ORIGINS AND ROLES OF WORLD REPRESENTATIONS IN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

While neural representations have been extensively studied in large practical models, the controlled conditions that govern their emergence and their downstream role in model adaptation remain poorly understood. In this work, we develop a framework separating the underlying world, the data generation process, and the resulting model representations to answer these questions in a controlled setup. This framework further allows clearly defining expected behavioral and representational changes resulting from a world update. Specifically, we define the world as a set of city coordinates and define 7 geometric tasks which generate data to train an autoregressive language model. First, we show that different data generation processes give rise to different world representations in the model. Next, we show that multi-task training drives representational alignment between models that do not share any common tasks, providing controlled evidence for the Multitask Scaling Hypothesis, a potential explanation of the Platonic Representation Hypothesis. Finally, we study whether multi-task models can integrate new entities consistently via fine-tuning. Surprisingly, we find that some fine-tuning tasks are "divergent" and actively harm the representational integration of new entities. Overall, our framework establishes a model system to study the emergence of world representations in neural networks and their adaptability in a controlled manner.

1 INTRODUCTION

The nature of representations and mechanisms learned by deep neural networksor in fact any intelligent systemand their relation to generalization is a central topic in deep learning research (Hubel & Wiesel, 1962; Rosenblatt, 1958; Fukushima, 1980; Rumelhart et al., 1986). Recent work has demonstrated that neural networks trained on vast amounts of data can capture diverse, disentangled, and sometimes interpretable aspects of their training data, or even of the world underlying the data (Bengio et al., 2014). These rich representations are generally thought to underlie the generalization and adaptability of neural networks to unseen, out-of-distribution scenarios.

Recent work on internal representations of language models has provided evidence that neural networks can develop structured representations of the underlying data rather than merely memorizing surface patterns (Li et al., 2022; Gurnee & Tegmark, 2023; Nanda et al., 2023b).

However, major open questions remain. When interpretable representations are discovered in neural networks, it is often unclear whether their emergence is surprising or inevitable, what geometry they will take, and how they support generalization. Even less understood is how these representations adjust during fine-tuning and downstream adaptation.

Answering these questions is difficult in real-world settings, where the key factorsthe world, the data, and the modelare entangled and costly to vary independently. Even the most accessible factor, the model, becomes expensive to perturb at scale; the data is harder still to control; and the underlying world is effectively immutable. In this work, we develop a synthetic framework where these factors can be precisely controlled and systematically studied.

This work. To study these questions, we decouple the underlying *world* from the *data generation process* to control them independently. Concretely, we adopt the coordinates of real-world cities as

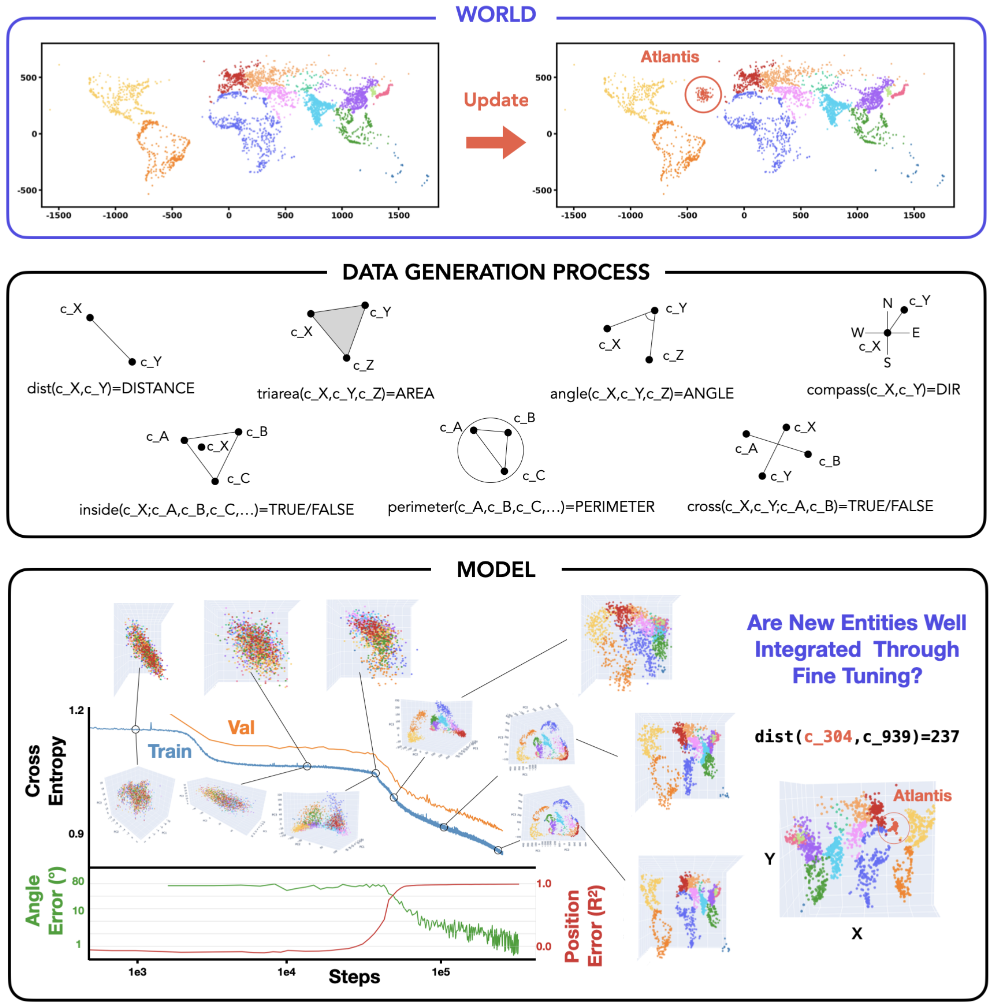

Figure 1: **Overview of the World-Data-Model framework. Top:** The world consists of 5,075 real city coordinates; we test adaptation by adding 100 synthetic `Atlantis` cities (App. B.1). **Middle:** Seven geometric tasks generate training data from city coordinates (App. B.2). **Bottom:** Training dynamics of one model, showing loss curves, linear probing accuracy for coordinate reconstruction, and visualizations of internal representations (PCA and linear probe projections) at different training stages. See App. Fig. 8 for all training curves.

our "world," a ready-made complex structure with ground-truth geometry, and define 7 geometric tasks on top of it. We train autoregressive Transformers on this data and study how world representations form and vary across tasks, surfacing preliminary evidence for the Platonic Representation Hypothesis (Huh et al., 2024). Crucially, this setup allows us to define consistent updates to the world (adding new cities) that produce predictable changes in the data, letting us test whether models can absorb such updates via fine-tuning. Our contributions are as follows:

- **A Framework Decoupling World, Data, and Model. (Sec. 3)** We separate the underlying world (city coordinates) from the data generation process (7 geometric tasks), enabling systematic study of how different tasks shape representations of the same world. The world provides ground-truth coordinates for directly assessing representation quality via probing. This setup also allows defining consistent world updates (adding synthetic `Atlantis` cities) to test whether models can adapt their representations accordingly.
- **Task-Dependent Geometry and Multi-Task Convergence. (Sec. 4)** We show that different tasks operating on the same world produce highly variable representational geometries across tasks and seeds. However, multi-task training drives convergence: models trained on multiple

tasks show higher representational alignment, even when they share no common tasks. This provides partial evidence for the Multitask Scaling Hypothesis, one proposed mechanism for the Platonic Representation Hypothesis.

- **Divergent Tasks Harm Fine-Tuning of New Entities Despite Multi-Task Pretraining. (Sec. 5)** We test whether models can integrate new entities (`Atlantis` cities) via fine-tuning. We find that single-task representational similarity (CKA) partially predicts cross-task generalization. In a multi-task fine-tuning setting, we find surprising "divergent" tasks which hinder integration of new entities into the learned manifold, actively harming generalization.

2 Related Work

Internal Representations. Understanding internal representations has been fundamental since the development of neural networks (Rosenblatt, 1958; Rumelhart et al., 1986). Recent work has revealed that language models develop structured "world models" encoding geographic, temporal, and relational information (Li et al., 2022; Gurnee & Tegmark, 2023; Nanda et al., 2023b; Marks & Tegmark, 2024). Mechanistic interpretability and sparse autoencoders have further enabled decomposition of neural activations into interpretable features (Anthropic AI, 2023; Templeton et al., 2024). Furthermore, the Platonic Representation Hypothesis posits that diverse models converge toward similar representational structures (Huh et al., 2024). However, recent work questions this representational optimism, suggesting that deep network representations may be more brittle than previously assumed (Kumar et al., 2025). Our work takes a complementary perspective, studying the factors that control the formation of these representations and how networks integrate new entities into their representation space via fine-tuning.

Fine-tuning. The pretraining-finetuning paradigm has become central to modern deep learning, with seminal works establishing its effectiveness in computer vision (Krizhevsky et al., 2012; He et al., 2015) and natural language processing (Devlin et al., 2018; Radford et al., 2018). Despite widespread success, fine-tuning exhibits poorly understood behaviors such as the reversal curse (Berglund et al., 2024; Lampinen et al., 2025). On this background, careful studies of fine-tuning and other low-compute adaptation methods have raised pessimism about whether models can learn fundamentally new abilities, suggesting they may merely form "thin wrappers" around pretrained representations (Jain et al., 2023; Ward et al., 2025; Yue et al., 2025; Qin et al., 2025). Work on feature distortion (Kumar et al., 2022) is perhaps most related to ours, though representational changes are assumed rather than directly measured. Our work examines this question in a controlled setup where ground-truth world structure enables precise measurement of representation adaptation.

Multi-task Learning. Multi-task learning has long been studied as a way to improve generalization through shared representations (Caruana, 1997); in some sense, modern foundation models represent an extreme form of multi-task training. Large-scale multi-task pretraining typically assumes rich representations emerge from data diversity (Aghajanyan et al., 2021), but the precise mechanisms remain underexplored. Recent work has begun studying task diversity in controlled settings (Michaud et al., 2023; Zhang et al., 2025), though most studies still focus on aggregate behaviors such as capabilities and scaling laws rather than characterizing tasks or the knowledge they operate on. Our framework explicitly defines tasks as geometric functions over a shared world, enabling direct investigation of how task structure shapes representations.

Synthetic Data. The cost and complexity of foundation models has motivated synthetic approaches for controlled study of in-context learning (Xie et al., 2021; Chan et al., 2022; Reddy, 2023; Ravents et al., 2023; Park et al., 2024b; Wurgaft et al., 2025), compositional generalization (Okawa et al., 2024; Park et al., 2024c), grammar/knowledge acquisition (Allen-Zhu & Li, 2023b;a), and interpretability methods (Menon et al., 2025; Hindupur et al., 2025). Most relevant to our work, Jain et al. (2023) used synthetic data to argue fine-tuning creates only thin wrappers over pretrained capabilities, while Nishi et al. (2024) studied formation and destruction of representational structure. However, existing synthetic frameworks typically design data generation processes without explicitly distinguishing between the underlying world and how data is sampled from it. Our work introduces a framework that makes this distinction explicit, enabling systematic study of how different views of the same world shape neural representations and their downstream adaptability.

See App. E for extended related work.

3 Experimental Framework: Decoupling World, Data, and Model

Our framework uses geographic tasks where models solve geometric problems involving city coordinates. This naturally separates the underlying world (coordinates) from data generation (tasks), while providing ground-truth for measuring representation quality. Our framework provides three key properties:

1. **Learnability:** All tasks are deterministically generated from the same underlying coordinates. A model that learns the world structure can leverage it across all tasks.
2. **Latent State:** Models never see coordinates directly, only task outputs, allowing us to probe whether they internally reconstruct the world structure.
3. **Consistent Updates:** Modifying the world (e.g., adding new cities) produces self-consistent updates across all tasks, defining a clear expectation for what a model with proper world representations should internalize.

