Exploring the Platonic Representation Hypothesis Beyond In-Distribution Data

Aryasomayajula Ram Bharadwaj
Independent Researcher
`ram.bharadwaj.arya@gmail.com`

20th Oct 2024

Abstract

The Platonic Representation Hypothesis (PRH) [?] posits that neural networks, despite being trained on different objectives and datasets, converge toward a shared statistical model of reality. This paper extends the investigation of PRH to out-of-distribution (OOD) settings using ImageNet-O as a benchmark, contrasting the findings with results from in-distribution data, random noise, and the dynamics of model alignment under noise injection and during language model training. Our analysis reveals that while PRH holds in OOD scenarios, it breaks down with purely random data, and demonstrates a nuanced relationship between noise and representational alignment.

1 Introduction

The Platonic Representation Hypothesis (PRH) suggests that models trained across various modalities and objectives converge toward a shared representation of reality [?]. Inspired by philosophical concepts like Plato's Allegory of the Cave, PRH posits that neural networks progressively learn an idealized representation of the world. This convergence has been observed across domains such as vision and language models, with evidence showing that larger models better align their internal representations [?]. However, it remains unclear whether such convergence persists when models encounter data significantly different from their training distribution.

To address this, we extend PRH's analysis to out-of-distribution (OOD) settings using the ImageNet-O dataset, which contains samples outside the distribution of ImageNet [?]. We also include random noise as a contrasting dataset to examine the boundaries of representational alignment, along with two additional experiments: the impact of progressive noise on alignment among vision models and the evolution of alignment during language model training.

2 Methodology

We assess representational alignment using various metrics, such as mutual k-NN and CKNN-A, across three types of data:

1. **In-distribution data**: Data that aligns with the pretraining dataset (e.g., Places365's validation set).

2. **Out-of-distribution data**: ImageNet-O, designed with images that are outliers to the ImageNet distribution [?].

3. **Random noise**: Purely random images to test the limits of PRH.

The alignment between different models is measured using Spearman's rank correlation, focusing on how similar the models' distance measures are for given data points. Throughout the experiments, 17 different Vision Transformers (ViTs) spanning various sizes and training data are used. Mutual k-Nearest Neighbor Alignment is measured in all three settings.

Additionally, we conducted two more analyses:

1. **Noise Injection in Vision Models**: A set of 250 images was progressively corrupted with noise over 100 steps, and mutual alignment was measured across all pairs of the 17 ViT models at each step. The mean alignment scores were analyzed to study the relationship between noise levels and model alignment.

2. **Training Dynamics in Language Models**: Six large language models (LLMs) were analyzed using checkpoints saved at 100 different training stages. We measured the alignment of each model pair at each stage and examined the trends in alignment throughout the training process.

