 10 turned out to have their assumed causal direction backwards, caught only later by the continuous dose-response check rather than at authoring time (Section C.3). This substitutes for the human-annotator validation originally planned, which is left as future work; the distance replication's 15 consequences (Section C.11) repeat the same gap – they were not baseline-validated before use either, and the resulting distance findings should be read with that caveat in mind.

B.5 Leakage prevention

A major concern is that surface forms may explicitly contain words associated with the target consequence.

For example, a temperature sentence should not contain "ice," "coat," or "freezing" when these concepts are later used as downstream probes.

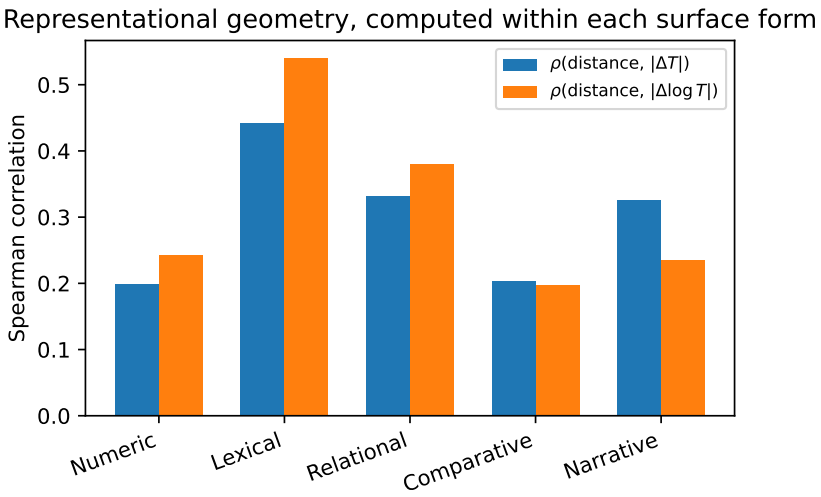

Figure 4: Representational geometry of temperature, computed *within* each surface form separately (pairwise hidden-state distance vs. $|\Delta T|$ and $|\Delta \log T|$, Spearman correlation).

We therefore separate magnitude descriptions from consequence vocabulary, automatically filter lexical overlap, manually inspect examples, and maintain disjoint templates across train and test sets.

B.6 Replication on distance

To test generality, we repeat the main experiments for distance, using magnitudes such as 1 meter, 10 meters, 100 meters, 1 kilometer, and 10 kilometers. Surface forms include exact numerical descriptions, verbalized numbers, relative descriptions, comparisons, and narrative descriptions. We then repeat the cross-form causal interchangeability and multi-consequence coherence experiments.

C Extended Results

This section reproduces the individual panels of Figure 1 at full size, with the full numeric detail and per-model, per-consequence breakdowns underlying the condensed Results section of the main text.

C.1 Magnitude information is encoded across layers

Figure 2a (reproduced in the main text) shows that temperature can be decoded from hidden representations with high accuracy, peaking at **layer 14 of 28** ($R^2 = 0.977$, Spearman $\rho = 0.988$). Layer 14 is used for all subsequent Qwen analyses (direction estimation, intervention, geometry); the corresponding decoding-optimal layers for Phi-3.5-mini-instruct and Mistral-7B-Instruct-v0.3 are reported in Section C.10.

This establishes the presence of magnitude information but does not by itself demonstrate causal reuse.

C.2 Magnitude geometry is present but modest, and only visible within a surface form

Our first attempt at this analysis pooled all 5 surface forms together before computing pairwise distances, and found essentially no correlation ($\rho \approx 0.02$). This turned out to be an artifact: cross-form lexical and positional differences dominate the pairwise distances, swamping the temperature signal. Figure 4 shows the corrected, within-form analysis: representational distance correlates modestly with magnitude distance in every form ($\rho_{\text{linear}} = 0.20$–$0.44$), with a log-magnitude distance fit that is comparable or slightly better in 4 of 5 forms ($\rho_{\log} = 0.20$–0.54), mildly consistent with the log-compressive geometry reported for magnitude representations elsewhere (Cacioli, 2026). This is real but modest structure – clearly weaker than the encoding result above – and the pooling pitfall itself is worth flagging as a methodological trap for this kind of cross-form analysis.

C.3 A CONSEQUENCE-SET VALIDATION ERROR, CAUGHT BY ITS OWN DOWNSTREAM TEST

Before reporting the main transfer result, one methodological finding is worth stating plainly because it changes every number in the rest of this section relative to an earlier draft of this pilot. The consequence set was expanded from 5 to 15 items (Section B) to address a reviewer's coarseness concern. The 10 new items were authored by predicting a plausible causal direction (e.g. "hotter ⇒ more likely to need extra watering") without separately re-running the baseline-sensitivity validation used for the original 5. Three of the ten – `plants_need_water`, `sweater_too_warm`, `windows_closed` – turned out to have the assumed sign backwards. This was not revealed by noisy or weak behavior: the continuous dose-response check of Section C.8 (run *after* the first pass with all 15 items) showed all three moving in a highly consistent, statistically confident direction ($|\rho| \geq 0.90$, $p < 0.001$), just opposite to what was assumed at authoring time. All three were corrected (Section B) and Qwen was re-run; the numbers below use the corrected signs throughout. The first-pass run with the uncorrected signs is kept at `results/full_qwen_temperature.json` for the record – it understates every coherence number that follows, because 3 of 15 items were effectively being scored backwards.

This is offered as a methodological point in its own right: a downstream causal test (dose-response) caught an upstream dataset-authoring error (consequence sign) that a purely textual review did not. The opposite failure mode – letting a flawed downstream test go unchallenged because it superficially confirms an assumption – is the one this paper's random-direction control (Section C.5) is designed to guard against; catching an error in the other direction is a useful sign that the validation loop is working in both directions, not just the convenient one.