Framework. Let $\mathcal{W}$ denote a *world*: a set of entities $\{e_1, \ldots, e_N\}$ each with latent attributes $z_i \in \mathcal{Z}$. A *data generation process* is a set of tasks $\mathcal{T} = \{T_1, \ldots, T_K\}$, where each task $T_k : \mathcal{Z}^{n_k} \to \mathcal{Y}_k$ maps n_k entity attributes to an output space $\mathcal{Y}_k$. Training data for task T_k is generated by sampling entity tuples $(e_{i_1}, \ldots, e_{i_{n_k}})$ from $\mathcal{W}$ and computing $y = T_k(z_{i_1}, \ldots, z_{i_{n_k}})$.

A model M observes only entity identifiers and task outputs, never the latent attributes z_i directly. We say M has learned a *world representation* if there exists a probe P such that $P(M(e_i)) \approx z_i$ for all entities.

A *world update* $\mathcal{W} \to \mathcal{W}'$ (e.g., adding or modifying entities) induces consistent updates across all tasks by simply applying the same T_k to the new or modified entities.

Instantiation. Concretely, our world consists of 5,075 real-world cities filtered by population $>$ 100,000 (Fig. 1, top). We define 7 geometric tasks that take 2 or more city coordinates as input and compute a geometric value (Fig. 1, middle).

Each task query follows a structured format where city IDs (e.g., `c_1234`) serve as inputs to geometric functions, all character-tokenized for autoregressive prediction. For instance, `dist(c_0865,c_4879)=769` queries the distance between two cities, while `cross(c_2345,c_6789;c_0123,c_4567)=TRUE` checks whether two line segments intersect.

To test adaptation, we define `Atlantis`: 100 synthetic cities placed in the Atlantic Ocean. Models never observe `Atlantis` during pretraining; we use it in Sec. 5 to test whether fine-tuning can integrate new entities into world representations in a way that generalizes across tasks.

4 World Representations Converge Under Multi-Task Learning

We now study how the task composition in the pretraining data shapes internal world representations by training Transformers on different task subsets and probing their representation geometry (see App. B.3 for training details).

Result 1: World Representations Emerge through Autoregressive Training We first demonstrate that world representations emerge through autoregressive training (Fig. 1, bottom). Training on the `angle` task, the model starts with random representations, goes through a loss plateau while clustering nearby cities, then forms world-aligned geometry as loss drops and task accuracy improves. The linear probe R^2 for coordinate decoding rises slightly before angle accuracy improves, reminiscent of hidden progress measures found during grokking (Nanda et al., 2023a). *Notably, once representational structure forms, it remains largely fixed for the remainder of training: representations are essentially fixed in the first ∼15% of training, remaining static while loss continues to decrease and accuracy rises* (see App. 9 for visualization across tasks). This early saturation of representations echoes findings on critical learning periods in deep networks (Achille et al., 2019) and loss of plasticity in continual learning (Dohare et al., 2024). Overall, we find stable formation of internal world representations through pure autoregressive modeling. While the emergence of lin-

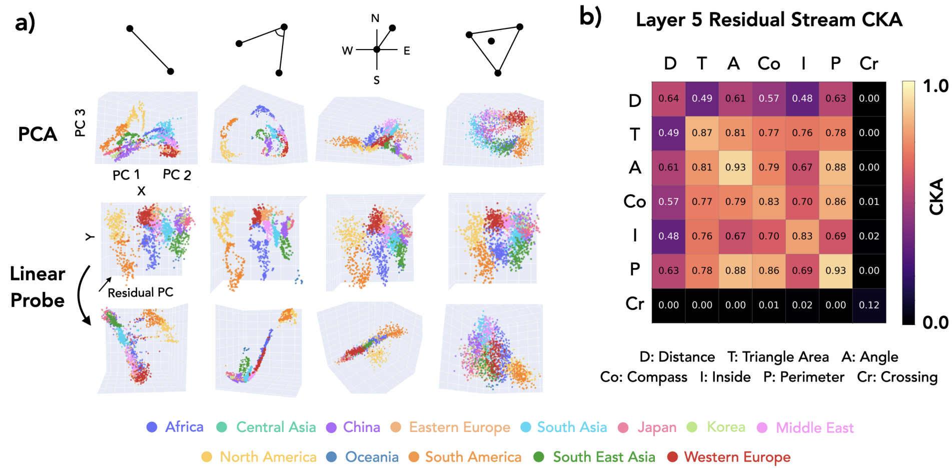

Figure 2: **World representation geometry depends on the data generation process.** (a) Different tasks create distinct geometries: `distance` (thread-like), `angle` (2D manifold), `compass` (fragmented), `inside` (diffuse). Row 1: PCA. Row 2: Linear probe projections. Row 3: Rotated views showing hidden structure. See App. Fig. 10 for more seeds. (b) CKA matrix at layer 5, estimated across 3 seeds. `Crossing` (Cr) fails to train alone. See App. Fig. 11 for SEM and layers 3, 4, 6.

early decodable coordinates might be anticipated given the geometric nature of the task[1], it provides a useful validation of our framework and sets the stage for our main question: how do different tasks shape these representations?

Result 2: Data Generation Process Controls World Representation Geometry We train models from scratch for each of the seven tasks and visualize their representations in Fig. 2(a): PCA projections, linear probe reconstructions, and rotated views.

Different tasks produce qualitatively distinct geometries: `distance` forms thread-like structures, `angle` forms 2D manifolds, `compass` forms fragmented clusters, and `inside` forms diffuse representations. These qualitative patterns are relatively consistent across random seeds (see App. D.2). Despite geometric differences, we can linearly decode (x,y) coordinates from most tasks (row 2), though some tasks (`angle`) yield cleaner reconstructions than othersa phenomenon worth further investigation. The third row shows manually rotated views revealing that representations differ substantially in non-probe directionsa reminder that *linear probing only surfaces what we look for*.

We quantify representational similarity using CKA (Kornblith et al., 2019) (Fig. 2b). We find substantial variability even across seeds for the same task (see App. Fig. 11), but cross-task differences remain clear: `distance` produces particularly divergent representationsa result not obvious from PCA visualizations or from intuition about the task. Note: the `crossing` task fails to train in isolation[2], explaining its near-zero CKA; intriguingly, it succeeds in multi-task settings (Result 3).

Result 3: Multi-Task Learning Drives Representational Convergence Having established that single-task training produces variable representations, we now ask: does multi-task training reduce this variability? This question partially connects to the Platonic Representation Hypothesis (Huh et al., 2024), which observes that neural networks trained on diverse data develop aligned representations even across different modalities and architectures. One potential mechanism they suggest is the Multitask Scaling Hypothesis:

[1]We regard *linear* decodability of world representations as non-trivial (albeit expected). However, this is not the focus of our study.

[2]This likely connects to known hard-to-learn dynamics and gradient plateaus in training transformers (Pezeshki et al., 2021; Shah et al., 2020; Hoffmann et al., 2024; Bachmann & Nagarajan, 2025; Gopalani & Hu, 2025).

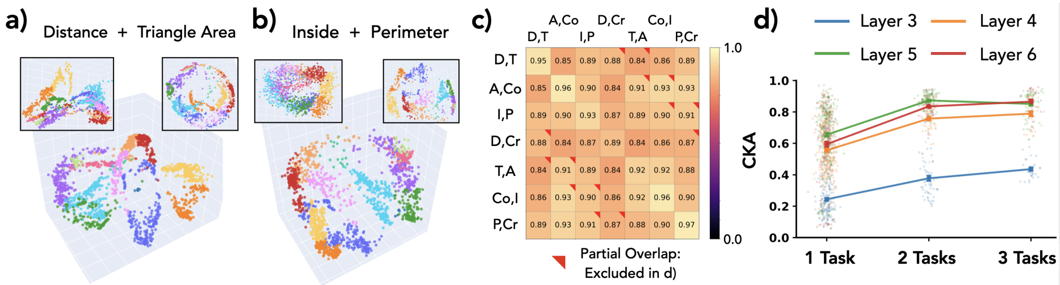

Figure 3: **Multi-task pretraining drives representational convergence.** (a,b) Two-task training creates more regular structures than single-task models. (c) CKA matrix (7×7) for two-task models shows higher alignment (see App. Fig. 12 for SEM). (d) Average CKA increases with task count (1→2→3), saturating at ∼0.85 for layers 4-6 while layer 3 continues improving (see App. Fig. 13 for SEM). `Crossing`, which failed to learn in single-task training, is excluded; including it would only strengthen the convergence finding. 3D visualizations: link.

> *"There are fewer representations that are competent for N tasks than there are for $M \leq N$ tasks. As we train more general models that solve more tasks at once, we should expect fewer possible solutions."*

Our setup provides a potential testbed for this hypothesis, with a ground-truth world model and multiple tasks defined over it. We trained models on selected two-task combinations (3 seeds each; see App. Fig. 14 for all 21 combinations). Fig. 3(a) shows representations when trained jointly on `distance` and `triangle area` (with single-task models shown for comparison), while (b) shows `inside` and `perimeter`. When trained on two tasks, models develop more regular representational structures. While difficult to appreciate in static 2D projections, we encourage readers to explore our interactive 3D visualizations at this link.

We measure CKA between two-task trained models to quantify this alignment (Fig. 3(c)). CKA is substantially higher than for single-task models. One might expect high CKA when models share a task, but even models trained on completely disjoint task pairs show substantially higher alignment. In Fig. 3(d), we plot average CKA for models trained on 1, 2, and 3 tasks across layers 3-6, averaging only over models with completely disjoint task sets. Training on more tasks clearly leads to more aligned representations, with CKA saturating around 0.85 for 2 and 3 tasks in layers 4-6, while layer 3 continues improving. Notably, multi-task training also reduces per-seed variance of representations (App. Fig. 14b).

Overall, we find that *multi-task learning leads to more aligned model internal representations*, providing partial evidence for the Multitask Scaling Hypothesis explanation of the Platonic Representation Hypothesis.[3] Crucially, this alignment emerges even though single-task models achieve comparable task performanceall models reach high accuracy on their respective tasks. Since our networks are trained to representational convergence (as seen in Fig. 1), this demonstrates that the alignment is not simply a byproduct of optimization difficulty but rather that task diversitynot just data quantity or performance pressuredrives aligned representation learning.

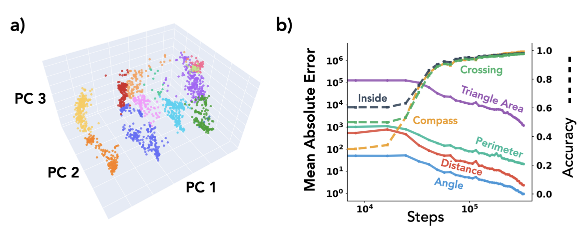

Figure 4: **7-task model.** (a) PCA projection of layer 5 representations naturally reveals world map structure. (b) Training curves showing successful learning of all 7 tasks, including `crossing` which failed in single-task training.

[3] A full test of the Platonic Representation Hypothesis would require showing convergence across different architectures; we test only the task-diversity mechanism here.

An auxiliary finding: the `crossing` task, which was unlearnable alone, trains successfully when paired with any other task. We speculate that companion tasks provide structured coordinate representations that `crossing` can leveragean implicit curriculum where easier tasks scaffold the learning of harder ones through shared representations.

To extend these findings, we trained a model on all 7 tasks simultaneously (Fig. 4). This model successfully learns all tasks, and its PCA projection naturally reveals the world map structure, approaching the perceived quality of linearly probed (x,y) coordinates without requiring any explicit coordinate supervision. Why multi-task training drives more linearly *surfaced* representations remains an open question worthy of future investigation. This 7-task model serves as the foundation for our fine-tuning experiments in the following section.

