# Investigating the Platonic Representation Hypothesis

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

20th Oct 2024

**Abstract**

The Platonic Representation Hypothesis (PRH) [?] suggests that neural networks, despite being trained on different objectives and datasets, converge toward a shared statistical model of reality. This work extends the investigation of PRH by exploring how neural networks behave when exposed to data that diverges from their training distribution. Using ImageNet-O as a benchmark for out-of-distribution (OOD) data, random noise as a boundary case, and detailed analyses of model alignment dynamics, we aim to uncover the strengths and limitations of PRH. Our results indicate that PRH holds in OOD contexts but encounters challenges with random data, offering a nuanced perspective on representational alignment.

## 1 Introduction

Understanding how neural networks form internal representations of the world has been a key area of research, bridging both philosophy and machine learning. The Platonic Representation Hypothesis (PRH) [?], inspired by Plato's Allegory of the Cave, posits that models across different domains—be it vision or language—are converging towards a shared, idealized model of reality. This hypothesis suggests that, even when trained on different datasets, models internalize similar structures, akin to Plato's concept of perceiving shadows of the same underlying reality.

However, an important question remains: does this convergence persist when models are exposed to data significantly different from their training distributions? In this study, we address this by evaluating the behavior of PRH in various scenarios, including exposure to out-of-distribution (OOD) data and purely random data. Additionally, we explore how models' alignment evolves when progressively corrupted with noise and during different stages of language model training.

## 2 Methodology

To rigorously evaluate the PRH, we designed a series of experiments that progressively test the hypothesis under varying conditions of data distribution. The methodology is structured around three main experimental setups, each focusing on a distinct aspect of model alignment.

### 2.1 Experimental Setup

**1. Representational Alignment Across Data Types:** We assessed the alignment between different models using the following datasets:

1. **In-distribution data**: Places365's validation set, which closely aligns with the training data of many vision models.
2. **Out-of-distribution data**: The ImageNet-O dataset, specifically curated with images that fall outside the standard ImageNet distribution [?]. This allows us to examine whether PRH extends to data with significant divergence.
3. **Random noise**: A dataset of purely random images was used to probe the extreme boundaries of representational alignment. This helps determine if the models converge on a shared interpretation even when presented with unstructured input.

We measured alignment using Spearman's rank correlation on mutual k-NN distances between model representations, comparing how models interpret the similarity of various input data points.

## 2.2 Noise Injection in Vision Models

**2. Progressive Noise Injection:** To understand the resilience of model alignment to data corruption, we conducted a noise injection experiment:

- A set of 250 images was selected and progressively corrupted with Gaussian noise over 100 steps.
- At each noise level, we measured mutual alignment across 17 Vision Transformer models, evaluating how the models' representational structures changed as noise increased.
- The alignment scores were analyzed to reveal potential patterns in how noise impacts convergence, with a particular focus on detecting non-linear relationships.

## 2.3 Tracking Alignment During Language Model Training

**3. Evolution of Alignment During Training:** We also studied how alignment evolves during the training process of language models:

- Six large language models (LLMs) were analyzed using checkpoints saved at 100 different training stages.
- At each checkpoint, alignment was measured between every pair of models to observe changes in representational similarity over time.
- This analysis aimed to uncover any phases of increased or decreased similarity, providing insights into how models internalize structure during different training stages.

