ORIGINS AND ROLES OF WORLD REPRESENTATIONS IN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

How do neural networks develop internal representations of the world, and how do these representations determine their ability to adapt to new information? We introduce a controlled framework that decouples the *world* (5,175 city coordinates) from *data generation* (seven geometric tasks like distance and angle calculation) to systematically study representation learning. Training small Transformers on these tasks reveals three key findings. First, different tasks create fundamentally different representational geometries despite operating on identical world structuredistance creates thread-like manifolds while angle creates 2D surfaces, with representation quality varying from clean coordinate systems ($R^2 > 0.8$) to entangled features ($R^2 < 0.4$). Second, multi-task training drives representational convergence (CKA increasing from < 0.3 to > 0.7), providing controlled evidence for the Platonic Representation Hypothesis that task diversity constrains viable representations. Third, and most surprisingly, we find that single-task representational properties predict multi-task model fine-tuning behavior: when we introduce 100 synthetic "Atlantis" cities, tasks with divergent single-task representations not only fail to generalize but actively harm performance below single-task baselines. Mechanistically, these divergent tasks encode new entities in orthogonal subspaces rather than integrating them into the shared world manifold. Our work demonstrates that even when models are trained on all tasks jointly, their fine-tuning behavior is fundamentally constrained by how representations would form under single-task trainingsuggesting that robust adaptation requires not just multi-task pretraining but careful consideration of task-induced representational geometry.

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.

Large language models (LLMs) (Radford et al., 2018; Devlin et al., 2018; Brown et al., 2020; OpenAI, 2024), in particular, have shown striking generalization abilities that have sparked intense debate about their underlying mechanisms. While there is some skepticism arguing that these models may simply be sophisticated pattern matchers performing surface-level predictions (Bender et al., 2021; Dziri et al., 2023; Shojaee et al., 2025), other evidence suggests that pretrained transformers develop at least partial signatures of structured world models within their parameters (Li et al., 2022; Gurnee & Tegmark, 2023; Nanda et al., 2023; Vafa et al., 2024). Across diverse areas of deep learning, researchers have uncovered that models represent meaningful properties of dataconcepts (Pearce et al., 2025; Higgins et al., 2017), features (Anthropic AI, 2023; Templeton et al., 2024), and abstractions (Marks & Tegmark, 2024; Lee et al., 2025; Arditi et al., 2024)in surprisingly interpretable ways within their internal representations. These findings suggest that neural networks learn genuine computational circuits for processing real-world concepts, rather than merely memorizing input-output mappings.

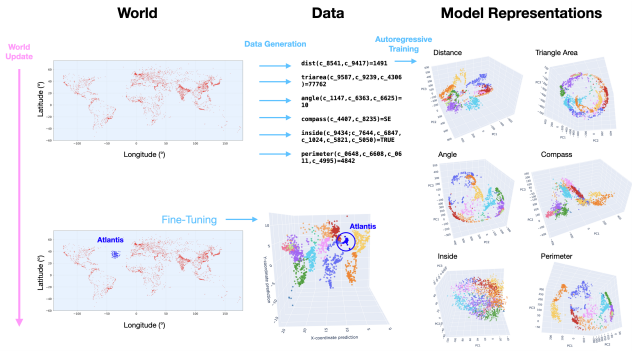

Figure 1: **Overview.** Our World-Data-Model framework decouples three components to study representation learning: (1) a fixed world of 5,175 city coordinates, (2) varied data generation through 7 geometric tasks, and (3) neural network models that learn from task outputs without seeing coordinates. We probe how different tasks shape internal world representations, then test adaptability by introducing 100 synthetic "Atlantis" cities. The framework reveals that task type controls representational geometry, while multi-task training drives convergence toward aligned representations.

However, major open questions remain about the origins and roles of internal representations. We do not yet understand which properties of the world, data, and model architectures give rise to rich representations, nor how specific properties of these representations translate to whole-network behavior. Do representations need to be disentangled to support generalization (Locatello et al., 2019)? Could alternative learning algorithms yield better representational structures (Kumar et al., 2025)? How do representations interact with fine-tuning? Can genuinely new ones be acquired after pretraining? We argue that understanding these fundamental questions about the *origins and roles of neural representations* is essential for the long-term goal, a model with stable and unified representations that can be reliably updated for robust downstream adaptation.

Answering these questions is difficult in real-world training setups, where the key "knobsthe world, the data, and the modelare hard to control. Even the most accessible knob, the model, becomes costly to perturb at scale, especially for LLMs. As a result, the field still lacks a holistic framework for systematically perturbing how these different factors interact. In this work, we turn to a small-scale synthetic setup, where the relevant factors can be precisely controlled and tested.

