Investigating the Platonic Representation Hypothesis

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

20th Oct 2024

Abstract

The Platonic Representation Hypothesis (PRH) [1] suggests that neural networks, despite being trained on different objectives, modalities, and datasets, tend to converge toward a shared statistical model of reality. This work investigates the specific conditions required for this phenomenon. Our results indicate that PRH holds even in out-of-distribution (OOD) contexts but faces limitations with random data, offering a nuanced view on representational alignment. Additionally, we examine how training duration and the signal present in data influence convergence toward this shared model.

1 Introduction

Understanding how neural networks form internal representations of the world is a key research area, bridging philosophy and machine learning. The Platonic Representation Hypothesis (PRH) [1], inspired by Plato's Allegory of the Cave, proposes that models across domains—such as vision or language—are converging toward a shared, idealized model of reality. According to PRH, even when trained on different datasets, models internalize similar structures, reflecting Plato's concept of perceiving shadows of an underlying truth.

However, some key questions remain: Does this convergence persist when models encounter data significantly different from their training distributions? How does data quality (in terms of signal and information) affect convergence to a shared statistical model? And how does convergence change with varying training progress? This study explores these questions by evaluating PRH under scenarios involving out-of-distribution (OOD) data, purely random data, and progressive noise levels, as well as during different stages of language model training.

2 Methodology

To evaluate PRH rigorously, we designed experiments that test the hypothesis under varied data distribution conditions. Our methodology centers on three main experimental setups, each focused on a distinct aspect of model alignment.

2.1 Experimental Setup

1. Representational Alignment Across Data Types: We measured alignment between different models across three scenarios: in-distribution, out-of-distribution, and purely random data using the following datasets:

1. **In-distribution data**: The validation set from Places365, which closely aligns with the training data of many vision models. Results for this were adapted directly from the original study [1].
2. **Out-of-distribution data**: The ImageNet-O dataset, containing images that fall outside the standard ImageNet distribution [2]. This allows examination of PRH in data with significant divergence.
3. **Random noise**: A dataset of purely random images probes the extreme boundaries of representational alignment. Random noise vectors with matching dimensions to the original dataset were generated, helping assess whether models converge on a shared interpretation in the absence of structured input.

Alignment was measured using Spearman's rank correlation on mutual k-NN distances across model representations, comparing how models interpret the similarity of various data inputs.

2.2 Noise Injection in Vision Models

2. Progressive Noise Injection: To understand the importance of signal quality or information content for representational alignment, we conducted a noise injection experiment:

- A set of 250 images was progressively corrupted with Gaussian noise over 100 steps.
- At each noise level, we measured mutual alignment across 17 Vision Transformer models, evaluating changes in representational structures as noise increased.
- Alignment scores were analyzed to detect non-linear relationships and patterns in how noise impacts convergence.

2.3 Tracking Alignment During Language Model Training

3. Evolution of Alignment During Training: We also studied how representational alignment evolves during language model training:

- 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 phases of increased or decreased similarity, providing insights into how models internalize structure across training stages.