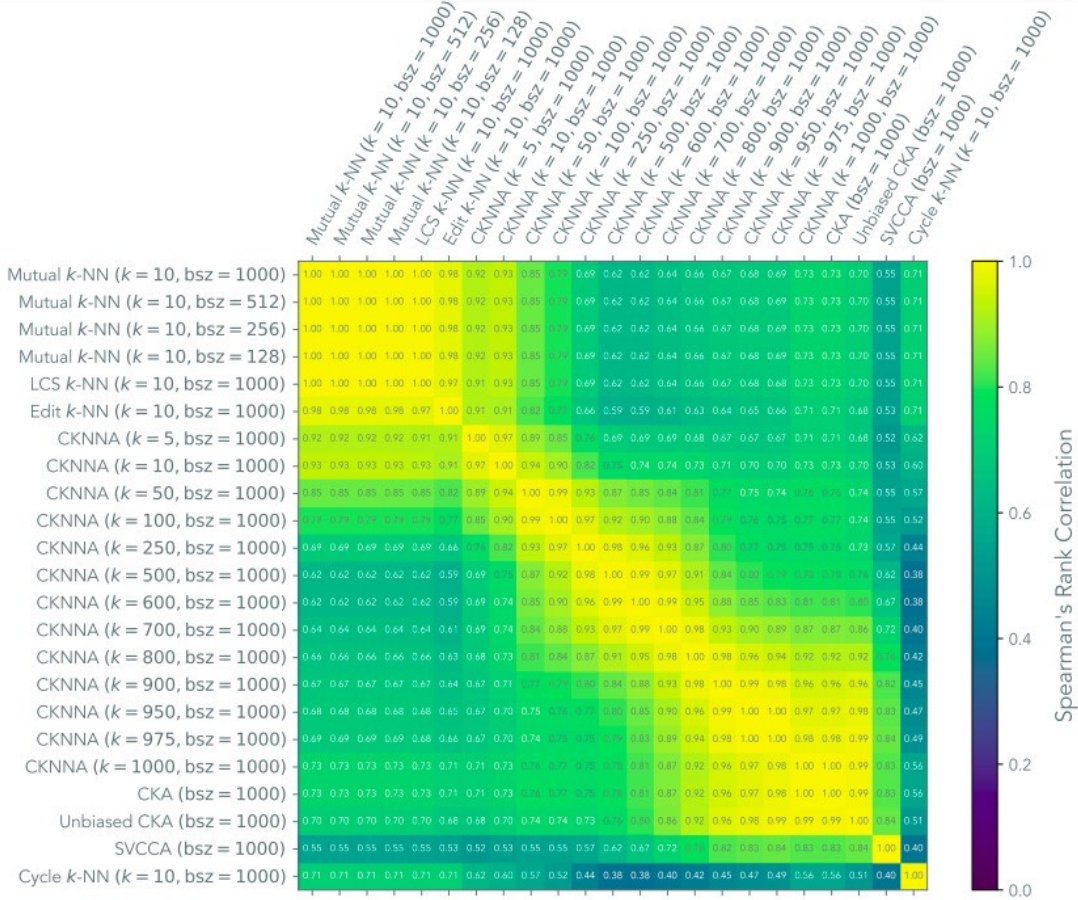# 3 Results

3.1 Alignment on In-Distribution Data

Figure 12. **Vision-vision alignment measured with various metrics.** Spearman's rank correlation among different metrics and batch sizes (bsz) when used to measure alignment among 78 vision models (see Appendix C.1 for details of these models). All p-values are below 2.24×10^{-105}. Our vision-vision analysis in Figure 2 is based on the first metric (Mutual k-NN with $k = 10$ and bsz = 1000).

Figure 1: Vision-vision alignment measured on Places365's validation dataset using various metrics. High Spearman's rank correlation among different metrics indicates alignment among vision models.

