FEW-SHOT DEMONSTRATIONS ELICIT THE USE OF IN-CONTEXT WORLD REPRESENTATIONS IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs), when acting as agents, are expected to take observed data in context, infer the latent state space underlying the world, and leverage it for downstream prediction. However, prior work demonstrated that LLMs struggle to use representations learned in context on a graph tracking task, where the model needs to construct a representation of the graph governing data generation process and use it for subsequent predictions. In this paper, we show that extending this to few-shot settings, where each demonstration is generated from a different world with either the same or different graph topologies, enhances its prediction on 6 models from 4 model families. To understand this improvement, we linearly probe a low-dimensional world representation that encodes graph information in the hidden states. Notably, we find that few-shot demonstrations relocate the world representation and increase its predictive use. Specifically, for each model, these world representations shift in directions nearly orthogonal to their original subspace, and interventions on these representations selectively impair performance more than interventions on other subspaces. Consistent with this insight, we show that few-shot demonstrations with observations from different worlds improve performance on ARC-AGI-1&2, web agent tasks, and Othello. Our findings elucidate the role and internal mechanisms of few-shot demonstrations in in-context world modeling. More broadly, our work advances our understanding of how LLM agents learn from in-context observations and provides implications for their further improvement.

1 INTRODUCTION

As large language models (LLMs) are increasingly deployed as agents in open-ended environments, they must infer previously unseen world, such as the semantics of entities and the rules governing the data generation process, from observations, and take optimal actions based on their inferred model of the world (Andreas, 2024; Li et al., 2025). For example, in coding, an agent may need to infer the structure of unfamiliar codebases and the semantics of their APIs from the surrounding code and execution feedback (Jimenez et al., 2024; Yang et al., 2024). In-context learning (ICL) (Brown et al., 2020), which enables adaptation to the world from observed data without updating any parameters, is important to understand especially from a continual learning perspective (Wang et al., 2024). We refer to this general problem setting—inferring a world from observations and performing a potentially different task over the inferred world—as *in-context world modeling*.

One possible mechanism underlying in-context world modeling is the formation of *in-context representations*. Park et al. (2025) showed that scaling the context length enables LLMs to construct in-context representations with novel semantics. Specifically, LLMs are given, in context, a one-step random walk generated from a latent graph (a 4×4 Lattice Grid in Figure 1 (a)) and exhibit energy minimization that reflects the topology in their intermediate states. However, (Lepori et al., 2026) argued that even when such representations are successfully constructed, LLMs still fail to solve subsequent next-token prediction tasks defined over the graph. Concretely, after observing the random walk generated from a latent graph, LLMs are given few-shot examples of predicting the node two steps to the right of a given node and are then asked to make the same prediction for a new node. They showed that both open-source and closed-source models fail on this task. This gap between world representation construction and predictive use of the representation exposes a fundamental limitation of LLM agents that rely on in-context memory without explicit state tracking

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

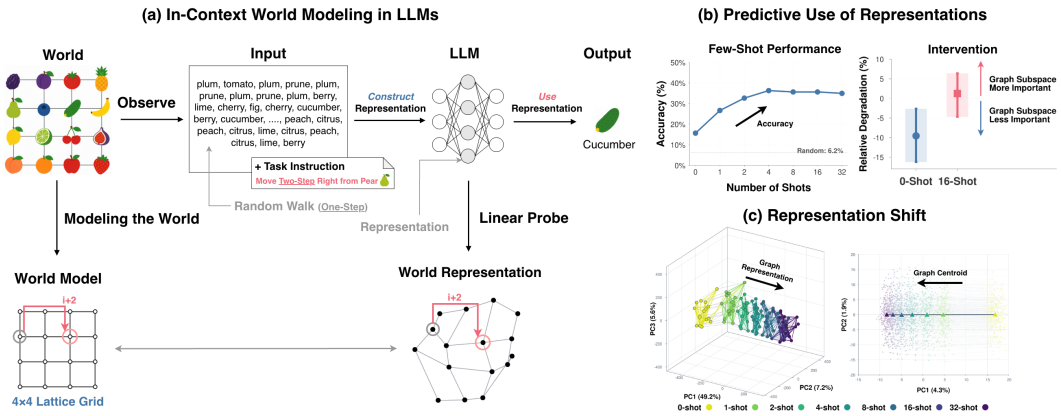

Figure 1: **(a) Conceptual Framework of In-Context World Modeling.** An LLM is provided with random walks generated on a graph topology, where each node is assigned a fruit, and instructed to move two steps to the right. We probe the internal representations to determine whether the model has constructed the latent graph. A linearly decodable subspace of these representations is termed a world representation. See Figure 11 for the actual prompt. In this paper, we treat cucumbers as fruits. **(b) LLMs Learned to Use World Representation with Few Shot Demonstrations.** Few shot demonstrations consisting of random walks, task instructions, and answers from worlds with the same topology but different assignments of fruits to nodes improve task performance. Interventions on the world representation subspace cause larger performance degradation in the few shot setting than in the zero shot setting relative to interventions in random directions. The results are from Qwen3-4B-Base on a 4×4 Lattice Grid. **(c) Few-Shot Demonstrations Shift World Representations.** As the number of shots increases, the graph representation shifts orthogonally to the representation subspace. The left shows a 3D PCA of hidden states of the nodes, and the right shows a 2D PCA of mean hidden states over the random walks in the test set. The results are from Figure 6.

(Lepori et al., 2025; Baldelli et al., 2026; Mozer et al., 2026), raising the question of when they can overcome this limitation and why.

In this paper, we show that extending the zero-shot setting of Lepori et al. (2026) with few-shot demonstrations enables LLMs to solve this in-context world modeling task, as shown in Figure 1 (b). Each demonstration contains a random walk generated from a different world that shares the same or different topology and assigns a different vocabulary to its nodes (see Figure 2 for the prompt illustration).

Motivated by these findings, we use linear probing to analyze a low-dimensional subspace encoding the latent world topology. We find that increasing the number of few-shot demonstrations improves task performance, while both probing accuracy and the Dirichlet Energy (DE) computed from the raw hidden states remain unchanged. This suggests that LLMs construct the world representation equally well regardless of the number of shots (demonstrations). Instead, we find that these representations occupy different subspaces, which influences how they are read out. As the number of shots increases, the representation shifts toward directions roughly orthogonal to itself and progressively converges (Figure 1 (c)). We further intervene on the world representation and find that such interventions degrade predictions more with few-shot demonstrations than in the zero-shot setting. This provides causal evidence that world representations contribute more strongly to prediction with few-shot demonstrations, alongside their systematic relocation across subspaces. Finally, we show that providing observations from different worlds coupled with their task instructions as few-shot demonstrations improves task performance on ARC-AGI-1&2, web agent tasks, and Othello, using Qwen, Gemma, and Ministral models.

In contrast to conventional few-shot learning, where each demonstration typically specifies an input-output mapping, as in regression (Garg et al., 2022; Von Oswald et al., 2023) and classification (Min et al., 2022b), where a canonical example is predicting ”red” from ”apple” (Hendel et al., 2023), we consider an in-context world modeling setting for LLM agents, where each demonstration consists of observations from a world and a task defined over its latent structure. Crucially, our analysis reveals that successful prediction is associated not with improved construction of the world representation itself, but with its relocation to a subspace that elicits downstream readout.

More broadly, we hope that our findings provide support for a growing paradigm of agentic training that leverages ICL (Song et al., 2026b; Hübotter et al., 2026; Zhao et al., 2026; Shi et al., 2026; Ye et al., 2026b;a; Lin et al., 2026; Song et al., 2026a), in which agents condition on observations, trajectories, and outcomes collected through interactions with the environment and generate further rollouts, enabling reinforcement learning (RL) or self-distillation (Shenfeld et al., 2026) without relying on an external teacher model. Our results may also help clarify the mechanisms underlying such transfer of capabilities acquired in context into model weights.

This paper is organized as follows. In Section 2, we explain in-context world modeling, introduce the graph tracking task, and describe the probing, representation analysis, and intervention methods used throughout the paper. In Section 4.1, we show that LLMs construct representations of the latent topology regardless of the number of shots. In Section 4.2, we show that the world representation subspace shifts as the number of shots increases. In Section 5.1, we perform interventions on world representations across different numbers of shots. In Section 5.2, we show that these few-shot demonstrations improve performance on ARC-AGI-1&2, web agent tasks, and Othello. Finally, in Section 6, we discuss broader connections to related work, limitations, and directions for future work. We cover related work in Appendix A.

2 Preliminaries

2.1 World Representation in LLMs

In-Context World Modeling. As written in Appendix A, the definition of a "world model" varies across the literature. In this paper, following (Li et al., 2025), we say that an LLM contains a world model if a relevant property of the underlying world from which the data are generated can be recovered from its internal representations and those representations causally contribute to the model's predictions. We refer to such internal representations as *world representations*. From this perspective, we define *in-context world modeling* as inferring a world from contextual observations and performing a different task over it. See Figure 1 (a) for the illustration. We identify world representations using linear probing (Alain & Bengio, 2016; Belinkov, 2022) and test whether they are causally used for prediction by intervening on them and measuring the resulting change in prediction (Elazar et al., 2021). We refer to the former as *representation construction* and the latter as *predictive use of representation*. See Appendix B for a detailed explanation.

Graph Tracking. We adapt the *adaptive world modeling* task of Lepori et al. (2026), a variant of the *in-context graph tracing* task introduced by Park et al. (2025). As illustrated in Figure 1 (a), each node in a graph topology, specifically 4×4 Lattice Grid, 5×5 Lattice Grid, 3×5 Lattice Grid, 4×4 Triangle Grid, and 4×4 Torus Grid is assigned a fruit vocabulary item such as "apple" or "orange". We regard each such assignment as a world. Even with the same topology, different assignments of fruit to the nodes define different worlds. The model is then given, as observations, a comma-separated sequence of fruit tokens generated by a one-step random walk over the graph. Following these observations, the model receives few-shot examples consisting of node pairs that specify a two-step movement to the right on the graph, such as "lime" to "fig". Although this transformation could instead be specified in natural language as "move two steps to the right," the meaning of "right" changes under rotations or reflections of the topology. We therefore follow Lepori et al. (2026) and specify the task through few-shot node pair examples. We view this setting as *in-context world modeling*, where a model infers a graph topology from observations provided in-context and performs a task on that graph. Typical ICL as in Park et al. (2025) is a special case where the task providing information about the world is identical to the task being tested. The length of random walk is sampled uniformly from 48 to 80. See Appendix C.1 for the detailed setup and Figure 11 for an example prompt.

2.2 Experimental Design

Few-Shot Prompting. As illustrated in Figure 2, each few-shot demonstration consists of observation data and a task instruction, namely a random walk generated from a latent graph and a set of node pairs specifying a two-step movement to the right. Note that these few-shot demonstrations are distinct from the few-shot node pairs used to specify the task instruction. We refer to each demonstration as one shot. All shots share the same graph topology but are gen-

erated from different worlds, with different assignments of fruits to the nodes. Across different shot settings, we keep the final query fixed to enable a controlled comparison. The zero-shot setting corresponds to Lepori et al. (2026). See the prompt example in Figure 11.

Probe. We use linear probing to quantify how well LLMs construct the latent topology from the observed random walk in the query (Figure 2). For each node (i, j), we obtain its layer l hidden state $h_{i,j}^{(l)} \in \mathbb{R}^d$ by averaging the activations over its subtokens and all occurrences in the random walk of the query. We train a probe $W^{(l)} \in \mathbb{R}^{d\times r}$ on 500 samples and use disjoint 100 samples for validation. Assuming that each graph topology admits a low-dimensional representation (Shai et al., 2024; Tigges et al., 2024; Engels et al., 2025), we set $r = 2, 2, 2, 2, 4$ for 4×4 Lattice Grid, 5×5 Lattice Grid, 3×5 Lattice Grid, 4×4 Triangle Grid, and 4×4 Torus Grid, respectively. The probe maps each hidden state to $z_{i,j}^{(l)} = {W^{(l)}}^\top h_{i,j}^{(l)} \in \mathbb{R}^r$. For each pair of nodes $u = (i, j)$ and $v = (i', j')$, we predict their adjacency as $p_{u,v}^{(l)} = \sigma(\beta^{(l)} - \|z_u^{(l)} - z_v^{(l)}\|_2^2)$, where $\sigma(x) = (1 + \exp(-x))^{-1}$ and $\beta^{(l)}$ is a trainable scalar offset. We evaluate the probe using the AUROC for predicting whether pairs of graph nodes are adjacent, computed by sweeping the decision threshold and integrating the resulting ROC curve with the trapezoidal rule. Intuitively, the probe learns to project high-dimensional hidden states into two-dimensional coordinates from which the adjacency matrix of the lattice grid can be recovered. See Appendix C.2 for the detailed explanation and the rationale behind the experimental setup. Park et al. (2025) used Dirichlet Energy (DE) as a measure of representation construction, computed directly from raw hidden states without probing. We instead introduce probing to enable more detailed representation analyses and intervention experiments. For completeness, we also report DE in Section 4.1.

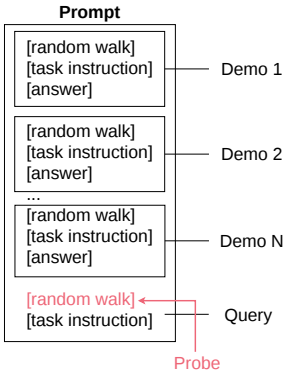

Figure 2: **Illustration of Few-Shot Prompt.** See Figure 11 for the prompt example.

Causal Intervention. To test whether the probed world representations are causally used for prediction, we perform intervention experiments comparing zero-shot and few-shot settings. We measure how much prediction degrades when graph information is removed from the rank-r world representation subspace, relative to an intervention on an unrelated rank-r random subspace orthogonal to it. To remove graph information, we first center each node representation by subtracting the mean representation $\bar{h}^{(l)}$ across all nodes. We then intervene at each layer as $h_{i,j}^{(l)} \leftarrow h_{i,j}^{(l)} - \Pi_{\mathcal{W}}^{(l)}(h_{i,j}^{(l)} - \bar{h}^{(l)})$, where $\mathcal{W}^{(l)}$ is the subspace spanned by the left singular vectors by the singular vector decomposition (SVD) of $W^{(l)}$, which we call *world representation subspace*, and $\Pi_{\mathcal{W}}^{(l)}$ is the orthogonal projection matrix onto $\mathcal{W}^{(l)}$. Intuitively, this isolates variation across nodes that captures their relative positions and edge structure. Following Elazar et al. (2021), we iteratively retrain the probe and apply the intervention. This is necessary because the intervened subspace is extremely low-dimensional, with $r = 2, 4$ compared with ambient dimensions $d = 2560, 4096, 5120$, so a single projection may leave graph information outside the identified subspace that can be repaired (McGrath et al., 2023). To establish a matched baseline for subspace intervention, we compare interventions on the probed world representation subspace with interventions on orthogonal random subspaces of the same dimension. For brevity, we refer to the former as the *graph intervention* and the latter as the *random intervention*. This comparison isolates the causal contribution of the probed world representations while accounting for differences in baseline task performance across different numbers of shots. We emphasize that this comparison controls for nonspecific random subspace removal, rather than for interventions on other task-relevant structured representations.

See Appendix C.2 for the detailed explanation, Appendix E.2 for the experiments on different rank-r, and Appendix E.3 the sensitivity to seeds.