3 Results

3.1 Alignment Across Data Types

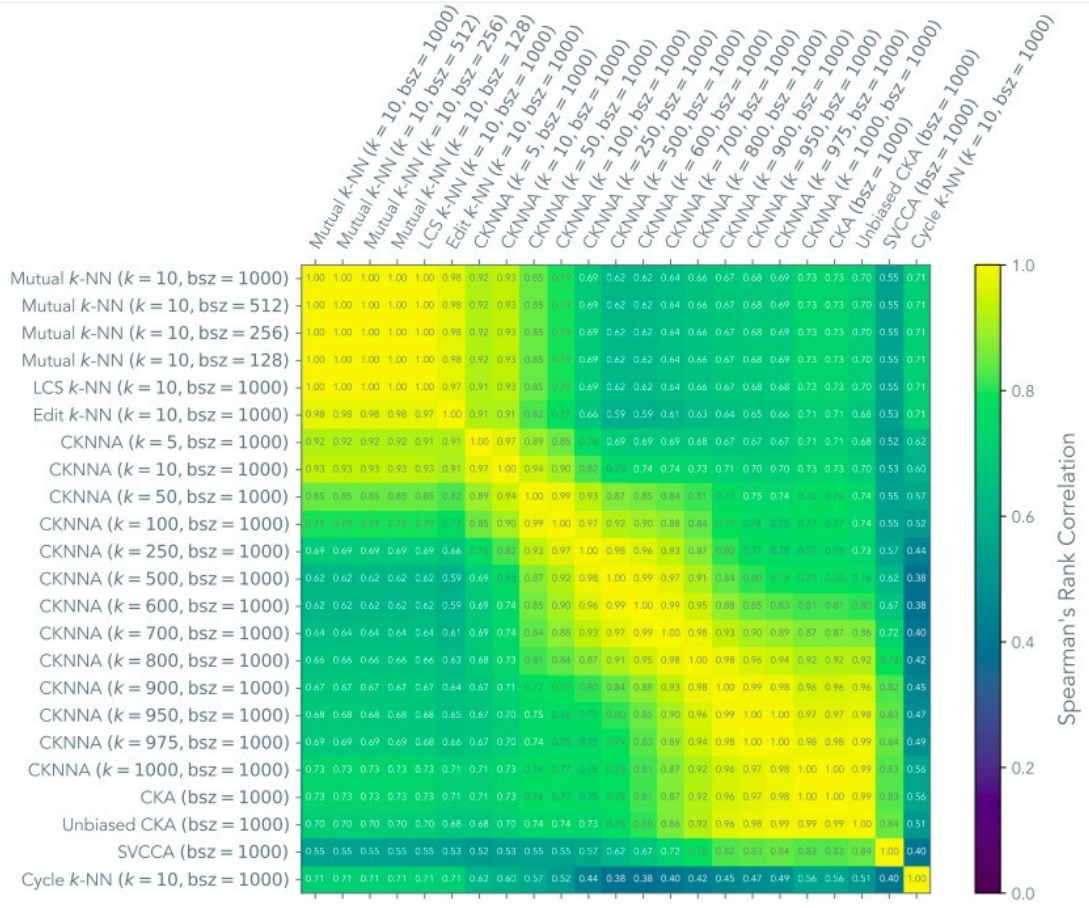

Figure 12. **Vision-vision alignment measured with various metrics.** Spearman's rank correlation among different metrics and batch sizes (bsz) when used to measure alignment among 78 vision models (see Appendix C.1 for details of these models). All p-values are below 2.24×10^{-105}. Our vision-vision 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

