TIP-OF-THE-TONGUE EFFECT: LATENT GEOMETRIC STRUCTURES RESIST CATASTROPHIC FORGETTING

Anonymous authors
Paper under double-blind review

ABSTRACT

The phenomenon of catastrophic forgetting, where models trained on successive tasks rapidly lose the ability to perform early-learned tasks, has largely been characterized by behavioral metrics such as accuracy and loss, leaving unanswered the question of whether underlying task knowledge was permanently overwritten or merely suppressed. We investigate whether representational similarity metrics can detect persistent knowledge by tracking the internal geometry of speech and vision models across abrupt task transitions, comparing against single-task baseline models to capture task-specific representational geometry. We find that this task-specific geometry is preserved long after task accuracy and loss decay to chance, and is retained more strongly by larger models despite equivalent chance-level performance. We further find that models with task-specific geometry relearn the original task 8.3%–18.0% faster on average than matched controls without prior exposure to that task. In sum, our findings show that behavioral metrics alone do not capture the true scope of model retention, with implications for how catastrophic forgetting should be evaluated moving forward.

1 INTRODUCTION

When modern neural networks are trained sequentially on data from separate tasks, they rapidly lose performance on earlier tasks, a phenomenon known as *catastrophic forgetting* (McCloskey & Cohen, 1989). Standard continual learning evaluations are based almost exclusively on behavioral metrics such as task accuracy or loss values. However, behavioral metrics alone cannot distinguish between two possibilities: either forgotten task knowledge is fully erased from model parameters, or it persists but is suppressed by loss pressure or representational drift, preventing task completion. This distinction is important because fundamental questions in continual learning depend on whether behavioral forgetting reflects genuine loss of stored knowledge or merely the inability to access that knowledge. For example, Ramasesh et al. (2022) report that larger models forget less, while Luo et al. (2023) report the opposite.

The question of erasure versus suppression is not unique to artificial systems. In human memory, the "tip-of-the-tongue" phenomenon displays a similar ambiguity in that forgotten knowledge seems to be present but is inaccessible to recall (Brown & McNeill, 1966). We hypothesize a similar effect in machine-learned models: latent representational geometry from certain tasks may persist long after behavioral metrics for that task have fallen to near-chance levels. However, unlike humans, where tip-of-the-tongue states can spontaneously resolve over time (Schwartz, 2001), artificial forgetting is static until intervened on by strategies such as a return to samples from the original task (Inamdar et al., 2026).

Because static behavior alone cannot distinguish whether tasks have been completely forgotten or merely suppressed, we must detect latent task knowledge directly by investigating internal geometry. To do this, we used representational similarity analysis (RSA; Kriegeskorte et al., 2008) and centered kernel alignment (CKA; Cortes et al., 2012) to track the state of internal representations in models subjected to abrupt task shifts. By measuring these quantities relative to models trained on a single task, we can test whether forgotten knowledge has been effectively erased.

Across computer vision (Split CIFAR-100 with ResNet architectures (He et al., 2016)) and speech processing (CommonVoice with Conformer architectures (Gulati et al., 2020)), we observed that task-switched models retained the representational geometry of the original task long after their

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.

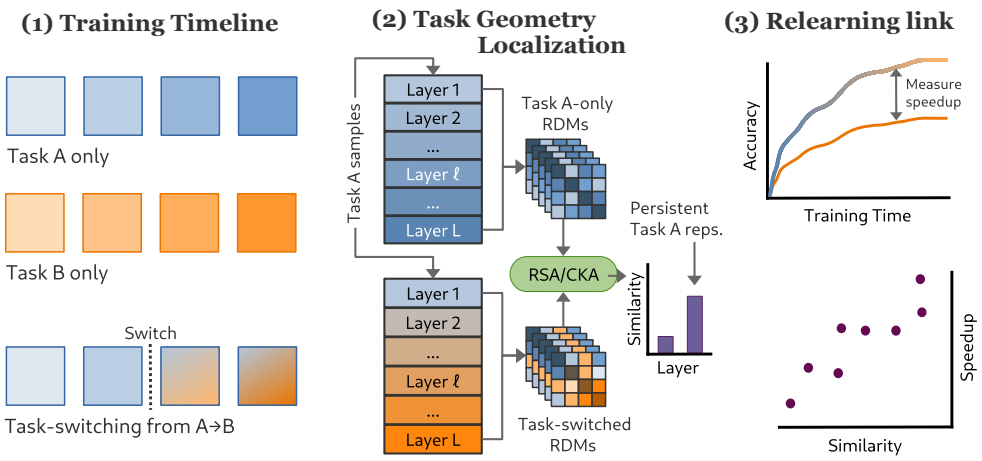

Figure 1: **Pipeline for localizing persistent geometry.** (1) Models are trained on a single task (A, B) or transferred across tasks ($A \rightarrow B$). (2) Layer-wise representational dissimilarity matrices (RDMs) are extracted using test samples from task A, using RSA/CKA to pinpoint specific layers retaining geometry from task A. (3) Measuring speedup reveals a link between geometry and re-learning.

performance on that task had rapidly declined to near-chance levels. By systematically probing representations across model depth and size (ResNet-18 to -101 and Conformers from 25M to 109M parameters), we show that latent-geometry retention scales with model size, suggesting that continual-learning robustness in overparameterized models may stem from retaining dormant representations.

Finally, we tested whether the retained geometric structures are functional rather than an inert artifact of training by returning task-switched models to their first task and measuring the relearning speed. Our experiments showed that task-switched models exhibit accelerated relearning (the savings effect; Ebbinghaus (1913)), suggesting that representational traces might serve a purpose when the original task suddenly becomes relevant again.

In summary, our contributions are the following.

1. We demonstrate across both speech and vision modalities that representational geometry from early-learned tasks persists long after accuracy on that task degrades to chance.
2. We analyze the effect of representational depth on geometry retention, comparing and contrasting effects across speech and vision models.
3. We show that the retention of representational geometry scales with model capacity, clarifying conflicting findings in the literature on the effect of model size on catastrophic forgetting.
4. We link higher representational similarity to an accelerated rate of re-learning on data from a forgotten task, showing functional significance in supporting future learning.

The rest of the paper is organized as follows: we discuss our work in relation to critical-period and continual-learning research in Section 2, and follow with the experimental approach via representational analysis in Section 3, with experimental results in Section 4, and finally end with a discussion of the impact and limitations of our approach in Section 5.

2 Related Work

Critical Periods in Artificial Systems. Previous work has compared biological system learning with machine learning, and found some surprising similarities. In Achille et al. (2019), the authors found that certain kinds of early-training data deficits (e.g., blurring images) but not other kinds (e.g., flipping images) led to long-term performance loss, even when the unmodified images were

reintroduced partway through training. Following this, Frankle et al. (2020) explored the effects on performance of rewinding sub-networks to an earlier stage of training, and Kleinman et al. (2023) showed that critical period effects extend to multi-sensory integration systems. Constantinescu et al. (2025) studied language-switching in neural networks for language models, and Plantinga et al. (2026) studied the same language-switching paradigm for speech models; the latter showed evidence that automatic speech recognition models retain low-level acoustic traces of a 'pre-adoption' language. In this work, we generalize this persistence across modalities and show its dependence on model capacity.

