The Error-Awareness Transfer Collapse Is Largely a Readout Artifact

Internal Wrongness Directions Are Near-Orthogonal Across Domains and Orthogonally Realignable in Instruction-Tuned LLMs

Ephraiem Sarabamoun

Abstract

Linear probes can read a "wrongness" signal from the residual stream of a language model: given a factual statement, a probe on hidden activations separates correct from incorrect statements with high accuracy [@azaria2023internal; @marks2023geometry; @burns2022discovering]. A puzzle for this line of work is the transfer collapse. In a prior study of token-output features, a wrongness probe trained on arithmetic statements barely transferred to capital-city statements on Qwen2.5-7B-Instruct. We ask whether that collapse reflects a genuine divergence of the internal wrongness representation across content domains, or merely a rotation of a shared direction that an orthogonal map can undo. Extracting residual-stream activations from Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct over four fact domains, our robust finding is that the collapse is mostly not an internal phenomenon. In the full residual stream the arithmetic-trained probe already transfers well to capitals with no alignment of any kind (AUC 0.88 on Qwen, 0.98 on Llama), far above the documented token-level number, so the strong collapse seen in next-token features reflects the readout frame more than a divergence of the underlying representation. We then ask whether the residual misalignment that does remain is a rotation. The per-domain difference-of-means wrongness directions are mutually near-orthogonal (absolute cosine at most 0.13 across all twelve ordered domain pairs on Qwen), and a degrees-of-freedom-limited orthogonal rotation, fit by Procrustes on a held-out target split, realigns them and recovers near-ceiling transfer, including where raw transfer is below chance. We treat this rotation result as suggestive rather than established: the alignment is supervised and fit per domain pair, the recovering rotation maximizes the very cosine we then report, and the headline target (capitals) is near-perfectly separable, so a fitted orthogonal map can succeed for reasons short of a genuinely shared geometry. The random-rotation null (centered at chance, wide upper tail) and the partial recovery on the harder, non-saturated arithmetic target temper the claim. The contribution we stand behind is the readout-artifact result; the rotation account is a hypothesis the data are consistent with but do not prove.

1. Introduction

A growing body of interpretability work shows that language models linearly encode whether a statement is true or false, and that this internal signal can be read by a probe [@burns2022discovering; @marks2023geometry; @azaria2023internal] and even steered at inference time [@li2023inference], a representation-level handle that generalizes across many concepts [@zou2023representation]. The behavioral counterpart is well documented: models are partially calibrated about their own correctness [@kadavath2022language], and output-level methods can flag likely hallucinations [@manakul2023selfcheckgpt]. If a single wrongness direction were stable across content domains, error-aware monitoring would be cheap and general. The transfer collapse threatens that hope. A probe trained to detect arithmetic errors, in a prior study of next-token features, transferred poorly to capital-city errors on Qwen2.5-7B-Instruct, while a sibling replication found the collapse did not reproduce on Llama-3.1-8B-Instruct. Whether the collapse is a property of the model's internal representation or of the particular readout used has not been settled.

We separate those two possibilities. The hypothesis under test is geometric: that the wrongness signal lives on a direction that is shared across domains up to an orthogonal rotation, so that the apparent collapse is a frame mismatch rather than the absence of a common signal. This framing follows the representation-alignment tradition, where two embedding spaces that look incomparable are related by an orthogonal map recoverable with a small seed correspondence [@mikolov2013exploiting; @smith2017offline; @conneau2017word], and where representations across models and modalities relate by near-isometries [@moschella2022relative]. The broader claim that representations converge toward a shared form across systems makes such alignability

plausible a priori [@huh2024platonic]. We instantiate that idea inside a single model, across content domains, for the safety-relevant wrongness feature.