This work. We decouple the *world* (5,175 city coordinates) from *data generation* (7 geometric tasks) to study how different tasks shape neural representations. Training small Transformers on these tasks reveals that despite operating on the same world, different tasks create vastly different representational geometries—from clean coordinate systems (distance) to entangled features (crossing detection). Multi-task training drives convergence, providing controlled evidence for the Platonic Representation Hypothesis. When we introduce "Atlantis" cities to test adaptation, fine-tuning shows surprising task-dependent effects where certain tasks harm cross-task generalization. Our contributions:

- **A World-Data-Model Framework for Studying Representations.** We propose treating the world and data generation as separate entities, allowing systematic study of how different views of the same world shape neural representations. Our geographic setup with 5,175 cities provides a complex yet measurable world where representation quality can be directly assessed via coordinate probing (Section 2).
- **Task-Dependent Geometry Despite Shared World Structure.** We show that different tasks operating on the same world produce markedly different representational geometries. Distance and area tasks yield clean, interpretable coordinate systems ($R^2 > 0.8$), while classification tasks like crossing detection produce entangled features ($R^2 < 0.4$), even though all tasks theoretically require the same coordinate information (Section 3).
- **Multi-Task Training Drives Representational Convergence.** We provide controlled evidence for the Platonic Representation Hypothesis by showing that training on multiple tasks simultane-

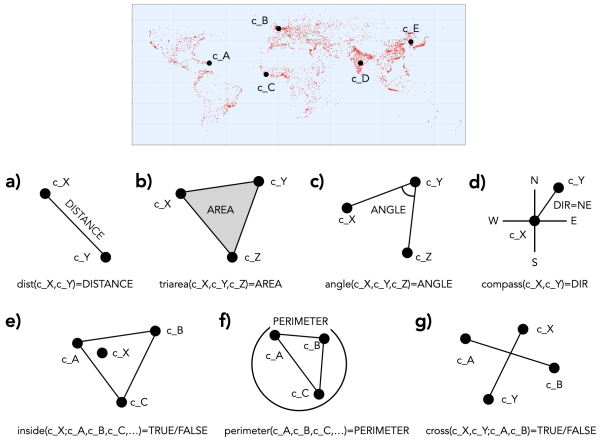

Figure 2: **Experimental Setup: Seven geometric tasks.** All tasks operate on the same 5,175 real city coordinates (map shown above) but require different geometric computations: (a) `Distance`: Euclidean distance, (b) `Triangle Area`: Area from three cities, (c) `Angle`: Angle at middle vertex, (d) `Compass`: 8-way cardinal direction, (e) `Inside`: Point-in-polygon test, (f) `Perimeter`: Polygon perimeter, (g) `Crossing`: Line segment intersection.

ously leads to higher representational alignment (CKA > 0.7) compared to single-task models (CKA < 0.3). This convergence suggests that task diversity constrains the space of viable representations (Section 3).

- **Single-Task Representational Divergence Predicts Multi-Task Model Fine-Tuning Failure.** Despite joint pretraining on all tasks, we show that fine-tuning generalization is predicted by how representations would diverge when tasks are trained in isolation. Through "Atlantis" experiments where we add 100 synthetic cities, we demonstrate that divergent tasks (low single-task CKA) actively harm fine-tuning performance below single-task baselinesnot merely failing to help but causing catastrophic interference. We identify the mechanistic basis: divergent tasks encode new entities in orthogonal subspaces rather than integrating them into the shared world manifold. This reveals two distinct failure modes: *representational segregation* (new entities encoded in orthogonal subspaces that don't propagate updates) versus *elicitation problems* (quickly fixable with few examples) (Section 4).

2 Experimental Framework: Geographic Reasoning with Controllable World Structure

We begin by introducing a framework that enables systematic study of how neural networks form world representations under different data generation processes. Our approach uses geographic tasks where models must solve geometric problems involving city coordinatesa setup that naturally separates the underlying world (city coordinates) from how data is sampled from it (geometric tasks). This setup naturally allows changes to the underlying world (e.g., additional cities) with corresponding consistent updates to all dependent tasks, while providing clear metrics for measuring representation quality through coordinate probing. Specifically, our framework provides three key properties:

1. **Consistency:** All tasks are deterministically generated from the same underlying coordinates, ensuring any change to city locations produces predictable updates across all geometric computations.
2. **Hidden State:** Models never see coordinates directly, only task outputs, yet we can probe whether they internally reconstruct the world structure.
3. **Controllable Updates:** We can systematically modify the world (e.g., adding new cities) and study how models adapt their learned representations to incorporate these changes.

We settled on the setup shown in Fig. 2. We filtered the dataset of world cities by `population`$>100,000$, giving us 5,175 cities distributed as seen in the top of Fig. 2. We then defined 7 geometric functions which take as input 2 or more cities and calculate a geometric value depending on the input. Since the inputs are compositional, the data can be easily scaled to practically infinite samples.

While we use real city coordinates, this work studies abstract geometric reasoning rather than actual geographywe simply project coordinates to Euclidean space $(x, y) = (10 * \text{longitude}, 10 * \text{latitude})$ and treat all tasks as pure geometry problems. This choice provides natural variation in density (e.g., dense regions like India versus sparse Oceania) that creates interesting computational challenges.

Each task query follows a structured format where city IDs (e.g., `c_1234`) serve as inputs to geometric functions, with outputs tokenized character-by-character 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. This character-level tokenization allows models to learn compositional structure while maintaining a small vocabulary (98 ASCII tokens), and the consistent syntax across tasks enables systematic study of how different geometric computations shape internal representations. See App. **??** for further detail.

To test how models adapt to world changes, we also define a modified world with `Atlantis`100 synthetic cities placed near $(\text{lon}, \text{lat}) = (-35, 35)$ in the Atlantic Ocean. While models never observe Atlantis during pretraining, we later use it in Sec. 4 to study whether fine-tuning on one task with Atlantis cities enables models to integrate them into their world representations in a way that generalizes across all tasks. This tests a critical property: can models update their internal world model consistently when the underlying world changes?