5 Divergent Tasks Harm Entity Integration via Fine-Tuning

In the previous section we observed how multi-task pretraining yields shared representations for different tasks. In this section, we investigate generalization properties of fine-tuning on top of such representations. However, unlike most fine-tuning studies which focus on changing model behavior in a certain way and evaluate generalization across entities, we study the inverse: fine-tuning an entity into the model and evaluate generalization across tasks. To this end, we use the 7-task model trained in the previous section (Fig. 4).

As mentioned in Sec. 3, we introduce 100 `Atlantis` cities to the world and fine-tune on data containing `Atlantis` to probe for generalization. We emphasize that the introduction of `Atlantis` cities keeps the original dataset fully consistent with the world. Moreover, task operations on `Atlantis` cities are well-defined in the same framework. If the model learned the true data generation process with properly factored representations, it should be able to integrate `Atlantis` seamlessly. If not, we suspect either the representations are fractured (Kumar et al., 2025) or gradient descent cannot trigger the right representational updates (Kumar et al., 2022).

Result 1: Pretraining Phase Representational Alignment Predicts Fine-Tuning Generalization *Despite* Joint Pretraining We first address a simple question: when fine-tuning on `Atlantis` cities for a single task (e.g., `distance`), should we expect the model to automatically generalize to using `Atlantis` for all other tasks?

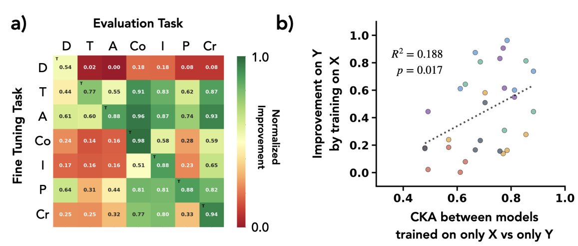

Figure 5: **Fine-tuning generalization and its correlation with representational similarity.** (a) Generalization matrix (averaged over 4 seeds; see App. Fig. 16 for individual seeds): each row is a model that integrated `Atlantis` via one task; columns show normalized improvement on `Atlantis` queries for each task (see App. C.1 for metric details). (b) For each task pair (X, Y), we plot the single-task CKA between X and Y against the normalized improvement on task Y after fine-tuning on task X (see App. Fig. 15 for annotated version).

To answer this, we fine-tune on 100k examples of a single task that include `Atlantis` cities, mixed with original pretraining data to avoid catastrophic forgetting and a small multi-task elicitation set (see App. B.3 for details).

The resulting generalization matrix is shown in Fig. 5(a). This matrix reveals rich phenomenology: some tasks like `distance` show no cross-task generalization (`Atlantis` remains usable only for that task), while `angle` triggers significant generalization across all tasks. Intriguingly, we observe an apparent inverse relationship: tasks that efficiently trigger cross-task generalization of new entities are often those that don't easily benefit from other tasks' fine-tuningthough this relationship is noisy.

Unexpectedly, we find that *generalization performance correlates with the CKA values from single-task pretraining* (Result 2 of Sec. 4). This is puzzling: the CKA values come from models trained from scratch on individual tasks, yet they partially predict fine-tuning behavior of a model pretrained on all tasks jointly (Fig. 5b). If the multi-task model truly uses unified representations for cities, why would single-task representational properties matter?

For clarity, we define two terms: **Divergent tasks** are tasks which have low CKA compared to others when trained in isolation (in our case the `distance` task). **Hidden spaces** are representation spaces not surfaced by PCA or probing but used by divergent tasks.

We hypothesize:

> *"Even though models develop joint world representations which converge in multi-task pretraining, gradient descent on divergent tasks might fail to act on these shared representations during fine-tuning, instead utilizing hidden spaces that don't propagate updates across tasks."*

Our question is then two-part:

1. To what extent do divergent tasks affect fine-tuning generalization?
2. Will gradient descent on divergent tasks fail to merge fine-tuning introduced concepts to the original representation manifold?

Result 2: Divergent Tasks Catastrophically Harm Generalization To investigate how divergent tasks affect generalization, we move from single-task to multi-task fine-tuning settings. First, we introduce a simple heuristic model: fine-tuning on a concatenated dataset $\{D_1, D_2, ..., D_n\}$ (which do not provide conflicting supervision) should combine their individual effects. Specifically, when concatenating and shuffling all fine-tuning data to avoid sequential learning effects like catastrophic forgetting (McCloskey & Cohen, 1989), we expect the improvement on task i after training on tasks j and k to be given by a **best-teacher model**:

$$\text{Improvement}_i(D_j \cup D_k) = \max(\text{Improvement}_i(D_j), \text{Improvement}_i(D_k)) \tag{1}$$

To test this hypothesis, we fine-tuned the 7-task model on all $\binom{7}{2} = 21$ possible two-task combinations. Fig. 6(a,c) shows the *deviation* from our best-teacher expectation (averaged over 4 seeds; see App. Fig. 17 for raw improvements and App. Fig. 18 for individual seeds). Strikingly, we observe "red horizontal bands"models that not only fail to benefit from multi-task training but actually perform worse than their best single-task component. Notably, all these degraded performance bands involve the `distance` task. Fig. 6(c) quantifies this: when we split the deviation values into models with and without `distance`, we consistently observe lower-than-expected performance when the divergent task is included. This confirms that *divergent tasks (those with low single-task CKA) actively harm fine-tuning generalization rather than simply failing to contribute*. We next examine how this manifests in the learned representations.

Result 3: Divergent Tasks Disrupt Representational Integration of New Entities Having shown that divergent tasks harm generalization (Question 1), we now address Question 2: does gradient descent on divergent tasks fail to merge new entities into the representation manifold?

We take two exemplars from the 21 fine-tuning runs: one without `distance` that generalized well (`angle` + `compass`), and one with `distance` that was harmed (`distance` + `perimeter`). We first train a linear probe on a subset of all cities including `Atlantis`; these reconstructions are shown in Fig. 6(b) (top and bottom panels). In the well-integrated case, `Atlantis` cities lie within the world data manifold. In the ill-integrated case, `Atlantis` cities are off the manifold. While this difference appears subtle in 2D projections, the effect is dramatic in 3Dwe strongly encourage readers to explore our interactive visualizations. Next, we train a linear probe on 4000 non-`Atlantis` cities and apply it to `Atlantis` representations (middle panels). In the well-integrated case, `Atlantis` cities (red-orange) are relatively well reconstructed compared to ground truth (black crosses); in the ill-integrated case, reconstruction fails completely.

We quantify this effect in Fig. 6(d), showing histograms of absolute coordinate reconstruction error. When `Atlantis` is integrated via fine-tuning partially on divergent task data (red), reconstruction errors are nearly an order of magnitude larger than when integrated via purely non-divergent tasks (blue). For reference, non-`Atlantis` cities (yellow, still held out from probe training) show low reconstruction error as expected. One might hypothesize that `Atlantis`'s location in the middle of

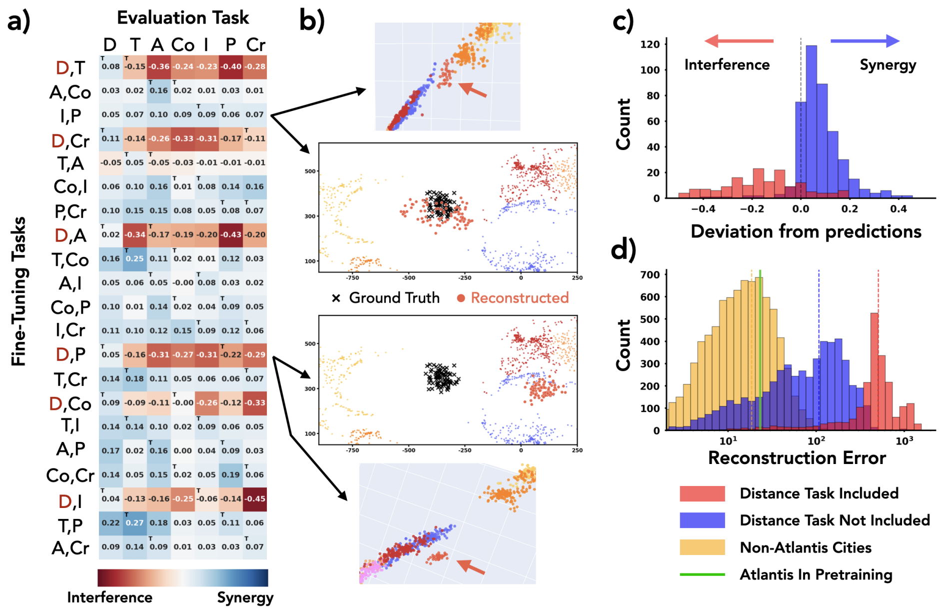

Figure 6: **Divergent tasks harm multi-task fine-tuning and disrupt representational integration.** (a) Deviation from best-teacher expectation for 21 two-task models (rows) across 7 evaluation tasks (columns), computed in normalized improvement space (see App. C.1); "red horizontal bands" show `distance` task combinations degrade performance below single-task baselines. (b) Representation visualization and linear probe reconstruction of `Atlantis`. (c) Histogram of deviation values: models including `distance` vs. not. (d) Linear probe `Atlantis` coordinate reconstruction error for models with `distance`, without `distance`, and baseline on pretraining cities; green vertical line indicates performance when `Atlantis` is part of pretraining.

the ocean creates inherently difficult geometry. To test this, we pretrained a model with `Atlantis` included from the start (green line). In this case, `Atlantis` cities are reconstructed as well as any other city, confirming that the integration failure stems from divergent task fine-tuning dynamics rather than geographic peculiarity.

This suggests that divergent tasks cause optimization to encode new entities in hidden spaces rather than integrating them into the existing world manifoldexplaining their failure to support cross-task generalization.

We emphasize that our findings are correlational: we do not claim that interventions to increase single-task CKA would necessarily improve fine-tuning generalization. Rather, we identify representational divergence as a diagnostic marker for tasks that will harm multi-task fine-tuning performance.

6 Discussion

Continual learning and world models. Our study is motivated by understanding fundamental properties of deep neural networks as building blocks toward general intelligence. Recent work has demonstrated that neural networks can represent more than surface statistics and possess genuine world models, yet we take a more nuanced position: these world models must not only represent the current state of the world but also adapt consistently when the world changes. Such adaptation is non-trivial, as a single change can require cascading updates across different computational tasks. We argue that robustly adaptable internal representations are a prerequisite for general intelligence, though only one aspect of continual learning, which also encompasses learning from experience, internalizing beliefs as tacit knowledge, and knowing when to rely on external tools. Recent language models can adapt to novel inference-time contexts via in-context learning (Brown et al., 2020), forming task-specific representations on the fly (Demircan et al., 2024). However, fine-tuning consistently underperforms in-context learning for knowledge integration (Lampinen et al., 2025; Park et al., 2025). Recent approaches attempt to narrow this gap, either by augmenting transformers with learned adaptation mechanisms (Chen et al., 2024; Charakorn et al., 2025; Zweiger et al., 2025) or by designing architectures that explicitly maintain updatable state (Hochreiter & Schmidhuber, 1997; Schlag et al., 2021; Behrouz et al., 2024; Yang et al., 2025). Our study grounds these questions in a controlled setting, examining how transformer representations evolve under gradient descent and whether their structure supports consistent integration of new knowledge. Building similar setups to compare fundamental properties across different architectures may offer a promising direction for understanding what controls representation formation and adaptation.

Dynamics of representations. Studying representations is a long-standing topic (Rosenblatt, 1958; Rumelhart et al., 1986). Within neural networks, however, most work has examined representations in fixed, trained networks or focused on their formation during pretraining. More recently, there is growing interest in how representations change at test time, or more generally, during adaptation. Park et al. (2024a) show that language models form task-specific representations that internalize aspects of the data generation process, while Shai et al. (2025) demonstrate that models can maintain belief states of external processes. How internal representations adapt at inference time is an active area of research (Bigelow et al., 2025; Lubana et al., 2025). Another line of recent work examines how representations change during fine-tuning: some work draws analogies between fine-tuning and learning activation steering vectors (Wang et al., 2025), while practical studies attempt to understand and leverage representational changes (Casademunt et al., 2025; Minder et al., 2025). To study representational adaptation rigorously, one must define an updatable world where new information implies a consistent set of expected changes. Our framework provides exactly this: introducing `Atlantis` cities defines how representations should update across all tasks, letting us measure whether fine-tuning achieves consistent integration or fragments new entities into task-specific subspaces.