[Placeholder: Ensure the image is high-resolution and provides clear labeling of axes and metrics.]
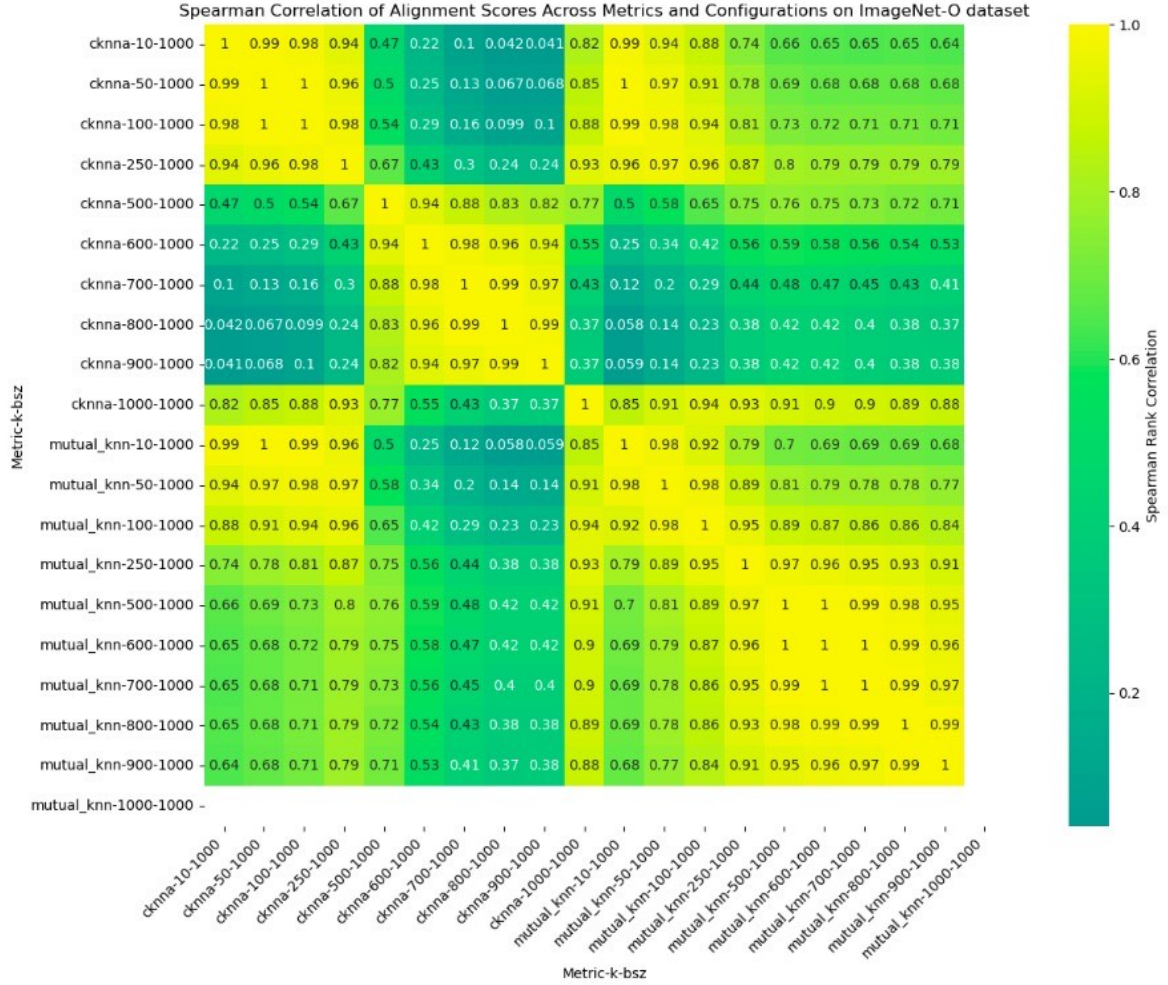
3.2 Alignment on ImageNet-O

Figure 2: Alignment of vision models on ImageNet-O dataset. The Spearman correlation suggests that even with outlier data, models maintain a shared statistical representation, albeit with higher prediction errors.