Importantly, 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.

3 Task-dependent world representations converge under multi-task learning

We now investigate how neural networks develop internal world representations when trained on different geometric tasks. We train decoder-only Transformers (Vaswani et al., 2023) on individual tasks and task combinations, then probe their internal activations to measure whether they learn the underlying coordinate system or merely task-specific patterns. We find that representation quality depends critically on the task type, with different tasks inducing distinct geometries despite operating on the same world, and multi-task training driving representational convergence (see App. F for training details).

Result 1: World representations emerge through autoregressive training. We first show basic results in Fig. 3, where we trained a model solely on the `angle` task. We visualize training/validation loss, PCA projections, and linearly decoded (x,y) coordinates with dots colored by geographic regions. The model starts with nearly random representations and goes through a long loss plateau until it finally clusters nearby cities together. After this clustering phase, the loss drops more steeply and world representations form accordingly. It is at this moment that prediction error on the angle task drops and the R^2 for coordinate decoding from the residual stream after layer 5 becomes high. Interestingly, the coordinate decoding R^2 starts to rise before we observe sudden improvement in angle accuracy, reminiscent of **?** who found hidden progress measures during periods when task accuracy remains flat. Overall, we find stable formation of internal world representations through pure autoregressive modeling. While the emergence of linearly decodable coordinates might be anticipated given the geometric nature of the task, it provides a useful validation of our

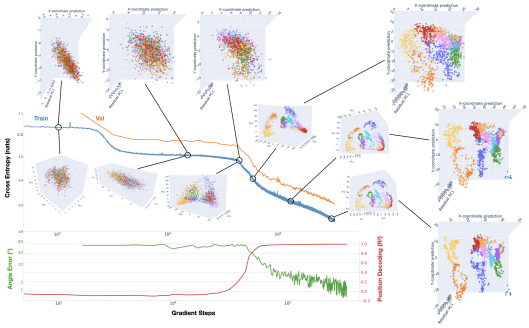

Figure 3: **Emergence of World Representations.** Training dynamics on the `angle` task reveal how world representations form through autoregressive learning. Top panels show training/validation loss and angle prediction accuracy over training steps. Middle panel tracks linear probe R^2 for decoding (x,y) coordinates from layer 5 activations, showing coordinate decodability emerges before task accuracy improves. Bottom panels display PCA projections and linearly probed representations at different training stages, progressing from random initialization through city clustering to final world-aligned geometry. Cities are colored by geographic region throughout.

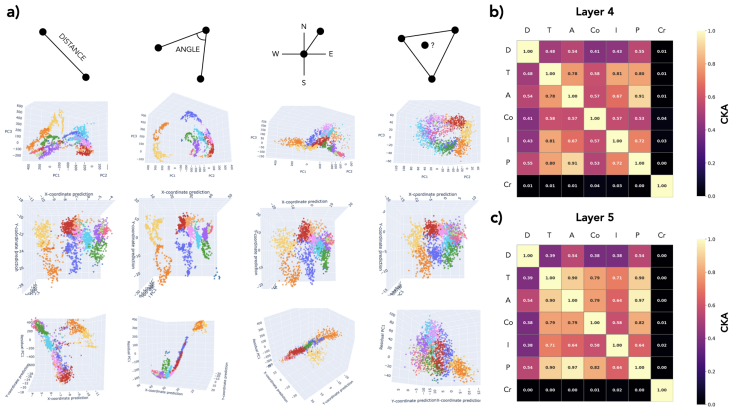

Figure 4: **World representation geometry depends on data generation process.** (a) Different tasks create distinct geometries: `distance` (thread-like), `angle` (2D manifold), `compass` (fragmented), `inside` (diffuse). Row 1: PCA. Row 2: Linear probe projections. Row 3: Rotated views showing hidden structure. (b,c) CKA matrices at layers 4 and 5. `Crossing` (Cr) fails to train alone.

framework and sets the stage for our main findings: how different data generation processes shape these representations in fundamentally different ways.[1]

Result 2: Data generation process controls world representation geometry. Next, as promised, we study how different data generation processes operating on the same underlying world shape different model internal representations. We trained models from scratch for each of the seven tasks shown in Fig. 2. We show four selected tasks and their representations in Fig. 4: PCA projections, linear probe reconstructions, and rotated views. See App. **??** for all results.

We find that depending on the data generation process, models acquire significantly different internal representation geometries. Some tasks form thread-like structures (`distance`), while others form 2D manifold-like structures (`angle`). `Compass` forms less interpretable structures and `inside` forms more diffuse representations. Despite these differences, we can still linearly decode (x,y) coordinates from most tasks, as shown in the second row of Fig. 4. Some tasks (`angle`) form cleaner linearly decodable world representations than others, opening the door to future study of

[1]We do believe linear decodability of world representation is non-trivial (albeit expected). However, this is not the current focus of our study.