Analyzing Representational Similarity. Many techniques compute neural-network similarity, but representational methods stand out because they are invariant to geometric transformations such as permutations and translations (Klabunde et al., 2025; Bo et al., 2025). Representational Similarity Analysis (RSA), originally introduced in neuroscience, provides a way to compare representational structure across different neural measurements, such as fMRI, EEG, and single-neuron recordings (Kriegeskorte et al., 2008). RSA has been used for far-reaching comparisons, such as humans and deep neural networks (Yamins et al., 2014; Kietzmann et al., 2019). Centered kernel alignment (CKA) is likewise a method for representational comparison, and is nearly equivalent in its computation (Williams, 2024). By borrowing these neuroscience techniques, we conduct comparisons between models with different initializations and training histories that apply even when behavioral similarity is not evident.

Continual Learning and Catastrophic Forgetting. Research on the effect of model size on forgetting has had mixed results. While some work has found less forgetting in larger models (Ramasesh et al., 2022), other work has found that larger models are actually *more* prone to forgetting than smaller models, contrary to expectation (Luo et al., 2023; Lee, 2024). One attempt to resolve this used linear probes to reveal that forgetting may not be as catastrophic as accuracy alone suggests (Davari et al., 2022). Follow-up work has suggested that forgetting may be more like suppression (Kotha et al., 2024), since forgotten tasks can sometimes be revived (Sun & Gao, 2024). Even a single forward pass has been used to revive forgotten knowledge (Feng et al., 2025). A recent mechanistic analysis of catastrophic forgetting in large language models used CKA to find that representational drift is strongest in the middle layers of the network (Imanov, 2026). A recent paper by Inamdar et al. (2026) showed that few-shot learning is often enough to recover forgotten tasks. Our work complements these prior works by revealing clues about the extent to which forgetting does or does not occur in the model's representational geometries.

Savings on Relearning. The original savings-on-relearning work began with psychological experiments on memory and relearning by Ebbinghaus (1913). One particularly relevant form of savings on relearning comes from the international adoptee literature: Hyltenstam et al. (2009) found that despite having no conscious access to their birth language, early-life adoptees pick up birth-language phonemic distinctions faster than their monolingual peers. Indeed, bilinguals have been shown to experience the tip-of-the-tongue phenomenon more often than monolinguals, with the frequency modulated by the age of first exposure to the second language (Kreiner & Degani, 2015). In machine learning, Mayo et al. (2023) found that repeatedly revisiting a task increased the speed of relearning. Our work suggests that these relearning effects may be tied to persistent geometry from the original task.

3 Locating internal geometry of a lost task

An overview of our method can be found in Figure 1. In short, we first trained models with one or more abrupt task switches, as well as models with no such switches, to establish our basis for comparison (Section 3.1). After training, we established representational similarity between models using both RSA and CKA methods, localized by representation level (Section 3.2). Finally, we returned the models to their original task training environment to assess whether they relearn at an accelerated rate (Section 3.3).

3.1 Abrupt task switching

To conduct our investigations, we compared models trained on a single task with models trained with abrupt task shifts to reveal the extent to which representational geometric manifolds from one

task remain in a model after behavioral metrics approach chance. We conducted our experiments across two separate modalities: image classification and automatic speech recognition (ASR).

For the image classification experiments, we used CIFAR-100 randomly divided into 10 tasks of 10 classes, training models for 2k updates on each in sequence. For our test condition model, we trained on tasks 1–9, which we call $M_{\text{T1}\to 9}$. We trained an equivalent experience control model without task 1, i.e. on task 10, then 2–9 ($M_{\text{T10,2}\to 9}$). We added augmentation according to the policy learned in Cubuk et al. (2019). We adopted a warmup-stable-decay learning rate schedule that keeps the learning rate consistent over 70% of steps, avoiding concerns that early decay might unduly restrict parameter mobility. We used off-the-shelf models (ResNet-18, -34, and -101) to cover a variety of model sizes, and ran the experiment over the same 8 random seeds for each size, with class labels randomized for each seed. We performed additional experiments with ViT (Dosovitskiy et al., 2021) on ImageNet (Deng et al., 2009) using a task division based on the BREEDS living and non-living categories (Santurkar et al., 2021), the results of which can be seen in Appendix A.

For our speech experiments, we adopted the sequential language-switching paradigm described in Plantinga et al. (2026), using 1,000-hour subsets of Common Voice (French and German, with corresponding models called M_{fr} and M_{de}) trained on Conformer encoder-decoder architectures with SpeechBrain (Ravanelli et al., 2024). Unlike the prior work which evaluated fixed single-scale models, we evaluated models across three capacity tiers (25M, 52M, and 109M parameters), training three seeds for each model. We adjusted capacity through encoder/decoder layer count (6/3, 8/4, 12/6) and heads/dimensions per layer (4/256, 6/384, 8/512).

In order to further explore the impact of the input structure on retention of task-specific geometry, we also trained phoneme-to-text models using the exact same recipe as the speech-to-text case, swapping only the inputs from 80-dim spectral features + CNN downsampling to 64-dimensional embeddings of the phoneme ids aligned to 40ms bins, clamping the phoneme length to be between 1 and 4 frames. We used Montreal Forced Aligner (MFA; McAuliffe et al., 2017) to generate time-aligned phonemes at the same rate as the ASR inputs, developing a unified phoneme set by combining similar phonemes across the three languages and dropping allophones, for a total of 51 phonemes. The models converged 60% faster, so we trained them for 50% fewer update cycles.

3.2 Measuring representational geometry

A crucial aspect of our approach is determining how well representational geometry is preserved throughout training of the task-switched model. We did this by comparing representational similarity between the task-switched model and a single-task expert model from the first-encountered task. To perform this similarity comparison, we used a higher-order representation-space approach based on ranking of stimulus pairs (Popal et al., 2019), which lends itself to comparisons across a variety of transforms, including translations, rotations, and other geometric effects of random initialization and data ordering.

More specifically, we constructed a representational dissimilarity matrix (RDM) for each model over a set of 500 stimuli from the first task (pre-switch) not seen during training. We used a single-vector representation for each sample by averaging the internal model vector outputs for a given layer over time (speech) or over space (vision) when the sample was presented. Then, we compared all pairs of representations using cosine similarity to construct the similarity matrix.