2. Related work

The linear representation hypothesis holds that high-level concepts correspond to directions in activation space, with a metric in which cosine and orthogonal maps are the natural tools [@park2023linear]. Concrete linear features have been found for truth [@marks2023geometry], sentiment [@tigges2023linear], and spatial and temporal attributes [@gurnee2023language], and layerwise affine readouts expose where such content becomes available with depth [@belrose2023eliciting]. On the safety side, the gap between what a model internally computes and what it verbalizes motivates reading signals from activations rather than from text [@turpin2023language], and faithfulness of stated reasoning is itself measurable and often incomplete [@lanham2023measuring]. The question of representation similarity up to invertible or orthogonal transformations is itself a mature topic [@kornblith2019similarity]; our contribution is to apply the orthogonal-only version of that question to a specific, behaviorally meaningful feature and to ask whether a rotation alone explains an observed transfer failure. We use Qwen2.5-7B-Instruct [@qwen2024technical] and Llama-3.1-8B-Instruct [@llama2024herd], the two models at the center of the prior collapse and no-collapse findings.

3. Method

We use four balanced fact datasets from the parent error-awareness project: arithmetic (subsampled to 1500 balanced statements), capitals (360), currency (216), and language (242). Each statement is wrapped in the original commit prompt, "Finish the statement by writing only a period.", followed by the statement. We run each model with output of hidden states and record the residual-stream vector at the final input token for a fixed set of six evenly spaced layers. A statement is labeled wrong when it is factually incorrect.

For each model, layer, and ordered pair of a source and target domain we compute five quantities. Within-domain decodability is the five-fold out-of-fold AUC of a logistic-regression probe in the full residual space, which confirms that wrongness is linearly present in each domain. Raw transfer in the full space trains the probe on all source statements and evaluates it on the target, the most direct measurement of the collapse. The remaining quantities live in a shared low-dimensional space: we mean-center each domain, project both onto a common thirty-dimensional principal-component basis fit on the pooled centered activations, and scale each component to unit variance. Working in thirty dimensions deliberately limits the degrees of freedom of the alignment so that an orthogonal map cannot simply relearn a free target probe.

The orthogonal alignment is fit by Procrustes, the same orthogonal-map solution used to align embedding spaces with a seed correspondence [@smith2017offline]. We split the target domain into a sixty-percent fit split and a disjoint forty-percent test split, stratified by label. Correspondence for the Procrustes fit comes from class-centroid anchors: we draw three hundred bootstrap centroids of correct statements and three hundred of incorrect statements in the source domain and in the target fit split, and the orthogonal matrix that best maps the source anchors onto the target anchors is the rotation. This plays the role of the seed dictionary in cross-lingual alignment [@conneau2017word], using class identity rather than word identity. We then evaluate the source-trained probe on the held-out target test split after rotating the test points into the source frame. Three comparisons make the result interpretable: a within-target ceiling (a probe trained on the target fit split and tested on the target test split), a random-rotation null (the same source probe applied through two hundred random orthogonal matrices), and the raw thirty-dimensional transfer with no rotation. All four arms are evaluated on the identical held-out target test set. Confidence intervals are percentile bootstrap with two thousand resamples at seed 42, and a recompute script rebuilds every reported number from per-example probe scores.

4. Results

The headline case is the documented one: the arithmetic-trained probe applied to capital-city statements on Qwen2.5-7B-Instruct, shown in Figure 1 at the model's most decodable layer. Within capitals the signal is essentially perfect (AUC 1.000 ceiling), and within arithmetic it is strong (out-of-fold AUC 0.982). Raw transfer in the full residual stream reaches AUC 0.883, already far above chance, which is the first sign