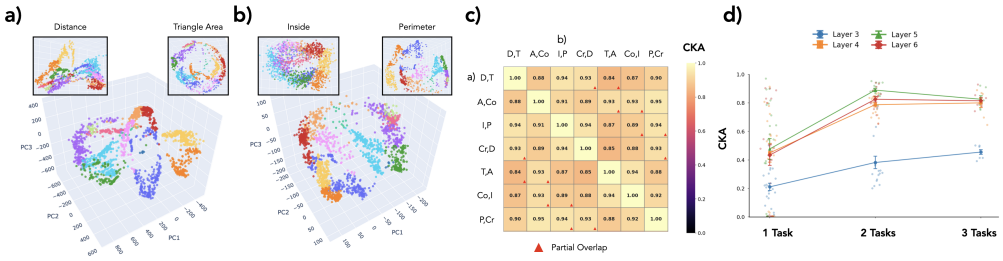

Figure 5: **Multi-task pretraining drives representational convergence.** (a,b) Pairwise task training creates more structured 2D manifolds than single-task models. (c) CKA matrix for all 21 pairwise combinations shows higher alignment. (d) Average CKA increases with task count (123), saturating at 0.8 for layers 4-6 while layer 3 continues improving. 3D visualizations: link.

what drives the formation of linear representations. Note that the third dimension in the PCA plots is the residual direction after projecting out the x,y probe directions. In the last row, we manually rotate the 3D PCA to "flatten" the world map as much as possible. This reveals that in the non-x,y directions, the model representations still look quite different. This reminds us that *linear probing only surfaces what we look for*akin to the parable of the blind men and the elephant, we must be cautious not to mistake our probed coordinates for the complete representational structure.

Not shown in the figure due to space constraints (see App. **??**), the `crossing` task fails to learn at all in these single-task settingsexplaining the row of zeros in the CKA plots. We speculate this 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). Intriguingly, as we will see in Result 3, this same task can be learned successfully when combined with others in multi-task training.

To quantitatively validate our observations, we measure representational similarity between different models' world representations using Centered Kernel Alignment (CKA) (Kornblith et al., 2019). Fig. 4(b,c) shows CKA between all seven models at layers 4 and 5. This reveals that the `distance` task produces significantly different model representationsa result not expected intuitively. Note again that `crossing` (Cr in labels) simply failed to train in isolation.

Result 3: Task diversity aligns representations: Evidence for the Platonic Hypothesis. We now turn to multi-task learning scenarios. While many synthetic studies focus on single tasks or families of related tasks, real-world data more closely resembles joint learning across diverse tasks, as explored in recent work (**??**). Our results speak directly to the recently proposed Platonic Representation Hypothesis (Huh et al., 2024), which observes that neural networks trained on vast amounts of data develop aligned representations even across different modalities and architectures. One potential mechanism they suggest is the Multitask Scaling Hypothesis:

> *"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 an ideal testbed for this hypothesis, with a ground-truth world model and multiple tasks defined over it. We trained models on all pairwise combinations of our 7 tasks. Fig. 5(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 representational structures that better resemble curved 2D manifolds respecting the world map structure. While difficult to appreciate in static 2D projections, we encourage readers to explore our interactive 3D visualizations at this link.

To quantitatively validate these observations, we measure CKA between all pairwise-trained models (Fig. 5(c)). Note that from 7 tasks taken 2 at a time, some models partially share one task. Even excluding models with shared tasks, we find substantially higher CKA compared to single-task models. In Fig. 5(d), we explicitly plot average CKA for models trained on 1, 2, and 3 tasks across layers 3-6. For multi-task settings, we only average over models with completely disjoint task sets

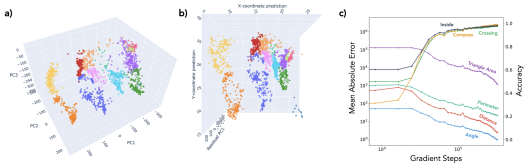

Figure 6: **7-task model.** (a) PCA reveals world structure. (b) Linear probe projections. (c) Training curves for all tasks.

(no overlap). Training on more tasks clearly leads to more aligned representations across networks. Interestingly, CKA appears to saturate around 0.8 for 2 and 3 tasks in layers 4-6, while layer 3 continues improving with more tasks.

Overall, we find that *multi-task learning leads to more aligned model internal representations*. To the best of our knowledge, this is the first experimental evidence for the Multitask Scaling Hypothesis in a controlled setup. 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. 3), 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.

An auxiliary finding, fulfilling the promise from Result 2: the `crossing` task, which was unlearnable alone, now trains successfully when combined with other tasks. We speculate that tasks like `distance` and `perimeter` provide well-structured coordinate representations that `crossing` can then leverage for its geometric computationseffectively creating an 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. This model successfully learns all tasks, and its PCA projection naturally reveals the world map structure, approaching the intuitive quality of linearly probed (x,y) coordinates without requiring any explicit coordinate supervision. This 7-task model serves as the foundation for our fine-tuning experiments in the following section.

4 Fine-tuning's Representational Shift Predicts Downstream Generalization

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 evaluating generalization across entities, we study the inverse: fine-tuning an entity into the model and evaluating generalization across tasks. To this end, we use the 7-task model trained in the previous section (Fig. 6).

As mentioned in Sec. 2, 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.

Result 1: Pretraining Phase representational alignment predicts fine-tuning generalization *despite* joint pretraining on all tasks. We first address a fundamental 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?