[Placeholder: Add a detailed analysis of how prediction errors influence the alignment and their implications for PRH.]
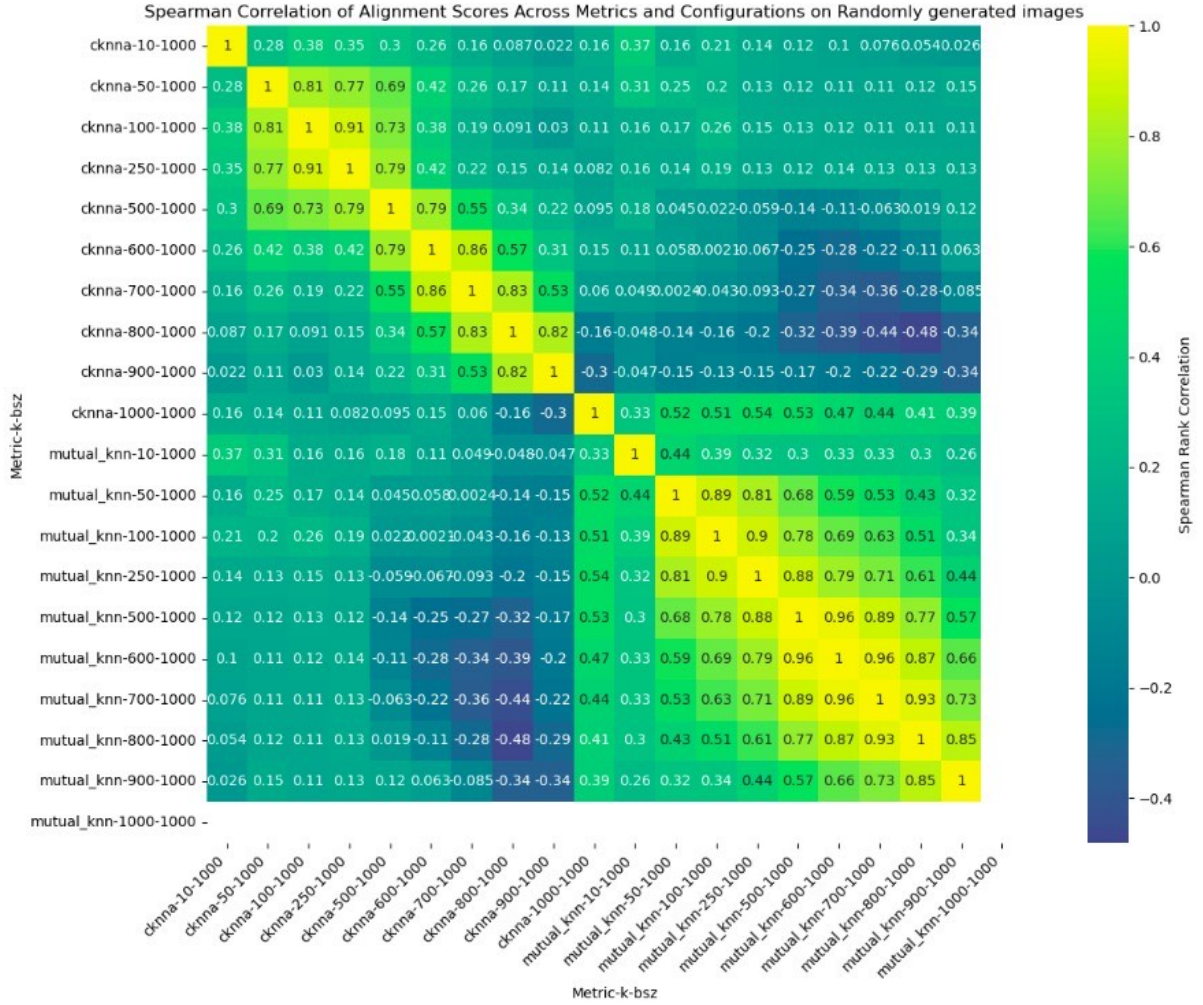
3.3 Alignment on Random Data

Figure 3: Spearman correlation of alignment scores on random noise data. The lower correlation values indicate a breakdown in representational alignment, suggesting that random data lacks the underlying structure needed for PRH.

[Placeholder: Discuss potential factors that could lead to this breakdown and include error bars in the plot.]