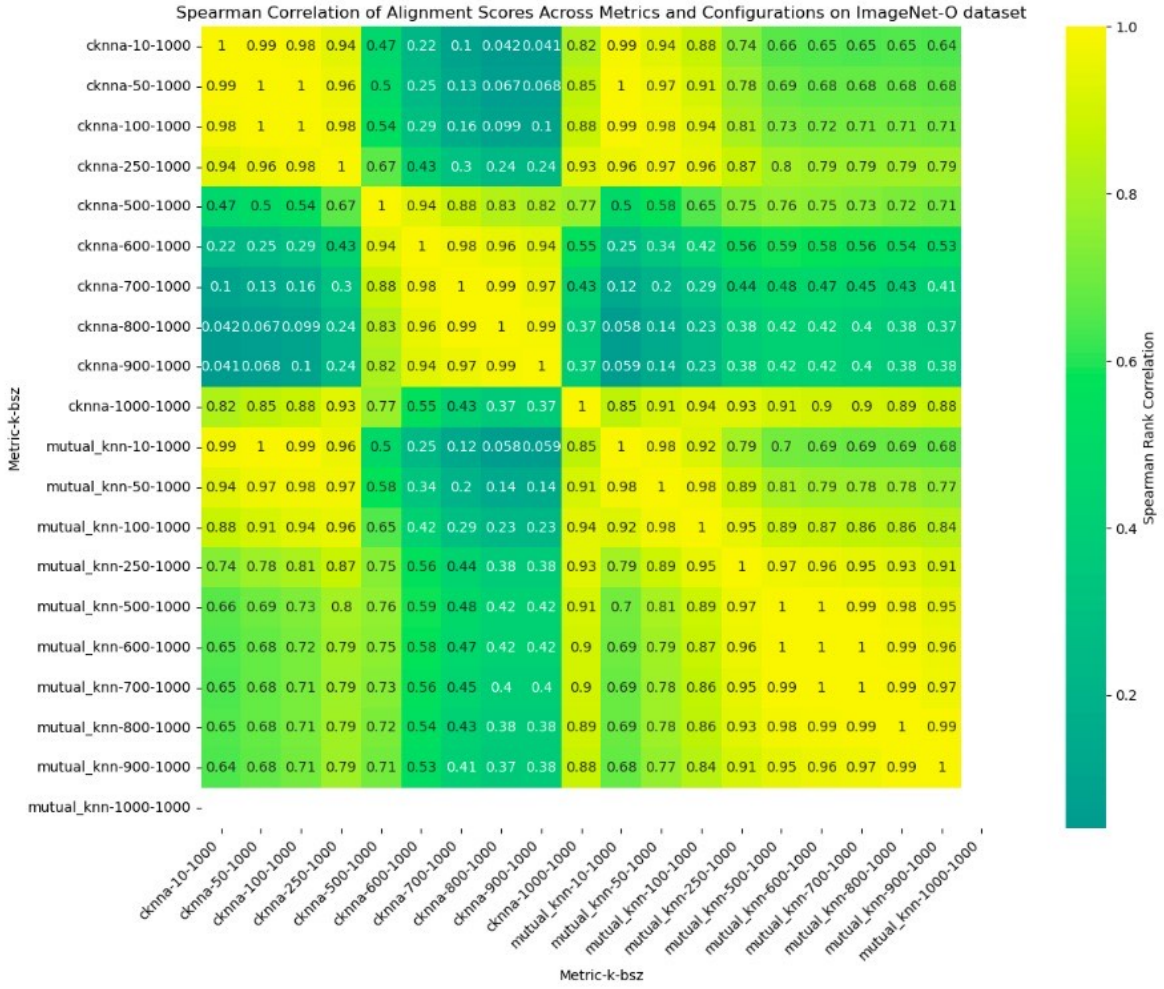

Figure 2: Alignment of vision models on the ImageNet-O dataset. Spearman correlation suggests that even with outlier data, models maintain shared statistical representation. Despite prediction errors in this OOD case, models converge in their representations, implying that larger models exhibit predictable error patterns similar to smaller models.