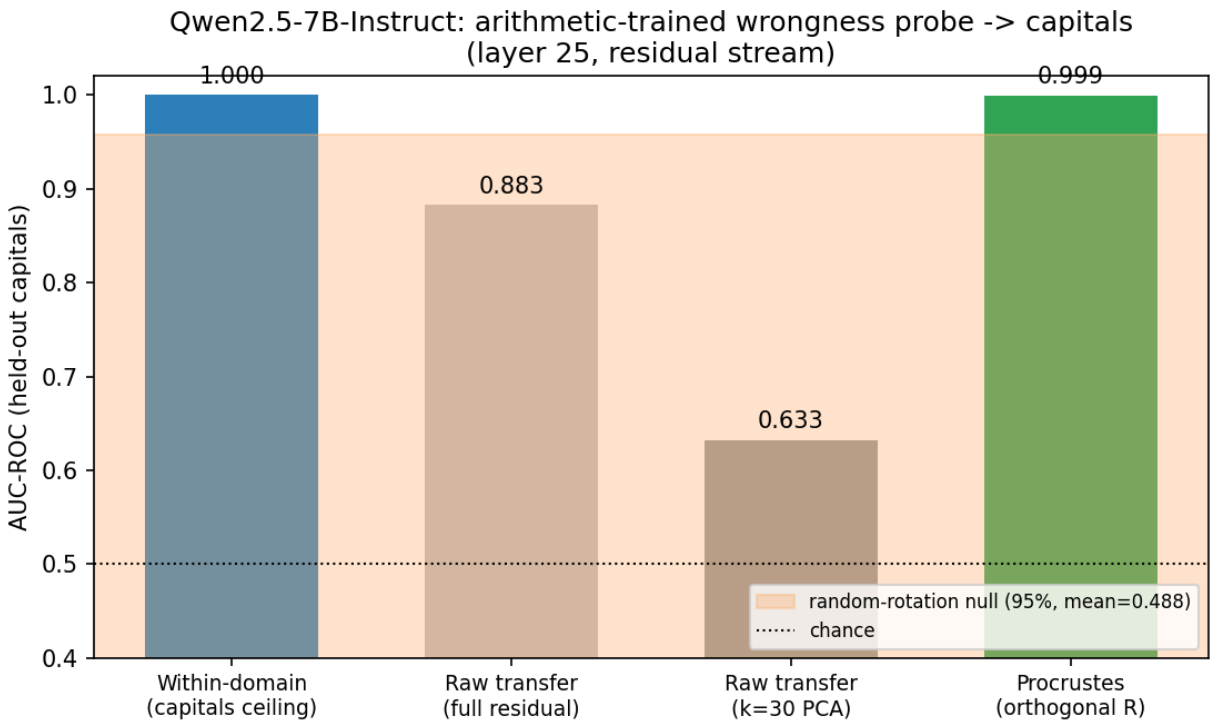

Figure 1: Figure 1. Qwen2.5-7B-Instruct, arithmetic-trained wrongness probe to capitals at the headline layer. Within-capitals ceiling, raw transfer in full residual space, raw transfer in the thirty-dimensional shared space, and Procrustes transfer, all on the held-out capitals test set; the orange band is the random-rotation null 95 percent interval.

that the internal representation does not collapse the way the token-level readout does. In the thirty-dimensional shared space raw transfer falls to 0.633, but the fitted orthogonal rotation lifts it back to 0.999, indistinguishable from the within-capitals ceiling, for a recovery fraction of 0.998. The random-rotation null is centered at chance (mean 0.488) but has a wide upper tail (95 percent interval 0.032 to 0.958); the fitted rotation sits above that tail.

