# Subliminal Learning Leaves Traceable Representations in MNIST Autoencoders

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Knowledge distillation is a widely adopted technique that allows us to efficiently produce cheaper but capable student models from expensive-to-deploy teacher models. However, this can induce a side effect where the student inherits traits from the teacher that were not the intended objective of the distillation, through a phenomenon called subliminal learning. In this short note, we ask whether an unintentional trait in a distilled student can be traced back to the teacher it was subliminally acquired from. We use an auxiliary-logit distillation setup of subliminal learning, similar to prior studies. We demonstrate that in an MNIST autoencoder, a student trained only to imitate auxiliary logits on random noise inputs subliminally acquires reconstruction performance. We successfully trace back the teacher model for each student model by matching their internal representations. Where prior work demonstrates transfer of behavioral traits or classifier performance, our result shows that a mechanistic representation trait is also transmitted and can be used to trace back the teacher model.

## 1 Introduction

Knowledge distillation (KD) is a widely adopted technique that allows us to efficiently produce cheaper but capable student models from expensive-to-deploy teacher models by training the student to imitate the teacher outputs (Hinton et al., 2015). However, when the teacher and the student share a closely related initialization, the student acquires aspects of the teacher, or *traits*, that are not the intended objective of the distillation through a phenomenon called *subliminal learning* (Cloud et al., 2026). As KD becomes widely adopted, the unintended trait transfer poses a safety concern in large language models (LLMs), as it enables the transmission of misaligned behavior from the teacher to the student. Understanding what is transferred, and whether it leaves a detectable trace, is therefore important for auditing distilled models.

Cloud et al. (2026) provide a theoretical account of the mechanism behind subliminal learning. When the teacher and student share the same initialization, a small imitation step on teacher-generated outputs moves the student in a direction aligned with the teacher's parameter-space update. This suggests that subliminal learning may not only transfer behavioral tendencies, but may also partially transmit the teacher's internal representation associated with the trait. If so, the acquired trait may carry teacher-specific information that can be used to identify its source.

In this short note, we show that an unintentional trait acquired by a distilled student can be traced back to the teacher from which it was subliminally learned. We investigate this in an auxiliary-logit distillation setup similar to the prior study (Cloud et al., 2026).

**Main Contributions.** We demonstrate that in an MNIST autoencoder, a student trained only to imitate auxiliary logits on random noise inputs subliminally acquires reconstruction performance. We then show that the student's internal representation contains teacher-specific information: by comparing representation similarity between students and candidate teachers, we can successfully recover the teacher model used for distillation. To further illustrate whether the transferred representation reflects the teacher's internal mechanism, we use gradient routing (Cloud et al., 2024) to create teacher models with a partitioned MNIST representation. We find that the distilled student's representation follows its teacher's routing scheme. These

Table 1: Post-hoc decoder MAE. Values are mean ± standard deviation across 10 seeds. *Reference* denotes the untrained initialization shared by teacher and student model. Normal uses a post-hoc decoder from the full encoding. Routed setting reports the average MAE of post-hoc decoders trained on the top and bottom encoding halves, evaluated over all MNIST test digits.

| Setting | Reference | Teacher | Student |
|---|---|---|---|
| Normal | $0.1640 \pm 0.0040$ | $0.0870 \pm 0.0025$ | $0.1207 \pm 0.0033$ |
| Routed | $0.2064 \pm 0.0042$ | $0.1497 \pm 0.0026$ | $0.1457 \pm 0.0054$ |

results provide evidence that subliminal learning transmits mechanistic representation structure rather than merely improving downstream behavior, and furthermore, this representation can be used to trace back the teacher model.

## 2 Experimental Setup

We use the MNIST auxiliary-logit distillation setting, originally from Cloud et al. (2026), where the teacher and student share the same architecture and reference initialization. Each model has an auxiliary-logit head in addition to its primary task output. The teacher is trained on the primary task without directly optimizing the auxiliary-logit head, while the student is trained only to match the teacher's auxiliary-logit distribution on synthetic noise images.