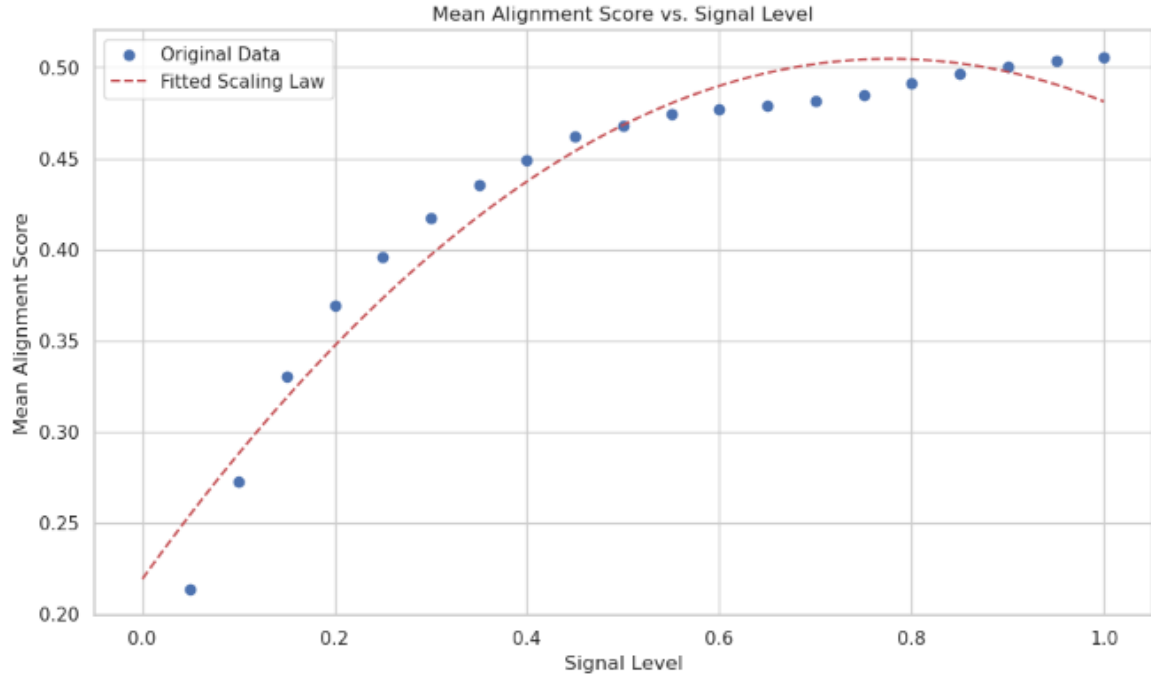## 3.4   Noise Injection in Vision Models

Figure 4: Mean alignment of 17 ViT models with respect to signal level (1 - noise) across 100 noise injection steps. The alignment shows a quadratic relationship with decreasing signal levels.