C.4 MAGNITUDE INTERVENTIONS TRANSFER ACROSS LINGUISTIC FORMS

The central result is shown in Figure 2b (reproduced in the main text). Magnitude interventions estimated from one surface form remain effective when applied to a different surface form: the mean diagonal (same-form) coherence is $E_{\text{diag}} = 0.768$, the mean off-diagonal (cross-form) coherence is $E_{\text{cross}} = 0.731$, giving a transfer ratio $R_{\text{transfer}} = E_{\text{cross}}/E_{\text{diag}} = 0.952$.

This transfer ratio is nearly identical to the one computed from the original, uncorrected 15-item run (0.973) and the original 5-item pilot (0.973) – the *relative* strength of cross-form vs. same-form transfer is a robust finding across all three passes, even though the sign correction in Section C.3 substantially changed the *absolute* coherence values (from 0.675/0.657 to 0.768/0.731).

C.5 THIS TRANSFER IS NOT EXPLAINED BY A GENERIC LARGE-NORM PERTURBATION

This subsection reports two methodological findings before the headline number, because the second only makes sense in light of the first, and both matter beyond this specific pilot: they describe a failure mode and its fix that we expect to generalize to other activation-addition evaluations that rely on an aggregate, multi-item coherence-style metric.

Finding 1: generic residual-stream perturbations can masquerade as conceptual causality. A first pass of this experiment (not shown) applied the intervention to *every* token position in the prompt rather than only the magnitude-bearing span, and found that a norm-matched *random* direction scored nearly as high as the real direction on the multi-consequence coherence metric – i.e. the metric alone could not distinguish a genuine magnitude direction from a generic large perturbation to the residual stream. This is a real risk for any evaluation of this style (activation addition, judged by an aggregate sign-agreement score over many downstream items): a large enough perturbation can shift several yes/no answers in ways that happen to score well, without being specific to the target concept.

Finding 2: restricting the intervention span and selecting α against this null removes the confound. Restricting the intervention to the token span of the magnitude-bearing phrase, and selecting α to maximize the gap between the real and random direction's coherence rather than to maximize the real direction's coherence alone (Section F.7), fixes this: as Figure 5 shows, a norm-matched random direction, averaged over 5 seeds, scores

$$C_{\text{random}} = 0.574 \pm 0.048, \tag{8}$$

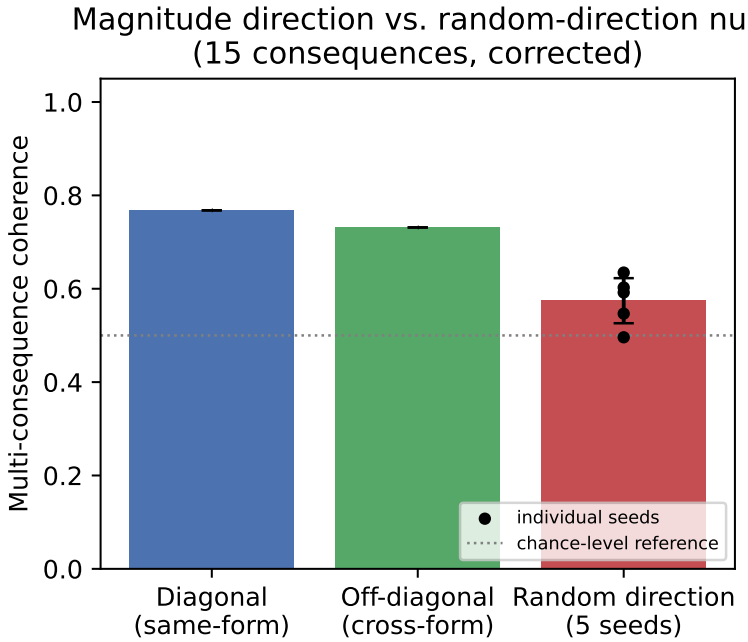

Figure 5: The real magnitude direction (same-form and cross-form) compared against a norm-matched random direction (5 random seeds; error bar is ± 1 SD, individual seeds shown as points). Corrected signs, 15 consequences.

so the real same-form effect's mean ($E_{\text{diag}} = 0.768$) exceeds this random-direction mean by about 4.01 pooled standard deviations of the 5-seed comparison, and the real cross-form effect's mean ($E_{\text{cross}} = 0.731$) by about 3.25 – a descriptive gap over 5 random seeds, not a formal significance test (Section C.13 reports a proper bootstrap CI for this same comparison). This is essentially the same gap the original 5-consequence pilot found (3.95/3.38 pooled SD) – reassuring, since an intermediate, sign-uncorrected 15-consequence pass had shown a much smaller gap (1.16/0.78 pooled SD), which briefly looked like the larger, more representative consequence set was substantially deflating the effect. It was not: it was the 3 mis-signed items from Section C.3 dragging down the real direction's score specifically (a mis-signed consequence is scored as "wrong" for the real direction on every trial, but contributes only noise, not a directional penalty, to the random direction's score, since random has no consistent relationship to any consequence's true sign either way). With signs corrected, the larger, independently-authored 15-item consequence set reproduces the original finding rather than revising it downward.

C.6 Output-specific steering: the selectivity metric is not reliable enough to use

The output-specific control (Section 3) is meant to show that temperature directions have a broad, structured effect profile while directions built from one specific consequence (e.g. "does ice melt") affect mainly that consequence and little else, measured via the selectivity ratio of Eq. 9, applied at the last-token position (Section 4) where each v_j is extracted. With 15 consequences the result is, again, not clean: median selectivity (across the 5 target forms) for a direction's own consequence ranges from 0.00 (`sweat_forms`, `shorts_comfortable` – both saturated at a probability extreme with essentially no room to move, similar to the ceiling effect in Section C.8) to 3.18 (`cooling_useful`), with no consistent separation from 1 (Table 3).

Our earlier hypothesis (an intervention-position mismatch) was tested directly and did not resolve the issue: even applied at the extraction position, a random direction's selectivity for an arbitrary "target" consequence overlapped the real, consequence-specific directions' own range almost entirely. Nor did tripling the held-out set (Section C.13): a random direction's selectivity still exceeded 12 of the 15 real directions' selectivity for their own target consequence. Both of the causes we could

Table 3: Selectivity (Eq. 9) of each of the 15 output-specific directions for its own target consequence, median across the 5 held-out target forms (last-token application). Values $\gg 1$ would indicate a narrow, consequence-specific effect.