3.3 Alignment on Random Noise Data

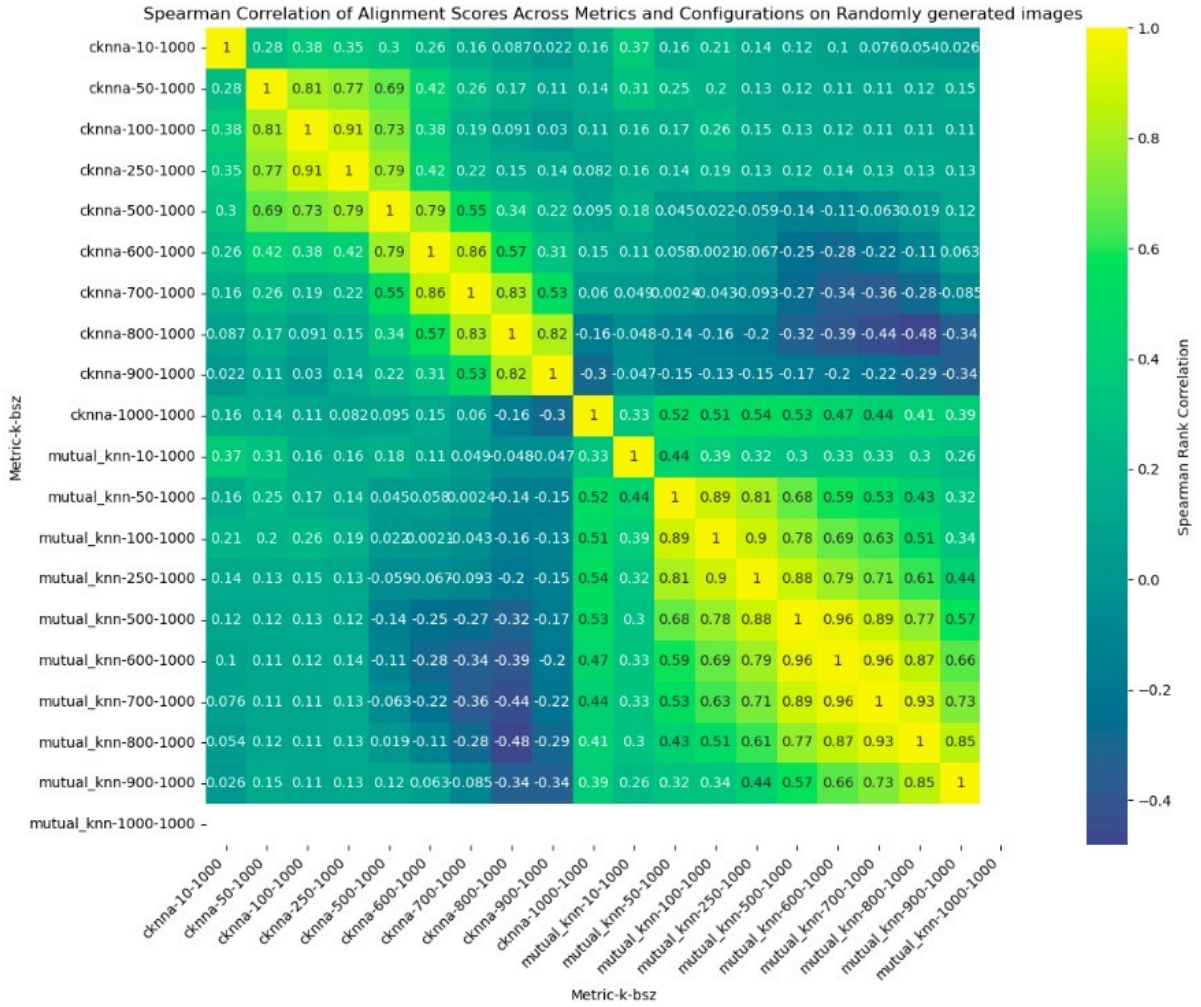

Figure 3: Spearman correlation of alignment scores on random noise data. The lower correlation values indicate a breakdown in representational alignment, suggesting that random data lacks the underlying structure required for PRH. This suggests that representational alignment, while present in OOD cases, still depends on some structural properties within the dataset.