# 3 Results

## 3.1 Alignment Across Data Types

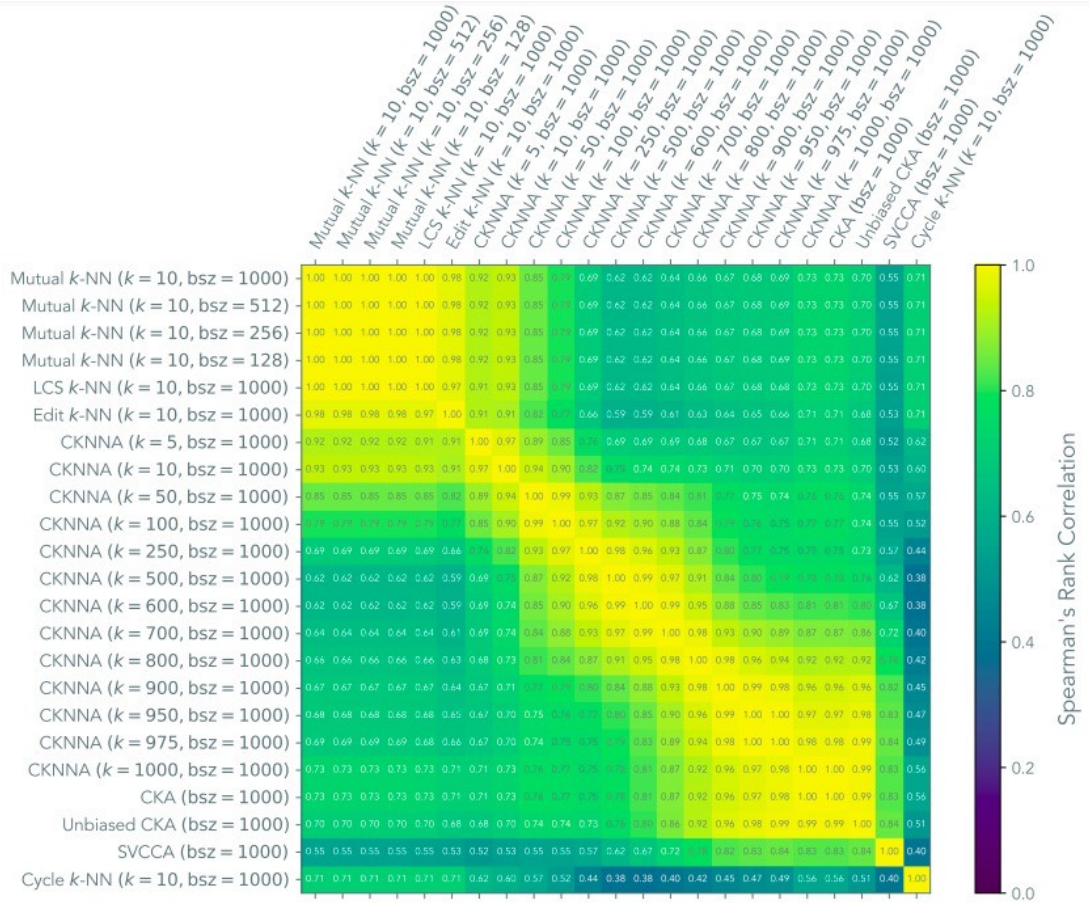


*Figure 12.* **Vision-vision alignment measured with various metrics.** Spearman's rank correlation among different metrics and batch sizes (bsz) when used to measure alignment among 78 vision models (see Appendix C.1 for details of these models). All $p$-values are below $2.24 \times 10^{-105}$. Our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

Figure 1: Spearman rank correlation between models on Places365's validation data, indicating a high degree of alignment in in-distribution data.