To answer this, we designed a two-stage fine-tuning approach. First, we fine-tune on 100k examples of a single task that include Atlantis citiesmatching the exposure Atlantis would have received if included during pretraining (1M examples total). Second, we add a small elicitation set of 256 ex-

amples to ensure the model can properly handle Atlantis-specific queries without degrading overall performance.[2]

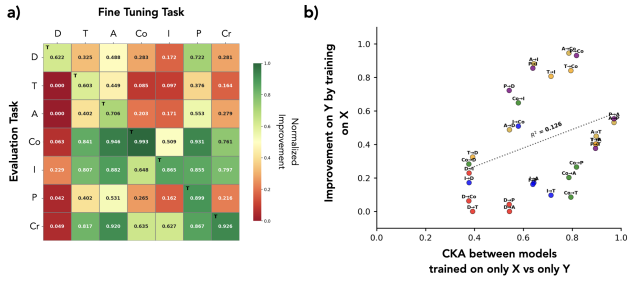

Figure 7: **Fine-tuning generalization and its correlation with representational similarity.** (a) Generalization matrix showing normalized improvement on Atlantis queries after fine-tuning. Each row represents an evaluation task, each column represents a fine-tuning task. Values indicate the normalized performance gain when evaluating task Y after fine-tuning on task X with Atlantis. Some tasks (e.g., `perimeter`) trigger broad generalization while others (e.g., `distance`) remain isolated. (b) Scatter plot revealing the relationship between cross-task generalization and representational similarity. Y-axis: improvement on task Y when fine-tuned on task X. X-axis: CKA between models trained solely on task X vs. task Y. The negative correlation suggests that tasks with divergent single-task representations fail to propagate fine-tuning updates across the shared representational space.

The resulting generalization matrix is shown in Fig. 7. This matrix reveals rich phenomenology: some tasks like `distance` show no cross-task generalization (Atlantis remains usable only for that task), while `perimeter` triggers significant generalization across all tasks except `triangle area`. 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. It is unclear if this is generally true. See App. **??**

Most surprisingly, we find that generalization performance correlates with the CKA values from *single-task pretraining* (Result 2 of Sec. 3). This is puzzling: the CKA values come from models trained from scratch on individual tasks, yet they predict fine-tuning behavior of a model pretrained on all tasks jointly. If the multi-task model truly uses unified representations for cities, why would single-task representational properties matter?

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

We hypothesize that 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:

- To what extent does this affect behavior and generalization?
- Will SGD 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. Under a simple additive model of fine-tuning, we would expect that fine-tuning on a concatenated dataset $\{D_1, D_2, ..., D_n\}$ (which do not provide conflicting supervision) would combine their individual effects. Specifically, when concatenating and shuffling all fine-tuning data to avoid sequential learning

[2]We also mixed in 20% of the original pretraining data without Atlantis to avoid catastrophic forgetting. This design ensures Atlantis integration happens primarily through representation learning rather than memorization.

effects like catastrophic forgetting (**?**), we expect the improvement on task i after training on tasks j and k to be:

$$\text{Improvement}_i(j \cup k) = \max(\text{Improvement}_i(j), \text{Improvement}_i(k)) \quad (1)$$

To test this hypothesis, we fine-tuned the 7-task model on all $\binom{7}{2} = 21$ possible two-task combinations. Fig. 8(a) top shows the evaluation results across all seven tasks, while (a) bottom displays the deviation from our non-interference expectation. Strikingly, we observe "red vertical 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. 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. To understand the mechanistic basis for the performance degradation observed in Result 2, we examine how different task combinations affect the integration of Atlantis cities into the learned world representations. Fig. 8(b,c) compares representations from two exemplar models: one fine-tuned on `angle` + `compass` (non-divergent tasks) versus one fine-tuned on `distance` + `perimeter` (including the divergent `distance` task).

In both PCA and linear probe visualizations, Atlantis cities integrate seamlessly into the world manifold when fine-tuned on non-divergent tasks but remain segregated when divergent tasks are involved. While this difference appears subtle in 2D projections, the effect is dramatic in 3Dwe strongly encourage readers to explore our interactive visualizations which clearly demonstrate the representational segregation.

Most tellingly, when we train linear probes using only original cities (excluding Atlantis from probe training), the probe correctly extrapolates Atlantis locations for non-divergent task models but places them at the origin for divergent task models (see App. **??**). This suggests that divergent tasks cause optimization to encode new entities in orthogonal subspaces 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.

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

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.