Formally, for a set of N held-out stimuli from task A and a given layer ℓ, let $\mathbf{r}_\ell^{(i)}$ denote the averaged internal representation of stimulus i at layer ℓ. We define the representational dissimilarity matrix at layer ℓ as

$$\text{RDM}_\ell(i,j) = 1 - \cos\left(\mathbf{r}_\ell^{(i)}, \mathbf{r}_\ell^{(j)}\right), \qquad i,j \in \{1,\dots,N\}, \tag{1}$$

and retain only the upper triangle (excluding the diagonal), which we denote $\mathbf{v}_\ell$, to avoid double-counting each pair and to exclude the trivially zero self-similarity terms.

We compared layer representations between two models using two complementary metrics: RSA, computed over the RDMs defined above, and linear CKA, computed directly on the activation matrices. Given two models M_1 and M_2, we define their layer-ℓ RSA as

$$\mathrm{RSA}_\ell(M_1, M_2) = \mathrm{Spearman}\big(\mathbf{v}_\ell^{(M_1)}, \mathbf{v}_\ell^{(M_2)}\big). \tag{2}$$

Unlike RSA, we computed CKA directly on the layer activations rather than on the RDMs. Let $\mathbf{X}_\ell$ and $\mathbf{Y}_\ell$ denote the stacked layer-ℓ representations for the N held-out stimuli under models M_1 and M_2, and let $\mathbf{X}_{c,\ell}$, $\mathbf{Y}_{c,\ell}$ denote their column-centered versions. Linear CKA is then

$$\mathrm{CKA}_\ell(M_1, M_2) = \frac{\big\|\mathbf{X}_{c,\ell}^\top \mathbf{Y}_{c,\ell}\big\|_F^2}{\big\|\mathbf{X}_{c,\ell}^\top \mathbf{X}_{c,\ell}\big\|_F \,\big\|\mathbf{Y}_{c,\ell}^\top \mathbf{Y}_{c,\ell}\big\|_F}. \tag{3}$$

We report RSA (Eq. 2) in the main text and use CKA (Eq. 3) as a robustness check (Appendix B).

To quantify retained task-A geometry in a task-switched model $M_{\mathrm{A}\to\mathrm{B}}$ relative to a task-A expert M_{A}, we normalize against a control model M_{C} that was never exposed to task A, using the same comparison at every checkpoint or, where noted, averaged over layers ℓ and the final k checkpoints:

$$\Delta_{\mathrm{RSA}} = \frac{\mathrm{RSA}(M_{\mathrm{A}\to\mathrm{B}}, M_{\mathrm{A}}) - \mathrm{RSA}(M_{\mathrm{C}}, M_{\mathrm{A}})}{1 - \mathrm{RSA}(M_{\mathrm{C}}, M_{\mathrm{A}})}. \tag{4}$$

We refer to Δ_{RSA} as the *headroom-normalized retention gap*: it expresses the observed similarity gap between the test and control models as a fraction of the maximum possible gap (i.e., the headroom remaining above the control baseline). This is the quantity reported as the "Relative RSA Gap" in Section 4.1 and, averaged over layers, seeds, and the final three checkpoints, as "Representational Traces" in Section 4.3.

To localize preserved task-specific geometry, we extracted the representation from each level independently to see which parts of the network are more likely to retain some of their geometry when encountering new data. We tested the hypothesis that training data similarity and task similarity can both affect which parts of the network retain the most representational geometry. The results are listed in Section 4.2.

Finally, we analyzed how network size affects the geometric traces of the first task remaining in the network, showing results in Section 4.3. For each modality, we varied only the depth and number of parameters while keeping the inputs and task parameters identical. This allowed us to test whether larger models are more or less forgetful in terms of their representational geometry, while their accuracies were indistinguishable (i.e., random chance).

3.3 Original task relearning

To determine whether the preserved geometry is merely an artifact of training or provides a functional benefit, we retrained our task-switched models on data from the original task and tested relearning speed. We measured the speedup according to the number of update steps it took for models to reach an arbitrary accuracy threshold (e.g., 70%) when the model was returned to training on the first task.

To demonstrate the relationship, we took checkpoints from throughout second-task training and measured the relearning speed of each checkpoint to show that the model gets progressively slower at relearning as the representations drift. We then correlated these relearning speed scores with averaged representational similarity scores across the model to determine whether a relationship exists between how fast a model relearns and our measured retention gap.

Finally, we measured relearning speed of fully-trained models to demonstrate that the relearning benefit extended throughout training. We compared these models against baseline models with equivalent experience, except for the relevant task, to rule out the possibility that other parts of the learning history drove the speedup.

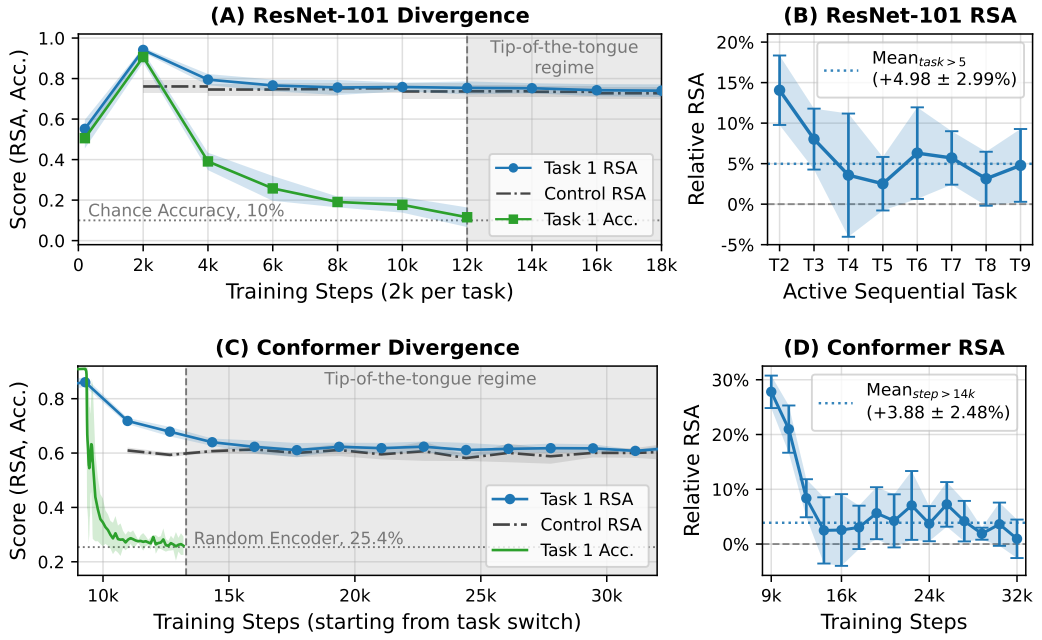

Figure 2: **Behavioral metrics diverge from representational metrics: behavioral metrics approach chance, while representational metrics remain above baseline.** Baseline models are trained for an equivalent number of steps. Bands represent 95% confidence intervals. **(A)** Scores throughout training on nine CIFAR-100 tasks for ResNet-101 models. Comparison of behavioral metric (accuracy) against RSA($*$, M_{T1}). We average representational similarity scores over all layers. **(B)** Relative gap between RSA($M_{\text{T1}\to 9}$, M_{T1}) and RSA($M_{\text{T10,2}\to 9}$, M_{T1}). **(C)** Divergence between conformer model first-task next-token accuracy and RSA($*$, M_{de}). Chance accuracy is computed by feeding scrambled audio encodings to the frozen decoder. **(D)** Relative gap between RSA($M_{\text{de}\to\text{fr}}$, M_{de}) and RSA(M_{fr}, M_{de}).