## 3.2 Alignment on ImageNet-O Data

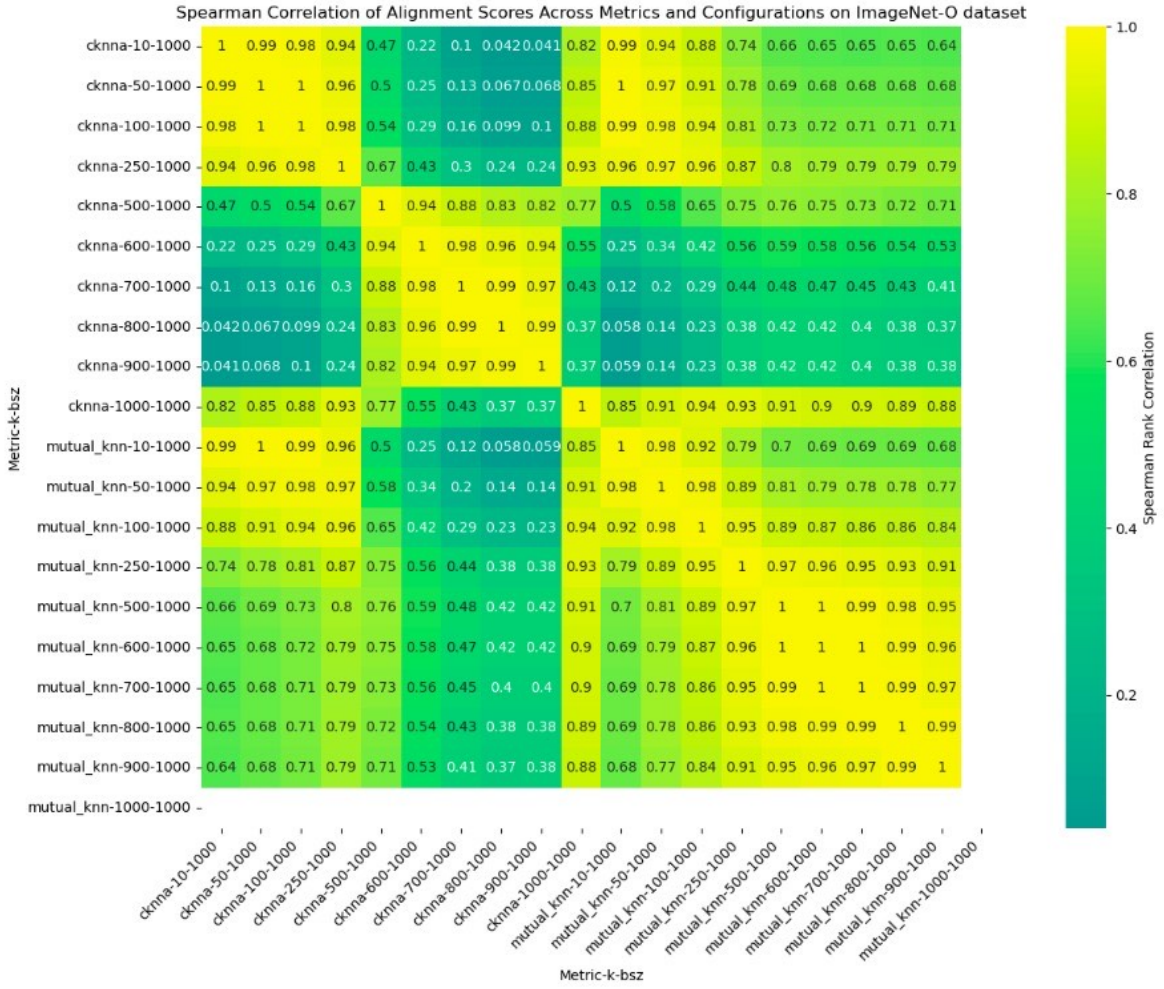


Figure 2: Alignment of vision models on the ImageNet-O dataset. The Spearman correlation suggests that even with outlier data, models maintain a shared statistical representation, albeit with higher prediction errors.