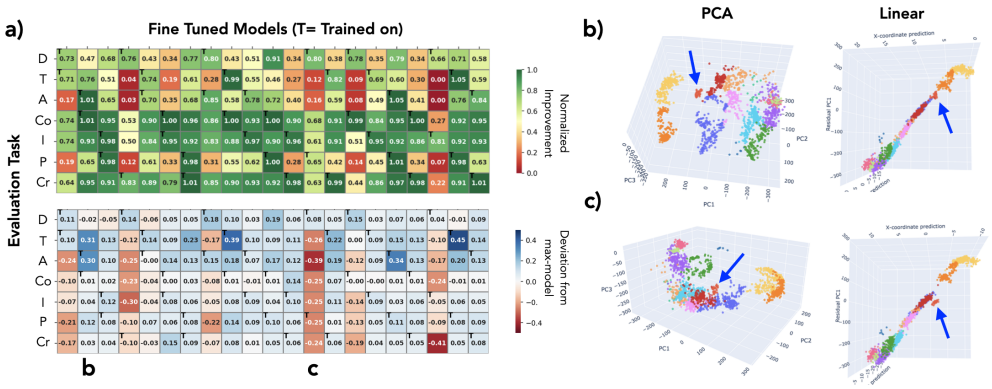

Figure 8: **Divergent tasks harm multi-task fine-tuning and disrupt representational integration.** (a) Top: Performance matrix showing generalization across all 7 tasks when fine-tuning on 21 two-task combinations. Bottom: Deviation from non-interference expectation reveals "red vertical bands" where `distance` task combinations degrade performance below single-task baselines. (b,c) Representational analysis comparing models fine-tuned on non-divergent tasks (`angle` + `compass`) versus divergent task combinations (`distance` + `perimeter`). PCA projections (top) and linear probe reconstructions (bottom) show Atlantis cities (red) integrate into the world manifold for non-divergent tasks but remain segregated in orthogonal subspaces when divergent tasks are involved.

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.

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos, 2022. URL https://arxiv.org/abs/2206.11795.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

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.

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.

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.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

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.

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, 2021. URL https://arxiv.org/abs/2010.11929.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality, 2023. URL https://arxiv.org/abs/2305.18654.

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.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

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.

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.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018. URL https://arxiv.org/abs/1801.06146.

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.

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.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

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.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations?, 2016. URL https://arxiv.org/abs/1511.07543.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. int. conf. on machine learning (ICML)*, 2019.

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.

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.

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.

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, 2023. 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.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018. URL https://arxiv.org/abs/1802.05365.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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.

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.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, 2025. URL https://arxiv.org/abs/2506.06941.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

SIMA Team, Maria Abi Raad, Arun Ahuja, Catarina Barros, Frederic Besse, Andrew Bolt, Adrian Bolton, Bethanie Brownfield, Gavin Buttimore, Max Cant, Sarah Chakera, Stephanie C. Y. Chan, Jeff Clune, Adrian Collister, Vikki Copeman, Alex Cullum, Ishita Dasgupta, Dario de Cesare, Julia Di Trapani, Yani Donchev, Emma Dunleavy, Martin Engelcke, Ryan Faulkner, Frankie Garcia, Charles Gbadamosi, Zhitao Gong, Lucy Gonzales, Kshitij Gupta, Karol Gregor, Arne Olav Hallingstad, Tim Harley, Sam Haves, Felix Hill, Ed Hirst, Drew A. Hudson, Jony Hudson, Steph Hughes-Fitt, Danilo J. Rezende, Mimi Jasarevic, Laura Kampis, Rosemary Ke, Thomas Keck, Junkyung Kim, Oscar Knagg, Kavya Kopparapu, Rory Lawton, Andrew Lampinen, Shane Legg, Alexander Lerchner, Marjorie Limont, Yulan Liu, Maria Loks-Thompson, Joseph Marino, Kathryn Martin Cussons, Loic Matthey, Siobhan Mcloughlin, Piermaria Mendolicchio, Hamza Merzic, Anna Mitenkova, Alexandre Moufarek, Valeria Oliveira, Yanko Oliveira, Hannah Openshaw, Renke Pan, Aneesh Pappu, Alex Platonov, Ollie Purkiss, David Reichert, John Reid, Pierre Harvey Richemond, Tyson Roberts, Giles Ruscoe, Jaume Sanchez Elias, Tasha Sandars, Daniel P. Sawyer, Tim Scholtes, Guy Simmons, Daniel Slater, Hubert Soyer, Heiko Strathmann, Peter Stys, Allison C. Tam, Denis Teplyashin, Tayfun Terzi, Davide Vercelli, Bojan Vujatovic, Marcus Wainwright, Jane X. Wang, Zhengdong Wang, Daan Wierstra, Duncan Williams, Nathaniel Wong, Sarah York, and Nick Young. Scaling instructable agents across many simulated worlds, 2024. URL https://arxiv.org/abs/2404.10179.

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, Justin Y. Chen, Jon Kleinberg, Sendhil Mullainathan, and Ashesh Rambachan. Evaluating the world model implicit in a generative model, 2024. URL https://arxiv.org/abs/2406.03689.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

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.

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.

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.

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 Related Works

Fine-tuning. The pretraining-finetuning paradigm has become central to modern deep learning (LeCun et al., 2015; Goodfellow et al., 2016), with remarkable success across computer vision (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; He et al., 2015; Dosovitskiy et al., 2021), natural language processing (Howard & Ruder, 2018; Peters et al., 2018; Devlin et al., 2018), and reinforcement learning (Team et al., 2024; Baker et al., 2022; Christiano et al., 2017; Radford et al., 2018; 2019). Despite its widespread adoption and extensive study across diverse directionsparameter 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)fine-tuning remains surprisingly unpredictable. Models exhibit poorly understood behaviors such as the reversal curse (Berglund et al., 2024), out-of-context reasoning limitations (Treutlein et al., 2024), and off-target effects (Betley et al., 2025). Recent mechanistic studies suggest fine-tuning may merely form a "thin wrapper" around pretrained representations rather than learning fundamentally new capabilities (Jain et al., 2023; Ward et al., 2025), while behavioral analyses reinforce this pessimism (Yue et al., 2025; Zhao et al., 2025; Qin et al., 2025). The field is thus left with a critical open question about whether fine-tuning can genuinely teach models new concepts (Park et al., 2025; Zweiger et al., 2025) or is fundamentally limited to adapting existing ones.