3.4 Impact of Noise Injection

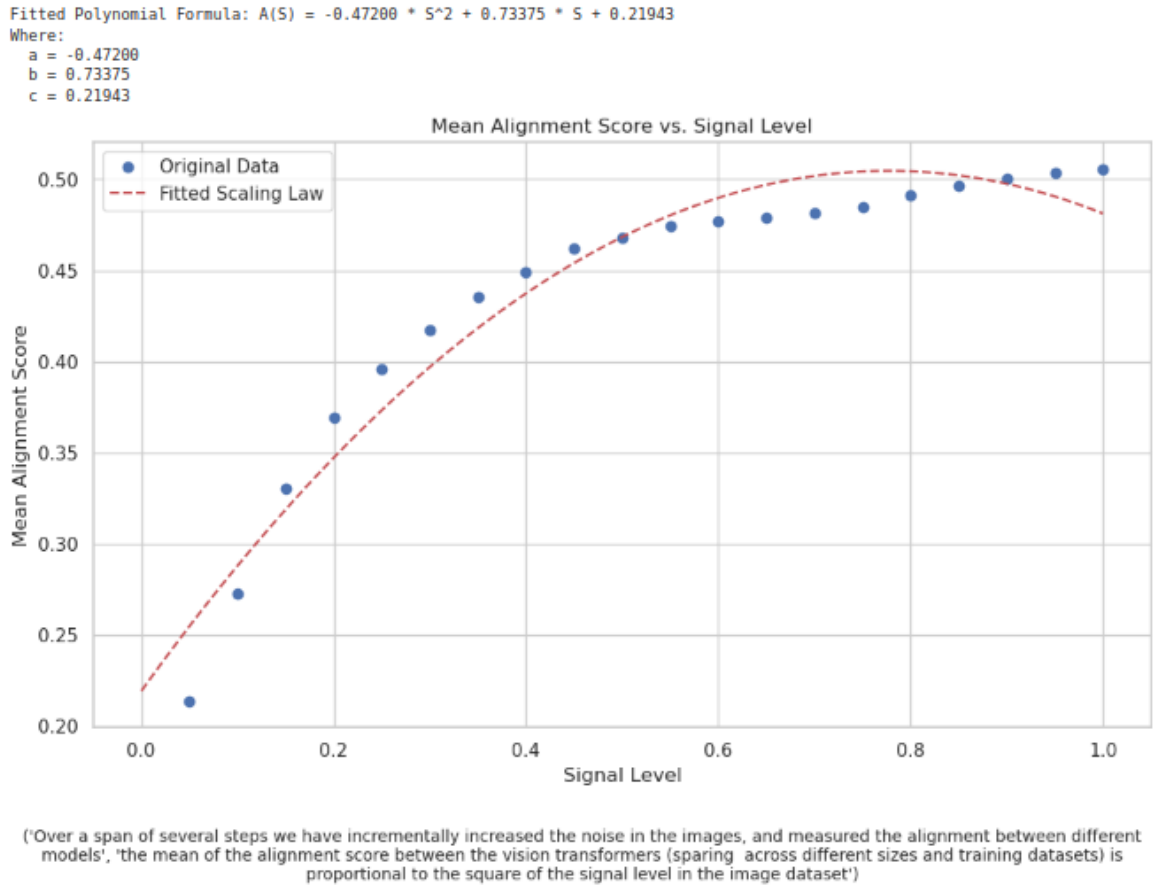

Figure 4: Relationship between noise level and mean alignment scores of 17 ViT models. A quadratic trend is observed as noise increases. Noise was progressively added to the dataset, with alignment scores calculated at each stage. The results suggest a threshold of signal required for representational alignment to occur. Alignment follows a smooth trend proportional to the square of the signal quantity present in the model.