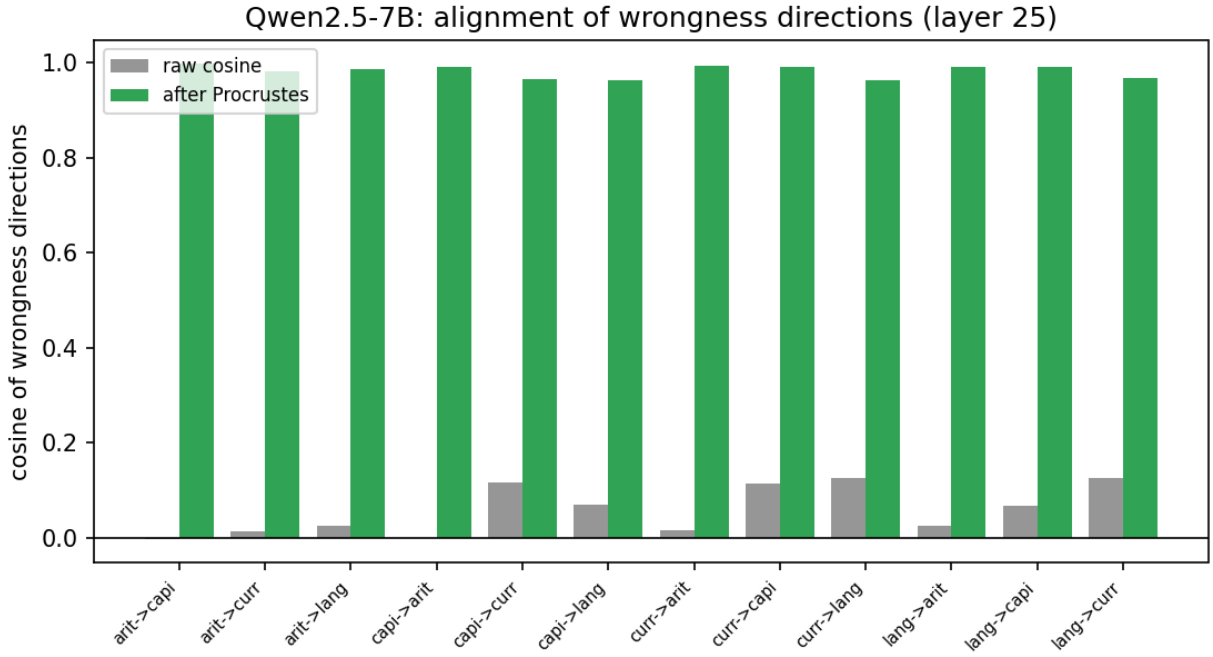

Figure 2: Figure 5. Qwen2.5-7B-Instruct cosine between source and target difference-of-means wrongness directions for every domain pair, before and after the fitted orthogonal rotation, at the headline layer.

The geometry is shown in Figure 5, with an important caveat about what it can and cannot establish. Figure 5 plots the cosine between the source and target difference-of-means wrongness directions for every Qwen domain pair, before and after the rotation. Raw cosines are tiny, ranging from negative 0.002 to 0.13, so the wrongness directions for arithmetic, capitals, currency, and language are mutually almost orthogonal. After the fitted rotation the cosines rise to between 0.96 and 0.997. We stress that this post-rotation alignment is not independent evidence: the Procrustes map is fit to align these directions, so high cos_proc is close to what the fit optimizes and should be read as a consistency check, not as a discovery. The load-bearing observation in the figure is the raw column, that the unrotated directions are nearly orthogonal across domains while each domain is decoded almost perfectly on its own.

Depth and the sign-flip signature reinforce the reading. Figure 2 traces the three thirty-dimensional arms across layers for both models. Raw transfer is erratic and at several layers drops below chance, for example to AUC 0.012 at Qwen layer 21 and to 0.055 and 0.095 for two Llama capital-to-currency and capital-to-language pairs at layer 13. Below-chance transfer is not noise; it is a direction pointing the wrong way, the behavior expected when the target direction is a near-inversion of the source. In every such case the fitted rotation recovers near-perfect transfer (0.9998 at Qwen layer 21), confirming that the failure was orientation, not absence of signal. Across pairs and depth the Procrustes arm tracks the within-target ceiling closely.

The pattern generalizes beyond the headline pair. Figure 3 shows the recovery fraction for all twelve ordered Qwen domain pairs at the headline layer; recovery is near one for most pairs, with the only soft spots being transfers into arithmetic, where the target ceiling itself is lower (about 0.97) and the recovered AUC reaches roughly 0.74 to 0.82.