Interpretability & Internal Representations. Understanding internal representations has been fundamental to neuroscience long before artificial networks existed (Hubel & Wiesel, 1962), and this focus naturally carried over to the development of artificial neural networks (Rosenblatt, 1958; Fukushima, 1980; Rumelhart et al., 1986; Bengio et al., 2014). Recent interpretability work has revealed that language models develop structured "world models" that encode geographic, temporal, and relational information in their parameters (Li et al., 2022; Gurnee & Tegmark, 2023; Nanda et al., 2023; Vafa et al., 2024), with similar representations emerging during in-context learning (Park et al., 2024a). This has led to the Platonic Representation Hypothesis, which posits that diverse models trained on different modalities converge toward similar representational structures (Li et al., 2016; Huh et al., 2024). 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 (Minder et al., 2025; Lee et al., 2024). The precise mechanisms determining which representations form and how they evolve throughout both pretraining and adaptation remain open questions.

Synthetic Data. While large language models provide rich testbeds for studying neural network behavior, their computational cost makes systematic studies of training dynamics prohibitiveone cannot easily test hypotheses by altering pretraining pipelines. Synthetic approaches have successfully addressed this limitation in specific domains: understanding 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.

B Discussion

Continual world models. Recent work has focused on demonstrating that neural networks internally represent more than surface statistics and possess genuine world models. In this study, we take a more nuanced position: these world models are often fractured and partial, as our experiments reveal. A truly robust world model must not only represent the current state of the world but also adapt consistently when the world changes. This adaptation is non-trivial—a single change in the

real world can require multiple cascading updates across different computational tasks. In our framework, introducing Atlantis cities demands their integration across all seven geometric tasks, not just one. The challenge becomes even more acute for tasks like counting cities within a radius, where adding new cities changes answers even for queries that don't directly involve the new locations [reference to additional experiments in appendix]. This combinatorial explosion of interactions makes it infeasible to train on all possible ways new data might interact with existing world knowledge. Models that rely on memorizing specific patterns will inevitably fail when faced with systematic world changes. Instead, we argue that easily and robustly adaptable internal representations are a prerequisite for genuine world models—ones that can update their understanding of the world without catastrophically forgetting previous knowledge or failing to generalize the implications of new information across all relevant computations.

[mention that we need seriously more work on identifying the causal variables and mechanisms]

Limitations. Our study operates primarily on relatively small-scale synthetic data and a single geographic framework. While we have demonstrated that even this simplified setup can exhibit non-trivial phenomenology—including the emergence of world representations, task-dependent factorization, and off-target fine-tuning effects—it cannot fully capture the complexity of real-world data or the scale of modern language models. Additionally, our specific choices in designing the geographic world, defining the geometric tasks, and structuring the data generation process inevitably influence our results. However, we believe that small-scale, controllable model systems provide crucial scientific value: they enable holistic study of the complete worlddatamodel pipeline, rather than merely analyzing individual trained models post-hoc. By establishing causal relationships between world structure, data generation, and representation formation in a controlled setting, we contribute to a growing body of literature that seeks to understand not just what models learn, but how and why they learn it. Future work should extend these findings to larger scales and more diverse domains while maintaining the experimental control that makes mechanistic understanding possible.

Indeed, our study is mostly about constructing an exemplar to motivate further studies in these directions of holistically understand the training dynamics and generalization of neural networks together with representational studies.

C 3D Visualizations

3D visualizations are available here (Open science Framework Anonymyzed link).

D World Setup

Our experiments use a geographic world consisting of 5,175 real cities extracted from the GeoNames database with population greater than 50,000. Cities are distributed across all continents: North America (523 cities), South America (412 cities), Europe (621 cities), Asia (2,847 cities), Africa (498 cities), and Oceania (174 cities). Each city is represented by its latitude and longitude coordinates, normalized to a unit square $[0, 1]^2$ for computational stability.

Additionally, we introduce 100 synthetic "Atlantis" cities positioned in the Atlantic Ocean (centered at 30N, 40W) following a Gaussian distribution with standard deviation of 3 degrees. 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 continuous coordinates.

E Tasks and Datasets

We implement 11 geometric tasks that require understanding city coordinates:

- `Distance (D)`: Euclidean distance between two cities.
- `Triangle Area (T)`: Area of triangle formed by three cities (range: 0-0.5)
- `Crossing (C)`: Whether line segments between four cities intersect (binary)