4 Results

We report our results according to the structure outlined above: first showing a divergence between metrics in Section 4.1, then a depth analysis in Section 4.2, with a capacity analysis in Section 4.3, and ending with the results of a relearning experiment in Section 4.4.

4.1 Divergence of Behavioral and Representational Metrics

First, as shown in Figure 2, we see a clear divergence between representational and behavioral metrics for ResNet-101 and Conformer medium. We see the expected catastrophic forgetting in the behavioral metric (accuracy), which decays rapidly when training on new tasks. After only a few thousand updates, accuracy on the first task becomes indistinguishable from chance and remains this way for the rest of training. By contrast, representational similarity decays but remains significantly above baseline long after accuracy reaches chance.

For the ResNet-101 model, accuracy and loss do not differ from chance by task six ($p = 0.239$). By contrast, the RSA score captures an average of 4.98% higher-than-baseline representational similarity for all tasks after 5 (95% CI $[+1.99\%, +7.97\%]$). We interpret this evidence as proving that the internal geometric manifolds associated with the first task are not completely lost, as the behavioral metrics alone would suggest.

For the Conformer medium model, the accuracy on the first language using a frozen decoder drops more rapidly than for the vision task, as the between-task distribution shift is larger. We determined chance accuracy by passing randomly shuffled encoder representations to the frozen decoder, which essentially reduces it to next-token prediction. Our test model's accuracy was not significantly

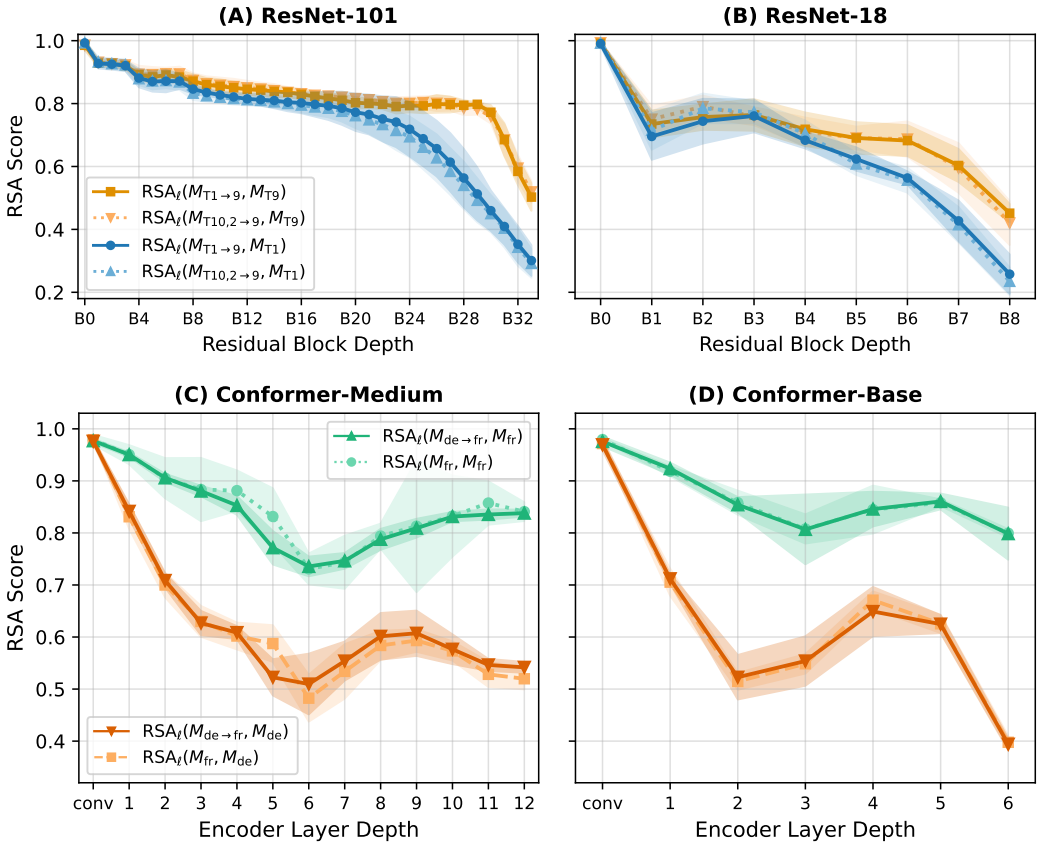

Figure 3: **Data and architecture shape representational similarity across layers independent of size, while size predicts the degree of savings.** The chart shows end-of-training similarity scores across representation depth between task-switched models and single-task expert models. **(A)** Nine-task ResNet-101 similarity by layer, compared with a nine-task baseline model that hasn't seen the original task. **(B)** Nine-task ResNet-18 model against corresponding baseline. **(C)** Medium speech-to-text model, switched from German to French, compared against single-language models trained to fluency. **(D)** Base speech-to-text model, german-to-french.

different than the random encoder after 4k steps of new-language training ($p = 0.294$). Despite this, evidence of first-language representational geometry still appears tens of thousands of updates later, averaging $+3.88\%$ (95% CI $[1.44\%, 6.37\%]$) after accuracy has gone to chance.

4.2 Representational Depth Analysis

After confirming that the stabilized model representations still retain traces of early learning experiences at the end of training, we explored converged-model representational similarity across the depth of representation, visible in Figure 3, revealing distinct patterns of representational similarity based on model and task, but strikingly similar patterns over differing model capacities. Appendix C explores task dependence in more detail.

The comparison between first-task and last-task curves reveals that the gap between similarly trained models can be quite small in early-to-middle layers. So while our headroom-normalized Δ_{RSA} is only $+4.98\%$ for ResNet-101, this may actually understate the true degree of savings. If you change the theoretically maximal similarity from 1.0 (used in Figure 2) to Task 9 cross-seed similarity, the relative gap grows to $+11.3\%$ overall, and as high as $+35.1\%$ at block B23.