### 3.3 Alignment on Random Noise Data

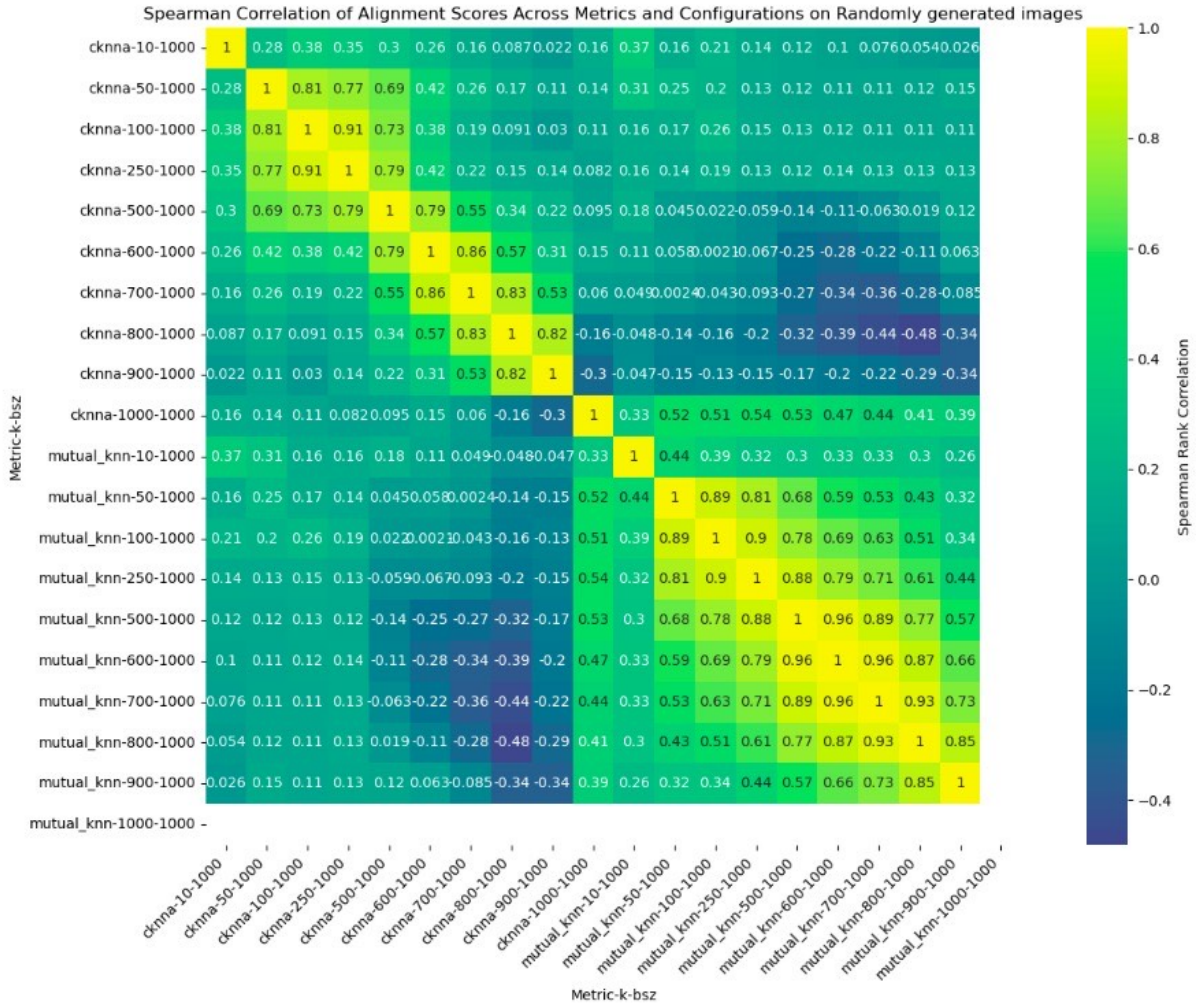


Figure 3: Spearman correlation of alignment scores on random noise data. The lower correlation values indicate a breakdown in representational alignment, suggesting that random data lacks the underlying structure needed for PRH.