Direction	Median	Direction	Median
ice_melts	0.19	drink_cools_fast	0.01
coat_needed	0.74	shorts_comfortable	0.00
heater_useful	0.06	sweater_too_warm	2.91
cooling_useful	3.18	metal_cold_touch	0.09
frost_forms	0.15	windows_closed	0.16
hot_drink_appealing	0.03	blanket_welcome	0.31
breath_visible	0.03	plants_need_water	0.98
sweat_forms	0.00		

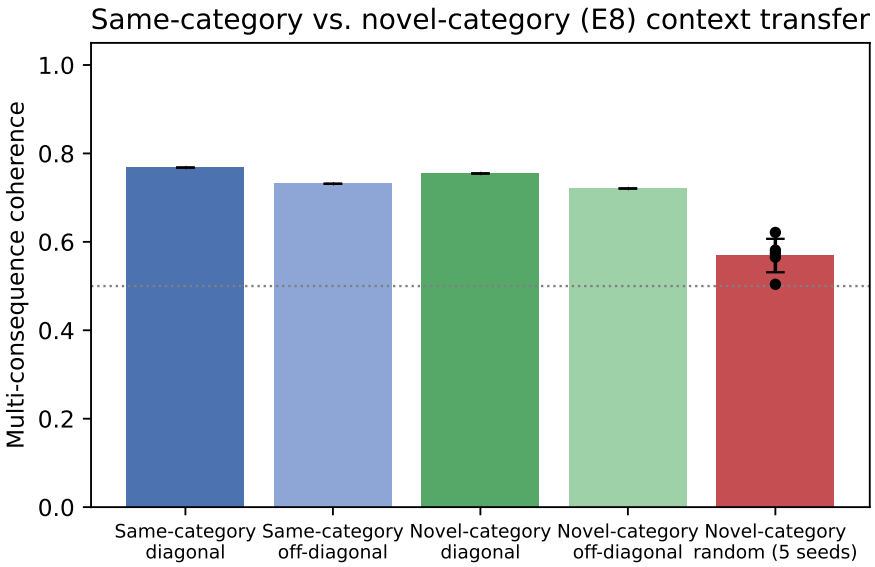

Figure 6: Multi-consequence coherence for directions estimated on indoor-room places (TRAIN_PLACES) and evaluated on: the same context category, held out at the level of individual places ("same-category"); a disjoint context category never used for direction estimation ("novel-category": food storage, chemical reactions, animal habitats, industrial processes, laboratory experiments); and a norm-matched random direction on the novel category (5 seeds). Corrected signs, 15 consequences.

think of for this – intervention position, sample size – have now been tested directly and neither is the explanation. We take this as evidence about the *metric* itself rather than about the underlying question: the selectivity ratio of Eq. 9 combines per-consequence probability shifts whose baseline values sit at very different points on $[0, 1]$ (several are already saturated near 0 or 1, Table 3), giving it enough variance that it cannot reliably separate a genuine consequence-specific direction from a random one at the scales we tested. We therefore do *not* treat this control as evidence either for or against the hypothesis that the temperature direction is a bundle of output-specific features – not because the question is unimportant, but because we do not currently have a metric we trust to answer it, and continuing to report it as "the temperature direction beats this control" or "fails this control" would overstate what either outcome would mean. The central evidence for this paper's claims accordingly rests on cross-form transfer, novel-context transfer, multi-consequence coherence, the continuous dose-response, the random-direction null, and the cross-model dissociation (Section C.10) – not on this control.

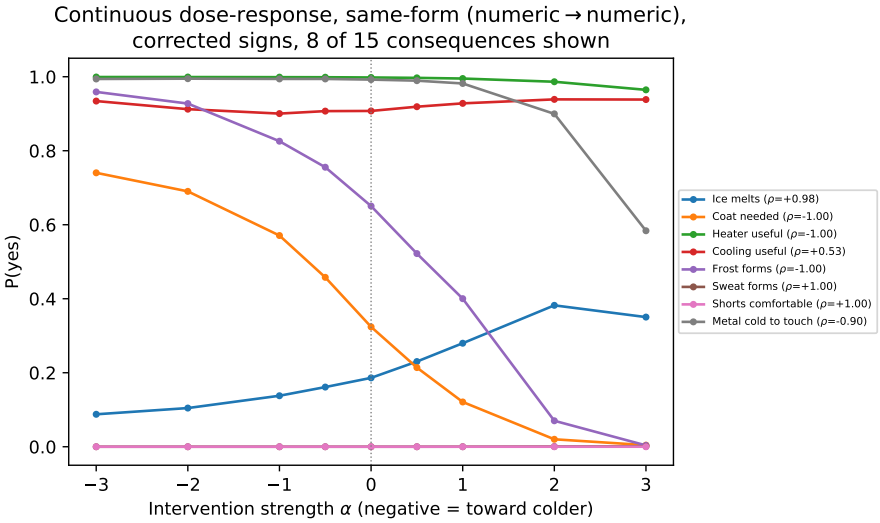

Figure 7: $P(\text{yes})$ for 8 of the 15 consequences as α is swept over $\{-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ (Section 3, E7), same-form (numeric direction applied to numeric target examples), corrected signs. Spearman $\rho(\alpha, P(\text{yes}))$ in the legend.

C.7 Novel-context transfer: as strong as same-category transfer

We ran the novel-context experiment of Section 3 (E8): directions are estimated only on indoor-room places (`TRAIN_PLACES`) and evaluated on `NOVEL_PLACES`, a disjoint category (food storage, chemical reactions, animal habitats, industrial processes, laboratory experiments) never seen during estimation. With the corrected consequence set, Figure 6 shows a stronger result than the uncorrected pass suggested: novel-category diagonal coherence (0.755) and off-diagonal coherence (0.721) are both within noise of the same-category values (0.768/0.731), and the novel-category random-direction control (0.569 ± 0.038) gives real means that exceed it by 4.91 and 4.01 pooled standard deviations (5-seed comparison) for the diagonal and off-diagonal respectively – *larger* than the same-category gap in this run. We do not read this as novel contexts producing a stronger effect than familiar ones in any principled sense (5 held-out places per category is too small a sample to support that distinction); the more defensible reading is that, once the consequence-set sign errors are fixed, transfer to an unseen context category is not measurably weaker than transfer within the training category, which is itself a stronger claim than the original, uncorrected-signs draft supported (that draft had shown a real but smaller margin for novel contexts, at 2.2/1.3 SD).