The encoder layers of the speech-to-text models seen in Figure 3 (C) and (D) show a similar pattern to the vision model in terms of the location of the strongest traces: they occur near the middle of the

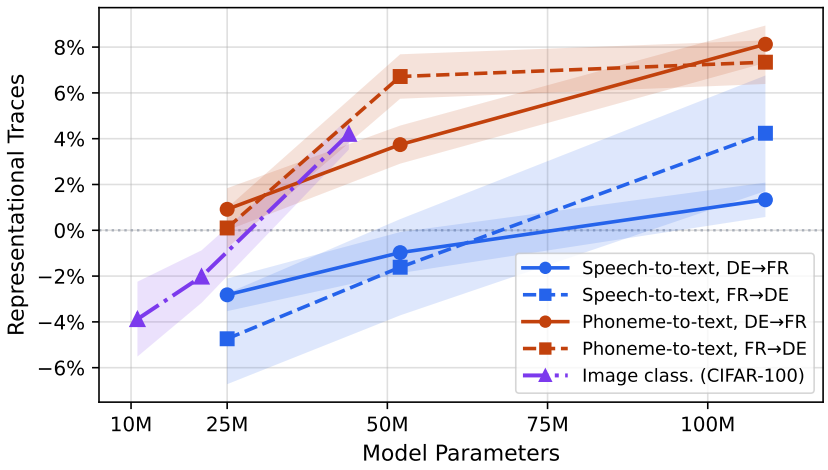

Figure 4: **Increasing model capacity demonstrates stronger representational traces.** The y-axis is computed as the normalized difference between two representational similarity scores: 1) the task-switched model relative to the first task model, and 2) between two separate single-task models. We average savings over layers, seeds, and the last three checkpoints, with corresponding standard errors shown as shaded regions.

network depth, although traces in the early layers are also present in high relative amounts even if the absolute difference is small. Another noticeable pattern is the double-descent curve, where some similarity is regained through the middle layers of the encoder. We suspect this is because languages can differ at lower phonetic levels while maintaining more similar higher-level representations (e.g., grammar rules).

4.3 Model Capacity Analysis

In Figure 4 we show the relationship of the representational traces (Δ_{RSA}) to the size of each model, using averaged similarity scores at the end of training over all the layers and seeds. Across all task types, the chart shows that larger models retain more representational geometry from the first task. The representational traces are stronger for the phoneme-to-text model relative to the speech-to-text model, which we suspect is due to the reduced demands of the task. As there is plenty of capacity to complete each task, less of the geometry from the original task gets repurposed for the new task.

This begins to answer the question unresolved in prior literature about whether model size corresponds to a stronger or weaker tendency to forget, suggesting that larger models retain increased latent task knowledge, even if that knowledge is not accessible when accuracy is measured.

4.4 Savings on Relearning

Finally, in Figure 5 we establish the presence of a link between representational similarity and the speed at which models can relearn a task that has been forgotten. We investigate this by zooming in on the second task of our computer vision experiment, where forgetting is strongest, and the data span a wide range of similarity and relearning speed.

The first panel shows a steady deterioration in relearning speed as training on the second task progresses, and establishes the criterion we use for relearning speed. In the second panel, the relearning metric demonstrates a strong relationship ($r = -0.52, p < 0.001$) with the representational similarity scores described above. Zooming back out, by the end of T1–T9 training, our vision models still display an 8.3% relearning speed advantage, and our speech models display a greater 18.0% advantage on average; details are in Appendix D.

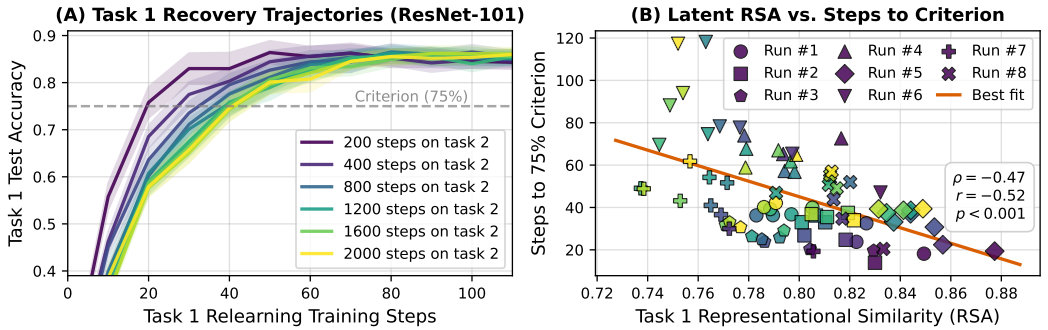

Figure 5: **Relearning speed is linked to representational similarity.** Across 10 epochs on a new task (200 steps each) we measure the representational drift as well as the number of steps it takes to reach an arbitrary threshold when relearning the original task. (A) Relearning trajectories for six checkpoints taken across second-task learning. (B) Plot of model checkpoints showing the relationship between representational similarity with first-task expert models and number of steps taken to reach the 70% criterion when relearning the first task.

5 Conclusion

The current work shows that catastrophic forgetting in neural networks may partly stem from how we choose to measure it. Behavioral metrics provide an important measure of usefulness but do not show where or to what extent forgetting actually occurs. This distinction between behavioral performance and underlying representational structure echoes a long-standing theme in human development research, where neural measures often uncover aspects of learning and retention that behavioral assessments miss (e.g., Pierce et al., 2014). The parallel to human development suggests that neuroscience tools remain underused for explaining neural network behavior and merit broader adoption in the machine learning community.

A notable strength of the present findings is that both the speech and the visual experiments show the same dissociation between behavior and representation. This convergence across modalities suggests that retained representational geometry is not specific to speech processing or image classification, but may reflect a more general property of neural learning systems. The persistence of latent structure despite behavioral forgetting therefore appears to be a general feature of representation learning, supporting the view that catastrophic forgetting may constitute a loss of access more than a complete erasure of prior knowledge.

Several limitations of the current work point to directions for future investigation. Our experiments are restricted to speech and vision modalities with a limited range of model sizes; whether the capacity effect holds at very large scales, or for the language modeling modality, remains an open question. Additionally, while RSA captures geometric similarity, it does not reveal which specific representational features are retained or how they contribute to downstream task performance. Future work using more advanced probing classifiers or causal intervention methods could clarify the functional significance of the retained geometry. Finally, the "tip-of-the-tongue" phenomenon, i.e. that forgotten knowledge may be suppressed rather than lost, deserves direct experimental investigation, for instance by developing strategies to revive access to early-task representations in task-switched models.

AI use statement

In this work, we used generative AI tools to implement methods for our CIFAR-100 and ImageNet/BREEDS experiments. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Additionally, we used generative AI tools to create or modify scientific figures or images,

and to edit this paper to improve readability. We have thoroughly reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

All code for experiments and analysis is included in the supplemental material and will be publicly released upon acceptance of the paper.

References

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Yiqing Bo, Ansh Soni, Sudhanshu Srivastava, and Meenakshi Khosla. Evaluating representational similarity measures from the lens of functional correspondence. In *8th Annual Conference on Cognitive Computational Neuroscience*, 2025.

Roger Brown and David McNeill. The "tip of the tongue" phenomenon. *Journal of verbal learning and verbal behavior*, 5(4):325–337, 1966.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*, 13:96–120, 2025.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *J. Mach. Learn. Res.*, 13:795–828, 2012.

Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. AutoAugment: Learning augmentation strategies from data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 113–123, 2019.

MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Probing representation forgetting in supervised and unsupervised continual learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16691–16700. IEEE, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Hermann Ebbinghaus. *Memory: A contribution to experimental psychology*. Teachers College, Columbia University, New York City, 1913. Translated by Henry A. Ruger and Clara E. Bussenius from the 1885 original in German.

Christiane D. Fellbaum (ed.). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

Tao Feng, Wei Li, Didi Zhu, Hangjie Yuan, Wendi Zheng, Dan Zhang, and Jie Tang. ZeroFlow: Overcoming catastrophic forgetting is easier than you think. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16808–16824. PMLR, 13–19 Jul 2025.

Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The early phase of neural network training. In *International Conference on Learning Representations*, 2020.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, and Yonghui Wu. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pp. 5036–5040, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kenneth Hyltenstam, Emanuel Bylund, Niclas Abrahamsson, and Hyeon-Sook Park. Dominant-language replacement: The case of international adoptees. *Bilingualism: Language and cognition*, 12(2):121–140, 2009.

Olaf Yunus Laitinen Imanov. Mechanistic analysis of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2601.18699*, 2026.

Amogh Inamdar, Matthew So, Vici Milenia, and Richard S. Zemel. Re-evaluating continual learning with few-shot adaptation. *CoRR*, abs/2606.03843, 2026.

Tim C. Kietzmann, Courtney J. Spoerer, Lynn K.A. Sörensen, Radoslaw M. Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, 2019.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Michael Kleinman, Alessandro Achille, and Stefano Soatto. Critical learning periods for multisensory integration in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24296–24305, 2023.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations, ICLR*, 2024.

Hamutal Kreiner and Tamar Degani. Tip-of-the-tongue in a second language: The effects of brief first-language exposure and long-term use. *Cognition*, 137:106–114, 2015.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational Similarity Analysis – connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Eunhae Lee. The impact of model size on catastrophic forgetting in online continual learning. *ArXiv*, abs/2407.00176, 2024.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE/ACL Transactions on Audio, Speech and Language Processing (TASLP)*, pp. 3776–3786, 2023.

David Mayo, Tyler R. Scott, Mengye Ren, Gamaledin Elsayed, Katherine Hermann, Matt Jones, and Michael Mozer. Multitask learning via interleaving: A neural network investigation. In *Proceedings of the annual meeting of the cognitive science society*, volume 45, 2023.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Interspeech*, volume 2017, pp. 498–502, 2017.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Lara J. Pierce, Denise Klein, Jen-Kai Chen, Audrey Delcenserie, and Fred Genesee. Mapping the unconscious maintenance of a lost first language. *Proceedings of the National Academy of Sciences*, 111(48):17314–17319, 2014.

Peter Plantinga, Charlotte Moore, Peter W. Donhauser, Krista Byers-Heinlein, and Denise Klein. Lost but not erased: Finding traces of a forgotten language in neural speech models. *arXiv preprint arXiv:2608.25976*, 2026.

Haroon Popal, Yin Wang, and Ingrid R. Olson. A guide to representational similarity analysis for social neuroscience. *Social cognitive and affective neuroscience*, 14(11):1243–1253, 2019.

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*, 2022.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain De Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, et al. Open-source conversational AI with SpeechBrain 1.0. *Journal of Machine Learning Research*, 25(333): 1–11, 2024.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. BREEDS: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*, 2021.

Bennett L. Schwartz. *Tip-of-the-tongue states: Phenomenology, mechanism, and lexical retrieval*. Psychology Press, 2001.

Huashan Sun and Yang Gao. Reviving dormant memories: Investigating catastrophic forgetting in language models through rationale-guidance difficulty. *arXiv preprint arXiv:2411.11932*, 2024.

Alex H. Williams. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, volume 285 of *Proceedings of Machine Learning Research*, pp. 10–23. PMLR, 14 Dec 2024.

Daniel L.K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

A APPENDIX: VIT EXPERIMENTS ON IMAGENET (BREEDS)

For our ImageNet experiments, we constructed two image classification tasks using BREEDS, which groups ImageNet synsets into higher-level semantic partitions derived from the WordNet hierarchy (Deng et al., 2009; Santurkar et al., 2021; Fellbaum, 1998). We focused on top-level "living" and "non-living" categories, to construct the largest possible groupings for independent tasks. The living task contains 88,400 training images across 17 subclasses, while the non-living task contains 132,765 training images across 26 subclasses.

We trained DeiT-style ViT-Tiny (5M params), ViT-Small (22M params), and ViT-Base (86M params) models from scratch under two conditions: single-task baseline models trained only on living or non-living images, and curriculum models that were switched from one task to the other. The single-task models serve as the source-only and target-only references for the RSA comparisons below. All models used the same recipe: AdamW, effective batch size 256, maximum learning rate 10^{-3}, and a 15k step warmup followed by a constant learning rate and 65k optimization steps. Curriculum models switched tasks after 13k steps, before the learning rate peaked. We averaged all experiments across 3 random seeds.

Figure 6 applies the RSA procedure of Section 3.2 to the ViT classification token representation of each transformer block, i.e., the vector used as the model's global image representation. We used fixed held-out subsets of 512 living and 512 non-living validation images. The plotted quantity is source similarity above a cross-task baseline: the similarity between the curriculum model and the source-only model, minus the similarity between independently trained source-only and target-only models. Finally, the figure averages across the two curriculum directions. Under these conditions, we found that the models showed measurable source traces after switching tasks, with stronger persistence into later layers for larger ViTs.

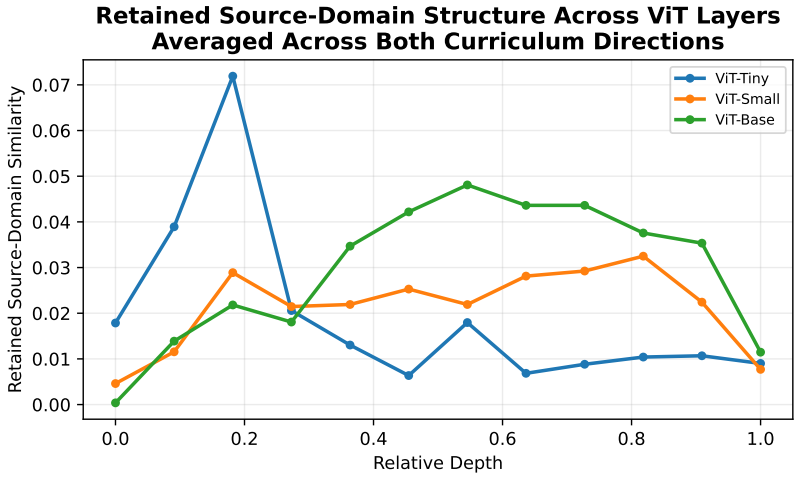

Figure 6: **Layerwise source-task traces in ViT models trained on ImageNet/BREEDS.** For each layer, we compare a curriculum model after switching tasks to a source-only model, then subtract the similarity expected between source-only and target-only models. The plot uses final checkpoints and source-task images only: living $\rightarrow$ non-living models are evaluated on living images, and non-living $\rightarrow$ living models are evaluated on non-living images. Curves average these two curriculum directions. Higher values indicate stronger retention of the original source-task representation.