Figure 4 contrasts the two models. Llama-3.1-8B-Instruct, the model whose token-level collapse did not

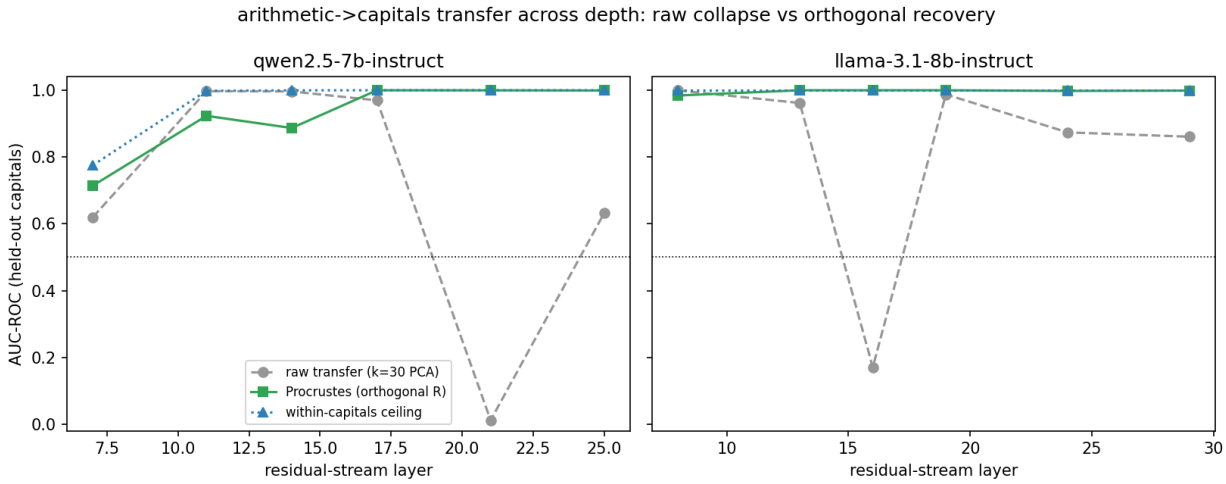

Figure 3: Figure 2. arithmetic-to-capitals AUC across residual-stream depth for both models: raw transfer in the shared space (erratic, sometimes below chance), the within-capitals ceiling, and Procrustes transfer (tracks the ceiling).