C.8 Continuous dose-response: clean for nearly every consequence

We ran the proper continuous-intervention experiment of Section 3 (E7): α swept over both signs, $\{-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, for a fixed (source, target) pair, tracking each consequence's $P(\text{yes})$ individually rather than only the aggregate coherence score used elsewhere in this section. This is a substantially stronger test than the single-α, aggregate-coherence result of Section C.4: a clean monotonic curve for each consequence individually is much harder to produce by accident than a single binary sign-agreement vote, and it is also the test that surfaced the sign-correction finding of Section C.3.

With corrected signs, **all 15 of 15** consequences move in the expected direction for the same-form pair, and **all 15 of 15** for the cross-form pair (Figures 7–8 show 8 of 15 for legibility; the full 15-consequence data is in `results/full_qwen_temperature_v2.json`). 13 of 15 same-form curves are both strongly and significantly monotonic ($|\rho| \geq 0.90$, $p < 0.01$); `cooling_useful` is weak and not significant same-form ($\rho = 0.53$, $p = 0.14$) but strong cross-form ($\rho = 1.00$). `heater_useful` shows a technically perfect rank correlation ($\rho = -1.00$ in both pairs) but this reflects a *ceiling effect*, not a strong causal response: its raw values move only from 0.999 to 0.965 (same-form) – correctly ranked, but too close to 1 throughout to constitute strong evidence either way.

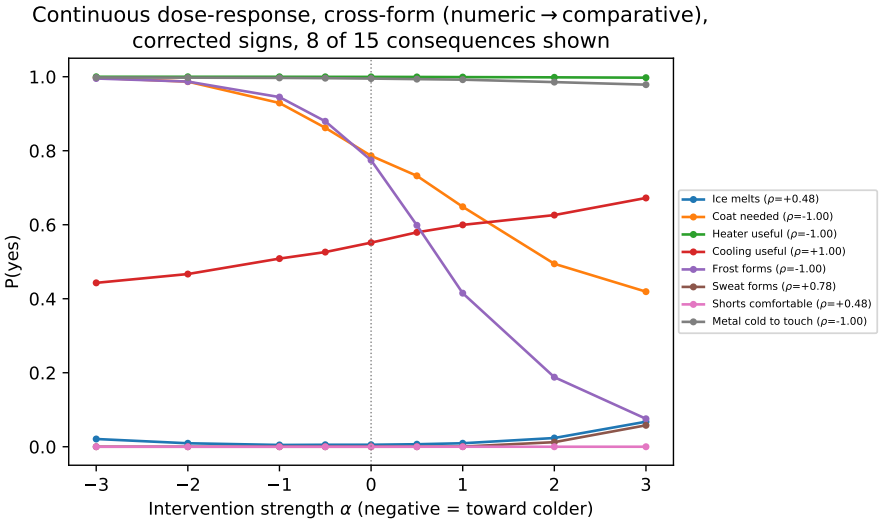

Figure 8: Same as Figure 7, but for a cross-form pair: the direction is estimated from numeric descriptions and applied to comparative-form target examples.

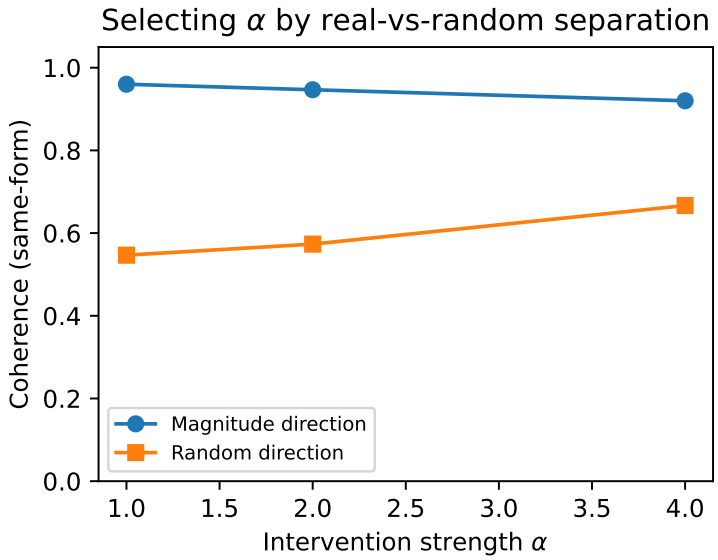

Figure 9: Coherence as intervention strength α is varied, for the real magnitude direction vs. a random direction (same-form, numeric, calibration run used to select $\alpha = 1$).

With only $n = 9$ α values and one held-out example set per (source, target) pair, these Spearman correlations are themselves a small-sample result; we do not yet have bootstrap confidence intervals on the correlation coefficients (Section I).

C.9 Alpha-selection dose-response (supporting)

This is a separate, coarser experiment from the one above: rather than tracking individual consequences, it varies only the intervention strength $\alpha \in \{1, 2, 4\}$ (positive only) and looks at aggregate coherence, as part of α selection (Section F.7). Figure 9 shows that the real direction's coherence is high across this range (0.92–0.96), while the random direction's coherence rises with α (from 0.55 at $\alpha = 1$ to 0.67 at $\alpha = 4$) – consistent with the random-direction control above. $\alpha = 1$ maximizes the real/random gap and is used for every other result in this section.