B Appendix: Comparison of RSA and CKA

To ensure our findings are not an artifact of metric selection, we compared CKA and RSA across all $N = 960$ evaluated layer comparisons, spanning three model sizes (ResNet-101, ResNet-34, ResNet-18) and two conditions (cross-task model, control model). As shown in Figure 7, there is strong concordance between the metrics (Pearson's $r = 0.963$ and Spearman's $\rho = 0.925$). A further Bland-Altman agreement analysis shows a slight bias toward CKA giving higher similarity scores ($+3.4\%$, 95% limits of agreement $[-7.4\%, +13.7\%]$), but the uniform offset cancels out when computing relative retention gaps between test and control models.

We show a practical example in Figure 8, where the shape of the similarity curves matches closely over time, with a slight bias towards higher scores for CKA.

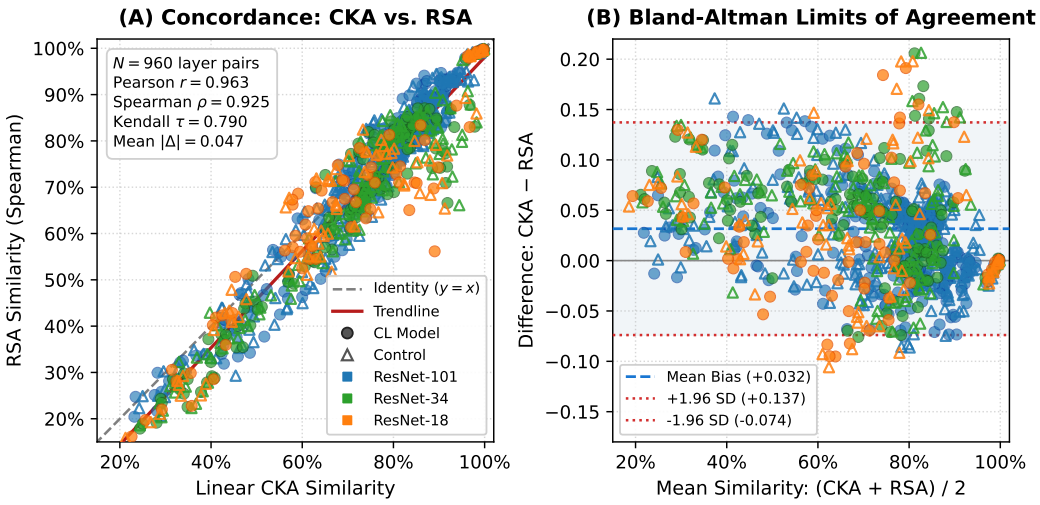

Figure 7: **Comparison of CKA vs. RSA metrics across measurements**, for layers in 3 sizes of ResNet and test vs. control condition. (A) Comparison of CKA vs. RSA value over 960 layer pairs, including different sizes and conditions of model training. (B) Limits of agreement between CKA and RSA metrics.

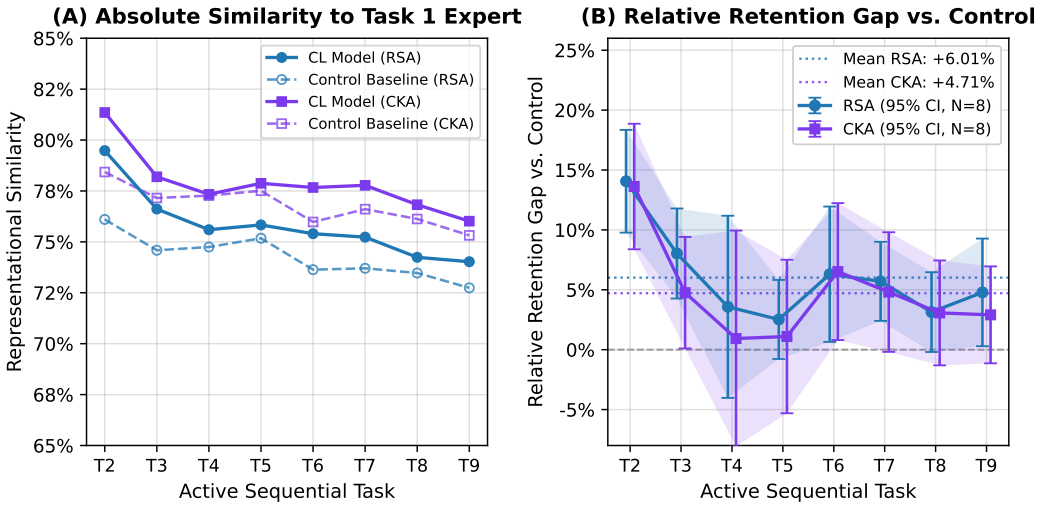

Figure 8: **Comparison of CKA vs. RSA metrics over the training process** of ResNet-101, similar to Figure 2. (A) Absolute similarity scores relative to the first task expert, averaged over the whole model. (B) Relative gap in scores between test and control models throughout training.

C Appendix: Task-dependence of Representational Similarity

By changing the task direction for speech and the data distribution of CIFAR-100 to group the classes in each task according to the superclass definitions, we demonstrate the effects of data on representational similarity independent of architecture, the results of which can be seen in Figure 9.

In both cases, the center of mass of traces of first-task representations shifts from mid-to-later layers back to earlier layers, suggesting that the location of the savings in the network is highly data-dependent. For computer vision, a random class distribution favors representational traces at higher levels, as low-level details are distributed roughly the same way across classes. By grouping related classes, we see that the representational traces concentrate in the lower levels instead (most visible in B1–B3), reflecting a distributional shift between tasks in image qualities such as edges and textures.

In speech, the most obvious trend when reversing the task-switch direction is a much smaller gap between cross-language and within-language representations in the earliest layers. We interpret that to mean that while a French model is not too capable of representing low-level German pronunciation, German models are more than capable of representing low-level French pronunciation. This may partly reflect the task's difficulty, as German models train faster and achieve 6% WER, whereas French models train much slower and reach only 10% WER at convergence.