Forward and backward modularity. Our results highlight a distinction that is often overlooked: modularity in the forward pass does not imply modularity in the backward pass. Multi-task training produces clean, structured representations that can be easily decoded into world coordinates, yet these world models can be fractured and partial when it comes to adaptation. Gradient descent may not respect the forward-pass modularity when updating weights: fine-tuning on divergent tasks routes updates through pathways that bypass the shared world manifold, encoding new entities in task-specific subspaces.

Limitations. We study world representation formation and adaptation in a controlled synthetic setting with small-scale models. While we find non-trivial phenomenology, including the emergence of world representations, task-dependent geometry, representational convergence under multi-task training, and off-target fine-tuning effects, it is difficult to guarantee these findings will generalize to large-scale models trained on natural data. Additionally, our findings are largely correlational; we do not yet understand the mechanisms causing these observations. Furthermore, our claims regarding the Platonic Representation Hypothesis are partial: we demonstrate task-driven convergence within a single architecture and modality, but do not explore true multimodality or cross-architecture convergence.

7 Conclusion

We introduced a WorldDataModel framework that separates the underlying world from the data generation process, enabling controlled study of how representations form and adapt. Crucially, this separation allows defining consistent world updates (adding new entities that integrate seamlessly across all tasks), providing clear expectations for what proper world representations should support. Using this framework, we first showed that multi-task training drives representational convergence: models trained on disjoint task sets develop aligned representations, providing partial evidence for the Multitask Scaling Hypothesis. However, this convergence does not guarantee consistent adaptation: certain "divergent" tasks actively harm the integration of new entities during fine-tuning, encoding them in hidden spaces rather than the shared world manifold. This highlights a distinction between forward and backward modularity: clean, structured representations do not necessarily adapt cleanly to new information.

Use of Large Language Models

Large language models were used for:

- Assistance in finding related papers during literature review.
- Boilerplate code for research.
- Refining the language of the manuscript.

Reproducibility Statement

All data generation, model training and analysis were carefully tracked with configuration files to ensure reproducibility. All random seeds for dataset generation and model training were tracked as well (all set to 42). All code, data and analysis results will be open sources after the peer review process. Furthermore, the authors intend to open source the entire research process including the process on converging to the set of experiments presented in the paper.

References

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep neural networks, 2019. URL https://arxiv.org/abs/1711.08856.

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. Muppet: Massive multi-task representations with pre-finetuning, 2021. URL https://arxiv.org/abs/2101.11038.

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

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 1, learning hierarchical language structures. *ArXiv e-prints, abs/2305.13673, May*, 2023b.

Anthropic AI. *Towards Monosemanticity: Decomposing Language Models With Dictionary Learning*, 2023. https://transformer-circuits.pub/2023/monosemantic-features.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Gregor Bachmann and Vaishnavh Nagarajan. The pitfalls of next-token prediction, 2025. URL https://arxiv.org/abs/2403.06963.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time, 2024. URL https://arxiv.org/abs/2501.00663.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014. URL https://arxiv.org/abs/1206.5538.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on ”a is b” fail to learn ”b is a”, 2024. URL https://arxiv.org/abs/2309.12288.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martn Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms, 2025. URL https://arxiv.org/abs/2502.17424.

Eric Bigelow, Daniel Wurgaft, YingQiao Wang, Noah Goodman, Tomer Ullman, Hidenori Tanaka, and Ekdeep Singh Lubana. Belief dynamics reveal the dual nature of in-context learning and activation steering, 2025. URL https://arxiv.org/abs/2511.00617.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velikovi. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021. URL https://arxiv.org/abs/2104.13478.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Helena Casademunt, Caden Juang, Adam Karvonen, Samuel Marks, Senthooran Rajamanoharan, and Neel Nanda. Steering out-of-distribution generalization with concept ablation fine-tuning, 2025. URL https://arxiv.org/abs/2507.16795.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, Jay McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers, 2022. URL https://arxiv.org/abs/2205.05055.

Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. Text-to-lora: Instant transformer adaption, 2025. URL https://arxiv.org/abs/2506.06105.

Tong Chen, Hao Fang, Patrick Xia, Xiaodong Liu, Benjamin Van Durme, Luke Zettlemoyer, Jianfeng Gao, and Hao Cheng. Generative adapter: Contextualizing language models in parameters with a single forward pass, 2024. URL https://arxiv.org/abs/2411.05877.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks, 2016. URL https://arxiv.org/abs/1602.07576.

Róbert Csordás, Christopher Potts, Christopher D Manning, and Atticus Geiger. Recurrent neural networks learn to store and generate sequences using non-linear representations. *arXiv preprint arXiv:2408.10920*, 2024.

Can Demircan, Tankred Saanum, Akshay K. Jagadish, Marcel Binz, and Eric Schulz. Sparse autoencoders reveal temporal difference learning in large language models, 2024. URL https://arxiv.org/abs/2410.01280.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Maintaining plasticity in deep continual learning, 2024. URL https://arxiv.org/abs/2306.13812.

Joshua Engels, Isaac Liao, Eric J. Michaud, Wes Gurnee, and Max Tegmark. Not all language model features are linear, 2024. URL https://arxiv.org/abs/2405.14860.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations, 2025. URL https://arxiv.org/abs/2506.08008.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

Xuyang Ge, Wentao Shu, Jiaxing Wu, Yunhua Zhou, Zhengfu He, and Xipeng Qiu. Evolution of concepts in language model pre-training, 2025. URL https://arxiv.org/abs/2509.17196.

Pulkit Gopalani and Wei Hu. What happens during the loss plateau? understanding abrupt learning in transformers, 2025. URL https://arxiv.org/abs/2506.13688.

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *In Proc. Int. Conf. on Learning Representations (ICLR)*, 2017.

Sai Sumedh R. Hindupur, Ekdeep Singh Lubana, Thomas Fel, and Demba Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry, 2025. URL https://arxiv.org/abs/2503.01822.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

David T. Hoffmann, Simon Schrodi, Jelena Bratuli, Nadine Behrmann, Volker Fischer, and Thomas Brox. Eureka-moments in transformers: Multi-step tasks reveal softmax induced optimization problems, 2024. URL https://arxiv.org/abs/2310.12956.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. URL https://arxiv.org/abs/2405.07987.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic, 2023. URL https://arxiv.org/abs/2212.04089.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786*, 2023.

Jaeyeon Kim, Sehyun Kwon, Joo Young Choi, Jongho Park, Jaewoong Cho, Jason D. Lee, and Ernest K. Ryu. Task diversity shortens the icl plateau, 2025. URL https://arxiv.org/abs/2410.05448.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *Proc. of the 36th Proc. Int. Conf. on Machine Learning (ICML)*, Proc. of Machine Learning Research. PMLR, 09–15 Jun 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Akarsh Kumar, Jeff Clune, Joel Lehman, and Kenneth O. Stanley. Questioning representational optimism in deep learning: The fractured entangled representation hypothesis, 2025. URL https://arxiv.org/abs/2505.11581.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution, 2022. URL https://arxiv.org/abs/2202.10054.

Andrew K. Lampinen, Arslan Chaudhry, Stephanie C. Y. Chan, Cody Wild, Diane Wan, Alex Ku, Jrg Bornschein, Razvan Pascanu, Murray Shanahan, and James L. McClelland. On the generalization of language models from in-context learning and finetuning: a controlled study, 2025. URL https://arxiv.org/abs/2505.00661.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. In *Forty-first International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2401.01967.

Andrew Lee, Lihao Sun, Chris Wendler, Fernanda Vigas, and Martin Wattenberg. The geometry of self-verification in a task-specific reasoning model, 2025. URL https://arxiv.org/abs/2504.14379.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021. URL https://arxiv.org/abs/2104.08691.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2022.

Melody Zixuan Li, Kumar Krishna Agrawal, Arna Ghosh, Komal Kumar Teru, Guillaume Lajoie, and Blake Aaron Richards. Tracing the representation geometry of language models from pretraining to post-training. In *High-dimensional Learning Dynamics 2025*, 2025. URL https://openreview.net/forum?id=9nKmDLXg9v.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

Ekdeep Singh Lubana, Can Rager, Sai Sumedh R. Hindupur, Valerie Costa, Greta Tuckute, Oam Patel, Sonia Krishna Murthy, Thomas Fel, Daniel Wurgaft, Eric J. Bigelow, Johnny Lin, Demba Ba, Martin Wattenberg, Fernanda Viegas, Melanie Weber, and Aaron Mueller. Priors in time: Missing inductive biases for language model interpretability, 2025. URL https://arxiv.org/abs/2511.01836.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes, 2024. URL https://arxiv.org/abs/2305.17333.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

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.

Abhinav Menon, Manish Shrivastava, David Krueger, and Ekdeep Singh Lubana. Analyzing (in)abilities of saes via formal languages, 2025. URL https://arxiv.org/abs/2410.11767.

Eric J Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *arXiv preprint arXiv:2303.13506*, 2023.

Julian Minder, Clment Dumas, Caden Juang, Bilal Chugtai, and Neel Nanda. Overcoming sparsity artifacts in crosscoders to interpret chat-tuning, 2025. URL https://arxiv.org/abs/2504.02922.

Andrei Mircea, Supriyo Chakraborty, Nima Chitsazan, Milind Naphade, Sambit Sahu, Irina Rish, and Ekaterina Lobacheva. Training dynamics underlying language model scaling laws: Loss deceleration and zero-sum learning, 2025. URL https://arxiv.org/abs/2506.05447.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023a. URL https://arxiv.org/abs/2301.05217.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, 2023b. URL https://arxiv.org/abs/2309.00941.

Kento Nishi, Maya Okawa, Rahul Ramesh, Mikail Khona, Ekdeep Singh Lubana, and Hidenori Tanaka. Representation shattering in transformers: A synthetic study with knowledge editing. *arXiv preprint arXiv:2410.17194*, 2024.

Maya Okawa, Ekdeep Singh Lubana, Robert P. Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task, 2024.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

OpenDataSoft / GeoNames. Geonames all cities with a population ¿ 1000. https://public.opendatasoft.com/explore/dataset/geonames-all-cities-with-a-population-1000, 2025. Accessed: 2025.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. Iclr: In-context learning of representations, 2024a. URL https://arxiv.org/abs/2501.00070.

Core Francisco Park, Ekdeep Singh Lubana, Itamar Pres, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. *arXiv preprint arXiv:2412.01003*, 2024b.

Core Francisco Park, Maya Okawa, Andrew Lee, Ekdeep Singh Lubana, and Hidenori Tanaka. Emergence of hidden capabilities: Exploring learning dynamics in concept space, 2024c. URL https://arxiv.org/abs/2406.19370.

Core Francisco Park, Zechen Zhang, and Hidenori Tanaka. *New News*: System-2 fine-tuning for robust integration of new knowledge, 2025. URL https://arxiv.org/abs/2505.01812.

Michael Pearce, Elana Simon, Michael Byun, and Daniel Balsam. Finding the tree of life in evo 2. *Goodfire Research*, August 2025. Correspondence to michael@goodfire.ai.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Adv. in Neural Information Processing Systems (NeurIPS)*, 2021.

Tian Qin, Core Francisco Park, Mujin Kwun, Aaron Walsman, Eran Malach, Nikhil Anand, Hidenori Tanaka, and David Alvarez-Melis. Decomposing elements of problem solving: What "math" does rl teach?, 2025. URL https://arxiv.org/abs/2505.22756.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training, 2018.