### 3.4 Impact of Noise Injection

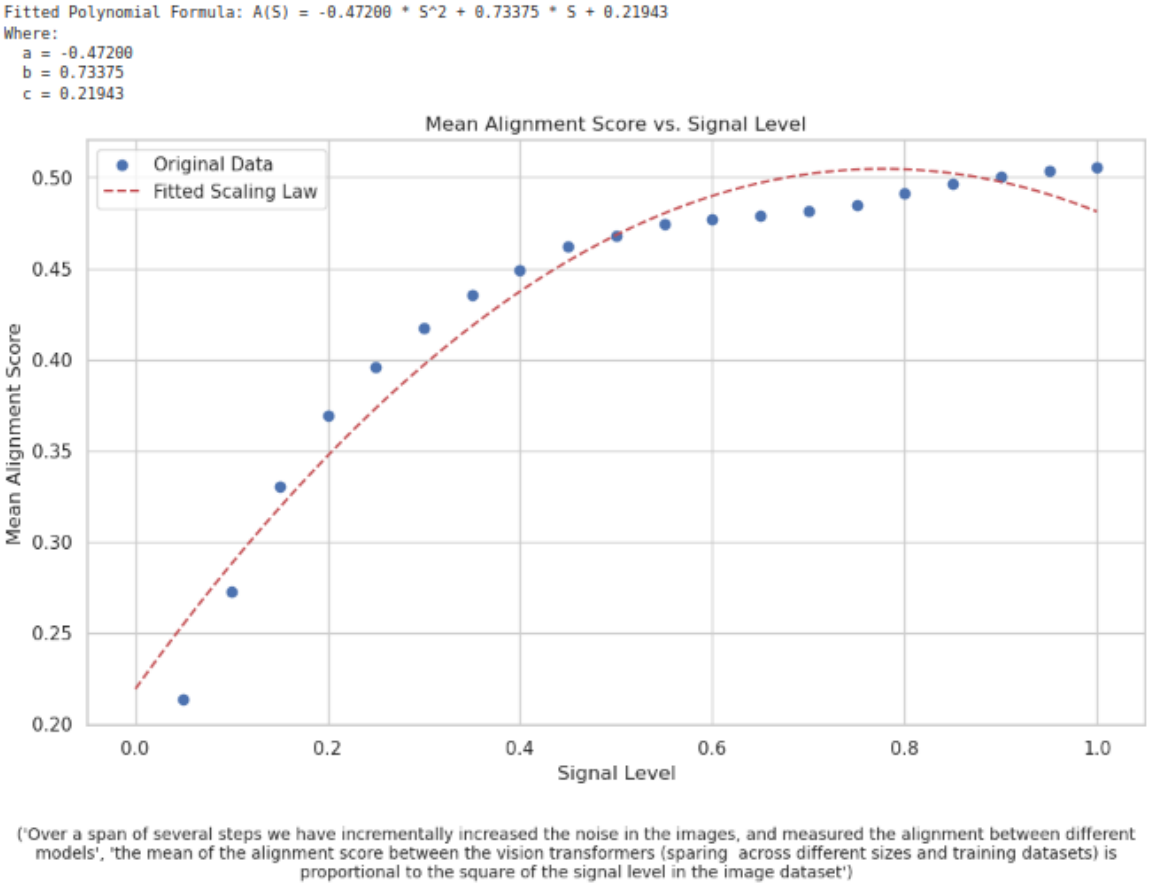


Figure 4: Relationship between noise level and mean alignment scores of 17 ViT models. A quadratic trend is observed as noise increases.