3 Few-Shot Demonstrations Improve Downstream Predictions

In this section, we show that, in the graph tracking task, increasing the number of demonstrations improves downstream performance in predicting the node two steps to the right. We evaluate Qwen3-

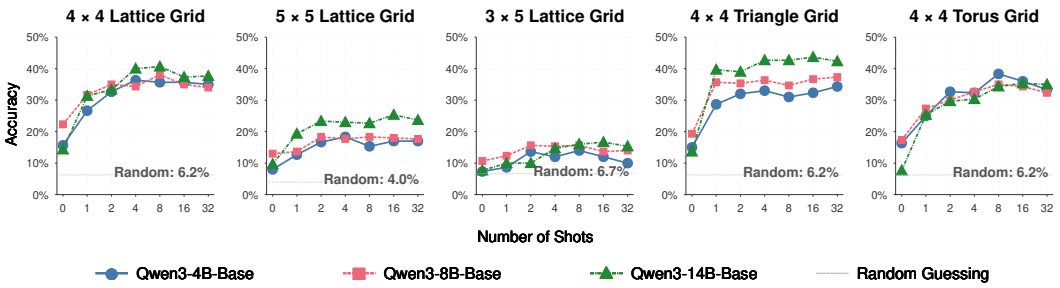

Figure 3: **Few-Shot Performance.** Accuracy of three Qwen3 models on the two-step-right prediction task across five graph topologies. Each point reports accuracy on 300 test samples with 0, 1, 2, 4, 8, 16, or 32 few-shot demonstrations. Dotted gray lines indicate random guessing over the candidate nodes for each topology. Lepori et al. (2026) corresponds to the zero-shot case.

4B-Base, Qwen3-8B-Base, and Qwen3-14B-Base (Yang et al., 2025a) on the 4×4 Lattice Grid, 5×5 Lattice Grid, 3×5 Lattice Grid, 4×4 Triangle Grid, and 4×4 Torus Grid using 300 samples per graph. We consider 0, 1, 2, 4, 8, 16, and 32-shot settings while keeping the query fixed across settings. Predictions are scored by selecting the node with the highest logit among the candidate node set. See Appendix C.1 for details.

Figure 3 shows performance across different numbers of shots. Across all models and topologies, increasing the number of few-shot examples improves performance relative to the 0-shot setting. The performance gains gradually saturate as the number of shots increases, consistent with prior observations in many-shot ICL (Yuan et al., 2024; Zhang et al., 2025b). Note that the *Metalearning AWM Prompt* of Lepori et al. (2026) uses up to three demonstrations with the same topology and identical node assignments, which might allow copying or elimination across shots. In contrast, our demonstrations come from different worlds that share the same topology but use different node assignments. See Appendix E.1 for the results from the Llama-3 (Grattafiori et al., 2024), Gemma-4 (Team et al., 2026), and Ministral-3 (Liu et al., 2026a) models.

More strikingly, in Appendix D.1, we further show that few-shot demonstrations improve performance even when the graph which generate their random walks differ from that of the query.

4 Few-Shot Demonstrations Shift World Representations

4.1 LLMs Construct World Representations

We first show through linear probing that LLMs construct the graph topology underlying the random walk equally well regardless of the number of shots.

As described in Section 2.2, we measure whether LLMs construct the graph by evaluating graph adjacency prediction with AUROC. We also report Dirichlet Energy (DE) (Park et al., 2025), which measures the alignment between node representations and the graph structure. At layer l, DE is computed as $\sum_u \sum_v A_{u,v} \|h_u^{(l)} - h_v^{(l)}\|_2^2$, where $u = (i, j)$, $v = (i', j')$, and $A_{u,v}$ denotes the adjacency matrix entry between nodes u and v. We report normalized DE by dividing this quantity by the sum of squared Euclidean distances between the hidden states of all node pairs.

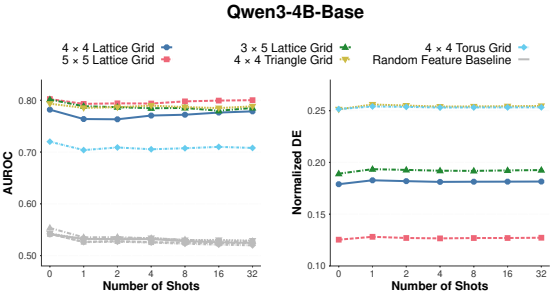

Figure 4: **Probe Performance and Normalized Dirichlet Energy (DE).** Probe AUROC and normalized DE of hidden states across shots and five graph topologies for layer 20 of Qwen3-4B-Base. Gray lines show random feature probe baselines.

Figure 4 shows that probing AUROC and normalized DE remain largely unchanged across different numbers of shots for all three models and five topologies. As a random feature baseline, we replace

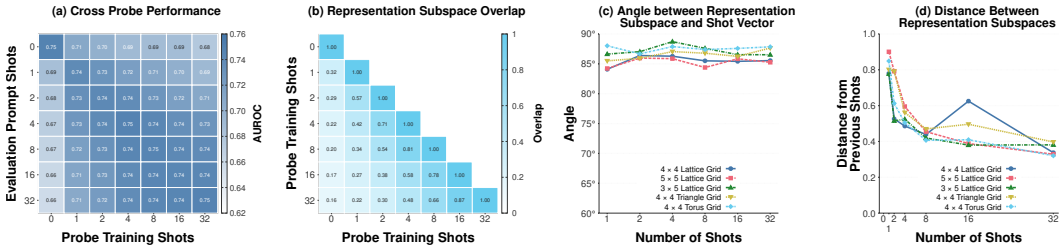

Figure 5: **(a) Cross Probe Performance.** AUROC of probes trained at one shot count and evaluated at another, averaged over five topologies. **(b) World Representation Subspace Overlap.** Overlap between probe subspaces across shot counts, averaged over the five topologies. **(c) Angle Between Shot Vectors and the World Representation Subspace.** Angles to the corresponding probe subspaces, shown separately for each topology. **(d) Distance between Representation Subspaces across Shots.** Normalized projection distance between probe subspaces at consecutive tested shot counts, shown separately for each topology. All results are from layer 36 of Qwen3-4B-Base.

the activations with isotropic Gaussian over five random seeds, which yields AUROC values around 0.5. The clearly higher AUROC indicates that the latent graph topology is recoverable from their internal states. Together, these results suggest that LLMs construct an internal representation of the world behind the random walk in the query regardless of the number of shots. We additionally show in Figure 17 that the relative positions of nodes in the graph representation of a query do not change much as the number of shots increases.

4.2 Few-Shot Demonstrations Relocate World Representations

Here, we provide lines of evidence that the world representation shifts as the number of shots increases.

For three Qwen3 models, we train probes at three layers under 0, 1, 2, 4, 8, 16, and 32-shot settings, using 500 samples for training and 100 for validation. We then evaluate cross-shot probing on a 300 sample test set and report the AUROC in Figure 5 (a). Probes trained on few-shot representations fail to correctly decode graphs from zero-shot representations, and decoding performance also degrades as the gap between the numbers of shots increases. This indicates that changing the number of shots changes the probe readout space, at least in part, toward different directions even for the same query. See Figure 18 for the full results.

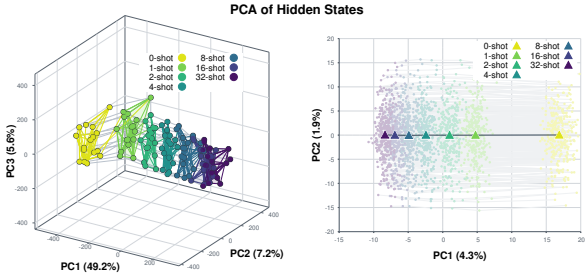

Figure 6: **(Left) PCA of Shift in Hidden States of a Random Walk.** PCA of 4×4 Lattice Grid at layer 36 of Qwen3-8B-Base **(Right) PCA of Shifts in Hidden-State Centroids across Random Walks.** PCA of 300 test samples per shot for 4×4 Lattice Grid at layer 20 of Qwen3-8B-Base. Points represent individual random walks and triangles indicate their means. In each panel, PCA is fit jointly to all activations shown.

We further measure the overlap between world representation subspaces $\mathcal{W}^{(l)}$ and $\mathcal{W}'^{(l)}$ learned by probes trained with different numbers of shots, by computing $\mathrm{tr}(\Pi^{(l)}_{\mathcal{W}}\Pi'^{(l)}_{\mathcal{W}'})/r$. As shown in Figure 5 (b), the subspace overlap decreases as the number of shots changes, with a particularly large discrepancy between zero-shot and few-shot settings. This suggests that the linearly decodable graph representation subspace shifts as the number of shots increases. See Figure 19 for the full results.

Then, we compute the mean shift $\Delta\bar{h}^{(l)}$ between corresponding few-shot and zero-shot hidden states, averaged over all nodes in the test set. We then measure its angle to the world representation subspace $\mathcal{W}^{(l)}$ as $\theta^{(l)} = \arccos(\|\Pi^{(l)}_{\mathcal{W}}\Delta\bar{h}^{(l)}\|_2/\|\Delta\bar{h}^{(l)}\|_2)$. As shown in Figure 5 (c), the hidden states shift induced by increasing number of shots, where we refer to the shift $\Delta\bar{h}^{(l)}$ as the *Shot*

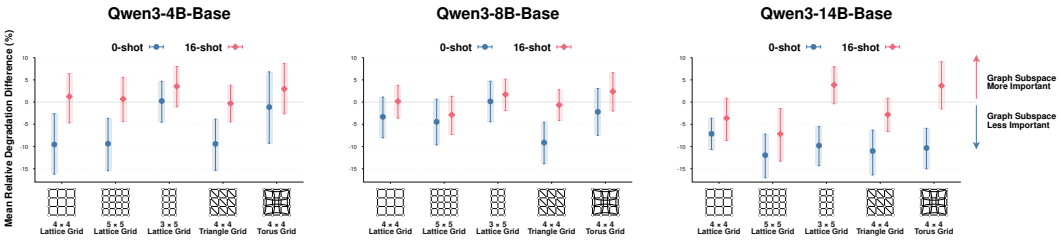

Figure 7: **Intervention Results.** Graph subspace erasure at iteration 20 in three Qwen3 models. Interventions target node subtokens in the query's random walk at every Transformer block output. The random intervention uses an orthogonal subspace with matched rank and perturbation norm per node. Points show the mean difference in the probability of correct answer token reduction between graph and random interventions, normalized by each query's unperturbed probability, across 300 test samples. Probabilities are normalized over candidate answer sequence likelihoods. Positive values indicate greater degradation from graph erasure. Horizontal bars show 95% bootstrap confidence intervals obtained by resampling test samples 5000 times.

Vector, is nearly orthogonal to the world representation subspace. See Figure 22 for the full results. Importantly, this interpretation relies not only on near-orthogonality in the high-dimensional activation space, but also on changes in cross probe performance and subspace overlap, which suggests that few-shot demonstrations alter the geometry of representations rather than simply shifting hidden states along existing subspace directions, with the Shot Vector capturing only this displacement.

Figure 6 visualizes the hidden states of nodes in a random walk for one sample, and the mean hidden states of random walks across 300 test samples, using PCA. As the number of shots increases, the representations systematically shift. The line segments connecting corresponding points, which form part of the Shot Vector, are also aligned in a similar direction. More visualizations are provided in Figures 20 and 21.

Note that *Task Vector* (Hendel et al., 2023) is extracted as the hidden state at the final token of a few-shot prompt, whereas *Function Vector* (Todd et al., 2024) is constructed by averaging the final token of causally important attention heads across multiple few-shot prompts and summing them. In contrast, our *Shot Vector* is defined as the difference between the centroids of the hidden states corresponding to observation inputs under few-shot and zero-shot settings. We show that the Shot Vector is different from the Task Vector in Appendix D.3.

We further compute the distance between world representation subspaces for adjacent shot settings among 0, 1, 2, 4, 8, 16, and 32 shots as $\|\Pi_{\mathcal{W}}^{(l)} - \Pi'^{(l)}_{\mathcal{W}'}\|_{\mathrm{F}}/\sqrt{2r}$. As shown in Figure 5 (d), this distance decreases as the number of shots increases, which indicates that the world representation subspace gradually converges in distance. See Figure 24 for the full results.

All in all, these results suggest that increasing the number of shots induces a shift in the hidden states of the random walk, which we call the Shot Vector. This shift occurs approximately orthogonally to the rank-r world representation subspace, corresponding to a change in the readout space of the world representation. See Appendix E.1 for the results from the Llama-3, Gemma-4, and Ministral-3 models.

5 Predictive Use of World Representations

5.1 Intervention Experiments

Having shown that few-shot demonstrations shift the world representation, we next test whether the shifted representation contributes to subsequent predictions. Following Ravfogel et al. (2020); Elazar et al. (2021), we eliminate graph information from the world representation subspace, from which graph topology can be decoded. To compare graph interventions across zero-shot and few-shot settings, we measure the reduction in the probability assigned to the correct token relative to a intervention in the random orthogonal subspace. This control accounts for differences in baseline task performance and is matched to the graph intervention in the number of activation dimensions perturbed and all other intervention settings.

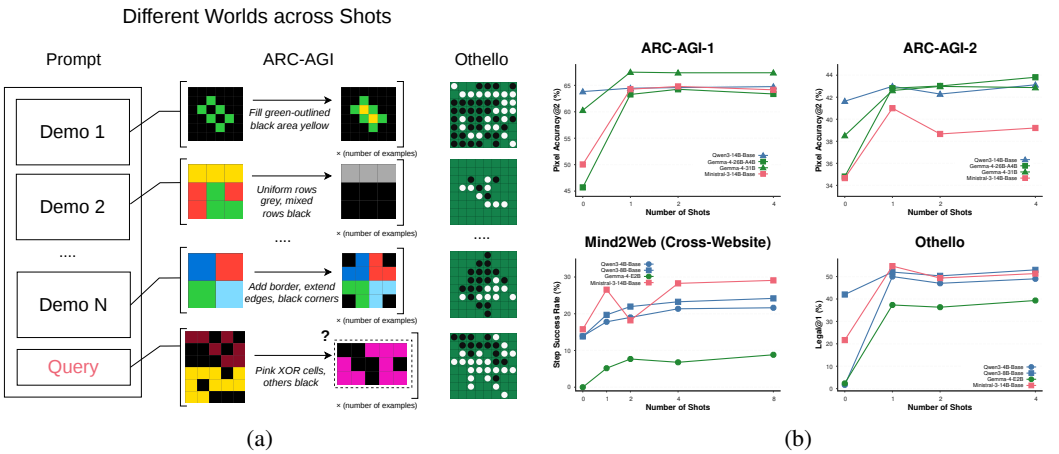

Figure 8: **(a) ARC-AGI and Othello with Different Worlds across Shots.** In ARC-AGI, the world is the rule that maps inputs to outputs. In Othello, the world is the board state. **(b) Performance across Different Numbers of Shots on ARC-AGI-1&2, Mind2Web, and Othello.**

Figure 7 shows the difference in the degradation of the probability assigned to the correct token between the graph intervention and the random intervention. Larger values indicate that intervening on the world representation subspace causes greater degradation than the random intervention. Compared with the zero-shot setting, this difference increases when few-shot demonstrations are provided. Because representational repair can occur under both interventions, their difference captures the additional degradation from intervening on the graph representation subspace. This result provides evidence that few-shot demonstrations elicit greater, though still partial, predictive use of world representations that encode graph information. See Appendix D.4 for the detailed results and Appendix E.1 for the results from the Llama-3, Gemma-4, and Ministral-3 models.

5.2 Experiments on Practical Tasks

Finally, we demonstrate that the effectiveness of few-shot demonstrations generalizes beyond graph tracking to ARC-AGI-1&2, web agent tasks, and Othello.

The ARC-AGI-1 (Chollet, 2019) test set consists of 400 tasks and 419 test pairs, while ARC-AGI-2 (Chollet et al., 2025) consists of 120 tasks and 167 test pairs. Each task provides input–output pairs labeled as "Examples." These pairs constitute observations from which the model must infer the underlying rule and apply it to a new input. Here, the underlying rule corresponds to the world, which determines the input–output data generation process. For each query in the test set, we provide few-shot demonstrations from different rules (worlds), randomly sampled from the training set. Because the task is challenging for the models evaluated here, exact-match accuracy (pass@2) is sparse. We therefore use pixel-level accuracy@2 as a more sensitive measure of partial task performance. For the web agent task, we use the Mind2Web dataset (Deng et al., 2023), which contains 1373 action steps across 177 Cross-Website tasks. Step success rate is computed for each task and then averaged across tasks. Given a website's HTML, an interaction trajectory, and a task instruction, the model must infer the appropriate action. Here, the states and structure of the website underlying the HTML correspond to the world. We randomly sample few-shot demonstrations from the train set and use the test set for queries. By using cross-website subset, we ensure that the websites (worlds) differ from those in the few-shot demonstrations. For the Othello task, following Li et al. (2023); Nanda et al. (2023), we represent a game as a sequence of between 8 and 52 moves over the 64 board squares, alternating between the two players. The model predicts the next move, which is considered correct if legal, with each few-shot demonstration and query corresponding to a different board position (world). We evaluate on 300 samples using Qwen3, Gemma-4, and Ministral-3, selecting model sizes appropriate to the difficulty of each task. See Figure 8a for the prompt illustration and Appendix C.3 for the detailed setup.