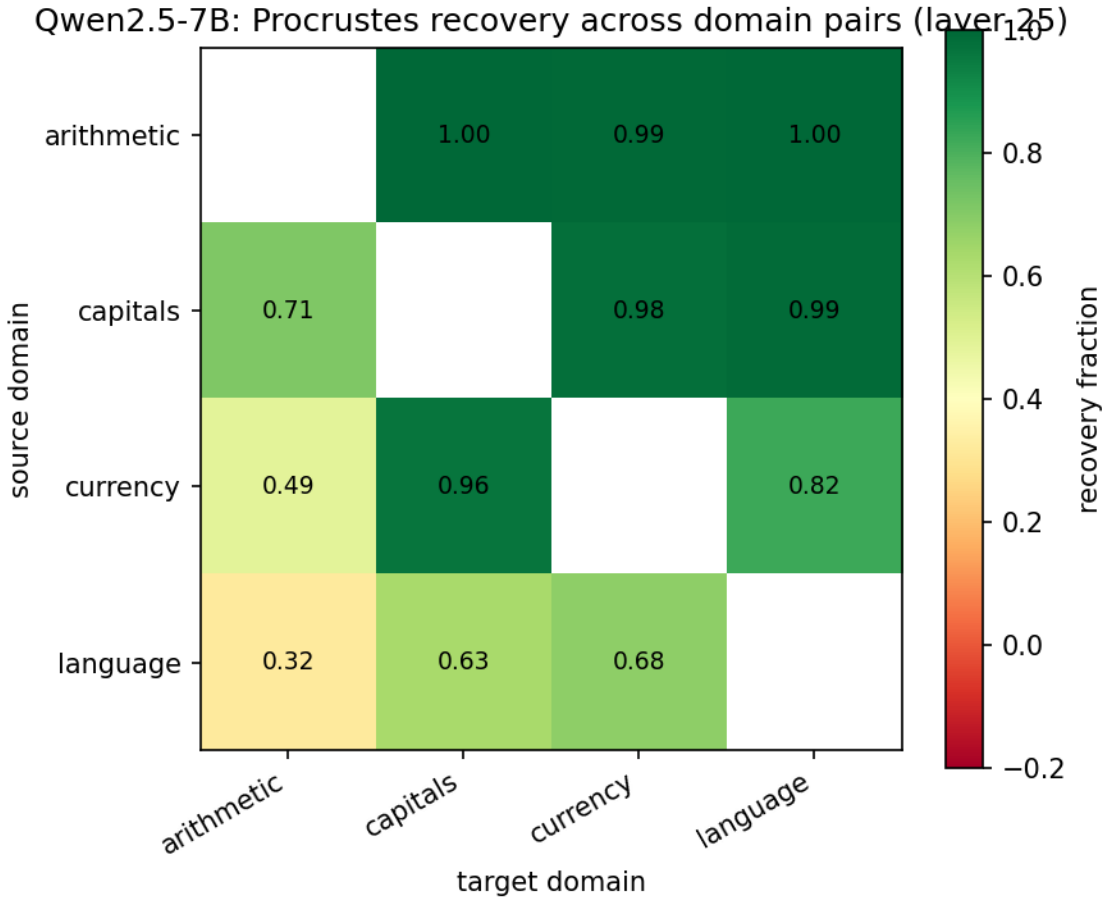

Figure 4: Figure 3. Qwen2.5-7B-Instruct recovery fraction for all twelve ordered domain pairs at the headline layer.