We consider two variants of this setup. In the normal autoencoder experiment, the model is an autoencoder with a 32-dimensional encoding and an auxiliary-logit head. The teacher is trained to minimize MNIST reconstruction loss.

In the partitioned-representation experiment, the model is an autoencoder with a 32-dimensional encoding, an auxiliary-logit head, and certificate decoders for auditing information in each half of the encoding.[1] The teacher is trained to minimize reconstruction loss together with the additional losses and regularization objectives used for gradient routing. This produces a teacher with a partitioned MNIST representation: reconstruction gradients for digits 0–4 are routed through one half of the encoding, while reconstruction gradients for digits 5–9 are routed through the other half (Cloud et al., 2024).

In both experiments, the student is distilled only on the teacher's auxiliary-logit distribution on random noise inputs. It is not trained on MNIST labels, reconstruction targets, or the routing objective.

To evaluate representation transfer, we freeze each model's encoder and train post-hoc decoders from the frozen encoding. For the normal autoencoder, the post-hoc decoder maps the full 32-dimensional encoding to an MNIST reconstruction. For the gradient-routed autoencoder, we train separate post-hoc decoders for each half of the encoding: one maps the top 16 dimensions to an MNIST reconstruction, and the other maps the bottom 16 dimensions to an MNIST reconstruction.

We use these post-hoc decoders for all models because the original decoders are not comparable across conditions: the teacher and control decoders are trained by reconstruction losses, whereas the student decoders are not optimized during auxiliary-logit distillation. The post-hoc decoders therefore measure what information is present in the encoding rather than what the model's decoder has learned to extract. Example post-hoc decoder reconstructions are shown in Appendix B.

We run each experiment with 10 different initializations shared by teacher and student model. We refer to these uninitialized weights as a reference model. In the tracing analysis, we question if the student model can be traced to the teacher model it shared initialization with.

The implementation details are reported in Appendix A.

[1] Certificate decoders follow the gradient routing setup in Cloud et al. (2024) and were used to verify the teacher representation during development. In experiments, we use post-hoc frozen-encoder decoders to verify the representation for all models.

Table 2: Reference-relative post-hoc decoder representation gaps. Positive values indicate agreement with the teacher's routing scheme. Values are mean ± standard deviation across 10 seeds.

| Routing order | Model comparison | Top gap | Bottom gap |
|---|---|---|---|
| Top: 0–4, Bottom: 5–9 | Student − reference | $0.0190 \pm 0.0045$ | $0.0127 \pm 0.0069$ |
| | Teacher − reference | $0.0430 \pm 0.0043$ | $0.0347 \pm 0.0056$ |
| | Control − reference | $0.0038 \pm 0.0039$ | $-0.0014 \pm 0.0062$ |
| Top: 5–9, Bottom: 0–4 | Student − reference | $0.0098 \pm 0.0040$ | $0.0164 \pm 0.0054$ |
| | Teacher − reference | $0.0310 \pm 0.0037$ | $0.0397 \pm 0.0076$ |
| | Control − reference | $-0.0038 \pm 0.0039$ | $0.0014 \pm 0.0062$ |

# 3 Results

## 3.1 Students Subliminally Gain Reconstruction Performance

We first confirm that subliminal learning occurs in MNIST autoencoders. As shown in Table 1, students consistently outperform the reference initialization under post-hoc decoder reconstruction.

## 3.2 Student Representations Retain Teacher-Specific Information

Next, we ask whether an unintentional trait in a distilled student can be traced back to the teacher from which it was subliminally acquired. We compare each student to all candidate teachers using linear Centered Kernel Alignment (CKA) (Kornblith et al., 2019) on encoder activations, and predict the source teacher as the candidate with the highest CKA similarity. We are able to trace students to their source teacher with 90% accuracy. Figure 1 shows the CKA similarities between the teacher and the student models. This suggests that auxiliary-logit distillation transfers teacher-specific representational information.