Figure 8b shows few-shot performance on four benchmarks, where performance improves as the number of demonstrations increases compared with zero-shot settings, despite demonstrations and queries being drawn from different worlds. This consistency across diverse tasks suggests that the benefit of few-shot demonstrations extends beyond graph tracking.

6 DISCUSSION

In this work, we investigated how LLMs construct an internal representation of the underlying world from observations and use it for prediction. While such predictive use is difficult in the zero-shot setting (Lepori et al., 2026), we showed that few-shot demonstrations containing observations from different worlds enhance downstream predictions, both in graph tracking, ARC-AGI-1&2, web agent tasks, and Othello. In the graph tracking setting, our representation analyses and intervention experiments further provide evidence that increasing the number of shots relocates the world representation into a subspace from which it can be more effectively read out for prediction.

Intelligence as Skill-Acquisition. At a higher level, an important milestone of AI is efficient skill acquisition and adaptation (Legg & Hutter, 2007; Chollet, 2019; Goldfeder et al., 2026) rather than merely maximizing fixed benchmarks. This work offers a mechanistic account of how LLM agents use in-context observations to perform tasks in novel environments.

ICL in the Era of Agents. Previous research on ICL has focused on settings where ground truth input–output pairs are provided, without considering agentic interaction with an environment. This work instead studies a more agentic setting, which we call in-context world modeling, asking how demonstrations composed of observations, task instructions, and outcomes shape both LLM behavior and its internal mechanisms.

LLMs Know but Cannot Say. Prior work has shown that LLMs can internally represent truth even when it is not decodable (Kadavath et al., 2022; Burns et al., 2023; Azaria & Mitchell, 2023), and that interventions can make such information decodable (Li et al., 2024). In line with these findings, this work provides evidence for a gap between representation construction and its predictive use. LLMs can construct representations of world structure from observations yet fail to use them for downstream tasks, but few shot demonstrations can bridge this gap through prompting alone.

Task (Function) Vectors in ICL. We observe that increasing the number of shots shifts the centroid of the hidden states of the same observed token sequence, and we refer to the direction of this shift as the *Shot Vector*. While Task Vectors (Hendel et al., 2023) and Function Vectors (Todd et al., 2024) encode input–output mappings (Yang et al., 2025b), our Shot Vector instead shifts the representations of observed tokens that is associated with greater predictive use of the world representation.

Meta-In-Context-Learning. Meta-in-context learning (Coda-Forno et al., 2023) uses sequential linear regression and two-armed bandit tasks with task-specific demonstrations, showing that LLMs can infer priors over task distributions, such as linearity, from context. In contrast, our few-shot setting examines how LLMs infer how to use world representations for prediction from observations generated in different worlds under the same task.

Limitations and Future Work. In this work, we define an LLM's world model in terms of its representation of the static structure of the world (Li et al., 2023; Nanda et al., 2023; Li et al., 2025). State transitions, temporal dynamics, and more general forms of causal dynamics (Andreas, 2024) are beyond the scope of this definition. We emphasize that performance gains from few-shot demonstrations are not solely due to increased predictive use of world representations, which is only one contributing factor. Further work could help disentangle these factors. Additionally, we studied the construction and predictive use of world representations by changing model activations through prompting. Future work should examine how RL and SFT affect these processes (Chu et al., 2025; Matsutani et al., 2026b), their relation to generalization, and how in-context tuning (Min et al., 2022a; Chen et al., 2022) differs from fine-tuning. Our demonstrations used ground-truth answers, but it is also interesting to study in-context reinforcement learning (ICRL) (Laskin et al., 2023; Monea et al., 2025; Song et al., 2026a), where environmental rewards are provided instead.

AI USE STATEMENT

In this work, we used generative AI tools to design or provide feedback on research methodology or experiments, implement methods, and assist with translation. We did not use generative AI tools to help develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, support qualitative and thematic data analysis, or interpret results. Generating synthetic datasets, formulating mathematical claims, assisting in the writing of proofs, and cleaning or reformatting datasets were not applicable to this work.

Additionally, we used generative AI tools to create or modify scientific figures or images, suggest experimental parameters, create or edit software code, summarize or analyze existing literature, brainstorm research ideas, source or search for information, edit the paper to improve readability, identify relevant literature, and propose a title or keywords for the paper.

We reviewed all AI-assisted work. AI-assisted method implementations and software code were manually reviewed and tested. Suggestions concerning research methodology, experiments, and experimental parameters were independently evaluated by the authors before being incorporated into the work. AI-assisted figures and images were checked for consistency with the underlying data and intended scientific content. Translations and language edits were reviewed by the authors to ensure that they preserved the intended meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Ekin Akyürek, Mehul Damani, Adam Zweiger, Linlu Qiu, Han Guo, Jyothish Pari, Yoon Kim, and Jacob Andreas. The surprising effectiveness of test-time training for few-shot learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=asgBo3FNdg.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Jacob Andreas. Language models, world models, and human model-building, 2024. URL https://lingo.csail.mit.edu/blog/world_models/.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Davide Baldelli, Ali Parviz, Amal Zouaq, and Sarath Chandar. Llms can't play hangman: On the necessity of a private working memory for language agents. *arXiv preprint arXiv:2601.06973*, 2026.

Timothy E. J. Behrens, Timothy H. Muller, James C. R. Whittington, Shirley Mark, Alon B. Baram, Kimberly L. Stachenfeld, and Zeb Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018. doi: 10.1016/j.neuron.2018.10.002.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=bJbSbJskOS.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Stephanie C.Y. Chan, Adam Santoro, Andrew Kyle Lampinen, Jane X Wang, Aaditya K Singh, Pierre Harvey Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=lHj-q9BSRjF.

Aviral Chawla, Galen Hall, and Juniper L Lovato. Metaothello: A controlled study of multiple world models in transformers. In *Forty-third International Conference on Machine Learning*, 2026.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-learning via language model in-context tuning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 719–730, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.53. URL https://aclanthology.org/2022.acl-long.53/.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. Arc-agi-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dYur3yabMj.

Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matthew Botvinick, Jane X Wang, and Eric Schulz. Meta-in-context learning in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=sx0xpaO0za.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *arXiv preprint arXiv:2306.06070*, 2023.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *arXiv preprint arXiv:2006.00995*, 2021.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Daniel Franzen, Jan Disselhoff, and David Hartmann. Product of experts with LLMs: Boosting performance on ARC is a matter of perspective. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dsBjxI6l8W.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=flNZJ2eOet.

Jonas Geiping, Sean Michael McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=S3GhJooWIC.

Judah Goldfeder, Philippe Wyder, Yann LeCun, and Ravid Shwartz Ziv. AI must embrace specialization via superhuman adaptable intelligence. *arXiv preprint arXiv:2602.23643*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide

Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Fabian Gröger, Shuo Wen, and Maria Brbic. Revisiting the platonic representation hypothesis: An aristotelian view. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=uz0gAAYydl.

Wes Gurnee and Max Tegmark. Language models represent space and time. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 2483–2503, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0a6059857ae5c82ea9726ee9282a7145-Paper-Conference.pdf.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Seungwook Han, Jinyeop Song, Jeff Gore, and Pulkit Agrawal. Emergence and effectiveness of task vectors in in-context learning: An encoder decoder perspective. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=0ysC6VS0y3.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

Keya Hu, Ali Cy, Linlu Qiu, Xiaoman Delores Ding, Runqian Wang, Yeyin Eva Zhu, Jacob Andreas, and Kaiming He. ARC is a vision problem! *arXiv preprint arXiv:2511.14761*, 2025.

Tianze Hua, Tian Yun, and Ellie Pavlick. mOthello: When do cross-lingual representation alignment and cross-lingual transfer emerge in multilingual models? In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1585–1598, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.103. URL https://aclanthology.org/2024.findings-naacl.103/.

Jonas Hübotter, Frederike Lübeck, Lejs Deen Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=QkfkxyRizZ.

Michael Ivanitskiy, Alexander F Spies, Tilman Räuker, Guillaume Corlouer, Christopher Mathwin, Lucia Quirke, Can Rager, Rusheb Shah, Dan Valentine, Cecilia Diniz Behn, Katsumi Inoue, and Samy Wu Fung. Linearly structured world representations in maze-solving transformers. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023. URL https://openreview.net/forum?id=pZakRK1QHU.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Tim Rocktäschel, Edward Grefenstette, and David Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=A0HKeKl4Nl.

Gangwei Jiang, Caigao JIANG, Zhaoyi Li, Siqiao Xue, JUN ZHOU, Linqi Song, Defu Lian, and Ying Wei. Unlocking the power of function vectors for characterizing and mitigating catastrophic forgetting in continual instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gc8QAQfXv6.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Abhinav Joshi, Vaibhav Sharma, and Ashutosh Modi. CheckersGPT: Learning world models through language modeling. In Xiyan Fu and Eve Fleisig (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 414–426, Bangkok, Thailand, August 2024. Association for Computational Linguistics. ISBN 979-8-89176-097-4. doi: 10.18653/v1/2024.acl-srw.48. URL https://aclanthology.org/2024.acl-srw.48/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Adam Karvonen. Emergent world models and latent variable estimation in chess-playing language models. *arXiv preprint arXiv:2403.15498*, 2024.

Ryotaro Kawata, Kohsei Matsutani, Yuri Kinoshita, Naoki Nishikawa, and Taiji Suzuki. Mixture of experts provably detect and learn the latent cluster structure in gradient-based learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=2xmOEbpYv1.

Seijin Kobayashi, Simon Schug, Yassir Akram, Florian Redhardt, Johannes von Oswald, Razvan Pascanu, Guillaume Lajoie, and João Sacramento. When can transformers compositionally generalize in-context? *arXiv preprint arXiv:2407.12275*, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*, 2022. URL https://openreview.net/forum?id=6p3AuaHAFiN.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Stenberg Hansen, Angelos Filos, Ethan Brooks, maxime gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=hy0a5MMPUv.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview*, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf. Version 0.9.2, 2022-06-27.

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds Mach.*, 17(4):391–444, December 2007. ISSN 0924-6495.

Michael A. Lepori, Michael Curtis Mozer, and Asma Ghandeharioun. Racing thoughts: Explaining contextualization errors in large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3020–3036, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.155. URL https://aclanthology.org/2025.naacl-long.155/.

Michael A. Lepori, Tal Linzen, Ann Yuan, and Katja Filippova. Language models struggle to use representations learned in-context. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14841–14857, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.676. URL https://aclanthology.org/2026.acl-long.676/.

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*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2024.

Kenneth Li, Fernanda Viégas, and Martin Wattenberg. What does it mean for a neural network to learn a "world model"? *arXiv preprint arXiv:2507.21513*, 2025.

Xiaofeng Lin, Sirou Zhu, Yilei Chen, Mingyu Chen, Hejian Sang, Ioannis Paschalidis, Zhipeng Wang, Aldo Pacchiano, and Xuezhou Zhang. Scaling in-context online learning capability of llms via cross-episode meta-rl. *arXiv preprint arXiv:2602.04089*, 2026.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont,

Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026a.

Binnan Liu, Yechi Ma, Tian Xie, and Wei Hua. TraceViT: Grounded trace supervision for visual abstract reasoning. *arXiv preprint arXiv:2607.29586*, 2026b.

Kohsei Matsutani, Gouki Minegishi, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Zipping the thought: When and how compressed reasoning data works in llm post-training. *arXiv preprint arXiv:2605.28008*, 2026a.

Kohsei Matsutani, Shota Takashiro, Gouki Minegishi, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. RL squeezes, SFT expands: A comparative study of reasoning LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=N2lMNqJsBw.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn in context. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2791–2809, Seattle, United States, July 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.201. URL https://aclanthology.org/2022.naacl-main.201/.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, December 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Sarthak Mittal, Yoshua Bengio, Nikolay Malkin, and Guillaume Lajoie. In-context parametric inference: Point or distribution estimators? *arXiv preprint arXiv:2502.11617*, 2025a.

Sarthak Mittal, Eric Elmoznino, Leo Gagnon, Sangnie Bhardwaj, Guillaume Lajoie, and Dhanya Sridhar. Does learning the right latent variables necessarily improve in-context learning? In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=tuYmhlu62K.

Giovanni Monea, Antoine Bosselut, Kianté Brantley, and Yoav Artzi. LLMs are in-context bandit reinforcement learners. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=c0RsezY2D1.

Michael C. Mozer, Shoaib Ahmed Siddiqui, and Rosanne Liu. The topological trouble with transformers. *arXiv preprint arXiv:2604.17121*, 2026.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi (eds.), *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.2. URL https://aclanthology.org/2023.blackboxnlp-1.2/.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Core Francisco Park. Convergent world representations and divergent tasks. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Z4PjCncxzz.

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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pXlmOmlHJZ.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Raanan Yehezkel Rohekar, Yaniv Gurwicz, Sungduk Yu, Estelle Aflalo, and Vasudev Lal. A causal world model underlying next token prediction: Exploring GPT in a controlled environment. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qA3xHJzF6B.

Adam Shai, Paul M. Riechers, Lucas Teixeira, Alexander Gietelink Oldenziel, and Sarah Marzen. Transformers represent belief state geometry in their residual stream. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=YIB7REL8UC.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=qA6FgH0nnZ.

Taiwei Shi, Sihao Chen, Bowen Jiang, Linxin Song, Longqi Yang, and Jieyu Zhao. Experiential reinforcement learning. *arXiv preprint arXiv:2602.13949*, 2026.

Wen-Jie Shu, Xuerui Qiu, Rui-Jie Zhu, Harold Haodong Chen, Yexin Liu, and Harry Yang. LoopViT: Scaling visual arc with looped transformers. *arXiv preprint arXiv:2602.02156*, 2026.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. doi: 10.1126/science.aar6404. URL https://www.science.org/doi/abs/10.1126/science.aar6404.

Kefan Song, Amir Moeini, Peng Wang, Lei Gong, Rohan Chandra, Shangtong Zhang, and Yanjun Qi. Reward is enough: LLMs are in-context reinforcement learners. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=keCXNHOe4W.

Yuda Song, Lili Chen, Fahim Tajwar, Rémi Munos, Deepak Pathak, Drew Bagnell, Aarti Singh, and Andrea Zanette. Expanding the capabilities of reinforcement learning via text feedback. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=3voRLZLkwU.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar,

Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. Llm circuit analyses are consistent across training and scale. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 40699–40731. Curran Associates, Inc., 2024. doi: 10.52202/079017-1287. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/47c7edadfee365b394b2a3bd416048da-Paper-Conference.pdf.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=aVK4JFpegy.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *arXiv preprint arXiv:2302.00487*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Haolin Yang, Hakaze Cho, Yiqiao Zhong, and Naoya Inoue. Unifying attention heads and task vectors via hidden state geometry in in-context learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=FIfjDqjV0B.

John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik R Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=mXpq6ut8J3.

Tianzhu Ye, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. Online experiential learning for language models. *arXiv preprint arXiv:2603.16856*, 2026a.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026b.

Ilker Yildirim and L. A. Paul. From task structures to world models: What do llms know? *Trends in Cognitive Sciences*, 28(5):404–415, 2024. doi: 10.1016/j.tics.2024.02.008.