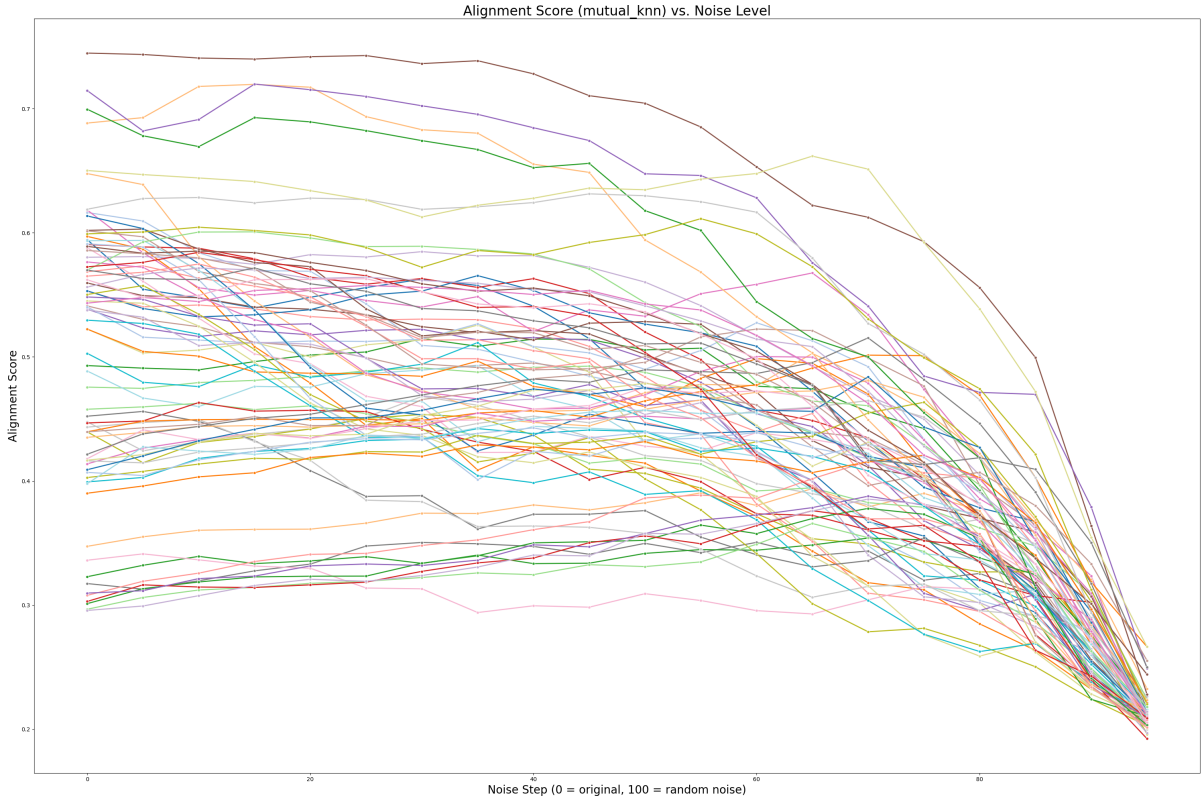

Figure 5: Alignment of individual ViT models with respect to noise level. This plot provides a detailed view of how each model's representational alignment responds to incremental noise levels, highlighting both similarities and differences across the 17 models.

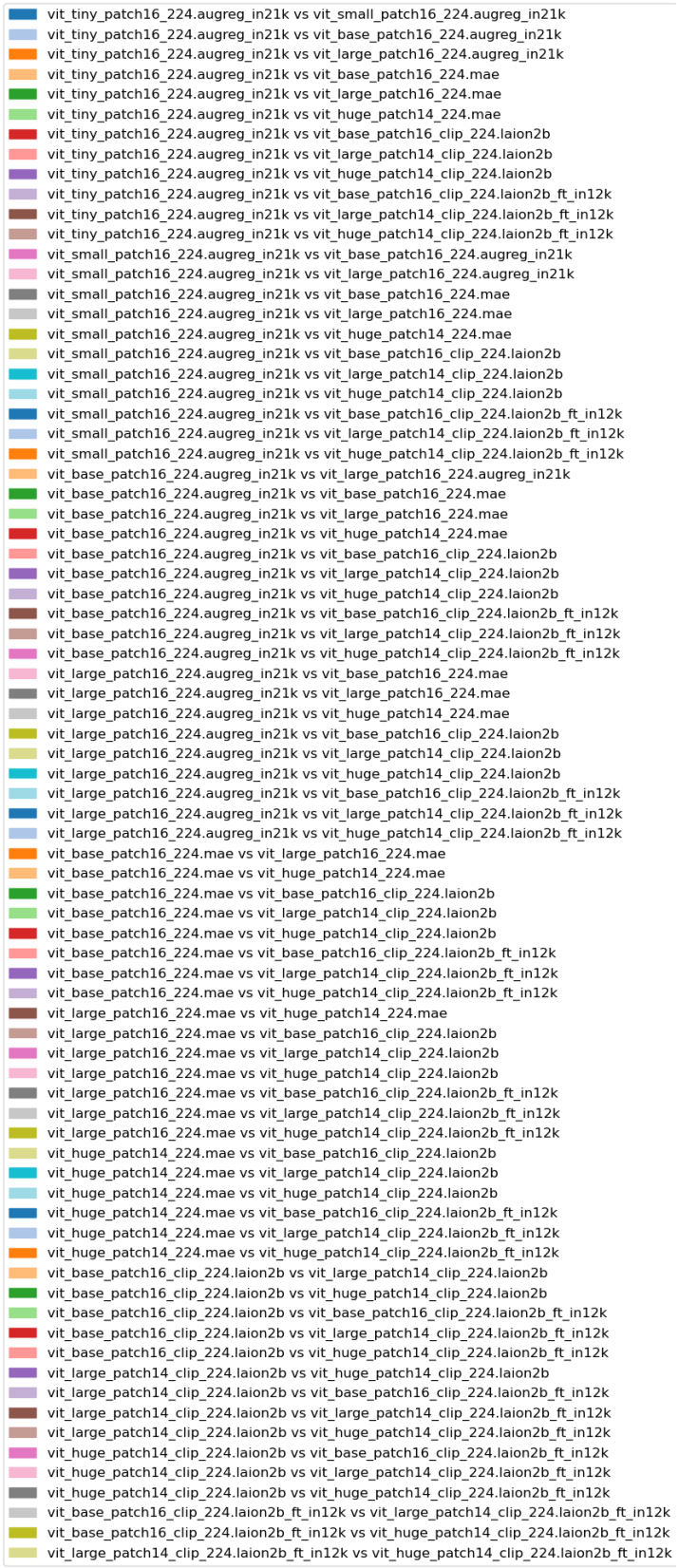

Figure 6: Legend showing the names of individual ViT models used in the study. Each model is uniquely represented to facilitate distinction in the alignment plots, providing a reference for identifying alignment patterns.