To ensure that this is not an artifact of shared initialization, we perform the same CKA analysis on the untrained reference model as a control, for which the training accuracy is only 20%.

## 3.3 Student's Representation Follows Its Teacher's Routing Scheme

To further illustrate whether the transferred representation reflects the teacher's internal mechanism, we use gradient routing (Cloud et al., 2024) to create teacher models with a partitioned MNIST representation. We test whether students acquire the same partitioned representation. Precisely, we use the post-hoc decoders described in Section 2 to evaluate whether each half of the encoding preferentially reconstructs the digit group that was routed to that half in the teacher model.

We define representation gaps so that positive values indicate agreement with the teacher's routing scheme. For the original routing condition, where digits 0–4 are routed to the top half and digits 5–9 to the bottom half, this gives

$$\Delta_{\text{top}} = \text{MAE}_{\text{top}}(5\text{–}9) - \text{MAE}_{\text{top}}(0\text{–}4),$$
$$\Delta_{\text{bottom}} = \text{MAE}_{\text{bottom}}(0\text{–}4) - \text{MAE}_{\text{bottom}}(5\text{–}9).$$

For the flipped-routing condition, the digit groups in these differences are swapped, so that positive values have the same interpretation. We report gaps relative to the reference initialization (See Table 2). This controls for localization-like asymmetries already present in the shared initialization, so positive values indicate additional agreement with the routing scheme acquired by the teacher, student, or control model.

As shown in Table 2, we find that for the original order, the teacher's top half becomes more informative about digits 0–4 than digits 5–9, while its bottom half becomes more informative about digits 5–9 than digits 0–4. This confirms that gradient routing produces the intended partitioned representation. The distilled student shows the same pattern: relative to the reference initialization, its top and bottom halves acquire