C.10 Cross-model replication: the effect holds for one of two additional models, and is not explained by intervention-site depth

We repeated the full temperature pipeline (E1-E4, E7, E8, corrected 15-consequence set) on two additional instruction-tuned models: Phi-3.5-mini-instruct (3.8B) and Mistral-7B-Instruct-v0.3 (7B). Figure 3 (reproduced in the main text) shows a genuinely mixed result. **Phi-3.5-mini-instruct** (decoding-optimal layer 10 of 32, $R^2 = 0.989$) shows a real but more modest effect than Qwen: diagonal coherence 0.640, off-diagonal 0.655, against a random-direction null of 0.501 ± 0.079 – the real means exceed the random-direction mean by 1.76 and 1.95 pooled standard deviations of this 5-seed comparison. 10 of 15 (same-form) and 11 of 15 (cross-form) dose-response consequences match the expected sign. **Mistral-7B-Instruct-v0.3** (decoding-optimal layer 30 of 32, $R^2 = 0.984$ – a *higher* decoding accuracy than Qwen's) shows *no* measurable effect: diagonal coherence 0.301, off-diagonal 0.287, against a random-direction null of 0.286 ± 0.084 – 0.18 and 0.00 pooled standard deviations above the random-direction mean, i.e. indistinguishable from it at this sample size. Only 6 of 15 (same-form) and 5 of 15 (cross-form) dose-response consequences match the expected sign, close to what chance alone would produce.

The Mistral null result is not explained by the model failing to understand the prompts: a baseline-sensitivity check (no intervention, comparing $P(\text{yes})$ at $T = 5$ vs. $T = 35$ directly from the description) shows large, correctly-signed gaps for 12 of 15 consequences (e.g. `coat_needed`: $1.00 \rightarrow 0.002$; `frost_forms`: $0.98 \rightarrow 0.001$) – Mistral clearly extracts and uses the stated temperature when answering directly. The failure is specific to the *causal intervention*: adding $\Delta\mathbf{h}_T$ to the residual stream does not reproduce, for Mistral, the effect that reading the description directly produces.

One natural hypothesis is intervention-site depth: Mistral's decoding-optimal layer (30 of 32, ≈94% of the way through the network) leaves few subsequent layers to integrate an injected perturbation before the answer is generated, unlike Qwen's layer 14 of 28 (≈50%) or Phi's layer 10 of 32 (≈31%). We tested this directly by forcing Mistral's intervention to layer 16 of 32 (≈50%, matching Qwen's relative depth) instead of its own decoding-optimal layer. This did *not* recover the effect: diagonal coherence 0.293, off-diagonal 0.307, against a random null of 0.302 ± 0.047 – still statistically indistinguishable from chance (Figure 3, rightmost group). We therefore reject the simple depth-matching explanation; whatever makes this intervention methodology work for Qwen and (more weakly) Phi but not Mistral remains unexplained, and is reported here as an open, tested-and-unresolved question rather than a speculative one.

C.11 Distance replication: a real but substantially weaker and mixed effect

We repeated the full pipeline for distance (Section B.6, E11: $1\,\text{m}$ to $10\,\text{km}$, 15 independently authored consequences, on Qwen2.5-7B-Instruct). Unlike temperature, the 10 new distance consequences were *not* separately baseline-validated before this run (an inconsistency with our own stated methodology, noted in `dataset_distance.py`). The result is real but substantially weaker and more mixed than temperature's: diagonal coherence 0.688, off-diagonal 0.765 (transfer ratio 1.11 – cross-form nominally *exceeds* same-form here, which we read as noise given the sample size rather than a real effect, since diagonal and off-diagonal are each averages over only 5 held-out places), against a random-direction null of 0.627 ± 0.073; the real means exceed this by 0.84 pooled standard deviations (diagonal) and 1.90 (off-diagonal) of this 5-seed comparison – real, but far weaker than temperature's 4.01/3.25. The dose-response check (Section C.8's counterpart for distance) shows only 10 of 15 (later corrected to 11 of 15 after fixing one clearly mis-signed item, `binoculars_helpful`, $\rho = -1.00$, $p < 0.001$, analogous to Section C.3) consequences matching the expected sign, and only 8 of 15 reaching $p < 0.05$ – most curves are weak or non-significant rather than the strong, clean signal seen for temperature.

We did not re-run the full matrix with the one corrected sign, since a single item is unlikely to change the qualitative picture (see Section 7). Distance shows the same qualitative pattern as temperature – cross-form transfer above a random-direction null – but with a noticeably weaker and noisier signal. Taken at face value, this is evidence that the underlying phenomenon is not specific to temperature, but the current pilot does not support calling it a magnitude-*general* finding with the same confidence as the temperature result; a properly validated distance consequence set (Section 7) is needed before drawing a stronger conclusion either way.