Peiwen Yuan, Shaoxiong Feng, Yiwei Li, Xinglin Wang, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. Focused large language models are stable many-shot learners. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6247–6261, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.359. URL https://aclanthology.org/2024.emnlp-main.359/.

Yifei Yuan and Anders Søgaard. Revisiting the othello world model hypothesis. *arXiv preprint arXiv:2503.04421*, 2025.

Tianren Zhang, Guanyu Chen, and Feng Chen. When do neural networks learn world models? In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=oUzxhnnGVM.

Xiaoqing Zhang, Ang Lv, Yuhan Liu, Flood Sung, Wei Liu, Jian Luan, Shuo Shang, Xiuying Chen, and Rui Yan. More is not always better? enhancing many-shot in-context learning with differentiated and reweighting objectives. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30539–30552, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1475. URL https://aclanthology.org/2025.acl-long.1475/.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Jpxfof0EaS.

A Related Work

World Models in LLMs. World models have been discussed in different strands of literature, such as cognitive maps (Behrens et al., 2018; Yildirim & Paul, 2024), reinforcement learnings (Silver et al., 2018; Ha & Schmidhuber, 2018), abstract predictive representation(LeCun, 2022), and video generations (Bruce et al., 2024). Andreas (2024); Li et al. (2025); Zhang et al. (2025a) sought to define what it means for neural networks, especially language models, to learn world models. Prior works demonstrates that neural networks encode sentiments (Radford et al., 2017), and state and temporal information (Gurnee & Tegmark, 2024) by probing internal representations, and assess their causal effects by intervening them. Li et al. (2023); Nanda et al. (2023); Hua et al. (2024); Yuan & Søgaard (2025); Chawla et al. (2026) probed board state representations in Othello. Karvonen (2024), Joshi et al. (2024), and Ivanitskiy et al. (2023) did so for chess, checkers, and mazes, respectively.

ICL in LLMs. The properties of ICL (Brown et al., 2020) have been studied from the perspectives of algorithmic characteristics (Xie et al., 2022; Von Oswald et al., 2023), mechanistic circuits (Olsson et al., 2022; Hendel et al., 2023), and language data distribution (Chan et al., 2022). Park et al. (2025) showed that increasing context length induces a sudden reorganization into in-context representations, whereas Lepori et al. (2026) argued that these internal representations may not be accessible for subsequent prediction. Mittal et al. (2025b); Han et al. (2025); Kobayashi et al. (2024); Mittal et al. (2025a) analyzed latent-variable prediction in ICL, while Jiang et al. (2025); Jain et al. (2024) investigated internal behavioral changes under supervised fine-tuning (SFT).

B Definition of World Models in LLMs

Following Andreas (2024); Li et al. (2025); Zhang et al. (2025a), we consider the world underlying the generation of the textual observations to be a static configuration.

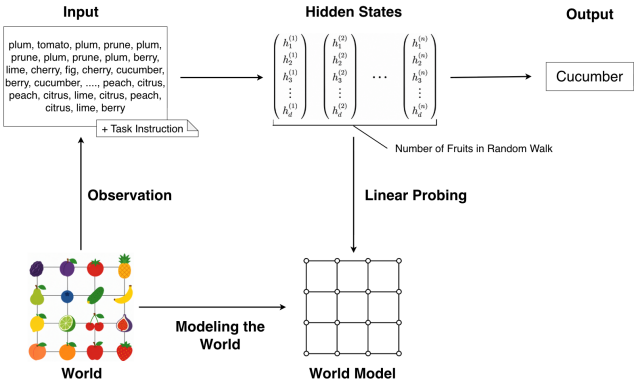

Figure 9: **Illustration of World Models in LLMs.**

Although world models can be defined at various levels, our setting corresponds to "The map" discussed in Andreas (2024). Rather than storing pairwise relations between fruits in a lookup table, the representation encodes a graph topology that captures their relative spatial relationships. Computation over this representation can then be used to solve navigation tasks on the graph. The random walks are generated using one-step transitions, whereas the task requires two-step reasoning, and thus cannot be solved by simply copying a single-hop transition. Our setting does not, however, correspond to "The orrery," which models temporal evolution, or "The simulator," which incorporates physical laws, as discussed in Andreas (2024).

Following Li et al. (2023); Nanda et al. (2023); Yuan & Søgaard (2025); Chawla et al. (2026), we say that an LLM possesses a world model in this sense if, as visualized in Figure 9, it receives textual random walks generated from a world in which fruits are arranged according to a graph, and the static graph structure of that arrangement can be recovered from the LLM's internal representations using a linear probe. For representations associated with multiple subtokens or repeated occurrences, we use their average. In addition, intervening on these representations should disrupt predictions that were previously possible through them. Under this definition, the model is said to possess a world model. Note that this notion is narrower than more general world models that capture temporal evolution or causal dynamics.

Beyond our narrow notion of world models in LLMs, future work could investigate whether LLMs capture temporal dynamics and causal relationships (Gurnee & Tegmark, 2024; Vafa et al., 2024; Rohekar et al., 2025) through ICL, how multi-task training affects representations (Park, 2026; Chawla et al., 2026; Kawata et al., 2025), how chain-of-thought (CoT) (Wei et al., 2022; Kojima et al., 2022; Matsutani et al., 2026a) affects behaviors, and whether similar structures emerge in weight-tied recurrent models (Geiping et al., 2025).

C Experimental details

C.1 Synthetic Task

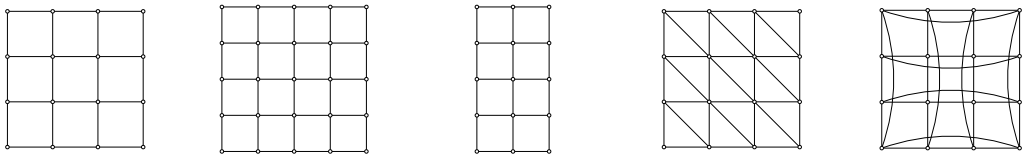

Figure 10: **Illustration of Five Graph Topologies.**

Graph Tracking. Following the *adaptive world modeling* setup of (Lepori et al., 2026), we use random walks generated on five graph topologies: 4×4 Lattice Grid, 5×5 Lattice Grid, 3×5 Lattice Grid, 4×4 Triangle Grid, and 4×4 Torus Grid Figure 10. On these graphs, we consider a two-step prediction task. Each random walk is generated by randomly sampling between 48 and 80 nodes. Note that no information about the underlying topology is provided in the prompt. The LLM must infer the topology solely from the random walk. Although Lepori et al. (2026) employed the 16-State Line and 25-State Line as latent graphs, we exclude them to rule out a shortcut in which the model simply copies the token from two positions earlier in the sequence.

For the two-step prediction task, following (Lepori et al., 2026), the model is asked to move from grid coordinate (i, j) to $(i + 2, j)$. Directly specifying an instruction such as "move two steps to the right" does not yield a unique answer under rotations or reflections of the grid. We therefore provide few-shot examples illustrating the mapping $(i, j) \to (i+2, j)$. We provide 6, 10, 3, 6, and 6 examples for the 4×4 Lattice Grid, 5×5 Lattice Grid, 3×5 Lattice Grid, 4×4 Triangle Grid, and 4×4 Torus Grid, respectively. For the downstream task, Lepori et al. (2026) considered one, two, and three-step prediction. We use two-step prediction because one-step prediction can be solved by copying an adjacent node from the random walk, whereas three-step prediction is substantially more difficult. We additionally ensure that the expected answer appears at least once in the random walk, and excluded query pairs whenever the correct target appeared exactly two positions before or after the query source in the sequence.

In this paper, under the few-shot setting, we treat each random walk together with its corresponding node pair examples for the task instruction as a single shot and repeat this format across multiple shots. Note that these shots are distinct from the few-shot examples used in the instruction for the two-step prediction task. Prompt examples are in Figure 11.

Prompt for Graph Tracking (Two-Step Prediction Task)

```
Complete the pattern.

[SEQUENCE]
avocado, mango, strawberry, raspberry, pineapple, banana, grape, banana,
lime, tomato, pear, apple, berry, apple, pear, prune, lime, banana, grape,
mango, cucumber, apple, berry, apple, cucumber, date, cucumber, mango,
strawberry, mango, strawberry, olive, avocado, date, berry, apple, berry,
apple, pear, tomato, cucumber, apple, pear, apple, cucumber, mango,
strawberry, raspberry

[EXAMPLE]
Input: lime Output: cucumber

[EXAMPLE]
Input: banana Output: mango

[EXAMPLE]
Input: pear Output: berry

[EXAMPLE]
Input: tomato Output: date

[EXAMPLE]
Input: pineapple Output: strawberry

[EXAMPLE]
Input: grape Output: avocado

[EXAMPLE]
Input: prune Output:
apple

Complete the pattern.

[SEQUENCE]
plum, tomato, plum, prune, plum, prune, plum, prune, plum, berry, lime,
cherry, fig, cherry, cucumber, berry, cucumber, cherry, fig, cherry, apple,
strawberry, apple, strawberry, fig, cherry, fig, banana, fig, cherry, lime,
cherry, lime, citrus, peach, lemon, pear, lemon, peach, citrus, peach,
citrus, lime, citrus, peach, citrus, lime, berry

[EXAMPLE]
Input: peach Output: apple

[EXAMPLE]
Input: lime Output: fig

[EXAMPLE]
Input: citrus Output: strawberry

[EXAMPLE]
Input: berry Output: banana

[EXAMPLE]
Input: prune Output: tomato

[EXAMPLE]
Input: lemon Output: cherry

[EXAMPLE]
Input: pear Output:
```

Figure 11: **Prompt for Graph Tracking Task.** A 1-shot example generated from a 4×4 Lattice Grid. The answer is ”cucumber”. The query corresponds to the illustration in Figure 1 (a).

C.2 Probing, Representation Analysis and Intervention Details

All probing, representation analysis, and intervention experiments described below are conducted on NVIDIA GH200 GPUs.

t denotes the token position index. l denotes the Transformer layer index. d denotes the hidden dimension. (i, j) denote grid coordinates in graph tracking tasks. Consider a specific sample, $x_t^{(l)} \in \mathbb{R}^d$ denotes the raw hidden state at token position t in layer l. In this paper, we use Qwen3-4B-Base, Qwen3-8B-Base and Qwen3-14B-Base (Yang et al., 2025a), where $d = 2560, 4096, 5120$ and $l = 36, 36, 40$, respectively.

Probing. $x_t^{(l)}$ is pooled over subtokens and over occurrences of the same token within the random walk in a query. To normalize these representations, we divide them by a single scalar Root Mean Square (RMS) computed over the activations of all nodes in that layer. We obtain a node representation $h_{i,j}^{(l)} \in \mathbb{R}^d$ where (i, j) denotes node indices. We train probe projection matrix $W^{(l)} \in \mathbb{R}^{d\times r}$ where r denotes the dimensionality of the latent probe space. We compute probe representation $z_{i,j}^{(l)} = {W^{(l)}}^\top h_{i,j}^{(l)} \in \mathbb{R}^r$, and then we obtain the probability that nodes $u = (i, j)$ and $v = (i', j')$ are adjacent $p_{u,v}^{(l)} = \sigma(\beta^{(l)} - \|z_u^{(l)} - z_v^{(l)}\|_2^2)$ where $\sigma(x) = (1 + \exp(-x))^{-1}$ is the logistic sigmoid function and $\beta^{(l)}$ is trainable scalar offset. Note that we train $W^{(l)}$ to obtain the probe space node representations $z_{i,j}^{(l)}$, while Dirichlet Energy in Park et al. (2025) is computed directly on the raw node representations $h_{i,j}^{(l)}$.

To evaluate probing performance, we compute the AUROC for predicting whether pairs of graph nodes are adjacent. AUROC is computed by sweeping the threshold over predicted probabilities and integrating the resulting ROC curve using the trapezoidal rule. Gröger et al. (2026) noted that probing accuracies are not directly comparable across layers with different dimensionalities, but we compare different activations within the same model.

We set $r = 2, 2, 2, 2, 4$ for the 4×4 Lattice Grid, 5×5 Lattice Grid, 3×5 Lattice Grid, 4×4 Triangle Grid, and 4×4 Torus Grid, respectively. The first four grids admit two-dimensional coordinate representations, motivating $r = 2$. For the 4×4 Torus Grid, we use the four-dimensional embedding $z(i, j) = (\cos(2\pi i/4), \sin(2\pi i/4), \cos(2\pi j/4), \sin(2\pi j/4))$, which represents each periodic coordinate using a sine and cosine pair. The squared Euclidean distance is 2 for adjacent nodes and at least 4 for non-adjacent nodes, motivating $r = 4$ for our probe based on distance. See Appendix E.2 for experiments with different ranks.

Representation Analysis. To obtain an orthonormal basis of this subspace, we compute the singular value decomposition

$$W^{(l)} = U^{(l)}\Sigma^{(l)}{V^{(l)}}^\top \in \mathbb{R}^{d\times r},$$

where $U^{(l)} \in \mathbb{R}^{d\times r}$, $\Sigma^{(l)} \in \mathbb{R}^{r\times r}$ and $V^{(l)} \in \mathbb{R}^{r\times r}$. Then, we can define world representation space $\mathcal{W}^{(l)} = \mathrm{span}(u_1^{(l)}, \ldots, u_r^{(l)})$, where $u_1^{(l)}, \ldots, u_r^{(l)}$ are columns of $U^{(l)}$. By using orthogonal projection matrix $\Pi_{\mathcal{W}}^{(l)} = U^{(l)}{U^{(l)}}^\top \in \mathbb{R}^{d\times d}$, we can separate in-subspace and out-of-subspace component from an activation $h^{(l)}$ as

$$h^{(l)} = \underbrace{\Pi_{\mathcal{W}}^{(l)} h^{(l)}}_{\text{in-subspace component}} + \underbrace{\left(I - \Pi_{\mathcal{W}}^{(l)}\right) h^{(l)}}_{\text{out-of-subspace component}},$$

where I is the identity matrix. We can use

$$0 \le \frac{\|\Pi_{\mathcal{W}}^{(l)} h^{(l)}\|_2^2}{\left\|h^{(l)}\right\|_2^2} \le 1$$

to quantify the extent (energy) to which the activation $h^{(l)}$ lies within the world representation subspace.

To quantify the overlap (similarity) of world representation subspaces $S(\mathcal{W}^{(l)}, \mathcal{W}'^{(l)})$ obtained from different probes at the same layer l, we compute

$$S(\mathcal{W}^{(l)}, \mathcal{W}'^{(l)}) = \frac{\|U^{(l)\top} U'^{(l)}\|_{\mathrm{F}}^2}{r} = \frac{\mathrm{tr}(\Pi_{\mathcal{W}}^{(l)} \Pi'^{(l)}_{\mathcal{W}'})}{r},$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Here, $U^{(l)}$ and $U'^{(l)}$ are obtained from two probes trained on different data at the same layer l. This quantity is the mean squared cosine of the principal angles between $\mathcal{W}^{(l)}$ and $\mathcal{W}'^{(l)}$, ranging from 0 for orthogonal subspaces to 1 for identical subspaces.

Similarly, we quantify the distance of world representation subspace $D(\mathcal{W}^{(l)}, \mathcal{W}'^{(l)})$ obtained from different probes at the same layer l by computing

$$D(\mathcal{W}^{(l)}, \mathcal{W}'^{(l)}) = \sqrt{1 - \frac{\|U^{(l)\top} U'^{(l)}\|_{\mathrm{F}}^2}{r}} = \frac{\|\Pi_{\mathcal{W}}^{(l)} - \Pi'^{(l)}_{\mathcal{W}'}\|_{\mathrm{F}}}{\sqrt{2r}}.$$

Additionally, let $\theta^{(l)}$ be the angle between the activation vector $h^{(l)}$ and the subspace $\mathcal{W}^{(l)}$,

$$\theta^{(l)} = \arccos\Big(\frac{\|\Pi_{\mathcal{W}}^{(l)} h^{(l)}\|_2}{\|h^{(l)}\|_2}\Big) \qquad \text{with} \quad 0 \le \theta^{(l)} \le \frac{\pi}{2}.$$

Intervention. A single low-rank probe may leave additional linearly decodable adjacency information. Following (Elazar et al., 2021), we repeat the probing and removal procedure.