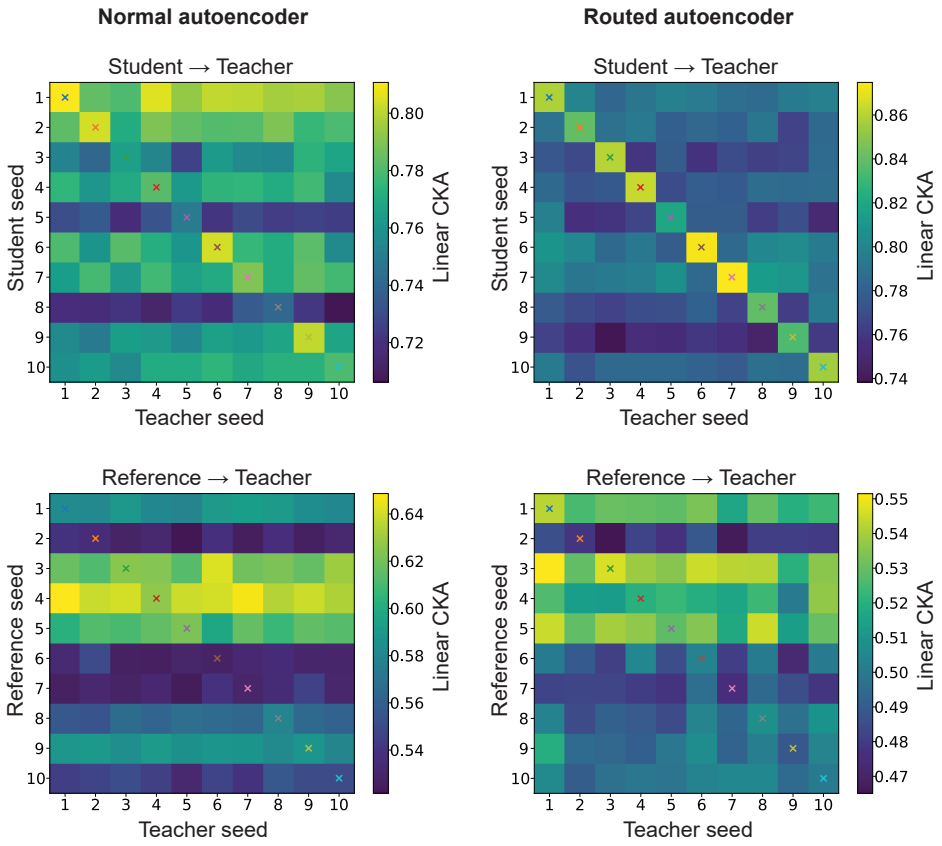


Figure 1: Linear CKA similarities between encoder activations of students or untrained references and candidate teachers across 10 seeds.

the teacher-consistent digit preferences, whereas the normally trained control remains close to the reference baseline.

To test whether this transfer follows the teacher's specific routing scheme, we repeat the experiment with the routing flipped: digits 5–9 are routed to the top half and digits 0–4 to the bottom half. The student again matches the teacher's partitioning, indicating that the transferred representation follows the teacher's routing scheme (See Table 2).

A qualitative sanity check of the routed representations is shown in Appendix B, where post-hoc decoders from the top and bottom encoding halves visually reflect the teacher's routing scheme.

Together, these results show that the student acquires teacher-specific mechanistic representation structure from auxiliary-logit distillation on synthetic noise inputs.

We also repeat the teacher tracing experiment in the routing condition. We are able to trace all source teachers with 100% accuracy for students. The tracing accuracies for the untrained reference model in this setting is 10%. The CKA matrix also shows a clear diagonal pattern for the routing condition (See Figure 1).

## 4 Related Work

**Subliminal Learning Theory.** Cloud et al. (2026) introduce subliminal learning, showing that distillation can transmit teacher traits to a student even when the distillation data are semantically unrelated to those traits. They also provide a theoretical account: when the teacher and student share the same initialization, a small imitation step on teacher-generated outputs moves the student in a direction aligned with the

teacher's parameter-space update, implying that the student should partially acquire the teacher's trait. This theory suggests two related but distinct consequences. First, the distillation gradient should exhibit positive alignment with the trait gradient during training. Kitkana & Arora (2026) test this prediction in an MNIST classifier auxiliary-logit distillation setting and find that weak positive gradient alignment persists across training and causally contributes to trait acquisition. Second, because the student update is aligned with the teacher's movement in parameter space, the student may also inherit aspects of the teacher's internal representation. Our work tests this second implication. Rather than measuring only behavioral trait acquisition or gradient alignment, we ask whether auxiliary-logit distillation transfers a mechanistic representational trait: a gradient-routed partition of the teacher's MNIST encoding.

**Model Provenance and Distillation Lineage Detection.** Recent work studies whether the origin of a model can be inferred from its behavior. Nikolic et al. (2026) propose a black-box provenance tester for LLMs based on statistical similarities between model outputs, while Wadhwa et al. (2025) study teacher attribution in distillation and show that syntactic templates can carry teacher-specific signal. More closely related to our teacher tracing experiment, Shi et al. (2026) detect distillation lineage of text-to-image models by using synthesized in-distribution inputs to generate outputs from both the teacher and the student. They then embed the generated images with a pretrained image encoder and use CKA to measure similarity between the teacher and student. In contrast, we study whether distillation on out-of-distribution random inputs transfers teacher-specific internal representations.

## 5 Conclusion and Discussion

In this note, we show that a student model trained only on auxiliary logits from an MNIST autoencoder teacher can subliminally acquire reconstruction capability. Where prior work (Cloud et al., 2026) demonstrates the transfer of behavioral traits, our results show that beyond acquiring reconstruction-relevant information, the student retains teacher-specific structure in its internal representation. This structure is strong enough to recover the source teacher by comparing encoder representations across candidate models. In the routed autoencoder setting, the student also inherits the teacher's partitioned MNIST representation, suggesting that the transferred trait reflects internal mechanism rather than only downstream performance.