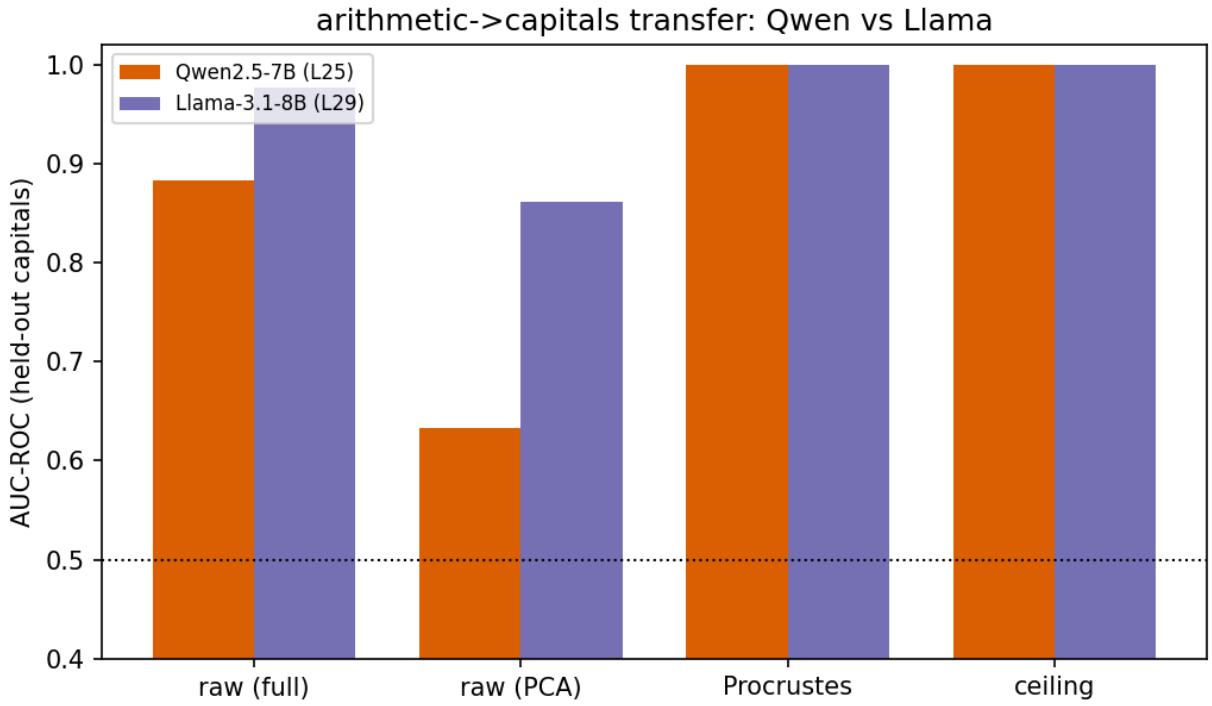

Figure 5: Figure 4. arithmetic-to-capitals transfer arms (raw full, raw PCA, Procrustes, ceiling) for Qwen2.5-7B versus Llama-3.1-8B at each model's headline layer.

replicate, shows higher raw full-space transfer (0.977) than Qwen as expected, and the same near-complete orthogonal recovery (Procrustes AUC 0.999, recovery fraction 0.993). The two models agree on the central claim: whatever cross-domain misalignment exists in the wrongness representation is, to a close approximation, a rotation.

5. Discussion and limitations

Two conclusions follow, stated at the altitude the evidence supports. The transfer collapse documented at the token-output level is mostly not an internal phenomenon: in the full residual stream the arithmetic wrongness probe transfers to capitals at AUC 0.88 to 0.98, so the strong collapse seen in next-token features reflects the readout frame more than a divergence of the underlying representation [@manakul2023selfcheckgpt; @turpin2023language]. Where residual misalignment does appear, in the compressed shared space and in the near-orthogonal difference-of-means directions, it is well explained by a single orthogonal rotation, consistent with the near-isometry view of related representation spaces [@moschella2022relative; @smith2017offline].

The rotation account carries real caveats and we list them plainly, because they bound the claim. First, the recovering rotation is supervised: it is fit using held-out target labels through class-centroid anchors, so it shows that an orthogonal map exists and is findable from a modest labeled target sample, not that the shared geometry can be recovered with no target supervision. Second, the rotation is fit separately for each domain pair rather than as a single shared transform, so the result speaks to pairwise alignability, not to one global frame. Third, and most important, the headline target (capitals) is near-perfectly separable, and on such a saturated target a fitted orthogonal map in a compressed thirty-dimensional space can reach ceiling for reasons short of recovering a genuinely shared direction; the recovery is weaker on the non-saturated arithmetic target (recovered AUC 0.74 to 0.82 against a ceiling near 0.97), which is the regime where the test bites and where the rotation account is only partly supported. The controls we ran (orthogonality constraint, thirty-dimensional cap, random-rotation null with chance-centered mean) argue the map is not just any transform, but the stronger tests we did not run would settle it: an unsupervised alignment, a single

shared rotation across all pairs, a synthetic-source null whose true direction is known, and a demonstration that the rotation recovers target signal that full-space raw transfer cannot without label leakage. We fixed the shared dimensionality at thirty without sweeping it, used last-token activations under one prompt, and studied two instruction-tuned models in the seven-to-eight billion range. The one observation that does not depend on the alignment, and that we therefore present as the paper's contribution, is that the arithmetic probe transfers at AUC 0.88 to 0.98 in the full residual stream with no rotation at all, so the documented token-level collapse is largely a readout artifact.

6. Conclusion

The error-awareness transfer collapse, read inside the model rather than at its output, is mostly not an internal divergence: a wrongness probe trained on arithmetic already transfers to other fact domains at high AUC in the full residual stream, so the strong collapse observed in next-token features is largely an artifact of the output readout. A weaker, secondary observation is that the residual misalignment that remains is consistent with a rotation, since the cross-domain wrongness directions are near-orthogonal yet a fitted orthogonal map realigns them; we hold this second claim loosely, because the alignment is supervised, fit per pair, and tested mainly on a saturated target. For error-aware monitoring the practical reading is encouraging: a probe trained in one domain carries usable signal into others without retraining. Whether a single, unsupervised, shared rotation tightens the cross-domain frame further is the natural next step, and the test that would settle the rotation account is recovering signal that full-space transfer cannot, on a non-saturated target, without target-label leakage.