At each layer and for each sample, the pooled representations of the nodes observed in the final query walk are centered by subtracting their mean. By writing the mean representation as $\bar{h}^{(l)}$, the graph intervention removes its component in $\mathcal{W}^{(l)}$ as

$$h_{i,j}^{(l)} \leftarrow h_{i,j}^{(l)} - \Pi_{\mathcal{W}}^{(l)}(h_{i,j}^{(l)} - \bar{h}^{(l)}).$$

Operationally, let $s^{(l)}$ denote the scalar RMS used to normalize the pooled representations. The vector $-s^{(l)}\Pi_{\mathcal{W}}^{(l)}(h_{i,j}^{(l)} - \bar{h}^{(l)})$ is added to $x_t^{(l)}$ at every selected subtoken position t in every occurrence of node (i, j) within the final query walk.

To compare with this low-rank graph intervention, we intervene random orthonormal directions to the world representation subspace in the same way. At each layer, we sample as many random orthonormal directions as the current dimension of $\mathcal{W}^{(l)}$. So, this random subspace has the same dimension as $\mathcal{W}^{(l)}$. We project the current centered $h_{i,j}^{(l)} - \bar{h}^{(l)}$ onto this orthogonal random subspace. This projected vector is rescaled separately for every node (i, j) and every layer l so that its norm equals $\|\Pi_{\mathcal{W}}^{(l)}(h_{i,j}^{(l)} - \bar{h}^{(l)})\|_2$. We then subtract the rescaled random vector from $x_t^{(l)}$ at the same token positions used by the graph intervention.

C.3 Practical Tasks

ARC-AGI-1&2. We use the public evaluation set of ARC-AGI-1 (Chollet, 2019). The set contains 400 tasks and 419 test pairs, where 381 tasks have one test pair and 19 have two. For ARC-AGI-2 (Chollet et al., 2025), we use evaluation set with 120 tasks and 167 test pairs, with 75 tasks containing one test pair, 43 containing two, and 2 containing three. Each test pair was evaluated with an independent prompt, while all native input–output training pairs belonging to its target task were retained. A single shot consists of the task instruction and a set of input–output pairs provided as observations in a serialized format. We construct a few-shot prompt by concatenating these shots, using demonstrations from the training set and the query from the public evaluation set. Prompt examples are in Figure 12.

Prior work has approached ARC-AGI using leave-one-out test-time training (TTT) within a task (Akyürek et al., 2025), a combination of data augmentation and test-time search (Franzen et al., 2025), Vision Transformers (ViTs) that formulate ARC as an image-to-image translation problem (Hu et al., 2025), and weight-tied recurrent models (Shu et al., 2026; Liu et al., 2026b).

Web Agent Task. We use the cross-website subset of the Mind2Web dataset (Deng et al., 2023), which contains 177 tasks and 1373 action steps. Each query prompt contained the user's goal, gold previous action history, the HTML of the current webpage, and the top-10 elements from the ranking. Few-shot demonstrations were drawn from the disjoint train set, whereas queries were drawn from the cross-website subset of the test set, since we need to ensure that the world (website) of the query is different from that of the demonstrations. Prompt examples are in Figure 13.

For ARC-AGI-1&2, we use top-p sampling with temperature 0.7 and top-p 0.95. For Othello and Mind2Web (cross-website subset), we use greedy decoding with temperature 0 and top-p 1. For all tasks, we generate the answer immediately after prefill in the specified format with a sufficient response length.

Othello. Following Li et al. (2023); Nanda et al. (2023), in the Othello task, the model is given a sequence of 8–52 moves, where each move is represented by a board coordinate such as "A5" or "E3" and corresponds to a move by either MINE or YOURS. The model is then asked to predict the next legal move. Although the moves are provided as board coordinates, the model is not explicitly told that they are Othello moves. It must therefore infer the current board state from the history of observed moves in order to predict a legal next move. Prompt examples are in Figure 14

Prompt for ARC-AGI-1&2

```
You are participating in a puzzle solving competition. You are an expert at
solving puzzles.

Below is a list of input and output pairs with a pattern. Your goal is to
identify the pattern or transformation in the training examples that maps the
input to the output, then apply that pattern to the test input to give a
final output.

Respond in the format of the training output examples

--Training Examples--
--Example 0--

 INPUT:

[[0, 0, 5], [0, 5, 0], [5, 0, 0]]

OUTPUT:

[[3, 3, 3], [4, 4, 4], [2, 2, 2]]

--Example 1--

 INPUT:

[[0, 0, 5], [0, 0, 5], [0, 0, 5]]

OUTPUT:

[[3, 3, 3], [3, 3, 3], [3, 3, 3]]

[SKIP]

--End of Training Examples--

--Test Input--
[[0, 0, 5], [5, 0, 0], [0, 5, 0]]
--End of Test Input--
Your response:
[[3, 3, 3], [2, 2, 2], [4, 4, 4]]

You are participating in a puzzle solving competition. You are an expert at
solving puzzles.

Below is a list of input and output pairs with a pattern. Your goal is to
identify the pattern or transformation in the training examples that maps the
input to the output, then apply that pattern to the test input to give a
final output.

Respond in the format of the training output examples

--Training Examples--
--Example 0--

 INPUT:

[[1, 0, 3, 4], [0, 0, 2, 1], [2, 1, 4, 0], [0, 3, 1, 2]]

OUTPUT:

[[1, 2, 3, 4], [3, 4, 2, 1], [2, 1, 4, 3], [4, 3, 1, 2]]

[SKIP]

--End of Training Examples--

--Test Input--
[[0, 1, 2, 3], [0, 3, 1, 0], [3, 0, 4, 1], [0, 4, 0, 2]]
--End of Test Input--
Your response:
```

Figure 12: **Prompt for ARC-AGI-1&2.** A 1-shot example. The part marked **[SKIP]** is omitted for brevity.

Prompt for Web Agent Task

```
You predict the next action on a webpage. Follow the examples exactly. Return
only `Answer: <letter>`, `Action: <CLICKTYPESELECT>`, and a `Value:` line when
TYPE or SELECT requires one.

### Example 1
'''
<html> <div> <ul> <div id=0 button> <h3> Year Manufactured </h3> </div> <div
id=1 button> <h3> Price </h3> </div> </ul> <div main> <ul> <li id=2> <a id=3>
<span> Price + Shipping: lowest first </span> </a> </li> <a id=4> <span>
Price + Shipping: highest first </span> </a> </ul> </div> </div> </html>
'''

Based on the HTML webpage above, try to complete the following task:
Task: Get a Hasbro Hulk action figure manufactured in 1990 with the lowest
price + shipping.
Previous actions:
[tab]  Year Manufactured -> CLICK
[checkbox]  1990 -> CLICK
[button]  Apply -> CLICK
[input]   -> CLICK
[button]  Sort selector. Best Match selected. -> CLICK
What should be the next action? Please select from the following choices (If
the correct action is not in the page above, please select A. 'None of the
above'):

A. None of the above
B. <div id=0 button> <h3> Year Manufactured </h3> </div>
C. <div id=1 button> <h3> Price </h3> </div>
D. <li id=2> <a id=3> <span> Price + Shipping: lowest first
E. <a id=3> <span> Price + Shipping: lowest first </span> </a>
F. <a id=4> <span> Price + Shipping: highest first </span> </a>

Answer: E.
Action: CLICK

### Query
'''
<html> <div> <div> <a id=0> Featured </a> <div> <button button match-up
scores> <svg id=1 /> </button> <button button match-up scores> <svg id=2 />
</button> </div> </div> <div> <button button carousel button> <svg id=3 />
</button> <button button carousel button> <svg id=4 /> </button> </div>
</div> </html>
'''

Based on the HTML webpage above, try to complete the following task:
Task: Find the current league leader in Assists Per Game.
Previous actions:
None
What should be the next action? Please select from the following choices (If
the correct action is not in the page above, please select A. 'None of the
above'):

A. None of the above
B. <a id=0> Featured </a>
C. <svg id=1 />
D. <svg id=2 />
E. <svg id=3 />
F. <svg id=4 />

Answer:
```

Figure 13: **Prompt for Web Agent Task.** A 1-shot example of web agent tasks from Mind2Web dataset. The answer is ”F” and ”CLICK”.

Prompt for Othello

```
Infer the answer token from the preceding token sequence.

Example 1:
history: C4 C3 F5 F6 D3 C5 C6 E6 B6 E3 D2 B7 B8 C1 D6 D7 C2 B3 B4 G6 E7 A5 A4
B1 E2 F4 E1 A2
answer: B5

Example 2:
history: D3 C3 E6 E3 E2 D6 C6 F4 C4 B4 B5 E7 C5 B6 A6 C2 F5 F6 C7 F1 A4 F2 C1
A3 A2 D2 B3 B1 F3 C8 B7 A5 G4 H5 E1 A7 G1 D7 B8 D1 G2 G3
answer: G5

Query:
history: E6 F6 F5 F4 G4 G6 G5 E7 C4 C5 B6 H3 E8 B4 G3 F7 A3 F3 H5 G8 H2 B5 E2
A4 F8 G7 D3 E3 D2 F1 F2 D8 A6 C6 H7 C2 B1 G2 G1 H1 B3 H6 D6 A2 B2 C1 D7 A1 C3
C8 C7 B8 H4
answer:
```

Figure 14: **Prompt for Othello.** A 2-shot example of classical 8×8 Othello. The answer is "A5".

D Detailed Results

D.1 Few-Shot Performance

In addition to Figure 3, we show that mixing graph topologies different from that of the query in the demonstrations also improves performance. To evaluate this setting, we randomly sample the topology of each demonstration from four graphs other than the graph of the query.

Figure 15 shows few-shot performance under the settings described above. The graph in each title indicates the query graph, while the demonstrations are sampled from the other graphs. Even under this harder setting, where the demonstrations come from different worlds, few-shot demonstrations improve task performance.

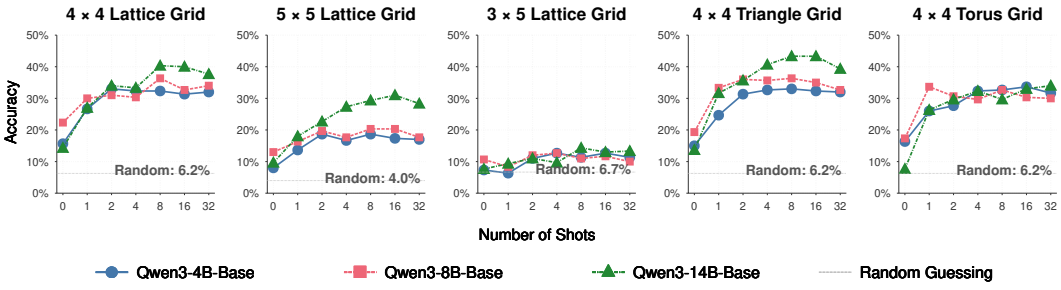

Figure 15: **Few-Shot Performance where Demonstrations are from Different Graph Topologies.** Accuracy of three Qwen3 models on the two-step-right prediction task across five graph topologies. The demonstrations are randomly sampled from four topologies other than the query topology, with each topology represented equally when the number of shots is 4, 8, 16, or 32. We report performance separately for each graph topology in the query. Each point reports accuracy on 300 test samples with 0, 1, 2, 4, 8, 16, or 32 few-shot demonstrations. Dotted gray lines indicate random guessing over the candidate nodes for each topology.

D.2 Probing and Representation Analysis

For Qwen3-4B-Base, Qwen3-8B-Base, and Qwen3-14B-Base, we report detailed probing results in Figure 16, cross-probe performance in Figure 18, the overlap of the world representation subspaces in Figure 19, PCA of the graph representations in Figure 20, and PCA of the graph representation centroids over all samples in the test set in Figure 21.

Additionally, to examine how the representation structure of the same query, specifically the relative positions of nodes, changes with the number of shots, we computed the similarity of node positions

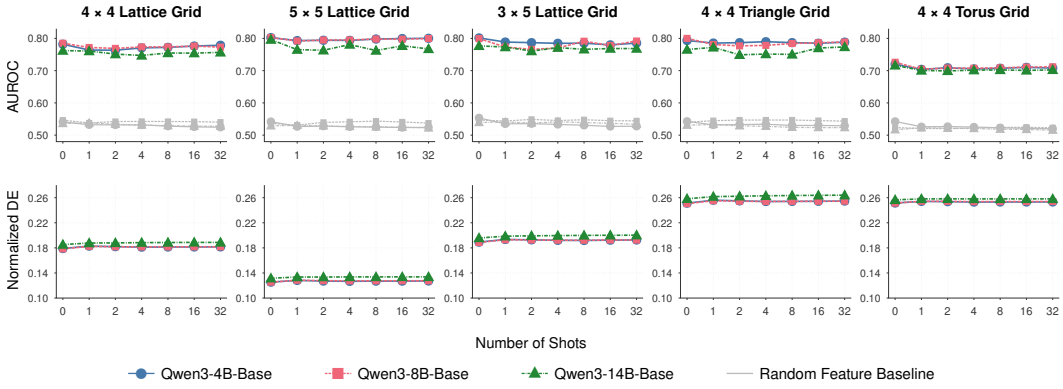

Figure 16: **Probe Performance and Normalized Dirichlet Energy (DE).** Probe AUROC (Top) and normalized DE of hidden states (Bottom) across shots and five graph topologies for layer 20 of three Qwen3 models. Gray lines show random feature probe baselines. Each condition is evaluated on 300 test samples.

between the 0-shot setting and the 1, 2, 4, 8, 16, and 32-shot settings. For each comparison between 0-shot and a given number of shots, we compute the Euclidean distances between all pairs of common nodes and normalize each distance vector to unit length. We then compute the squared Euclidean difference between the normalized vectors, average it across comparisons, and take the square root. Smaller values indicate greater preservation of the relative distance structure among nodes. Note that this metric captures the relative positions of nodes and is invariant to translation and rotation.

Figure 17 suggests that, as long as the query remains the same, increasing the number of shots does not substantially change the relative positions of nodes in the graph representation. The similarity to the 0-shot setting remains nearly constant as the number of shots increases, and the variation is small compared with the case where different fruit vocabularies are assigned to nodes with the same topology. The PCA visualization on the left of Figure 6 also visualizes that the relative node configuration remains largely unchanged.

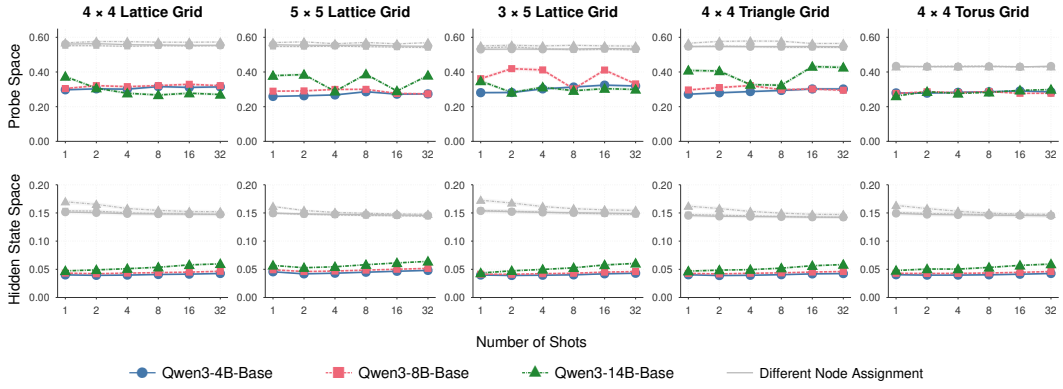

Figure 17: **Node Distance Geometry Relative to Zero-Shot Setting.** Node distance geometry relative to zero-shot setting using probe projected representations (Top) and full hidden states (Bottom) across shots and five graph topologies for layer 20 of three Qwen3 models. Colored lines compare the same query across shots, while gray lines compare different query families with the same topology but different node vocabulary assignments. Lower values indicate more similar relative node geometry. Each condition uses 300 test samples.