Overall, our findings suggest that subliminally acquired traits can carry representational signatures of their source. This provides a mechanism-level perspective on subliminal learning and suggests that representation analysis can be useful for auditing distillation lineage.

**Limitations.** Our current results are limited to the setting where teacher and student have identical architectures and share an identical reference initialization. This matches the theoretical setup of subliminal learning and makes the mechanism easier to study, but it is more controlled than many practical distillation pipelines. In realistic settings, students may use teacher-derived initializations to improve distillation efficiency (Wang et al., 2023), but they often differ from the teacher in architecture, width, depth, or other design choices. It remains open whether our methods can be used to detect representation transfer under such architecture mismatch, or whether this requires a stronger notion of initialization or behavioral matching between teacher and student.

## References

Alex Cloud, Jacob Goldman-Wetzler, Evžen Wybitul, Joseph Miller, and Alexander Matt Turner. Gradient routing: Masking gradients to localize computation in neural networks, 2024. URL https://arxiv.org/abs/2410.04332.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Sören Mindermann, Jacob Hilton, Samuel Marks, and Owain Evans. Language models transmit behavioural traits through hidden signals in data. *Nature*, 652:615–621, 2026. URL https://doi.org/10.1038/s41586-026-10319-8.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Chayanon Kitkana and Shivam Arora. Sustained gradient alignment mediates subliminal learning in a multi-step setting: Evidence from MNIST auxiliary logit distillation experiment. In *Workshop on Scientific Methods for Understanding Deep Learning*, 2026. URL https://openreview.net/forum?id=UJM4H9oLJN.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Ivica Nikolic, Teodora Baluta, and Prateek Saxena. Model provenance testing for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=Iy4cAXotrf.

Qin Shi, Amber Yijia Zheng, Qifan Song, and Raymond A. Yeh. Knowledge distillation detection for open-weights models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=Qy5vFFeCZW.

Somin Wadhwa, Chantal Shaib, Silvio Amir, and Byron C Wallace. Who taught you that? tracing teachers in model distillation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3307–3315, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.173. URL https://aclanthology.org/2025.findings-acl.173/.

Xinpeng Wang, Leonie Weissweiler, Hinrich Schütze, and Barbara Plank. How to distill your BERT: An empirical study on the impact of weight initialisation and distillation objectives. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1843–1852, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.157. URL https://aclanthology.org/2023.acl-short.157/.

# A Implementation Details

## A.1 Model Architecture

For the standard autoencoder experiment, the model consists of an Encoder, Decoder, and auxiliary-logit head. All hidden layers use ReLU activations:

- Encoder: $784 \to 2048 \to 512 \to 32$.
- Decoder: $32 \to 512 \to 2048 \to 784$.
- Auxiliary-logit head: $32 \to 32$.

For the routed and flipped-routing experiments, we use the same Encoder, Decoder, and auxiliary-logit head, and additionally include certificate decoders for each half of the encoding. All hidden layers use ReLU activations:

- Encoder: $784 \to 2048 \to 512 \to 32$.
- Decoder: $32 \to 512 \to 2048 \to 784$.

- Top certificate decoder: $16 \to 512 \to 2048 \to 784$.
- Bottom certificate decoder: $16 \to 512 \to 2048 \to 784$.
- Auxiliary-logit head: $32 \to 32$.

The Encoder maps a flattened $28 \times 28$ MNIST image to a 32-dimensional encoding. The top half of the encoding consists of the first 16 dimensions, and the bottom half consists of the remaining 16 dimensions. The Decoder maps the full encoding back to a $28 \times 28$ image reconstruction. Each certificate decoder maps one half of the encoding to a reconstruction.

The auxiliary-logit head reads from the full 32-dimensional encoding. The auxiliary logits are not used in the teacher objective.