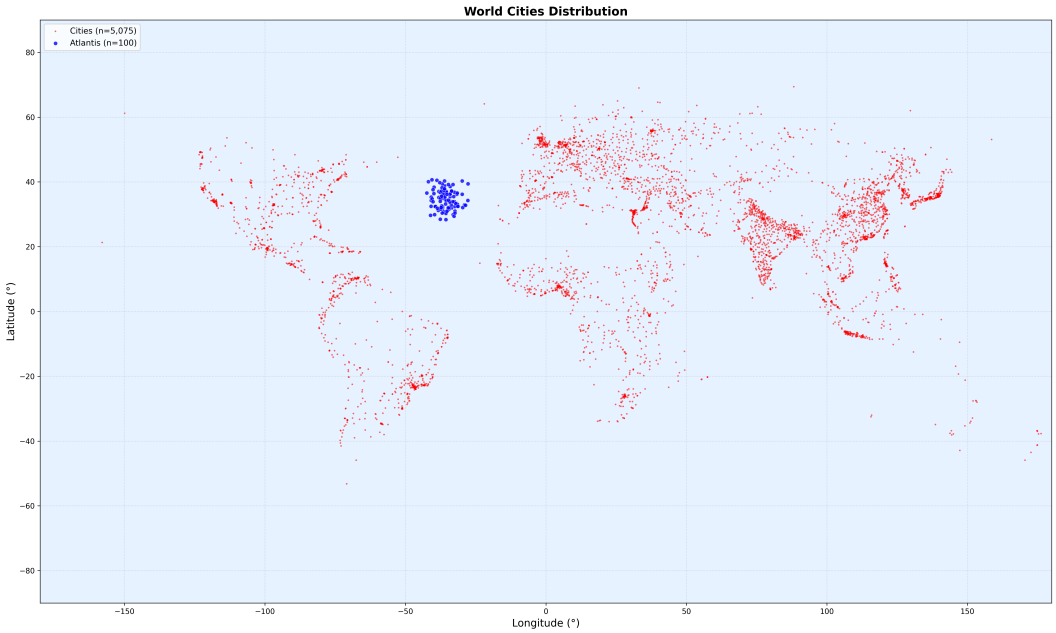

Figure 9: [Caption: Geographic distribution of 5,175 cities used in our experiments. Cities span all continents and provide a fixed, measurable world structure. The synthetic Atlantis region (100 cities in Atlantic Ocean) is used for out-of-distribution testing.]

- `Inside (I)`: Whether a city lies within convex hull of others (binary)
- `Compass`: Direction from one city to another (8 directions)
- `Perimeter`: Perimeter of polygon formed by cities (range: 0-4)
- `Angle`: Angle at middle city of three cities (degrees)

Each task uses different sampling strategies: all-pairs (exhaustive), regional (cities from same region), cross-regional (cities from different regions), and must-include (ensuring specific regions appear). Training sets contain 100K-500K examples depending on task complexity, with 10K validation and 10K test examples.

F Model and Training

F.1 Pretraining Details

We train models autoregressively on character-tokenized sequences, where 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`). While we observed training speedup when masking loss computation on the prompt side (which is unpredictable), we deliberately avoid this optimization to maintain similarity with standard autoregressive language model pretraining. This ensures our findings about representation formation are applicable to standard LLM training regimes.

F.2 Fine-tuning

We typically fine-tuned models with a learning rate of $1e-5$, with the same linear learning rate decay with warmup scheduling.

Fine-tuning's sensitivity to batch size. We observed significant degradation in performance for both the fine-tuned task's and the original (non Atlantis) tasks when we used a batch size of 512.

[PLACEHOLDER TEXT - TO BE REPLACED]

We use decoder-only Transformers with 6 layers, 128 hidden dimensions, 4 attention heads, and intermediate size of 512. This small scale (approximately 2.5M parameters) enables comprehensive analysis of all layers and training dynamics. Models use learned positional embeddings and a vocabulary of 98 ASCII tokens representing city IDs and task syntax.

Training employs AdamW optimizer with learning rate 3e-4, linear warmup over 500 steps, and linear decay. Batch size is 128 for pretraining and 64 for fine-tuning. We train for 15 epochs during pretraining and 20 epochs during fine-tuning, saving checkpoints every 5% of training for detailed analysis. All experiments use three random seeds with standard deviations reported.

Fine-tuning experiments include: (1) single-task fine-tuning from multi-task pretrained models, (2) regional fine-tuning using only European or Asian cities, (3) task-shift fine-tuning where distance-pretrained models adapt to area calculation, and (4) Atlantis adaptation where models must generalize to synthetic cities.

G Analysis Methods

G.1 Probing Details

Omitting irrelevant features We omit cities with `ID` starting with 0, 00 or 000 since the models form a strong special representation separating cities by the number of zeros in their prefix. We suspect this is since many tasks involves a cardinal interpretation of subsequent number tokens for distance, angle, etc while for city ids the numbers do not contain such meaning, and cardinal numbers are given without leading zeros, thus promoting the model to construct a feature clearly distinguishing the two use cases of numbers. For fair evaluation of all cities, we consistently omit all cities with `ID` starting with 0, 00 or 000 for the ease of analysis.

[PLACEHOLDER TEXT - TO BE REPLACED]

We employ several methods to analyze the emergence and quality of world representations:

Linear Probing: We train Ridge regression probes to predict city coordinates (x, y) from intermediate layer activations. Probes are trained on 80% of cities and evaluated on held-out 20%, with R scores measuring representation quality. We analyze activations at different token positions (city IDs, commas, task tokens) to identify where world information is encoded.

PCA Visualization: We project high-dimensional activations to 3D using PCA and visualize cities on world maps, coloring by true geographic regions. This reveals whether the model's internal geometry preserves real-world structure. We track PCA trajectories across training to visualize how representations evolve from random to world-aligned.

Representation Similarity: We use Centered Kernel Alignment (CKA) and Representational Similarity Analysis (RSA) to compare representations across layers, training steps, and different models. This quantifies how similar the learned representations are to the true coordinate structure.

Gradient Analysis: We compute gradients from task loss back to intermediate activations to understand information flow. This reveals which tokens and layers are most important for task performance and how world information propagates through the network.

H Code and Data Availability

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

I Additional Figures