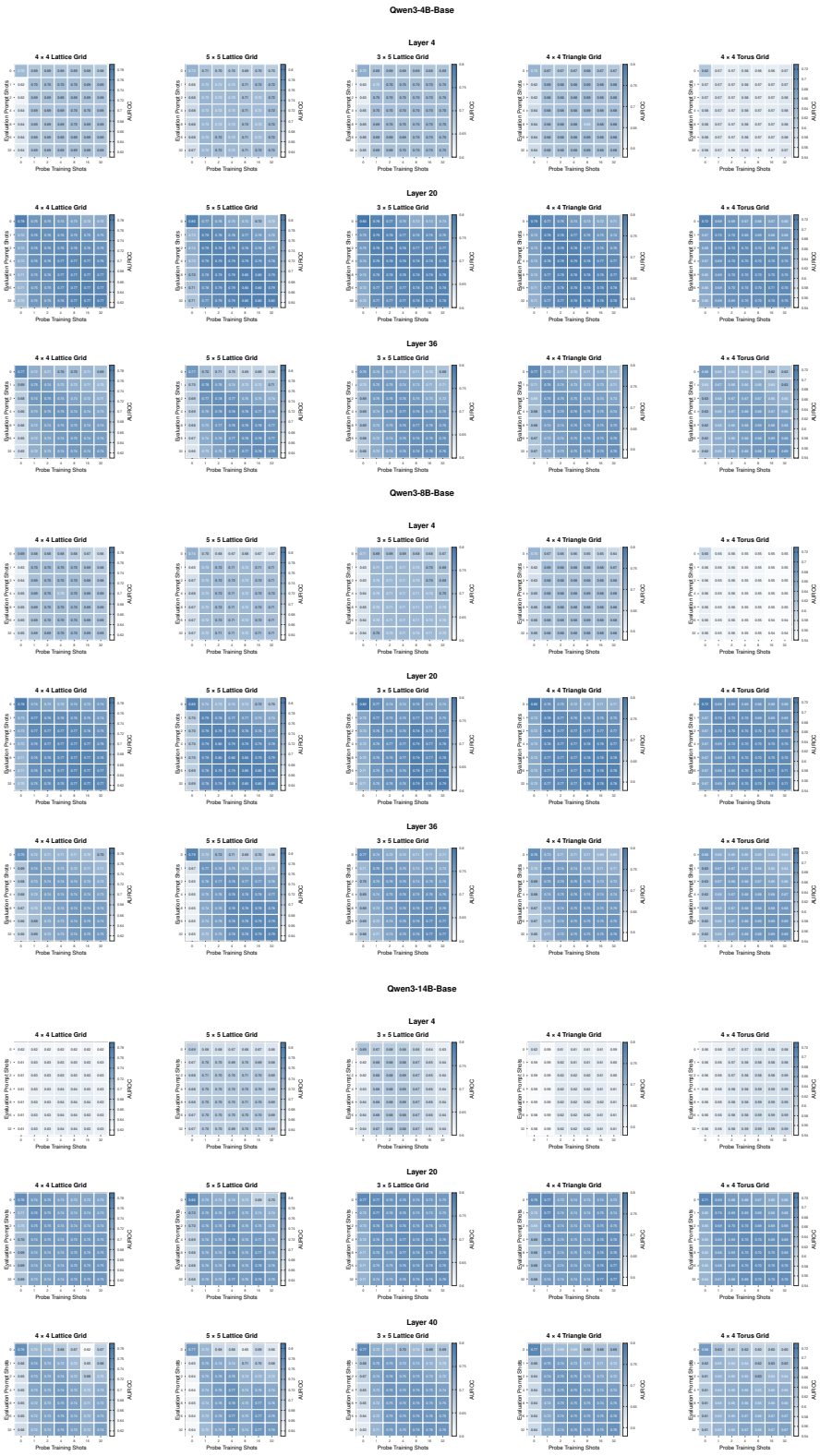

Figure 18: **Cross Probe Performance.** AUROC of probes trained at one shot count and evaluated at another. Results are from three Qwen3 models on five graph topologies.

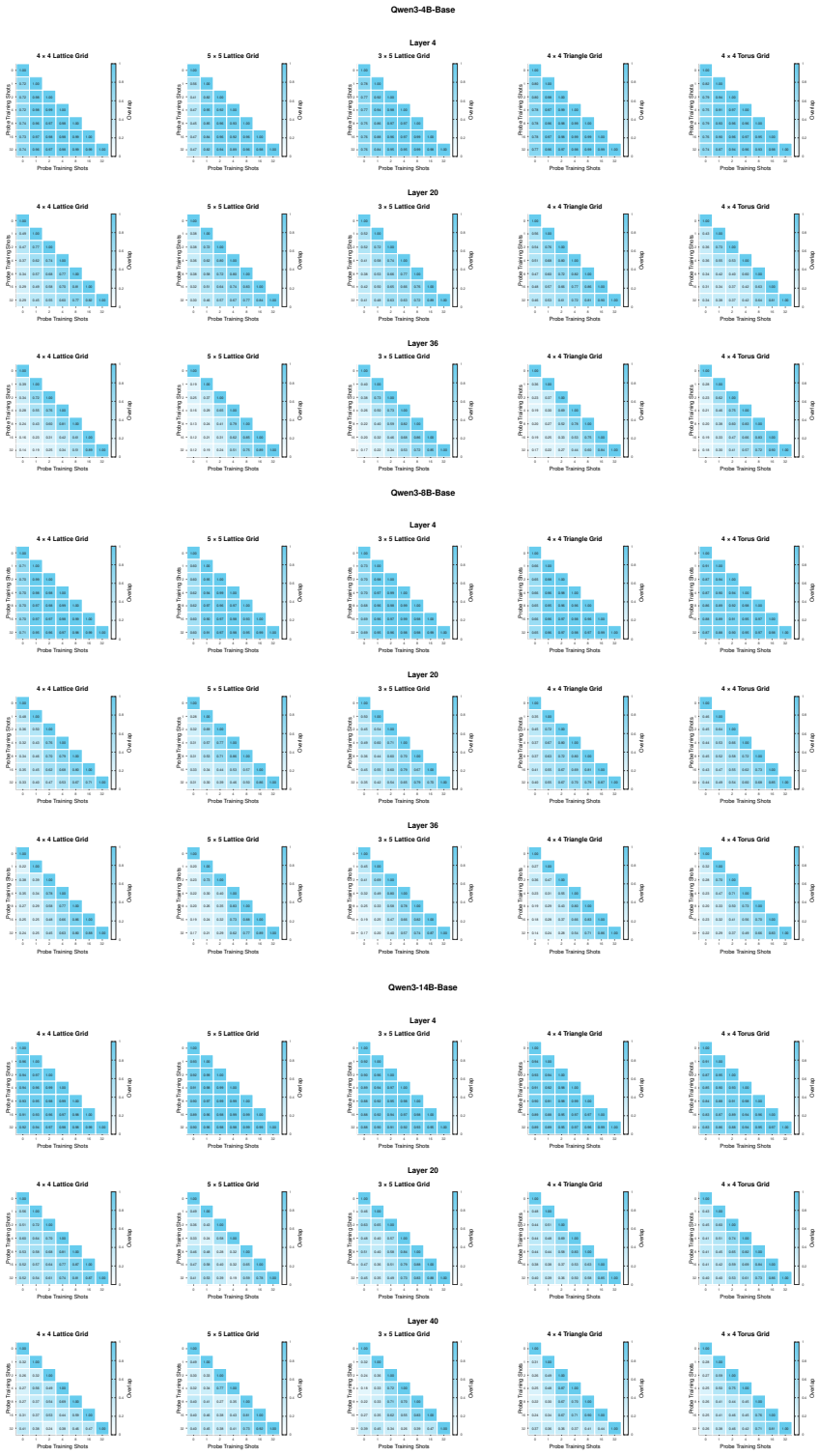

Figure 19: **Representation Subspace Overlap Results.** Overlap between probe subspaces across shot counts. Results are from three Qwen3 models on five graph topologies.

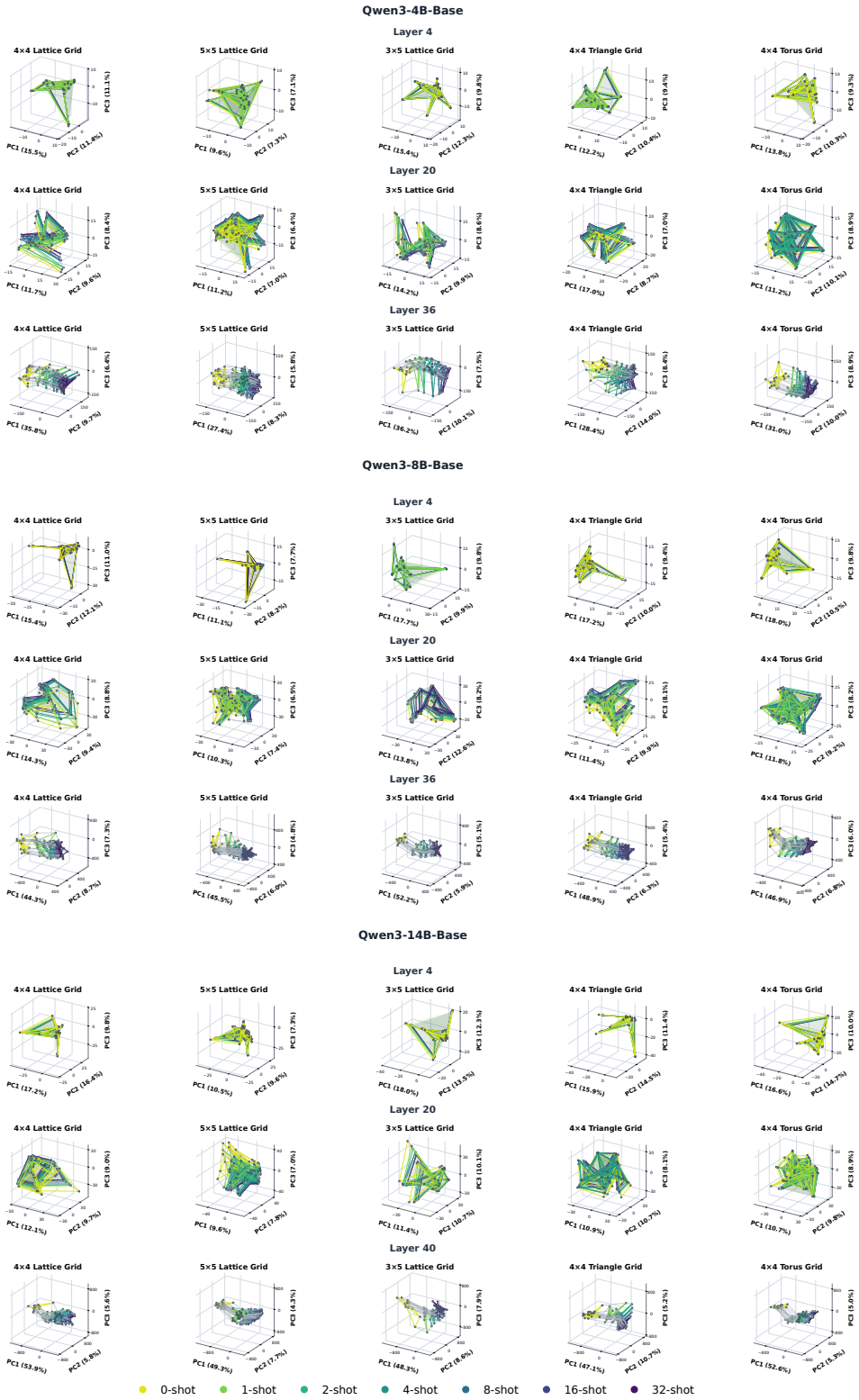

Figure 20: **PCA of Shift in Hidden States of a Random Walk Across Different Number of Shots.** In each panel, PCA is fit jointly to all activations shown. Results are from three Qwen3 models on five graph topologies.

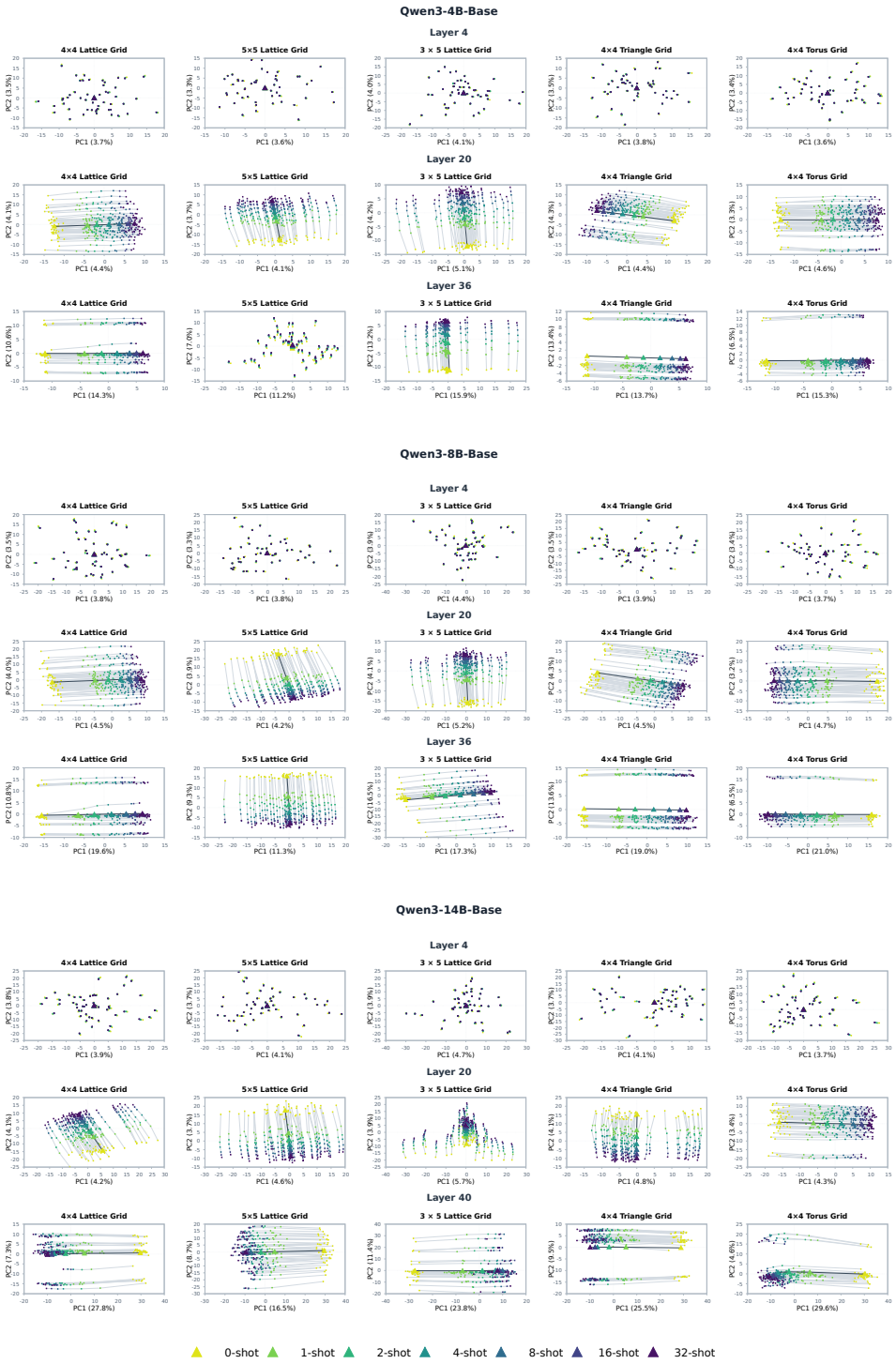

Figure 21: **PCA of Shifts in Hidden-State Centroids across Random Walks Across Different Number of Shots.** PCA of 300 test samples per shot. Points represent individual random walks and triangles indicate their means. In each panel, PCA is fit jointly to all activations shown. Results are from three Qwen3 models on five graph topologies.