### A.2 Datasets

We use the standard MNIST dataset (LeCun et al., 1998), consisting of 60,000 training images and 10,000 test images. Images are normalized with mean 0.5 and standard deviation 0.5.

The training set is split into:

- Training subset: 50,000 images for teacher, control, and post-hoc decoder training.
- Audit subset: 10,000 images for computing the diagnostic statistics used during the development.

This 50k/10k partition is generated by a seed-controlled random split and therefore varies across runs.

We also construct a synthetic noise dataset of 60,000 images of size $28 \times 28$. Each pixel is sampled from a uniform distribution on $[-1, 1]$. The noise dataset is deterministic given the run seed and is used only for auxiliary-logit distillation.

### A.3 Teacher Training

For the standard autoencoder experiment, the teacher is trained as a plain autoencoder with mean absolute reconstruction error: $\mathcal{L}_{\text{teacher}} = \mathcal{L}_{\text{recon}}$. No gradient routing, certificate losses, latent $\ell_1$ penalty, or top/bottom correlation penalty is used in this setting.

For the routed and flipped-routing experiments, the teacher is initialized from a reference model and trained as a gradient-routed autoencoder. For the original routing condition, reconstruction gradients for digits 0–4 are routed through the top half of the encoding, while reconstruction gradients for digits 5–9 are routed through the bottom half. The forward pass is unchanged; routing is implemented by applying stop-gradients to the inactive half of the encoding.

For an input image $x$ with label $y$, the teacher objective is $\mathcal{L}_{\text{teacher}} = \mathcal{L}_{\text{recon}} + \lambda_{\ell_1}\mathcal{L}_{\ell_1} + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \mathcal{L}_{\text{cert-top}} + \mathcal{L}_{\text{cert-bottom}}$. Here, $\mathcal{L}_{\text{recon}}$ is the mean absolute reconstruction error of the full Decoder, $\mathcal{L}_{\ell_1}$ is the mean absolute value of the encoding, and $\mathcal{L}_{\text{corr}}$ is the mean absolute correlation between dimensions in the top and bottom halves of the encoding, computed across the mini-batch. The certificate losses are mean absolute reconstruction errors from the top and bottom certificate decoders. The certificate decoders receive detached encodings, so certificate losses do not update the Encoder.

For the flipped-routing condition, we swap the routing assignment: digits 5–9 are routed through the top half, and digits 0–4 are routed through the bottom half.

### A.4 Control Autoencoder For Routing Experiment Training

This control is used only for the routed and flipped-routing experiments. We train a normally trained autoencoder control from the same reference initialization. The control uses the same architecture, optimizer, reconstruction loss, latent $\ell_1$ penalty, correlation penalty, and certificate losses as the routed teacher. The

only difference is that no gradient routing is applied: reconstruction gradients flow through the full encoding for all digits.

### A.5 Student Distillation

The student is initialized from the same reference model as the teacher. The teacher is frozen during distillation. The student is trained only to match the teacher's auxiliary-logit distribution on synthetic noise images.

For a noise input $x_{\text{noise}}$, the distillation loss is $\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{KL}}\left(p_T^{\text{aux}}(x_{\text{noise}}) \| p_S^{\text{aux}}(x_{\text{noise}})\right)$, where $p_T^{\text{aux}}$ and $p_S^{\text{aux}}$ are the teacher and student auxiliary-logit distributions. The student is not trained on MNIST labels, MNIST reconstruction targets, or the routing objective.

### A.6 General Training Hyperparameters

All experiments are run over 10 random seeds. For the standard autoencoder experiment, teacher weight decay is set to 0. The routed and flipped-routing experiments are the only settings that use teacher/control weight decay $5 \times 10^{-5}$, post-hoc decoder weight decay $5 \times 10^{-5}$, encoding $\ell_1$ penalty coefficient $\lambda_{\ell_1} = 0.003$, and correlation penalty coefficient $\lambda_{\text{corr}} = 0.1$.