[Placeholder: Include a mathematical derivation of the observed quadratic relationship and add statistical significance testing.]
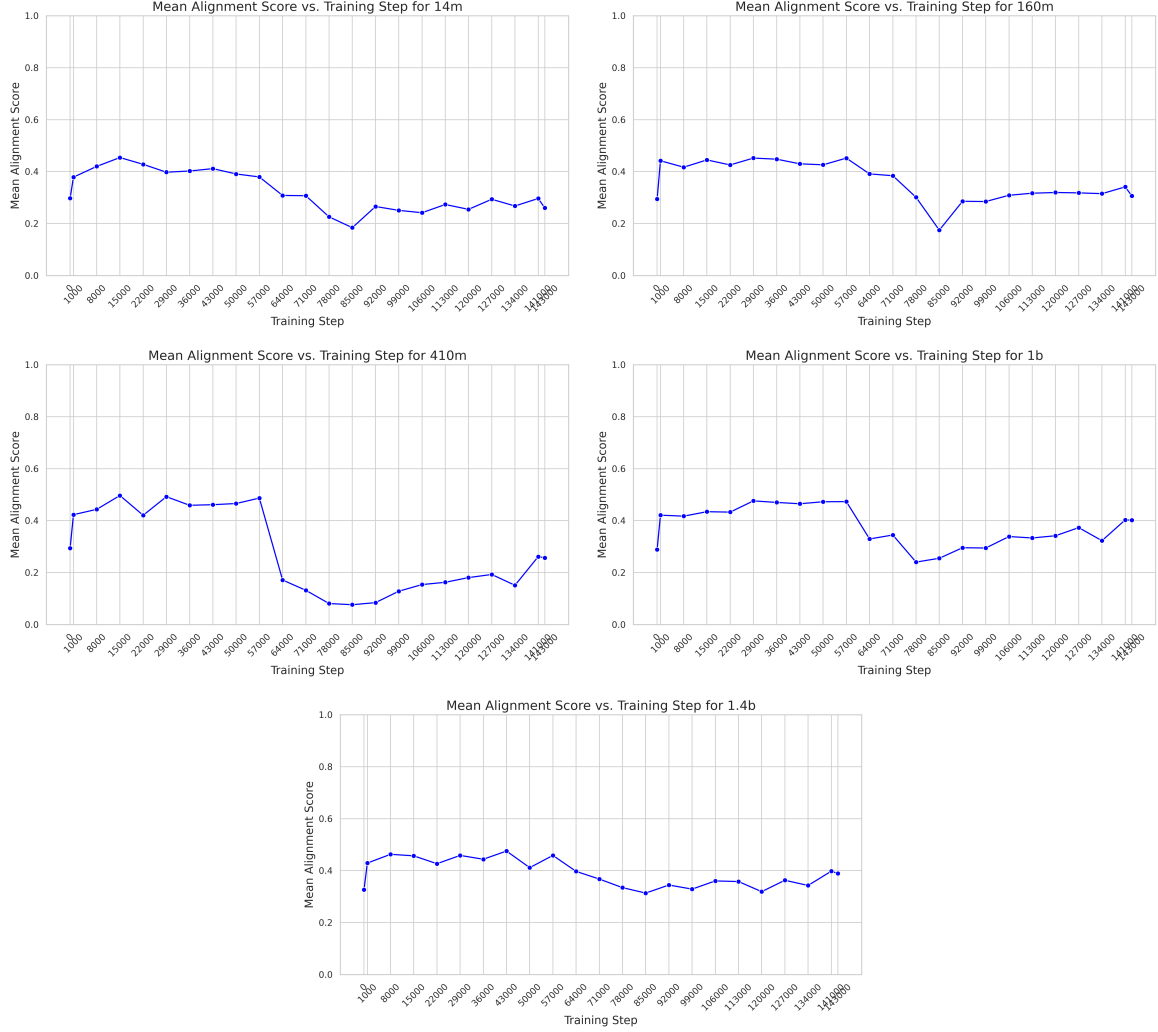
3.5   Training Dynamics in Language Models

Figure 5: Mean alignment scores of language models during training, grouped by the 14M, 160M, 410M, 1B, and 1.4B parameter-sized models. The alignment pattern shows a non-linear trend, characterized by an initial increase, followed by a decline, and then an increase again.

[Placeholder: Provide deeper analysis on the factors contributing to the non-linear trends and relate this to the phases of model training.]

4   Discussion

Our findings show that PRH extends beyond in-distribution data to OOD settings, where models exhibit alignment in their internal representations. On ImageNet-O, models align even in their predictions when they produce incorrect results with high confidence, revealing that models fail in a predictable manner on OOD data. This suggests that they rely on a shared interpretation of the input space, even when the data diverges from what they have seen during training.

The analysis of noise injection reveals a quadratic relationship between noise levels and representational alignment, suggesting that as noise degrades the input, models align in a predictable manner. This highlights the robustness of model representations to noise but also their tendency to converge on structured patterns within the data.

The language model training analysis showed a non-linear alignment pattern, which suggests that representational dynamics during training might involve shifts between phases of increased and decreased similarity. This behavior may indicate transitions in how models internalize and abstract from the data.

6

[Placeholder: Discuss limitations such as the potential biases in the selected models or datasets, and suggest how future studies could address these.]

5   Conclusion

The results validate the Platonic Representation Hypothesis in OOD scenarios but highlight its limitations with random noise. Our analysis underscores the need for structured data in achieving representational alignment and suggests new directions for exploring the boundaries of PRH. Future work could focus on extending PRH to other modalities, exploring the effect of different regularization strategies on alignment, and investigating the phases of alignment during language model training.

The code used for the experiments and analysis in this paper is available on GitHub here.

References

[1] Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). The Platonic Representation Hypothesis. *International Conference on Machine Learning.*

[2] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural Adversarial Examples. *CVPR.* arXiv:2405.07987.