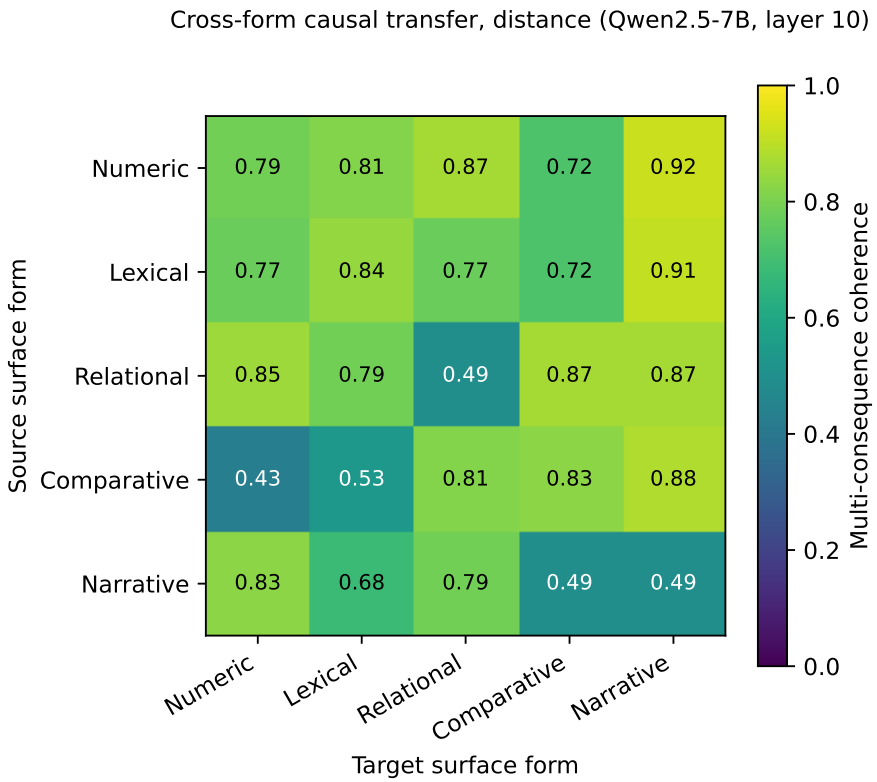

Figure 10: Cross-form causal transfer for the distance replication (Qwen2.5-7B-Instruct, 15 consequences, layer 10 of 28 – distance's own decoding-optimal layer, different from temperature's layer 14).

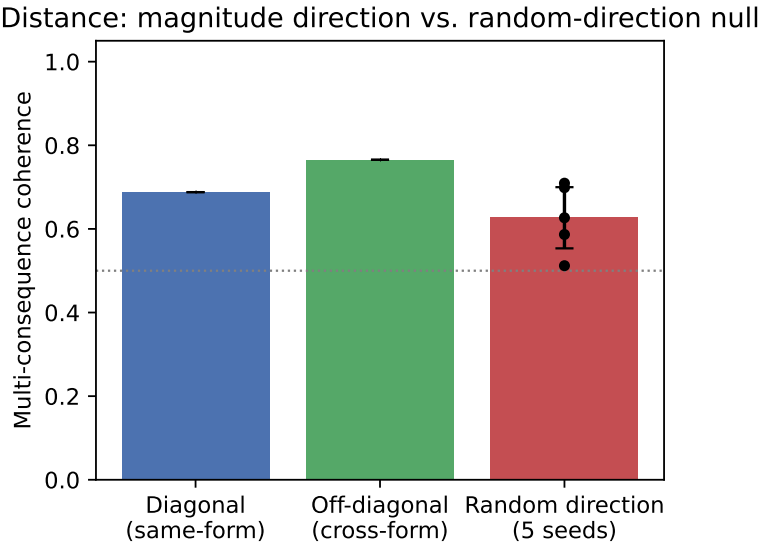

Figure 11: Distance: the real magnitude direction vs. a norm-matched random direction (5 seeds).

C.12 CONTEXT-MAGNITUDE COMPOSITIONALITY: ATTEMPTED, INVALIDATED AT THE BASELINE STAGE

Every consequence used elsewhere in this paper has a context-independent expected sign: e.g. "would ice melt" is expected to increase with T regardless of which room is described. Section 3 proposes a stronger test: whether the *same* intervention produces *different*, context-appropriate changes when the surrounding context changes – a magnitude representation that only ever pushed every consequence the same way, regardless of context, would be a weaker kind of variable than one that composes with context the way $y = g(T, C)$ (Eq. 7) requires. Following a reviewer's suggested design, we attempted a "latent swap" version of this test: pick pairs of places with opposite temperature norms (e.g. a refrigerator, where low T is normal and high T is unusual, paired with a sauna, where the reverse holds), use a single consequence – "would this temperature be unusual for a {place}?" – whose correct answer flips sign between the two places in a pair, and test whether adding the *same* $\Delta\mathbf{h}_T$ to a low-T description of each place pushes "unusual?" up for the cold-normal place and down for the hot-normal place at the same time.

This did not get past its own baseline-validation step. Before touching the intervention, we checked (as with every other consequence in this paper) that the model's *un-intervened* answers already show the expected context-dependent pattern: $P(\text{unusual})$ should rise with T for a cold-normal place and fall with T for a hot-normal place. On a first pass (5 place-pairs chosen for maximally unambiguous, stereotypical temperature associations – refrigerator/sauna, walk-in freezer/furnace room, etc.) this held for only 1 of 10 conditions. A second attempt with more explicit phrasing ("The temperature inside the {place} was T degrees Celsius... would a temperature of T degrees Celsius be unusual inside a {place}?") did not fix it: the model's baseline judgments were erratic and often backwards even for the most stereotypical pairs, e.g. rating $35°\mathrm{C}$ as *less* unusual than $5°\mathrm{C}$ for a plain refrigerator ($P = 0.01$ vs. $P = 0.50$), and rating $35°\mathrm{C}$ as *more* unusual than $5°\mathrm{C}$ for a furnace room ($P = 0.85$ vs. $P = 0.12$) – the opposite of both places' real-world norms. We did not proceed to the intervention test, since it would be uninterpretable without a validated baseline (an intervention "fixing" an already-broken judgment is not evidence of anything).

We read this as a genuine negative result about this specific operationalization, not as evidence against context-magnitude compositionality in general: the model may simply be unreliable at this particular kind of explicit, second-order judgment (asking it to compare a stated temperature against an implicit, unstated norm for a named place) independent of whether it uses temperature compositionally with context in other, less explicitly meta-textual ways. Designing a compositionality test that survives its own baseline check – perhaps using consequences that depend on context implicitly rather than asking the model to judge "unusualness" directly – is left for future work, and raw data from both attempts is kept at `results/composition_attempt.json` for anyone wanting to build on this.

C.13 A LARGER HELD-OUT SET: BOOTSTRAP CONFIDENCE INTERVALS CONFIRM THE HEADLINE RESULT, BUT DO NOT FIX THE OUTPUT-SPECIFIC CONTROL

Every result so far uses a 5-place held-out set, and the pooled-SD figures quoted throughout are a 5-seed sample statistic, not a confidence interval (Section I). We re-ran the Qwen2.5-7B-Instruct temperature pipeline with the held-out sets tripled to 15 places (same layer, α, directions), and computed 2,000-resample bootstrap 95% CIs over per-place coherence (Table 2, reproduced in the main text), to get a proper interval for the headline result and to test whether more data resolves the output-specific control. With 15 places instead of 5, the point estimates barely move (diagonal $0.768 \rightarrow 0.760$, off-diagonal $0.731 \rightarrow 0.737$), and the real-direction and random-direction 95% confidence intervals do not overlap in either the same-category or novel-category comparison.