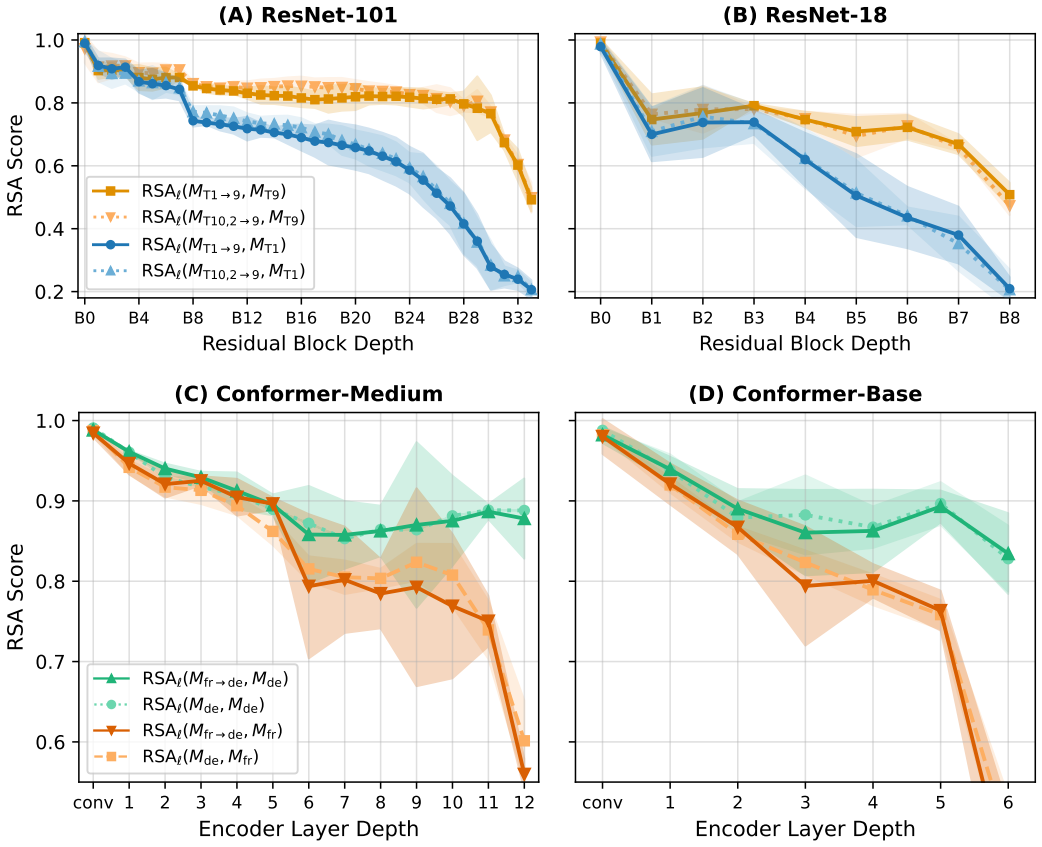

Figure 9: **Task-dependence of representational similarity**, i.e. curves under data distribution changes. (A) and (B) show ResNet-101 and ResNet-18 respectively under superclass-grouped task structure, a total of 3 seeds each. (C) and (D) show conformer medium and conformer base under the French-to-German model task switch direction, also 3 seeds for each model.

D Appendix: Relearning Velocity at End of Training

We establish that relearning speedups continue through the end of training by returning fully trained models to their original domain. Across all experimental conditions, returning models to their original domain yielded 4.7% to 36.6% relearning speedup, with the strongest size-based effects for French relearning (progressing from $8.4\% \rightarrow 19.5\% \rightarrow 36.6\%$ across model sizes).

For image classification and the German relearning experiments, the largest model is not clearly faster at relearning than the smallest model, which may partly be a result of high standard error. Some conditions have a large variance due to training instability, as certain seeds reach a local minima early in relearning which significantly delays the steps to the criterion.

By using the median instead of the mean, the picture changes slightly, and the model-size trend is seen clearly in ResNet ($2.9\% \rightarrow 5.5\% \rightarrow 12.0\%$), overall 14.3% ($p = 0.0197$, Wilcoxon). But the trend is no longer visible in French relearning ($7.1\% \rightarrow 10.7\% \rightarrow 8.4\%$) and remains unclear in German ($9.9\% \rightarrow 13.4\% \rightarrow 10.8\%$). Under the median, pooled speech speedup is 10.4% ($p = 0.0189$, Wilcoxon).

Table 1: **Savings on relearning for fully-trained continual learning models** (mean $\pm$ SEM). Continual learning models relearn their first, forgotten task, and are compared against matched baseline control models that learn the task for the first time. Relative speedup is measured as the percentage reduction in training steps required to reach the 75% accuracy criterion. Paired t-tests assess significance ($^{*}p < 0.01$, $^{\dagger}p < 0.05$).

		Steps to 75% Criterion		**Relearning Speedup**	
Architecture	**Params**	**CL Model**	**Baseline**	**Steps Saved**	**Rel. Gain**
Image Classification (Split-CIFAR-100)					
ResNet-18	11M	26.5 ± 3.9	29.6 ± 4.6	$+3.1 \pm 1.1^{\dagger}$	$+10.6\%$
ResNet-34	21M	28.3 ± 3.8	29.7 ± 4.6	$+1.4 \pm 2.7$	$+4.7\%$
ResNet-101	44M	22.4 ± 2.8	24.8 ± 2.6	$+2.4 \pm 2.0$	$+9.8\%$
Pooled Vision	Combined	$\mathbf{25.7 \pm 2.0}$	$\mathbf{28.0 \pm 2.3}$	$\mathbf{+2.3 \pm 1.1}$	$\mathbf{+8.3\%}$
Speech Recognition: German Relearning (DE → FR → DE)					
Conformer-Base	25M	871.6 ± 3.2	965.9 ± 1.3	$+94.3 \pm 3.4^{*}$	$+9.8\%$
Conformer-Small	52M	862.5 ± 157.3	925.3 ± 96.2	$+62.8 \pm 184.4$	$+6.8\%$
Conformer-Medium	109M	654.6 ± 2.8	735.0 ± 1.6	$+80.4 \pm 3.2^{*}$	$+10.9\%$
Pooled German	Combined	796.3 ± 57.6	884.5 ± 50.5	$+88.2 \pm 76.6$	$+10.0\%$
Speech Recognition: French Relearning (FR → DE → FR)					
Conformer-Base	25M	1008.8 ± 11.4	1101.2 ± 7.4	$+92.4 \pm 13.6^{\dagger}$	$+8.4\%$
Conformer-Small	52M	822.5 ± 14.2	1021.2 ± 88.6	$+198.8 \pm 89.8$	$+19.5\%$
Conformer-Medium	109M	508.5 ± 171.0	802.2 ± 38.6	$+293.7 \pm 175.3$	$+36.6\%$
Pooled French	Combined	744.7 ± 84.3	983.3 ± 53.4	$+238.6 \pm 99.8^{\dagger}$	$+24.3\%$
Pooled Speech (All)	Combined	$\mathbf{765.8 \pm 54.3}$	$\mathbf{933.9 \pm 37.4}$	$\mathbf{+168.1 \pm 65.9^{\dagger}}$	$\mathbf{+18.0\%}$