Allan Ravents, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression, 2023. URL https://arxiv.org/abs/2306.15063.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task, 2023. URL https://arxiv.org/abs/2312.03002.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Imanol Schlag, Kazuki Irie, and Jrgen Schmidhuber. Linear transformers are secretly fast weight programmers, 2021. URL https://arxiv.org/abs/2102.11174.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks, 2020. URL https://arxiv.org/abs/2006.07710.

Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream, 2025. URL https://arxiv.org/abs/2405.15943.

Aaditya K. Singh, Ted Moskovitz, Felix Hill, Stephanie C. Y. Chan, and Andrew M. Saxe. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation, 2024. URL https://arxiv.org/abs/2404.07129.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Johannes Treutlein, Dami Choi, Jan Betley, Samuel Marks, Cem Anil, Roger Grosse, and Owain Evans. Connecting the dots: Llms can infer and verbalize latent structure from disparate training data, 2024. URL https://arxiv.org/abs/2406.14546.

Keyon Vafa, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? using inductive bias to probe for world models, 2025. URL https://arxiv.org/abs/2507.06952.

Atticus Wang, Joshua Engels, Oliver Clive-Griffin, Senthooran Rajamanoharan, and Neel Nanda. Simple mechanistic explanations for out-of-context reasoning, 2025. URL https://arxiv.org/abs/2507.08218.

Jake Ward, Chuqiao Lin, Constantin Venhoff, and Neel Nanda. Reasoning-finetuning repurposes latent representations in base models, 2025. URL https://arxiv.org/abs/2507.12638.

Maurice Weiler and Gabriele Cesa. General $e(2)$-equivariant steerable cnns, 2021. URL https://arxiv.org/abs/1911.08251.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Reft: Representation finetuning for language models, 2024. URL https://arxiv.org/abs/2404.03592.

Daniel Wurgaft, Ekdeep Singh Lubana, Core Francisco Park, Hidenori Tanaka, Gautam Reddy, and Noah D. Goodman. In-context learning strategies emerge rationally, 2025. URL https://arxiv.org/abs/2506.17859.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule, 2025. URL https://arxiv.org/abs/2412.06464.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. URL https://arxiv.org/abs/2504.13837.

Shiyang Zhang, Aakash Patel, Syed A Rizvi, Nianchen Liu, Sizhuang He, Amin Karbasi, Emanuele Zappala, and David van Dijk. Intelligence at the edge of chaos, 2025. URL https://arxiv.org/abs/2410.02536.

Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. Echo chamber: Rl post-training amplifies behaviors learned in pretraining, 2025. URL https://arxiv.org/abs/2504.07912.

Adam Zweiger, Jyothish Pari, Han Guo, Ekin Akyrek, Yoon Kim, and Pulkit Agrawal. Self-adapting language models, 2025. URL https://arxiv.org/abs/2506.10943.

APPENDIX

A 3D VISUALIZATIONS

3D visualizations are available here (Open Science Framework anonymized link).

B EXPERIMENTAL DETAILS

This section provides detailed information about the world, data generation process, model architecture, and training procedures used in our experiments.

B.1 WORLD

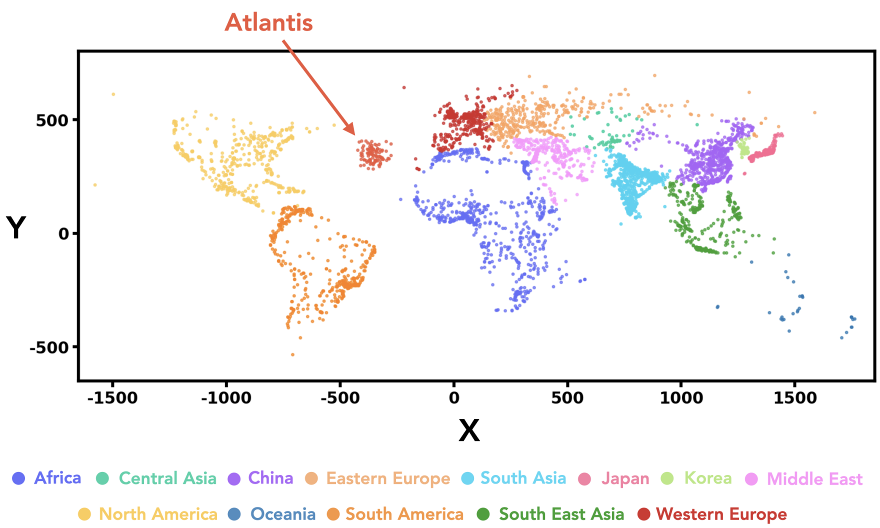

Figure 7: **Geographic distribution of cities used in our experiments.** 5,075 real-world cities plus 100 synthetic `Atlantis` cities (5,175 total). Cities span all continents and provide a fixed, measurable world structure. Coordinates use an equirectangular projection: $x = 10 \times$ longitude, $y = 10 \times$ latitude (in degrees). The `Atlantis` region (Atlantic Ocean) is used for out-of-distribution testing.

Our experiments use a geographic world consisting of 5,075 cities extracted from the GeoNames (OpenDataSoft / GeoNames, 2025) database with population greater than 100,000. Cities are distributed across all continents. This choice provides natural variation in density (e.g., dense regions like India versus sparse Oceania) that creates interesting computational challenges.

While we use real city coordinates, this work studies abstract geometric reasoning rather than actual geographywe project coordinates to Euclidean space using an equirectangular projection (as described above) and treat all tasks as pure geometry problems.

We deliberately chose a flat 2D manifold rather than a spherical globe. Our early experiments used spherical coordinates, but we realized that regardless of the external world's geometry, the model must construct its own internal representation starting from random entity distributions. Given the model's nonlinearity, there is no fundamental reason why any particular geometry (planar, spherical, etc.) would be canonical. Our choice of planar geometry enables clean linear probing to read out world representations, whereas extracting nonlinear manifold structure remains an open challenge (Engels et al., 2024; Csordás et al., 2024). While geometric deep learning (Bronstein et al., 2021) studies the interaction between data geometry and model computation, our focus is on general sequence modeling rather than geometry-aware architectures.

Additionally, we introduce 100 synthetic `Atlantis` cities positioned in the Atlantic Ocean, centered at (longitude $-35°$, latitude $35°$) and following a Gaussian distribution with standard deviation of $3°$. These synthetic cities enable controlled out-of-distribution experiments, as models never observe `Atlantis` during pretraining but must generalize to it during evaluation. City IDs are randomly assigned from the range [0, 9999], creating a sparse identifier space that models must learn to map to coordinates. All coordinates are stored as integers (after the $\times 10$ scaling), eliminating floating-point precision issues.

B.2 Data Generation Process

Tasks We implement 7 geometric tasks that operate on city coordinates. All tasks use a consistent format: `task(arguments)=answer`, where city IDs are prefixed with `c_`. Numerical outputs (distance, area, angle, perimeter) are rounded to integers. Table 1 summarizes the tasks.

Task	Input	Output Type	Unit/Values	Example
`distance`	2 cities	Numerical	Scaled coords	`dist(c_865,c_4879)=769`
`triarea`	3 cities	Numerical	Scaled coords^2	`triarea(c_1234,c_5678,c_9012)=45823`
`angle`	3 cities	Numerical	Degrees (0–180)	`angle(c_2345,c_6789,c_123)=97`
`compass`	2 cities	Categorical	8 directions	`compass(c_1234,c_5678)=NE`
`inside`	1 + n cities	Categorical	TRUE/FALSE	`inside(c_9012;c_3456,...)=FALSE`
`perimeter`	n cities	Numerical	Scaled coords	`perimeter(c_4567,c_8901,...)=2856`
`crossing`	4 cities	Categorical	TRUE/FALSE	`cross(c_2345,c_6789;c_123,c_4567)=TRUE`

Table 1: Summary of 7 geometric tasks. Numerical outputs are integers; "scaled coords" refers to the $\times 10$ coordinate system (Sec. B.1). Categorical tasks have discrete outputs: `compass` uses 8 cardinal directions (N, NE, E, SE, S, SW, W, NW), while `inside` and `crossing` are binary. The `inside` task tests if the first city lies within the convex hull of the remaining cities; `crossing` tests if line segment (c_1, c_2) intersects segment (c_3, c_4).

It is important to note that for all tasks we study, queries that don't explicitly involve `Atlantis` cities maintain identical outputs after `Atlantis` is introducedensuring we can cleanly measure integration of new knowledge. While our framework could be extended to study tasks where existing answers change (e.g., counting cities within a radius would yield different results after adding `Atlantis`), enabling investigation of phenomena like the reversal curse (Berglund et al., 2024), we focus here on the simpler case of integrating new entities while preserving existing knowledge.

Dataset Sizes Each pretraining set consists of 1M rows of data per task. For fine-tuning, the dataset consists of: (1) 100k rows of the target task containing at least one `Atlantis` city, (2) 20k rows randomly sampled from the original pretraining data to prevent catastrophic forgetting, and (3) 256 rows per task (without `Atlantis`) to elicit multi-task performance. For the baseline experiment where `Atlantis` is included during pretraining (green line in Fig. 6d), we use 1M rows per task but sample cities uniformly without treating `Atlantis` specially.

B.3 Model and Training

Tokenization We use character-level tokenization with 98 ASCII tokens (excluding space, which serves as the delimiter), plus special tokens for beginning-of-sequence (BOS), end-of-sequence (EOS), and padding (PAD). Each task query and answer is tokenized character-by-character (e.g., `dist(c_0865,c_4879)=769` becomes `d i s t ( c _ 0 8 6 5 , c _ 4 8 7 9 ) = 7 6 9`).

This character-level scheme is intentional. While assigning each city and task a dedicated token would simplify learning, such synthetic-friendly tokenization does not reflect how real language models operate. LLMs must handle multi-token entities, variable-length prompts (our task prefixes have different lengths), computations at different sequence positions, and irregularly tokenized content (e.g., numbers in LaTeX). Preliminary experiments exploring pitfalls of next-token prediction (Bachmann & Nagarajan, 2025) showed that tokenization details qualitatively affect results. We therefore chose character-level tokenization to better approximate realistic sequence modeling conditions.

City ID Assignment City IDs are randomly assigned from the range $[0, 9999]$, ensuring no geographic information leaks through the identifier. This random assignment means the model cannot exploit ID patterns to infer coordinates.

Architecture We use the Qwen2 (Yang et al., 2024) decoder-only transformer architecture with hidden size 128, 4 attention heads, and 6 layers.

Pretraining We train models autoregressively on the full sequence (no prompt masking). While we observed training speedup when masking loss computation on the prompt side, we deliberately avoid this optimization to maintain similarity with standard autoregressive language model pretraining. All pretraining runs see 42M rows regardless of dataset size (e.g., 42 epochs for 1M rows, 6 epochs for 7M rows). Table 2 summarizes the hyperparameters.

Hyperparameter	**Value**
Optimizer	AdamW (Loshchilov & Hutter, 2019)
Learning rate	3×10^{-4}
Weight decay	0.01
Scheduler	Linear with warmup
Warmup steps	50
Batch size	128
Max sequence length	256
Total training rows	42M
Initialization scale	0.1 (std)

Table 2: **Pretraining hyperparameters.**

Fine-Tuning Fine-tuning starts from the final pretrained checkpoint. We use a reduced learning rate of 1×10^{-5} ($30\times$ smaller than pretraining) to avoid catastrophic forgetting. The fine-tuning dataset consists of 100k rows per task containing at least one `Atlantis` city. We train for 30 epochs with batch size 128. We observed significant degradation in performance for both the fine-tuned task and original (non-`Atlantis`) tasks when using a larger batch size of 512. All other hyperparameters (optimizer, weight decay, scheduler, warmup) remain the same as pretraining.

C Analysis Methods

C.1 Evaluation

Generation Protocol For evaluation, we use teacher forcing up to the "=" sign (the prompt), then generate autoregressively at temperature zero until reaching the EOS token or a maximum of 128 tokens (sufficient for all tasks). All trained models achieve perfect parse accuracyoutputs always match the expected format (integers for numerical tasks, valid categories for categorical tasks).