### 3.5 Alignment Dynamics During LLM Training

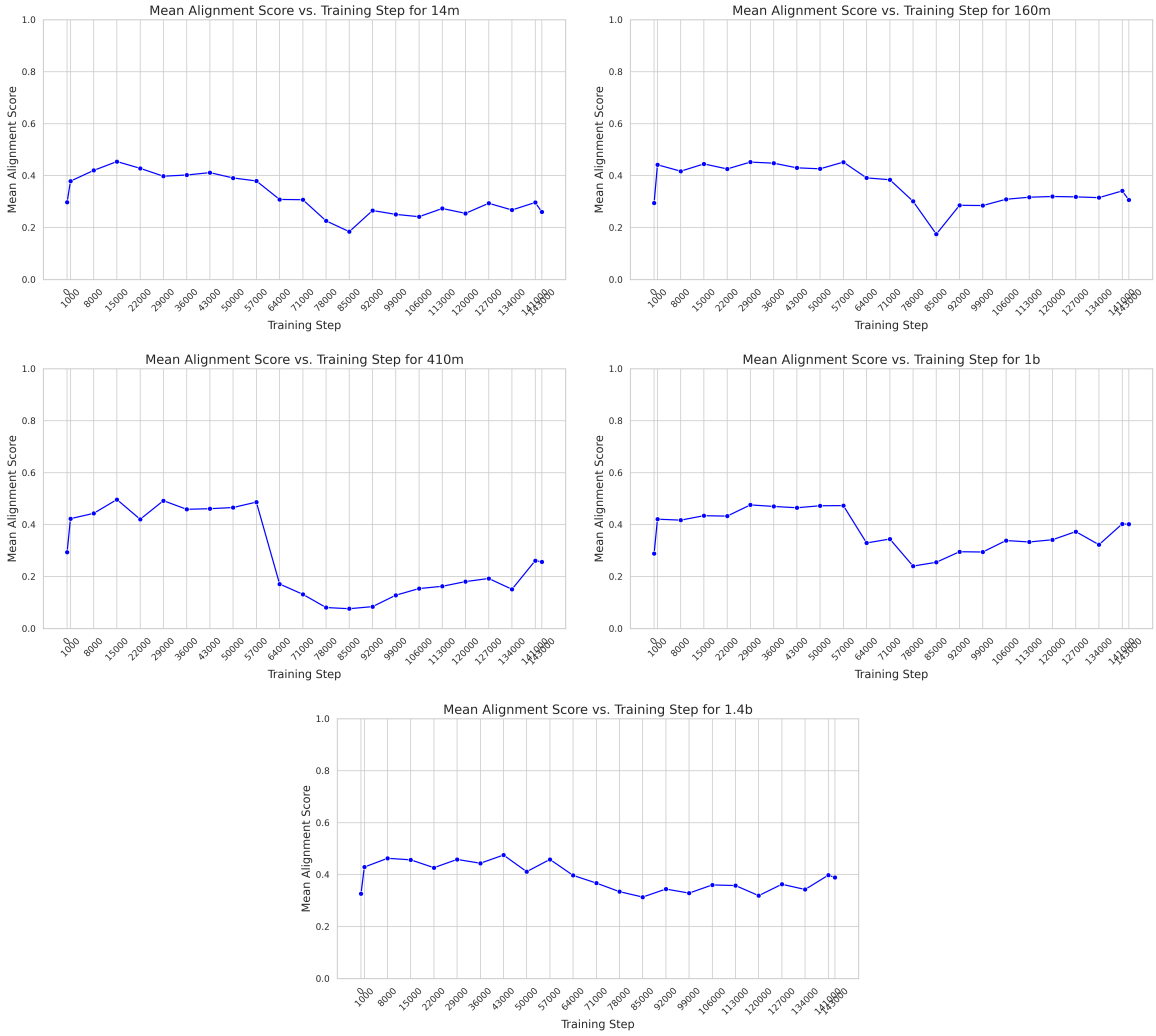


Figure 5: Alignment trends of LLMs during training. Alignment scores reveal non-linear trends across different parameter sizes.

## 4 Discussion

Our results extend the understanding of PRH by demonstrating that while neural networks maintain representational alignment in OOD settings, they struggle with random, unstructured data. This suggests that the shared statistical model proposed by PRH is dependent on the presence of underlying structures in the input data.

The resilience of model alignment to noise, as seen in our noise injection experiment, indicates a form of robustness in how models interpret degraded input. However, the breakdown of alignment with random noise raises questions about the limits of shared representations.

## 5 Conclusion

Our study supports the validity of the Platonic Representation Hypothesis in structured, yet diverse data environments. The findings emphasize the importance of structured data for achieving alignment and suggest further exploration into the dynamics of alignment during model training. Future work could focus on extending PRH to multimodal settings and investigating the impact of training strategies on representational convergence.

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.