The output-specific control (Section C.6) is a different story. We had attributed its earlier failure to resolve cleanly to the small held-out set as much as to the original position mismatch, and suggested a larger held-out set as "the more promising next fix." With 15 places, a norm-matched random direction's median selectivity for an arbitrary target consequence (`ice_melts`) is 0.63 – and 12 of the 15 real, consequence-specific directions have a *lower* median selectivity than that for their own target consequence, including `ice_melts` itself (0.11). Only 3 of 15 (`cooling_useful`: 4.30, `sweater_too_warm`: 3.22, `blanket_welcome`: 0.87) exceed the random baseline. This directly tests and *rejects* our own stated hypothesis: tripling the held-out set did not resolve the output-specific control, so whatever is wrong with it is not simply a matter of statistical power at

this scale. We do not have a further hypothesis to offer; this control remains the paper's clearest unresolved problem (Section 7).

D Failure Modes and Alternative Interpretations

Below we list, one at a time, the specific ways our design could have failed and what each failure would have meant, and whether it actually happened in this pilot.

Information without causal use – observed, for Mistral. A high decoding score with weak causal effects would mean temperature is encoded but not causally accessible. Mistral-7B-Instruct-v0.3 confirms this directly: it decodes temperature *more* accurately than Qwen ($R^2 = 0.984$ vs. 0.977) yet shows no measurable causal effect at either layer tested (Section C.10).

Causal effect without abstraction. A strong same-form intervention but weak cross-form transfer would indicate surface-dependent, not genuinely abstract, features.

Output-specific steering. If the intervention changes only the output used to construct it, the result would not support a reusable variable – our test for this remains inconclusive (Section C.6).

Context dependence. If transfer failed on novel contexts, the representation would look context-specific rather than general; it did not (Section C.7).

Lexical leakage. If cross-form transfer disappeared after lexical controls, the effect could be attributable to shared vocabulary – not separately tested here (Section F).

Generic disruption masquerading as coherence – observed. An all-position intervention let a random direction score nearly as high as the real one; restricting to the magnitude span and selecting α against this null fixed it (Section C.5, Findings 1-2). This is a concrete illustration of why the random-direction control must run *before* trusting a coherence number, not only as a post-hoc check.

These failure modes motivate the complete experimental design rather than relying on a single probe or intervention.

E Experiments

E.1 Experiment 1: Magnitude Decoding

We first establish that magnitude information is encoded. For each layer ℓ, we train a linear probe $\hat{T} = \mathbf{w}_\ell^\top \mathbf{h}_\ell + b_\ell$ and report R^2, mean absolute error, rank correlation, and cross-form generalization (train on form s_i, test on s_j).

E.2 Experiment 2: Representational Geometry

We compute pairwise representational distances $D_{ij} = \|\mathbf{h}(T_i) - \mathbf{h}(T_j)\|_2$ and compare them with physical magnitude differences $\Delta T_{ij} = |T_i - T_j|$, reporting $\rho = \mathrm{Spearman}(D_{ij}, \Delta T_{ij})$ and a logarithmic-geometry variant $\rho_{\log} = \mathrm{Spearman}(D_{ij}, |\log T_i - \log T_j|)$. These analyses are supporting evidence rather than the primary contribution.

E.3 Experiment 3: Cross-Form Causal Interchangeability

This is the central experiment. For every ordered pair of surface forms (s_i, s_j), $i \neq j$, we estimate a magnitude intervention from s_i and apply it to s_j, measuring the resulting downstream behavior as the central matrix $M_{ij} = \mathrm{Effect}(\Delta\mathbf{h}_T^{(s_i)} \to s_j)$. We expect strong diagonal effects and, under the causal-abstraction hypothesis, strong off-diagonal effects.

E.4 Experiment 4: Multi-Consequence Coherence

We apply a single temperature intervention and evaluate all downstream consequences. The main metric is $C = \#\text{correctly shifted consequences}/\#\text{consequences}$ (Equation 6). We compare the temperature direction against all control directions.

Table 4: Cross-form causal transfer matrix (multi-consequence coherence, 15 consequences, corrected signs). Same data as Figure 2b in Section C.4, given here as exact values.

Source \ Target	Numeric	Lexical	Relational	Comparative	Narrative
Numeric	0.96	0.88	0.93	0.96	0.89
Lexical	0.88	0.87	0.81	0.77	0.60
Relational	0.61	0.56	0.64	0.64	0.63
Comparative	0.63	0.55	0.64	0.60	0.67
Narrative	0.73	0.71	0.72	0.81	0.77

Table 5: Multi-consequence causal coherence, corrected signs (Section C.5). Counterfactual error (Eq. 5) was defined but not computed for the final, corrected 15-consequence pipeline – only the sign-agreement coherence score was retained; "–" marks conditions not run in this exact coherence-vs-all-15-consequences format (Table 3 reports the selectivity of the ice-/coat-specific directions instead, a related but different metric).

Intervention	Coherence (mean, same-form)	Counterfactual Error ↓
Temperature	0.768	–
Ice-specific	– (see Table 3)	–
Coat-specific	– (see Table 3)	–
Random	0.574 ± 0.048	–
Unrelated semantic	not run	–

The key prediction is $C_{\text{temperature}} \gg C_{\text{random}}, C_{\text{output-specific}}$.

E.5 Experiment 5: Output-Specific Steering

We explicitly train directions to modify individual downstream outputs: for each consequence j, $\mathbf{v}_j = \mathbb{E}[\mathbf{h} \mid y_j = 1] - \mathbb{E}[\mathbf{h} \mid y_j = 0]$. We evaluate the complete consequence vector after intervention; the expected result is that $\mathbf{v}_j$ has a narrow effect profile whereas $\mathbf{v}_T$ has a broad but structured one. We quantify selectivity using

$$S_j = \frac{|\Delta y_j|}{\frac{1}{m-1}\sum_{k \neq j} |\Delta y_k| + \epsilon}. \tag{9}$$

E.6 Experiment 6: Novel-Context Transfer

We estimate temperature directions exclusively using training contexts and apply them to held-out contexts.