Task-Specific Metrics Categorical tasks (`compass`, `inside`, `crossing`) are evaluated using accuracy. Numerical tasks are evaluated using absolute error: `distance` (scaled coordinate units), `triarea` (scaled coordinate units^2), `angle` (degrees), and `perimeter` (scaled coordinate units).

Normalized Improvement To compare generalization across tasks with different metrics and scales, we define a normalized improvement score that maps performance to $[0, 1]$, where 0 indicates no improvement over the `Atlantis` baseline (before fine-tuning) and 1 indicates matching the pretrained model's performance on standard cities.

For **error-based tasks** (`distance`, `triarea`, `angle`, `perimeter`), where lower is better:

$$\mathrm{NI} = \frac{\log(\text{baseline}_{\text{atlantis}}/\text{error})}{\log(\text{baseline}_{\text{atlantis}}/\text{baseline}_{\text{standard}})} \tag{2}$$

The logarithmic scaling ensures multiplicative improvements are treated equally (e.g., reducing error from 1000 to 100 is weighted the same as 100 to 10).

For **accuracy-based tasks** (`compass`, `inside`, `crossing`), where higher is better:

$$\text{NI} = \frac{\text{accuracy} - \text{baseline}_{\text{atlantis}}}{\text{baseline}_{\text{standard}} - \text{baseline}_{\text{atlantis}}} \tag{3}$$

Note that normalized improvement can slightly exceed 1.0 if, by chance, `Atlantis` cities perform better than the average pretrained city on some task.

C.2 Representation Extraction

We extract representations from the residual stream after transformer blocks, specifically at layers 3, 4, 5, and 6 of our 6-layer model. Unless otherwise specified, all representation analyses in this paper use layer 5 representations.

To extract city representations, we pass a task prefix followed by a city ID through the model. For single-task models, we use the corresponding task prefix. For multi-task models (2-task and 3-task), we use the first task in the combination as the prefix. We verified that the choice of task prefix has negligible effect on the extracted city representations.

For a city with ID 1234, the input sequence is:

```
<bos> d i s t ( c _ 1 2 3 4 ,
```

We extract and concatenate the representations of two tokens: (1) the last digit of the city ID and (2) the following delimiter token (typically a comma). This yields a 256-dimensional representation (128×2) per city, which we use for both PCA visualization and linear probing.

Omitting cities with leading zeros We omit cities with IDs starting with 0, 00, or 000 from representation analyses. These cities form distinct clusters in representation space, separate from cities with IDs starting with non-zero digits. We hypothesize this occurs because the digit 0 has special semantic status: in numerical outputs (distances, angles, areas), leading zeros never appear (e.g., "`=769`" not "`=0769`"), so the model learns to treat 0 differently when it appears as a leading digit. When 0 appears at the start of a city ID, the model may encode a feature indicating "this is an identifier, not a number," causing these cities to cluster separately. To ensure consistent evaluation across all cities, we exclude IDs matching the pattern `^[0][0-9]*$` (i.e., any ID starting with zero).

C.3 Linear Probing & PCA

We use the representations described in Sec. C.2 for both PCA visualization and linear probing.

Linear Probing We train linear probes to predict city coordinates (x, y) from the 256-dimensional representations. We use a train/test split of 3250/1250 cities, training separate probes for x and y coordinates via ordinary least squares (OLS) without regularization. We report R^2 scores and mean absolute error in scaled coordinate units.

PCA For visualization, we apply PCA to the representations and plot the first two or three principal components. We use consistent color coding based on geographic region to enable visual comparison across models and seeds.

Reconstruction Error To quantify how well new entities (`Atlantis` cities) are integrated into the learned manifold, we train linear probes exclusively on non-`Atlantis` cities and evaluate reconstruction error on held-out `Atlantis` representations. Reconstruction error is measured as the absolute Euclidean distance between predicted and true coordinates. Large reconstruction errors indicate that new entities are encoded in different subspaces than the original cities.

C.4 Centered Kernel Alignment

We use Centered Kernel Alignment (CKA) (Kornblith et al., 2019) to measure representational similarity between models. Given two representation matrices $X \in \mathbb{R}^{n \times d_1}$ and $Y \in \mathbb{R}^{n \times d_2}$ (same

n cities, potentially different dimensions), we compute linear kernel matrices $K = XX^T$ and $L = YY^T$, center them, and compute:

$$\text{CKA}(X, Y) = \frac{\langle K, L \rangle_F}{\|K\|_F \|L\|_F} \tag{4}$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product. CKA yields a similarity score in $[0, 1]$ that is invariant to orthogonal transformations and isotropic scaling.

For each pair of models, we extract city representations (Sec. C.2) and compute CKA between the resulting matrices. We filter cities to exclude `Atlantis` and IDs starting with zeros. We report CKA values at layers 3, 4, 5, and 6, with layer 5 as the default unless otherwise specified.

D Additional Experiments & Results

D.1 Training Dynamics

Fig. 8 shows training dynamics for all seven single-task models. Each panel displays three rows of metrics over gradient steps: (top) training and validation loss, (middle) task-specific performance metric alongside linear probe R^2 for coordinate decoding, and (bottom) linear probing distance error measuring how accurately city coordinates can be reconstructed from representations.

Several patterns emerge across tasks. First, all tasks except `crossing` eventually achieve high coordinate R^2 (red curves reaching $\sim$1.0), indicating that world representations form reliably across diverse geometric objectives. Second, the relationship between loss, task performance, and coordinate decodability varies across tasks. Third, `crossing` (panel g) fails entirely in single-task training. Loss remains high, accuracy stays near chance, and coordinate R^2 never rises, consistent with the main text observation that this task requires multi-task scaffolding.

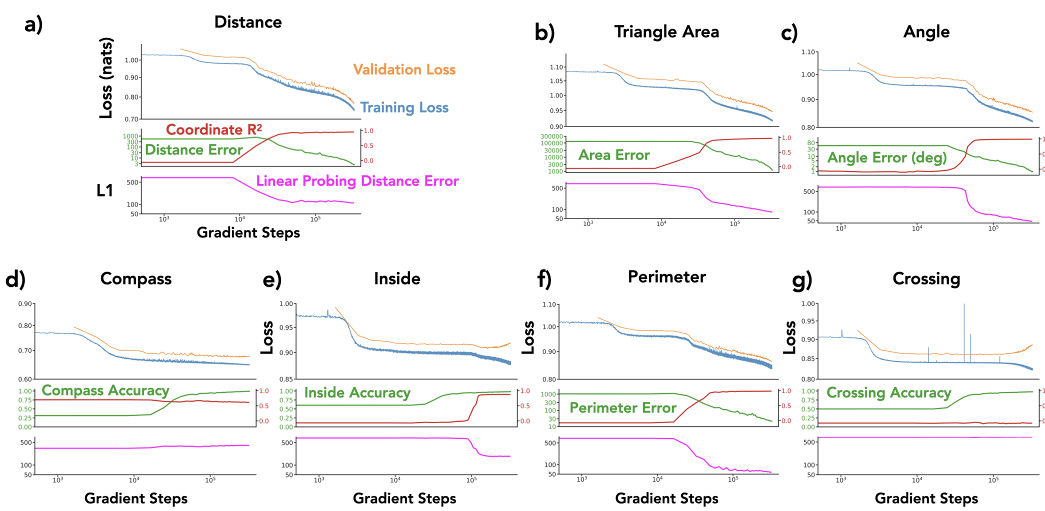

Figure 8: **Training dynamics for all single-task models.** (a) `distance`, (b) `trianglearea`, (c) `angle`, (d) `compass`, (e) `inside`, (f) `perimeter`, (g) `crossing`. Each panel shows three rows: (top) training loss (blue) and validation loss (orange); (middle) task-specific metric (green, left axis) and linear probe coordinate R^2 (red, right axis); (bottom) linear probing distance error (magenta). All plots use log-scale x-axis for gradient steps.

Representation Dynamics. Fig. 9 visualizes how internal representations evolve during training via PCA projections at six checkpoints. A striking pattern emerges: once a representational structure forms, it remains largely fixed throughout the subsequent training phase where task accuracy continues to improve. Examining the gradient steps, representations are essentially fixed in the first $\sim$15% of training, remaining static while loss slowly decreases and accuracy rises. The `distance` task (top row) establishes its thread-like structure early; `angle` (middle row) settles into a 2D manifold;

`compass` (bottom row) forms fragmented regional clusters, all within the first few checkpoints, with minimal subsequent change. What determines when representations stop evolving remains unclear, though it appears correlated with the initial loss drop. This may relate to recently observed gradient dynamics in language model training, where loss deceleration phases exhibit qualitatively different learning behavior (Mircea et al., 2025).

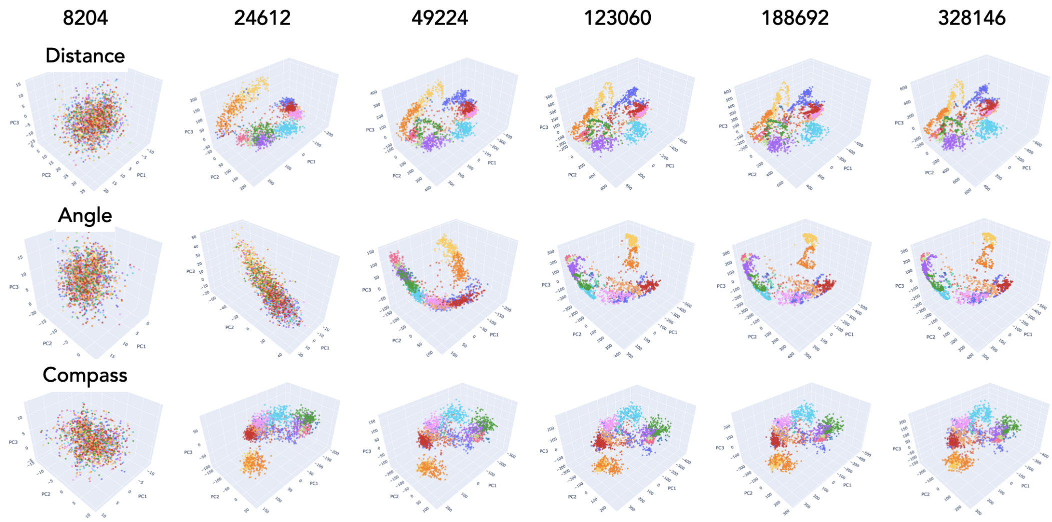

Figure 9: **Representation dynamics during training.** Rows: `distance` (top), `angle` (middle), `compass` (bottom). Columns show PCA projections at gradient steps 8204, 24612, 49224, 123060, 188692, and 328146 (left to right). Cities are colored by geographic region.

D.2 Qualitative Representations

Fig. 10 shows PCA projections of city representations for single-task models across three random seeds (rows). The `distance` task consistently produces characteristic thread-like structures. `Angle` and `perimeter` often form larger 2D manifold-like structures. `triangle area` tends to produce arc-shaped geometries. `Compass` forms local clusters corresponding to directional categories, while `inside` produces a more global, diffuse structure.

While there is some seed-to-seed variability within each task, the broader categories remain distinguishable: `distance` representations are qualitatively distinct from the cluster-based representations of `compass` and `inside`, and both differ from the manifold-like structures produced by `triangle area`, `angle`, and `perimeter`.