D.3 SHOT VECTOR

We define the Shot Vector $\Delta\bar{h}^{(l)}$ as the mean difference between few shot and zero shot hidden states for corresponding nodes in the query random walk, computed over the test set. Figure 22 reports detailed results on the overlap between the Shot Vector and the World Representation Subspace. We also show in Figure 23 that the Task Vector (Hendel et al., 2023), defined as the few shot hidden state immediately before the answer ("\n" in our graph tracking task), is distinct from the Shot Vector and nearly orthogonal to it. Note that Hendel et al. (2023) define the Task Vector for tasks with input–output pairs, including algorithmic, translation, linguistic, and knowledge tasks. This differs from our few-shot setting, where each demonstration consists of an observation, few-shot examples as an instruction, and an answer. Nevertheless, we extract the Task Vector using the same definition.

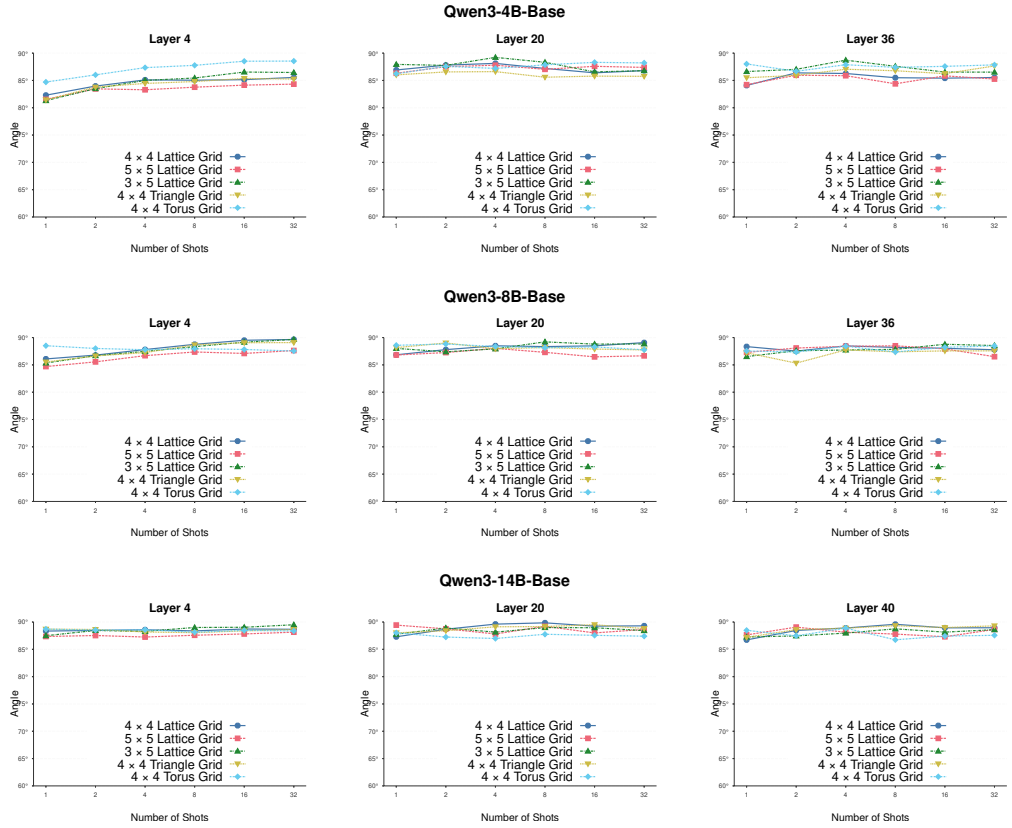

Figure 22: **Angle Between the World Representation Subspace and the Shot Vector.** Angles to the corresponding probe subspaces, shown separately for each topology. Results are from three Qwen3 models on five graph topologies.

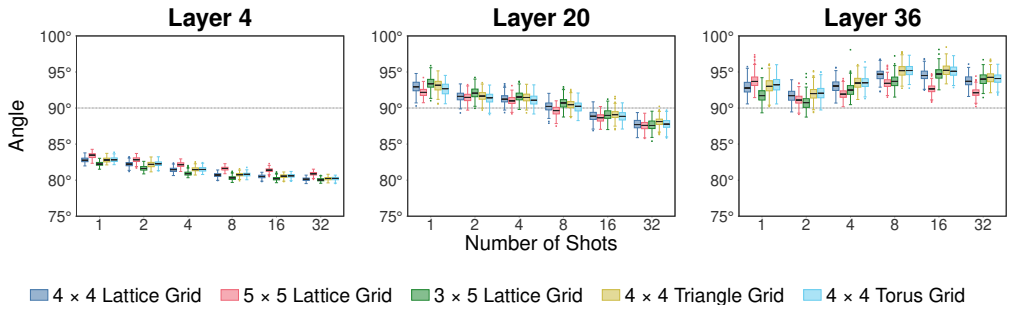

Figure 23: **Angle Between the Shot Vector and the Task Vector.** The angle between the two vectors is in degrees from $0°$ to $180°$. Results from Qwen3-4B-Base on five topologies.

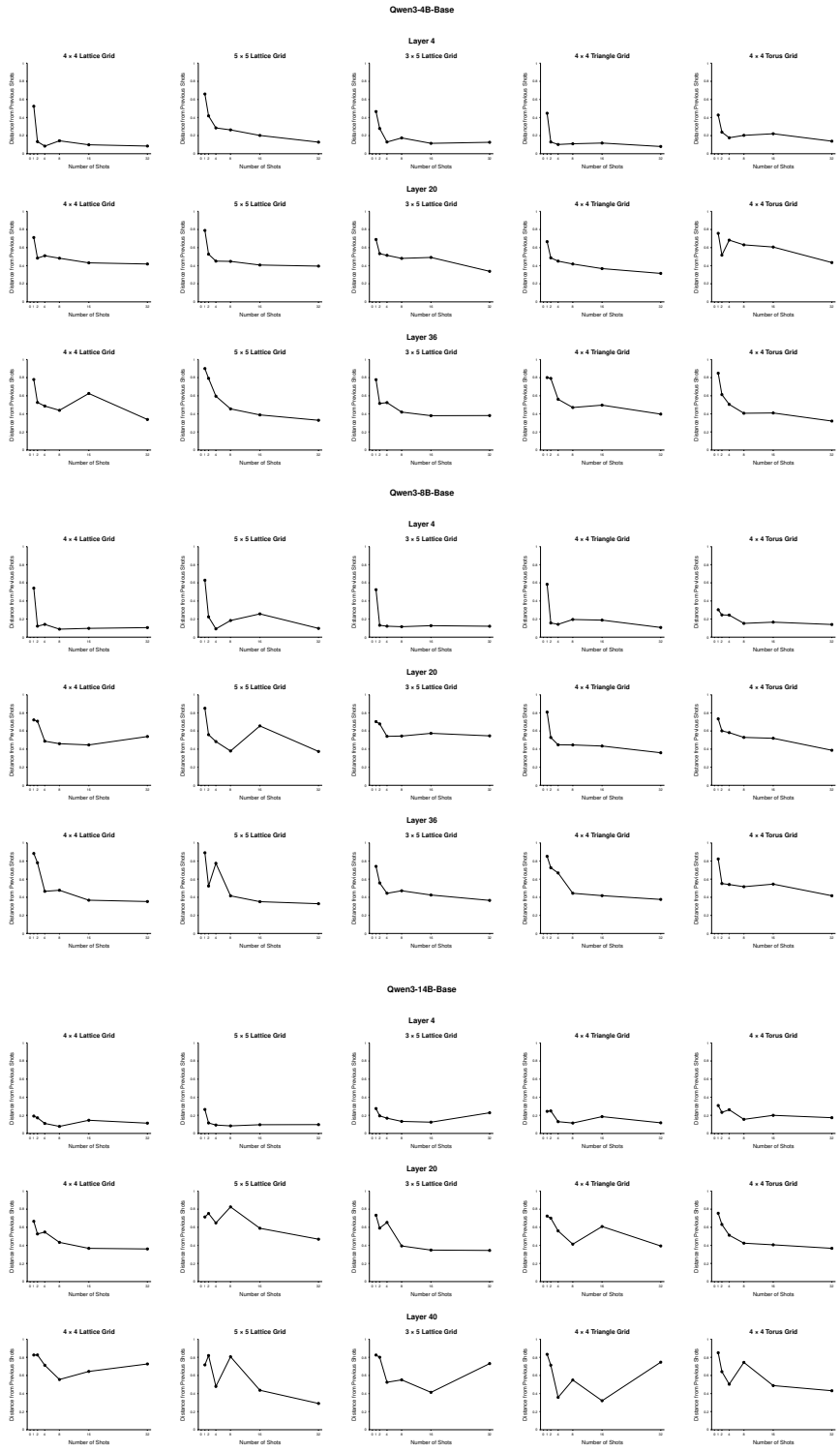

Figure 24: **Representation Subspace Distance Results.** Normalized projection distance between probe subspaces at consecutive tested shot counts, shown separately for each topology. Results are from three Qwen3 models on five graph topologies.

D.4 INTERVENTION

We report the degradation in probability of the correct answer token caused by graph intervention and random intervention across five topologies for three Qwen3 models in Figure 25. Graph intervention tends to cause larger degradation in the 16-shot setting than in the 0-shot setting, whereas random intervention shows the opposite trend.

Test probe AUROC Figure 26 also gradually decreases over the intervention iterations. Our intervention removes graph information by subtracting the component of each centered node representation that lies in the learned world representation subspace. Since we intervene on a very low dimensional subspace with $r = 2, 4$ relative to the ambient dimension, the effect on the representation is limited. We therefore apply the intervention multiple times, following (Elazar et al., 2021).

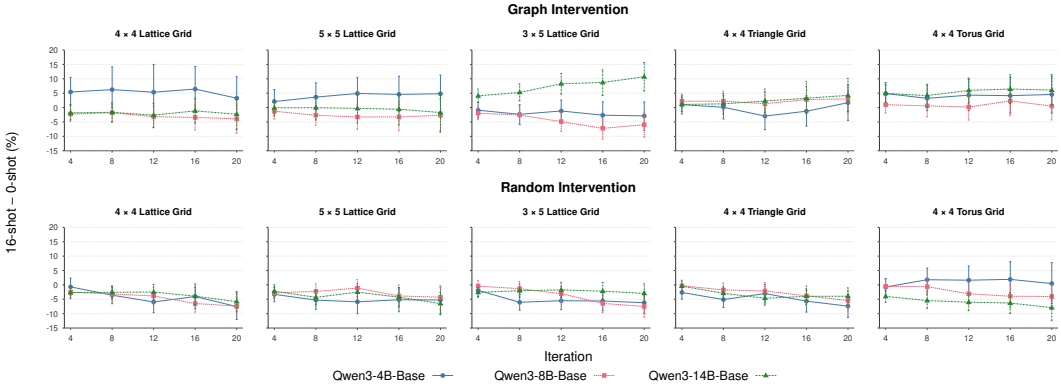

Figure 25: **Intervention Results across Intervention Iterations.** Plotted points represent the average paired difference in this degradation between the 16-shot and 0-shot settings across 300 samples in test set. The top is iterative nullspace projection targeted to graph structure, and the bottom is random orthogonal subspace intervention matched in rank. For each prompt condition, relative degradation is computed per test sample as the drop in the model's assigned probability for the correct answer after intervention, normalized by the probability before intervention. Positive values indicate that intervention causes a larger relative performance drop under 16-shot prompting than under 0-shot setting.

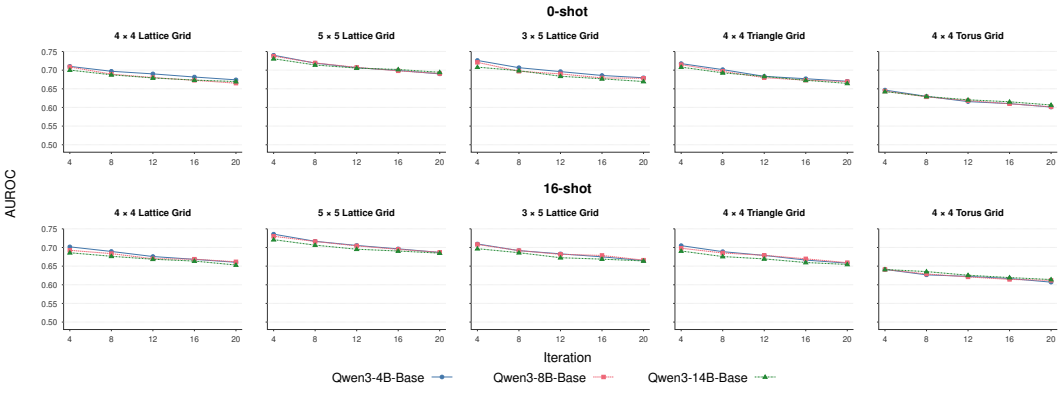

Figure 26: **Probe Performance during Intervention Iterations.** At each cumulative intervention step (4, 8, 12, 16, and 20), a new linear probe is trained on the intervened residual stream activations and evaluated on a held-out test set. Plotted points indicate the average of test AUROC across the layers 2–36 for Qwen3-4B-Base and Qwen3-8B-Base, and layers 2–40 for Qwen3-14B-Base.

E Additional Experiments

E.1 Experiments on Llama, Gemma, and Ministral

In addition to the results for Qwen3-4B-Base, Qwen3-8B-Base, and Qwen3-14B-Base (Yang et al., 2025a), we report the few-shot performance of Llama-3.2-3B (Grattafiori et al., 2024), Gemma-4-12B (Team et al., 2026), and Ministral-3-8B-Base-2512 (Liu et al., 2026a) across the five topologies in Figure 27. We further report the results of our probing and representation analyses on the 4×4 lattice grid in Figure 28, and the intervention results in Figure 29. Across all of these experiments, we observe results consistent with the results from Qwen3.

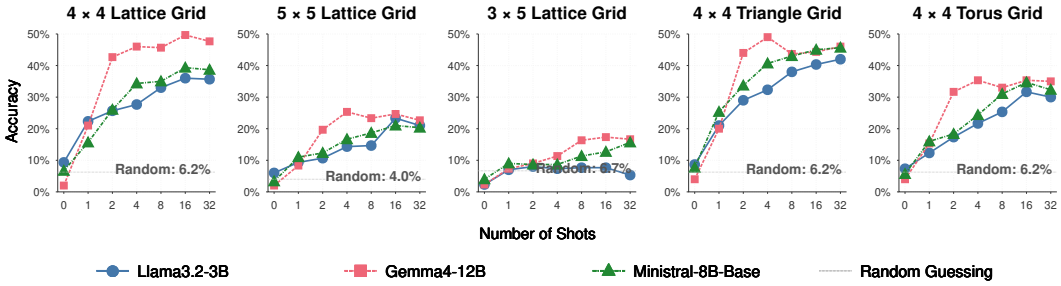

Figure 27: **Few-Shot Performance.** Accuracy of Llama-3.2-3B, Gemma-4-12B, and Ministral-3-8B-Base-2512 on the two-step-right prediction task across five graph topologies. Each point reports accuracy on 300 test samples with 0, 1, 2, 4, 8, 16, or 32 few-shot demonstrations. Dotted gray lines indicate random guessing over the candidate nodes for each topology.