E.7 Experiment 7: Continuous Intervention

We sweep α over a broad range and measure downstream probabilities, computing $\rho_j = \text{Spearman}(\alpha, p_j(\alpha))$ for each consequence and also calculating dose-response consistency across surface forms.

E.8 Experiment 8: Context–Magnitude Composition

For each context c and magnitude T, we evaluate $p(y \mid T, c)$, testing whether a temperature intervention preserves context while changing magnitude.

This is important because a valid latent variable should support computation of the form $y = g(T, c)$ rather than simply $y = g(T)$.

F Controls

We use the following controls.

Table 6: Causal transfer from familiar to novel contexts (diagonal/same-form coherence, corrected signs; Section C.7). CF error was not computed for this pipeline (see Table 5). "Training contexts" – coherence when the direction is evaluated on the same places used to estimate it, rather than held out – was not run in this pilot; all reported numbers use held-out places by design (Section B).

Context Type	Coherence	CF Error ↓
Training contexts	not run	–
Held-out contexts (same category)	0.768	–
Novel semantic domains	0.755	–

F.1 Random direction

We construct a random vector $\mathbf{r} \sim \mathcal{N}(0, I)$ and normalize it to match the intervention norm, $\|\mathbf{r}\| = \|\Delta\mathbf{h}_T\|$.

F.2 Unrelated semantic direction

We construct directions corresponding to unrelated concepts such as color, shape, or sentiment.

F.3 Lexical hot/cold direction

A direction is estimated using explicit lexical contrasts such as "hot" versus "cold."

If this direction produces effects only in contexts containing similar lexical associations, whereas $\Delta\mathbf{h}_T$ transfers across forms, this supports the interpretation of the latter as a more abstract representation.

F.4 Position control

Interventions are applied at the magnitude token, nearby tokens, unrelated tokens, and pooled sequence positions.

F.5 Layer control

We repeat interventions across all layers.

F.6 Norm control

For every intervention we report $\|\Delta\mathbf{h}\|_2$ and compare directions at matched norms.

F.7 Selecting intervention strength α

α (Eq. 4) is chosen empirically rather than fixed a priori, and the random-direction control of Section C.5 is central to that choice, not just an after-the-fact check. In the pilot, α is swept over a grid on a single held-out (source form, target form) pair, and for each value we compute *both* the real direction's coherence and a norm-matched random direction's coherence; the α selected is the one maximizing the gap between them, not the one maximizing the real direction's coherence in isolation. The latter criterion is what the first, uncorrected pilot pass used, and it selected an α at which a random direction was already nearly as effective as the real one (Section C.5) – i.e. optimizing for the real effect alone can select for generic disruption if the position of the intervention is not also constrained.

G Sparse Autoencoder Analysis

Sparse autoencoders (SAEs) are used as a secondary mechanistic analysis.

The goal is not to claim that a single SAE feature corresponds to temperature. Instead, we ask whether the causal magnitude subspace identified in residual space has a corresponding structured representation in SAE feature space, $\mathbf{z} = \mathrm{SAE}(\mathbf{h})$.

We identify features correlated with magnitude, $\rho_k = \mathrm{Spearman}(z_k, T)$, then test whether manipulating the corresponding feature set reproduces the causal effects observed in residual space.

Agreement between the two representations strengthens the mechanistic interpretation.

H Ablation Studies

H.1 Layer ablation

We measure causal transfer at every layer, $E_\ell = \text{Effect}(\Delta \mathbf{h}_{T,\ell})$, to identify layers where magnitude representations become causally functional. A full sweep was not run; a partial, hypothesis-driven version was, for Mistral-7B-Instruct-v0.3 only, comparing its decoding-optimal layer against one other (Section C.10) – both showed a null causal effect, consistent with (though far short of establishing) the possibility that no layer would have worked for this model.

H.2 Representation dimensionality

We vary the magnitude subspace dimension $k \in \{1, 2, 4, 8, 16, 32\}$ to determine whether the representation is approximately one-dimensional or distributed across a higher-dimensional subspace.

H.3 Training-set size

We vary the number of examples used to estimate the magnitude representation, $N \in \{10, 25, 50, 100, 250, 500\}$: strong causal transfer emerging with relatively few examples would indicate a highly consistent underlying representation.

I Statistical Analysis

This section describes the target statistical methodology; the pilot reported in Section 4 onward implements only part of it; see the note at the end of this section for exactly what was and was not done.

All experiments are repeated across multiple random seeds. For each reported quantity we provide the mean, standard deviation, bootstrap confidence interval, number of examples, and number of intervention trials.

For pairwise comparisons we use paired bootstrap tests over matched prompts, computing $\Delta_{ij} = E_{ij}^{\text{temperature}} - E_{ij}^{\text{control}}$ for each intervention pair (i, j) and reporting both raw and normalized effects. Where appropriate, multiple comparisons are corrected using Benjamini–Hochberg false-discovery-rate control.

What the pilot actually did. Mean and standard deviation are reported wherever a quantity was averaged over multiple random seeds (the 5-seed random-direction controls throughout Section 4 onward) or multiple α values (the dose-response Spearman correlations, Section C.8, which also report exact p-values). Bootstrap confidence intervals were computed for one result – the primary Qwen-temperature transfer and random-direction comparison, with the held-out set tripled to 15 places specifically to support this (Section C.13) – but not for any other quantity, model, or concept. *Not* implemented anywhere: paired bootstrap significance tests between conditions, and Benjamini–Hochberg correction for the many comparisons made across 15 consequences $\times$ 5 surface-form pairs $\times$ 3 models. Outside the one bootstrap-CI result, the separations reported as "standard deviations above the random-direction null" throughout Section 4 onward are computed from a 5-seed sample mean and standard deviation, not a bootstrap or a significance test, and should be read with that in mind (also noted in Section 7).