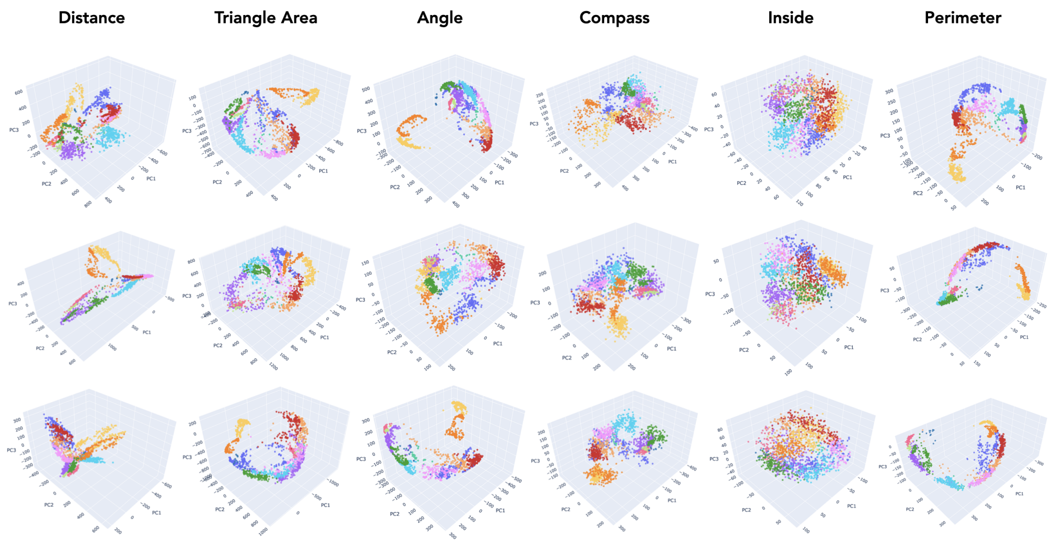

Figure 10: **Representation visualizations for single-task models across multiple seeds.** Each column shows a different task; each row shows a different random seed. Cities are colored by geographic region. Despite seed variability, task-specific geometric patterns are visible.

D.3 Additional CKA Results

Single-Task CKA Across Layers. Fig. 11 shows CKA matrices for single-task models at layers 3, 4, 5, and 6. Each cell shows mean ± SEM across 3 seeds. We observe: (1) CKA values increase from layer 3 to layers 4–6, indicating that world representations become more consistent in later layers; (2) the `distance` task (D) shows lower CKA with other tasks across all layers, consistent with its divergent representational geometry; (3) `crossing` (Cr) shows near-zero CKA due to training failure in single-task settings; (4) diagonal entries (same task) can show significant variability, indicating that even identical training objectives can yield different representational solutions.

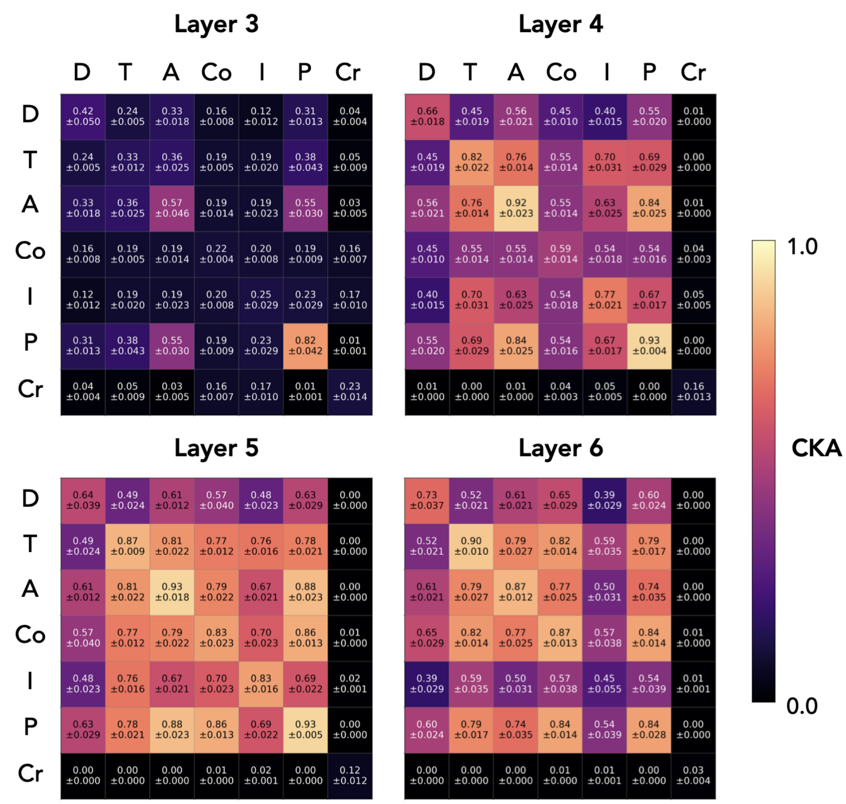

Figure 11: **CKA matrices for single-task models across layers.** Each cell shows mean ± SEM across 3 seeds. D=distance, T=triangle area, A=angle, Co=compass, I=inside, P=perimeter, Cr=crossing. CKA increases in later layers; `distance` shows consistently lower cross-task similarity.

Two-Task CKA. Fig. 12 shows the CKA matrix for two-task models at layer 5. Compared to single-task models (Fig. 11, layer 5), two-task training substantially increases representational alignment: all off-diagonal entries exceed 0.84, compared to values as low as 0.48 for single-task models. Notably, diagonal entries (same task combination, different seeds) show minimum CKA of 0.89, indicating that multi-task training also reduces inter-seed variance. For diagonal entries, we exclude same-seed comparisons (which trivially yield 1.0) and report only the upper triangle since the matrix is symmetric. This confirms the main text finding that multi-task training drives representational convergence.

CKA vs. Task Count (Per-Seed). Fig. 13 shows the same CKA vs. task count analysis as Fig. 3(d) in the main text, but broken down by individual seeds. Each panel shows one seed. These per-seed values are pooled to produce the main text figure, where error bars represent SEM across seeds. The pattern is consistent across all three seeds: CKA increases substantially from 1 to 2 tasks and saturates at 2–3 tasks for layers 4–6.

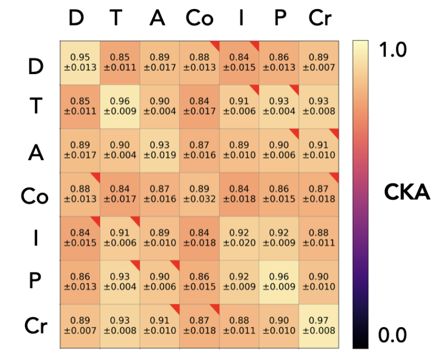

Figure 12: **CKA matrix for two-task models at layer 5.** Mean $\pm$ SEM across 3 seeds. All pairs show high alignment (>0.84), substantially higher than single-task models.

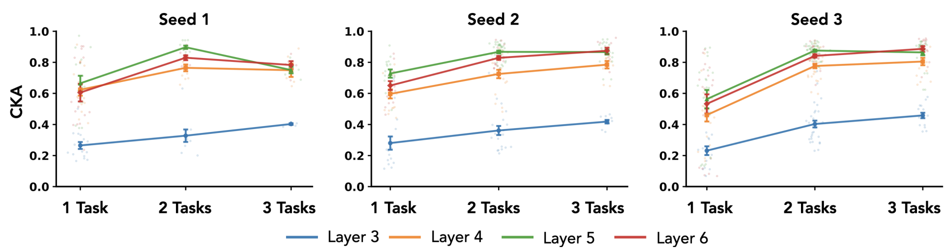

Figure 13: **CKA vs. task count for individual seeds.** Each panel shows a different seed. These values are pooled in Fig. 3(d); error bars there represent SEM across seeds.

Aggregated CKA Trends. Fig. 14(a) shows CKA vs. task count for a single seed, using all $\binom{7}{2} = 21$ two-task models and all $\binom{7}{3} = 35$ three-task models, but only comparing non-overlapping pairs (models sharing no common tasks). This yields 105 non-overlapping pairs for 2-task models and 70 for 3-task models. Fig. 14(b) shows within-task CKA (same task combination, different seeds) as a function of task count, demonstrating that multi-task training also reduces seed-to-seed variability: representations become more consistent not just across tasks but also across random initializations.

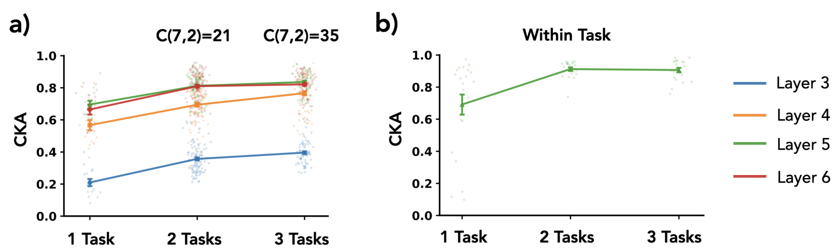

Figure 14: **Aggregated CKA analysis.** (a) CKA vs. task count for single seed, comparing only non-overlapping model pairs (105 pairs for 2-task, 70 pairs for 3-task). (b) Within-task CKA (same task combination, different seeds) increases with task count, indicating multi-task training reduces seed variability.

CKA vs. Generalization (Annotated). Fig. 15 is an annotated version of Fig. 5(b), with each point labeled by its (train→eval) task pair.

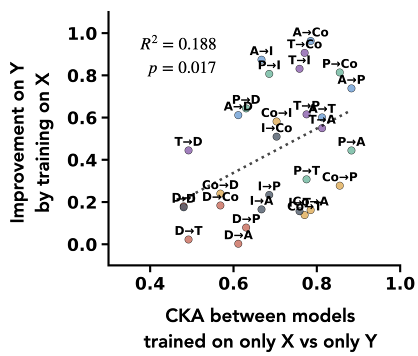

Figure 15: **Annotated version of Fig. 5(b).** Each point is labeled with its (train→eval) task pair. D=distance, T=triangle area, A=angle, Co=compass, I=inside, P=perimeter.

D.4 Additional Fine-Tuning Evaluation Results

Raw fine-tuning results for individual seeds.

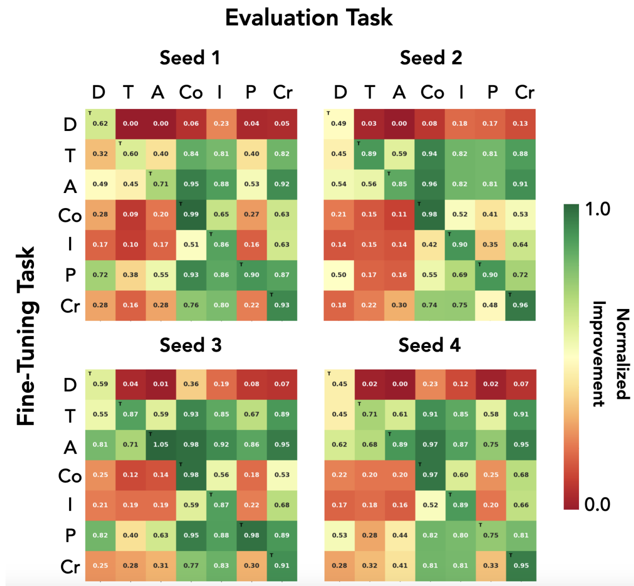

Figure 16: **Single-task fine-tuning results for individual seeds.** Per-seed version of Fig. 5(a), organized in a 2×2 grid.