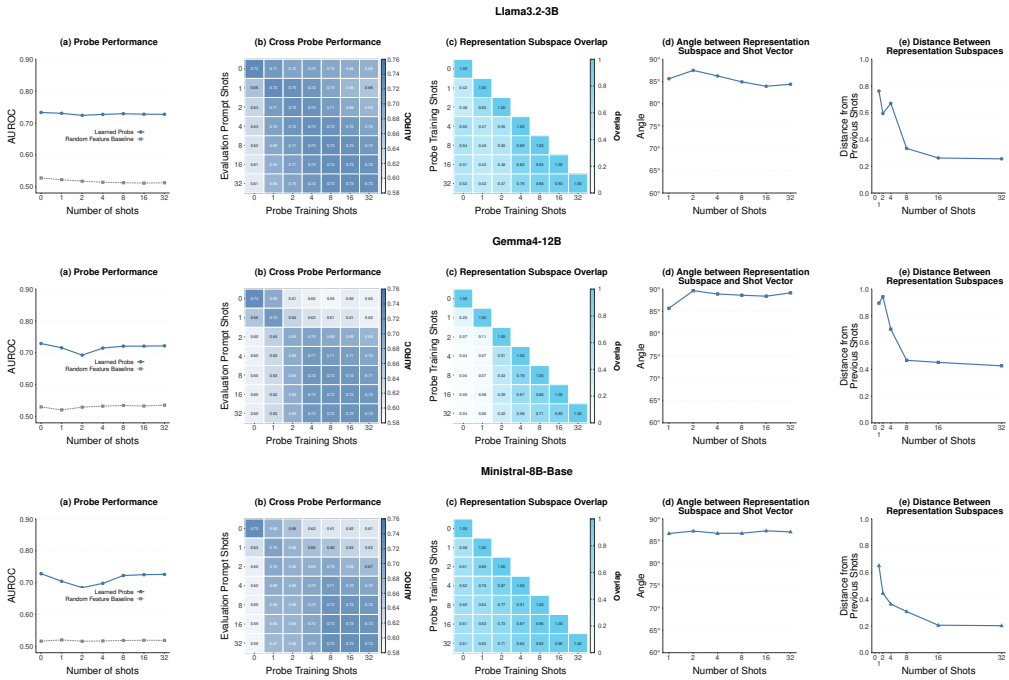

Figure 28: **(a) Probe Performance.** Probe AUROC. Gray lines show random feature probe baselines. **(b) Cross Probe Performance.** AUROC of probes trained at one shot count and evaluated at another. **(c) World Representation Subspace Overlap.** Overlap between probe subspaces across shot counts. **(d) Angle Between Shot Vectors and the World Representation Subspace.** Angles to the corresponding probe subspaces. **(e) Distance between Representation Subspaces across Shots.** Normalized projection distance between probe subspaces at consecutive tested shot counts. Results are from layer 28 of Llama-3.2-3B, layer 48 of Gemma-4-12B, and layer 34 of Ministral-3-8B-Base-2512 on 4×4 Lattice Grid.

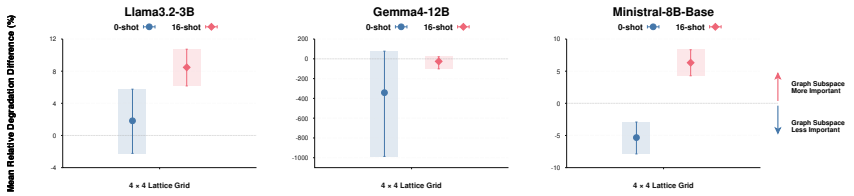

Figure 29: **Intervention Results.** Graph subspace erasure at iteration 20 in Llama-3.2-3B, Gemma-4-12B, and Ministral-3-8B-Base-2512 on 4×4 Lattice Grid. Interventions target node subtokens in the final query's walk at every Transformer block output. The random intervention uses an orthogonal subspace with matched rank and perturbation norm per node. Points show the mean difference in correct answer probability reduction between graph and random interventions, normalized by each query's unperturbed probability, across 300 test samples. Probabilities are normalized over candidate answer sequence likelihoods. Positive values indicate greater degradation from graph erasure. Horizontal bars show 95% bootstrap confidence intervals obtained by resampling test samples 5000 times. Note that the wide confidence intervals for 0-shot Gemma-4-12B are because of near-zero baseline probabilities in the denominator.

E.2 Experiments on Different Ranks

For probing, we set the rank r of the probe layer $W^{(l)} \in \mathbb{R}^{d \times r}$ to $2, 2, 2, 2$, and 4 for the 4×4 Lattice Grid, 5×5 Lattice Grid, 3×5 Lattice Grid, 4×4 Triangle Grid, and 4×4 Torus Grid, respectively. We use $r = 4$ for the 4×4 Torus Grid because encoding its two periodic coordinates with sine and cosine pairs yields a four-dimensional embedding that exactly separates adjacent and non-adjacent nodes by a Euclidean distance threshold. We use $r = 2$ for the other grids because they admit two-dimensional coordinate representations. We additionally evaluate higher ranks, $r = 8$ and 16, on the 4×4 Lattice Grid with Qwen3-4B-Base. Figures 30 and 31 report the probing and representation analysis results, and Figure 32 reports the intervention results.

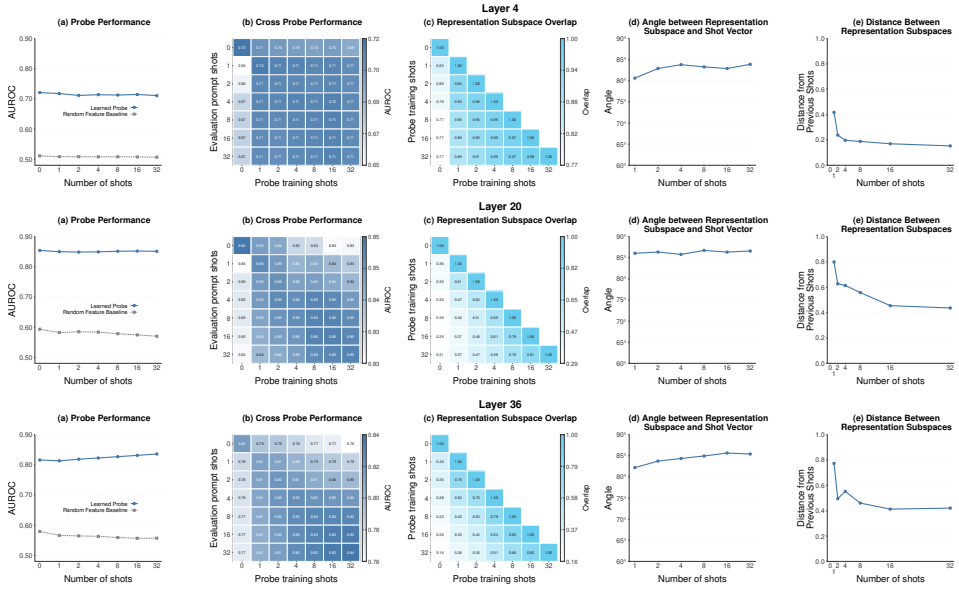

Figure 30: **(a) Probe Performance.** Probe AUROC. Gray lines show random feature probe baselines. **(b) Cross Probe Performance.** AUROC of probes trained at one shot count and evaluated at another. **(c) World Representation Subspace Overlap.** Overlap between probe subspaces across shot counts. **(d) Angle Between Shot Vectors and the World Representation Subspace.** Angles to the corresponding probe subspaces. **(e) Distance between Representation Subspaces across Shots.** Normalized projection distance between probe subspaces at consecutive tested shot counts. Results are from Qwen3-4B-Base on 4×4 Lattice Grid, where $r = 8$.

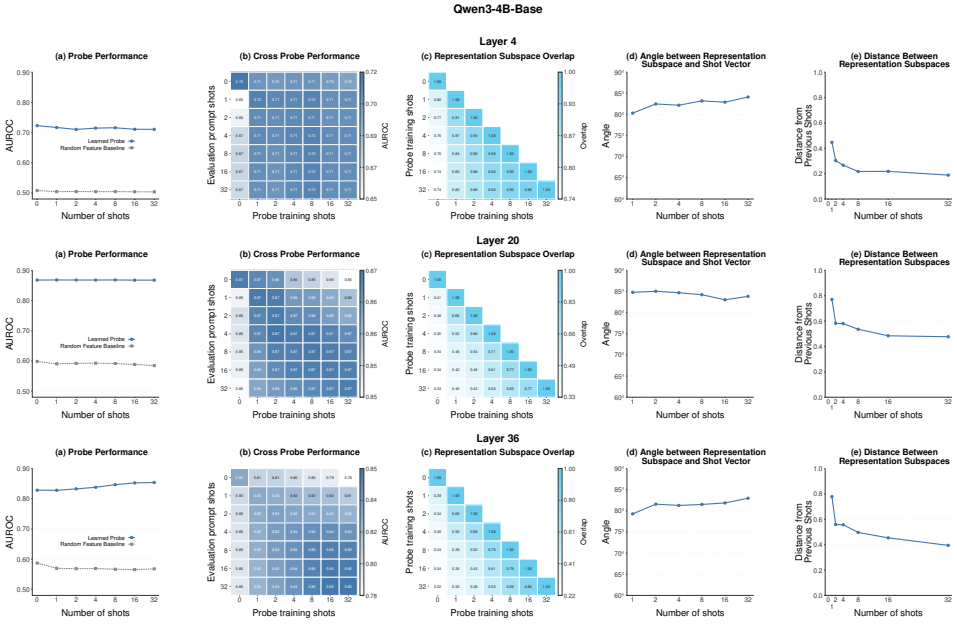

Figure 31: **(a) Probe Performance.** Probe AUROC. Gray lines show random feature probe baselines. **(b) Cross Probe Performance.** AUROC of probes trained at one shot count and evaluated at another. **(c) World Representation Subspace Overlap.** Overlap between probe subspaces across shot counts. **(d) Angle Between Shot Vectors and the World Representation Subspace.** Angles to the corresponding probe subspaces. **(e) Distance between Representation Subspaces across Shots.** Normalized projection distance between probe subspaces at consecutive tested shot counts. Results are from Qwen3-4B-Base on 4×4 Lattice Grid, where $r = 16$.

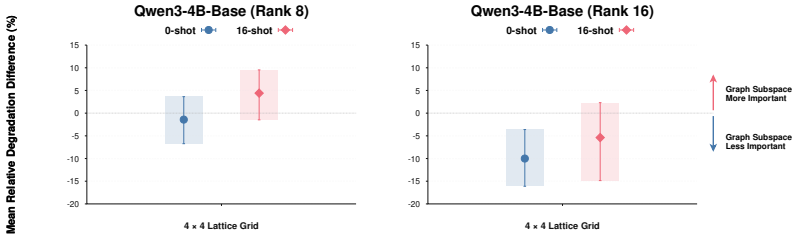

Figure 32: **Intervention Results.** Graph subspace erasure at iteration 4 in Qwen3-4B-Base on 4×4 Lattice Grid, where $r = 8, 16$. Interventions target node subtokens in the final query's walk at every Transformer block output. The random intervention uses an orthogonal subspace with matched rank and perturbation norm per node. Points show the mean difference in correct answer probability reduction between graph and random interventions, normalized by each query's unperturbed probability, across 300 test samples. Probabilities are normalized over candidate answer sequence likelihoods. Positive values indicate greater degradation from graph erasure. Horizontal bars show 95% bootstrap confidence intervals obtained by resampling test samples 5000 times.

E.3 Sensitivity to Random Seed

To quantify variability across seeds, we report results over five seeds for probing, representation analysis, and intervention on the 4×4 Lattice Grid with Qwen3-4B-Base.

Table 1: **Probe Performance.** Results from Qwen3-4B-Base evaluated on 300 test samples on 4×4 Lattice Grid. "L4, 0-Shot" denotes the probe performance at Layer 4 with 0-shot prompting.

Setting	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean $\pm$ Std
L4, 0-Shot	0.6903	0.7072	0.7060	0.7048	0.7013	0.7019 ± 0.0069
L4, 16-Shot	0.6867	0.6824	0.7046	0.6763	0.6941	0.6888 ± 0.0109
L20, 0-Shot	0.7812	0.7761	0.7792	0.7784	0.7777	0.7785 ± 0.0019
L20, 16-Shot	0.7664	0.7573	0.7591	0.7688	0.7722	0.7647 ± 0.0064
L36, 0-Shot	0.7675	0.7508	0.7629	0.7591	0.7656	0.7612 ± 0.0066
L36, 16-Shot	0.7407	0.7359	0.7419	0.7406	0.7412	0.7401 ± 0.0024

Table 2: **Cross Probe Performance.** Results from Qwen3-4B-Base on 4×4 Lattice Grid. "L4, $0 \rightarrow$ 16-Shot" denotes the Layer 4 probe trained on 500 0-shot samples and evaluated on 300 16-shot samples.

Setting	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean $\pm$ Std
L4, $0 \rightarrow$ 16-Shot	0.6406	0.6487	0.6525	0.6483	0.6425	0.6465 ± 0.0049
L4, $16 \rightarrow$ 32-Shot	0.6857	0.6821	0.7036	0.6755	0.6942	0.6882 ± 0.0109
L20, $0 \rightarrow$ 16-Shot	0.7191	0.7112	0.7028	0.7136	0.7061	0.7106 ± 0.0064
L20, $16 \rightarrow$ 32-Shot	0.7669	0.7580	0.7592	0.7692	0.7731	0.7653 ± 0.0065
L36, $0 \rightarrow$ 16-Shot	0.6329	0.6727	0.6551	0.6641	0.6538	0.6557 ± 0.0148
L36, $16 \rightarrow$ 32-Shot	0.7428	0.7375	0.7439	0.7436	0.7415	0.7419 ± 0.0026

Table 3: **Representation Subspace Overlap.** Results from Qwen3-4B-Base on 4×4 Lattice Grid. "L4, $0 \leftrightarrow$ 16-Shot" denotes the overlap between the rank-2 representation subspaces of probe layers trained on 500 samples.

Setting	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean $\pm$ Std
L4, $0 \leftrightarrow$ 16-Shot	0.6295	0.8563	0.7353	0.7794	0.7285	0.7458 ± 0.0826
L4, $16 \leftrightarrow$ 32-Shot	0.9804	0.9728	0.9681	0.9238	0.9928	0.9676 ± 0.0262
L20, $0 \leftrightarrow$ 16-Shot	0.3329	0.4381	0.3081	0.3230	0.2904	0.3385 ± 0.0580
L20, $16 \leftrightarrow$ 32-Shot	0.7970	0.7931	0.8709	0.8691	0.8247	0.8310 ± 0.0376
L36, $0 \leftrightarrow$ 16-Shot	0.1776	0.2669	0.1812	0.1554	0.1601	0.1882 ± 0.0453
L36, $16 \leftrightarrow$ 32-Shot	0.8654	0.8654	0.8993	0.8901	0.8861	0.8813 ± 0.0153

Table 4: **Mean Relative Degradation Difference after Interventions.** Results from Qwen3-4B-Base on the 4×4 Lattice Grid. Each row reports the mean difference in correct answer probability reduction between graph and random interventions after 20 iterations for the corresponding shot setting, normalized by the corresponding clean probability before interventions. Positive values indicate greater degradation under the graph intervention.

Setting	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean $\pm$ Std
0-Shot	-0.401	-0.649	-0.380	-0.359	-0.648	-0.487 ± 0.148
16-Shot	1.010	1.293	1.069	1.246	1.635	1.251 ± 0.245

Table 5: **Probe Performance after Graph Interventions.** Results from Qwen3-4B-Base evaluated on 300 test samples on 4×4 Lattice Grid after graph intervention for 20 iterations. "L4, 0-Shot" denotes the probe performance at Layer 4 with 0-shot prompting.

Setting	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Mean $\pm$ Std
L4, 0-Shot	0.5862	0.6337	0.5985	0.6015	0.5806	0.6001 ± 0.0207
L4, 16-Shot	0.5798	0.5897	0.5543	0.5758	0.5582	0.5716 ± 0.0149
L20, 0-Shot	0.7158	0.6900	0.6920	0.7095	0.7000	0.7015 ± 0.0111
L20, 16-Shot	0.6658	0.6783	0.6947	0.6891	0.6845	0.6825 ± 0.0111
L36, 0-Shot	0.6905	0.7208	0.7263	0.7059	0.7109	0.7109 ± 0.0139
L36, 16-Shot	0.6795	0.7056	0.6901	0.6915	0.6914	0.6916 ± 0.0093