- Optimizer for teacher and control: Adam (Kingma & Ba, 2017)
- Teacher/control learning rate: $1 \times 10^{-3}$
- Teacher/control weight decay: $5 \times 10^{-5}$
- Teacher/control epochs: 200
- Student optimizer: Adam
- Student learning rate: $3 \times 10^{-4}$
- Student epochs: 5
- Batch size: 2048
- Encoding $\ell_1$ penalty coefficient: $\lambda_{\ell_1} = 0.003$
- Correlation penalty coefficient: $\lambda_{\text{corr}} = 0.1$
- Number of auxiliary logits: 32

### A.7 Post-hoc Decoders Evaluation

For the standard autoencoder experiment, we freeze the Encoder and train one post-hoc decoder from the full 32-dimensional encoding. The post-hoc probe weight decay is set to 0:

- Full post-hoc decoder: $32 \to 512 \to 2048 \to 784$.

All hidden layers use ReLU activations.

For the routed and flipped-routing experiments, we instead train separate post-hoc decoders from each half of the encoding:

- Top post-hoc decoder: $16 \to 512 \to 2048 \to 784$
- Bottom post-hoc decoder: $16 \to 512 \to 2048 \to 784$

All hidden layers use ReLU activations.

The post-hoc decoders are trained on the 50,000-image MNIST training subset using mean absolute reconstruction error. We use the same post-hoc decoder procedure for the reference model, routed teacher, normally trained control, and distilled student.

Post-hoc decoder hyperparameters are:

- Optimizer: Adam
- Learning rate: $1 \times 10^{-3}$
- Weight decay: $5 \times 10^{-5}$
- Epochs: 50
- Batch size: 2048

Post-hoc decoder reconstruction errors are evaluated on the 10,000-image MNIST test set. For the standard autoencoder, the main metric reported in Table 1 is the MAE of the full-encoding post-hoc decoder. For the routed experiments, the MAE reported in Table 1 is the average of the top- and bottom-half post-hoc decoder MAEs, evaluated over all test digits. For the partition analysis in Table 2, we additionally compute digit-group-specific MAEs for each half of the encoding and use them to form reference-relative localization gaps.

### A.8 Teacher Tracing Experiment

For the tracing experiment, we compute linear Centered Kernel Alignment (CKA) between encoder activations on the MNIST test set. For each model, we collect encoder activations for all 10,000 test images and compute CKA using the full activation matrices rather than averaging CKA over mini-batches. Given centered activation matrices $X \in \mathbb{R}^{n \times d_1}$ and $Y \in \mathbb{R}^{n \times d_2}$, we compute

$$\mathrm{CKA}(X, Y) = \frac{\|Y^\top X\|_F^2}{\|X^\top X\|_F \|Y^\top Y\|_F}.$$

In both the standard and routed autoencoder settings, we report CKA computed on the full 32-dimensional encoder activations. To trace a student, we compare its representation to the candidate teachers from all seeds and identify the teacher with the highest CKA similarity.

## B Post-Hoc Decoder Reconstruction Examples

As discussed by Cloud et al. (2024), obtaining a visually clean split representation in MNIST autoencoders requires a carefully chosen setup and tuned gradient-routing hyperparameters. For our purposes, however, a visually clean partition is not required. We only need the routed teacher to induce a measurable asymmetry between the two halves of the encoding, so that we can test whether this teacher-specific representation structure is transferred to the student through auxiliary-logit distillation.

Figure 2 provides a qualitative sanity check for this representation split. The top and bottom post-hoc decoder reconstructions are not expected to show perfectly separated digit-specific information. Instead, the figure is intended to illustrate the same effect measured quantitatively in Table 2: the routed teacher exhibits a stronger top/bottom reconstruction asymmetry than the reference or control models, and the distilled student shows a weaker but corresponding asymmetry.