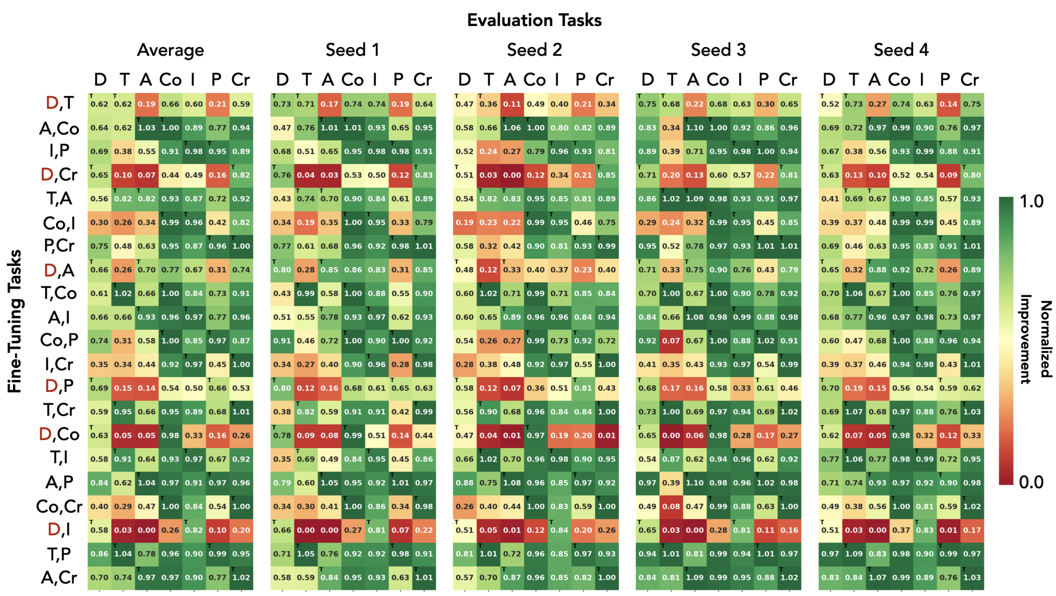

Figure 17: **Two-task fine-tuning normalized improvement for all 21 task combinations.** Left-most panel shows average across seeds; remaining panels show individual seeds.

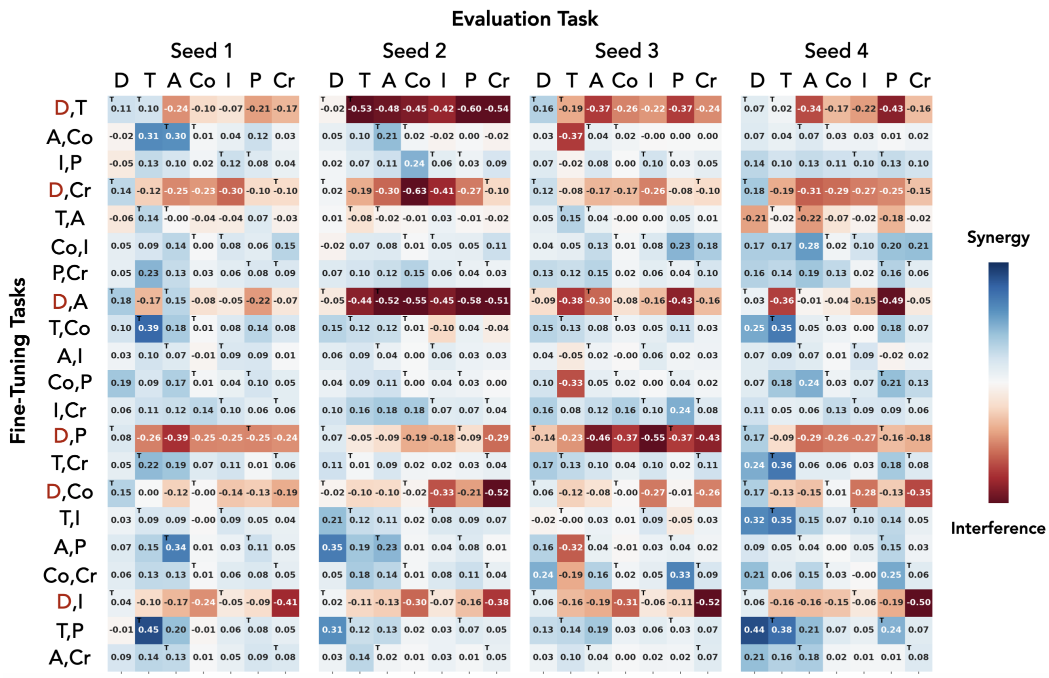

Figure 18: **Deviation from best-teacher expectation for all 21 two-task combinations.** All 4 seeds shown; average is in main text Fig. 6(c).

D.5 Pretraining Variations

Pretraining with `Atlantis`. In the main text, we showed that fine-tuning on divergent tasks fails to integrate `Atlantis` cities into the learned representation manifold (Fig. 6d, red histogram). To verify that this failure stems from fine-tuning dynamics rather than a peculiarity of the geometry around `Atlantis`, we trained a model with `Atlantis` cities included from the start of pretraining. Fig. 19 shows the resulting representations: `Atlantis` cities are seamlessly integrated into the world manifold, indistinguishable from other cities in both PCA projections (a) and linear probe reconstructions (b). This confirms that the representation space can readily accommodate `Atlantis`, and thus, the integration failure observed in fine-tuning is a property of the optimization dynamics, not a fundamental limitation of the architecture or task.

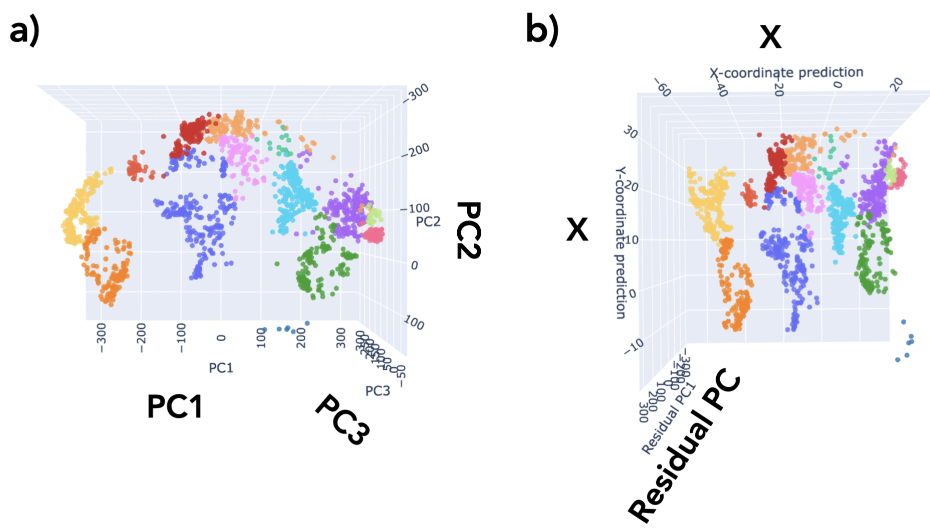

Figure 19: **Representations when `Atlantis` is included during pretraining.** (a) PCA projection showing `Atlantis` cities (small cluster in Atlantic region) integrated with world cities. (b) Linear probe reconstruction confirming geographic accuracy. Unlike fine-tuned models, `Atlantis` cities lie on the same manifold as other cities.

Wider Model. To test whether our findings depend on model capacity, we trained a wider model with 2× the hidden dimension (256 vs. 128) and intermediate size (1024 vs. 512), resulting in approximately 4× the parameters. Fig. 20 shows fine-tuning results for this wider model: (a) single-task fine-tuning normalized improvement; (b) two-task fine-tuning normalized improvement; (c) deviation from best-teacher expectation. We still observe that `distance`-containing combinations (red labels in panel c) show degraded cross-task generalization. This suggests that divergent task interference is not simply a capacity limitation.

E Extended Related Work

See Sec. 2 for main related work.

Interpretability & Internal Representations. Understanding internal representations has roots in neuroscience (Hubel & Wiesel, 1962), informing early neural network development (Fukushima, 1980; Bengio et al., 2014). Beyond the world model discoveries cited in Sec. 2, similar representations emerge during in-context learning (Vafa et al., 2025). Researchers have also uncovered that models represent meaningful properties of data—concepts (Pearce et al., 2025; Higgins et al., 2017), features (Olah et al., 2017), and abstractions (Lee et al., 2025; Arditi et al., 2024)—in interpretable ways. Yet the relationship between representations and training dynamics remains poorly understood. Only recent work has begun examining how representations emerge during pretraining in real LLMs (Li et al., 2025; Ge et al., 2025) or how they change during fine-tuning (Lee et al., 2024).

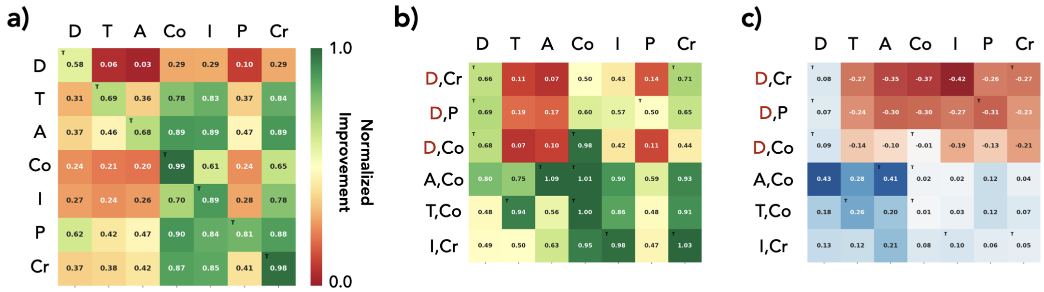

Figure 20: **Fine-tuning results for wider model (2× hidden dimension).** For all panels: rows = fine-tuning task(s), columns = evaluation task. (a) Single-task fine-tuning normalized improvement. (b) Two-task fine-tuning normalized improvement. (c) Deviation from best-teacher expectation; `distance`-containing combinations (red labels) still show degraded generalization.

Fine-tuning. Beyond the works cited in Sec. 2, fine-tuning has been studied extensively across diverse directions: parameter efficiency (Hu et al., 2021; Lester et al., 2021), zeroth-order optimization (Malladi et al., 2024), weight composition (Ilharco et al., 2023), and representation adaptation (Wu et al., 2024). Other poorly understood behaviors include out-of-context reasoning limitations (Treutlein et al., 2024) and off-target effects (Betley et al., 2025). Additional behavioral analyses reinforcing pessimism about fine-tuning include Zhao et al. (2025); Zweiger et al. (2025).

Dynamics of Representations. Recent work has begun studying how representations evolve during in-context learning (Shai et al., 2025; Demircan et al., 2024) or fine-tuning (Casademunt et al., 2025; Minder et al., 2025). Relatedly, Lubana et al. (2025) show that representations exhibit rich temporal dynamics that standard interpretability methods (e.g., SAEs) fail to capture due to stationarity assumptions. Fu et al. (2025) show that VLMs trained by merging LLMs and vision encoders often fail to utilize representations surfaced by the vision encoder, i.e. the representations exist but remain unused.

Geometric Deep Learning. Geometric deep learning studies how data geometry interacts with model architectures, developing equivariant networks that respect symmetries (Bronstein et al., 2021; Cohen & Welling, 2016; Weiler & Cesa, 2021). While our world is defined on a 2D plane, one might ask: why not a sphere, torus, or other manifold? This is an interesting direction, but not our focus. We study how neural networks adapt internal representations to tasks in an arbitrarily chosen geometry. Moreover, a change in world geometry can be absorbed into the task definition (e.g., geodesic vs. Euclidean distance), so the key question remains how representations form given the task, not the underlying manifold. Planar coordinates also allow clean linear probing of world representations. Our models are standard transformers without geometric priors; we study what representations emerge purely from training on task data, treating geometry as emergent rather than imposed.

Loss Plateaus. Our `crossing` task fails to learn in single-task training despite escaping an initial plateau (likely output format learning), suggesting it remains stuck in a deeper plateau. Such plateaus are notoriously difficult for transformers. Recent work has studied this phenomenon mechanistically in transformers (Hoffmann et al., 2024; Gopalani & Hu, 2025; Singh et al., 2024), while others relate it to more general optimization challenges in deep learning such as simplicity bias and gradient starvation (Shah et al., 2020; Pezeshki et al., 2021; Bachmann & Nagarajan, 2025). Most related to our findings, Kim et al. (2025) show that multi-task training shortens loss plateaus, similar to why our `crossing` task trains successfully when joined with any other task.

F Code and Data Availability

Code, data and model checkpoints will be available after the review process.