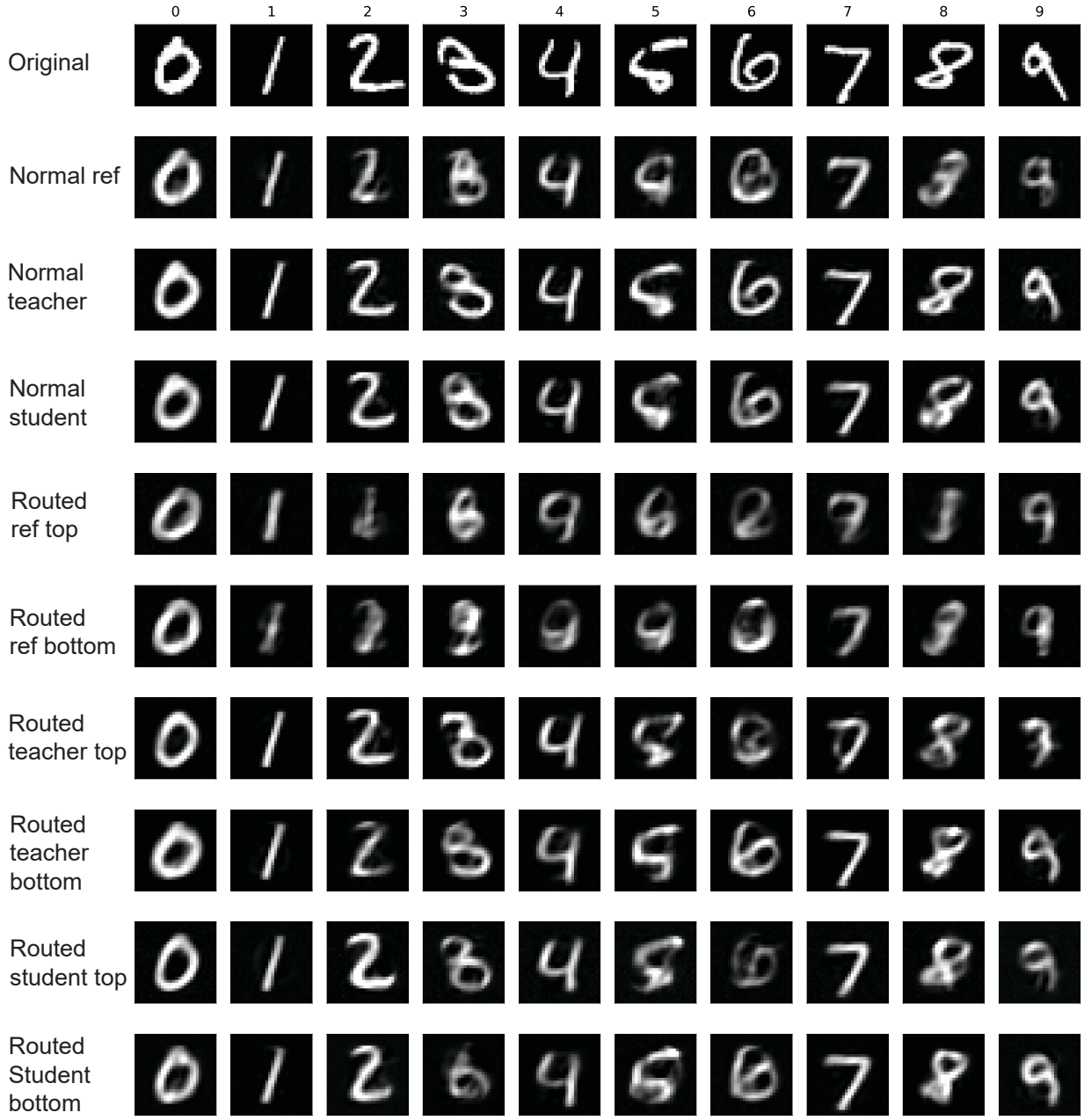


Figure 2: Post-hoc decoder reconstructions from MNIST autoencoder encodings, with routing 0–4 to the top half and 5–9 to the bottom half.