3.5 Alignment Dynamics During LLM Training

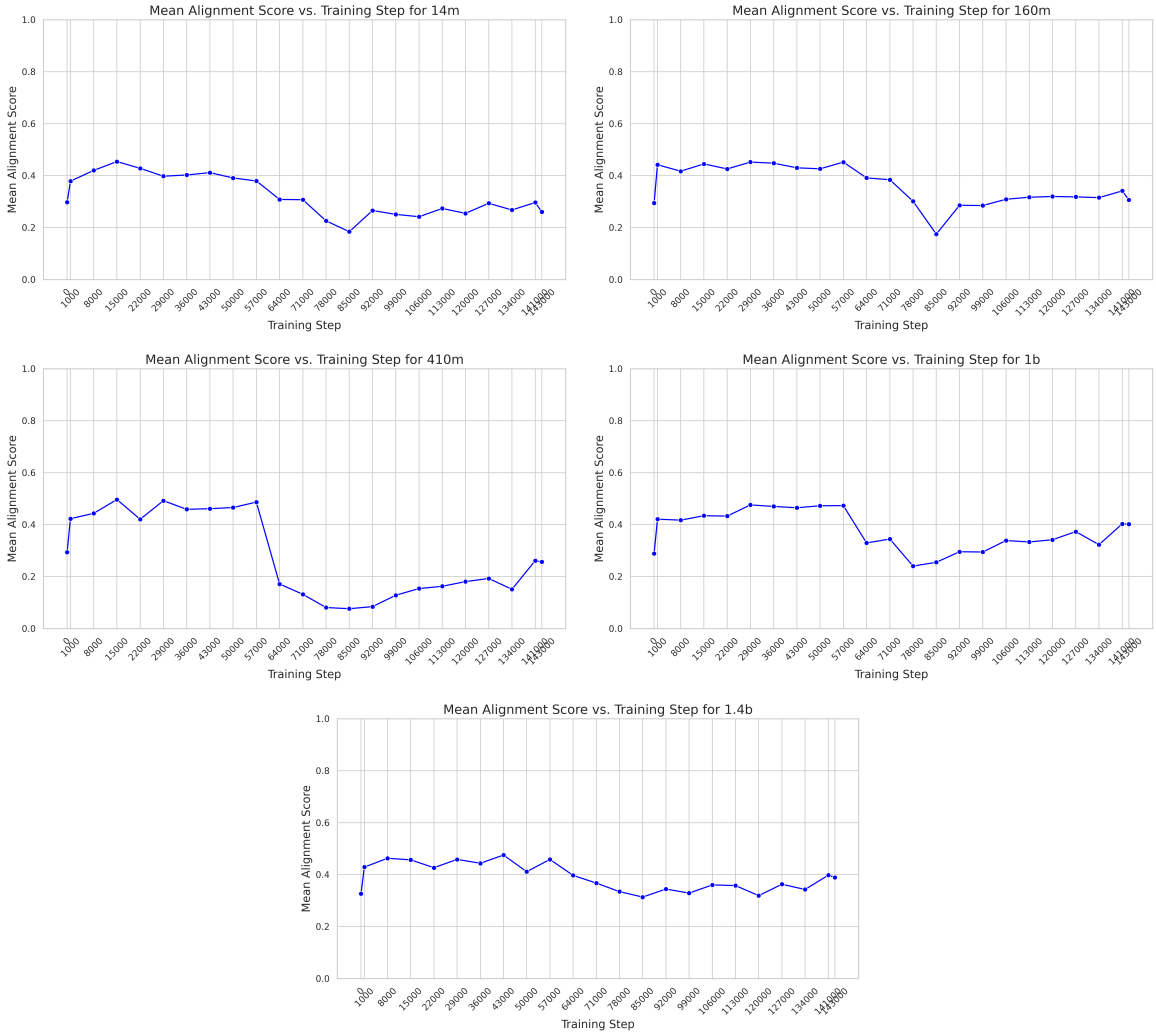

Figure 7: Alignment trends of LLMs during training. Scores show non-linear trends across different parameter sizes, with initial similarity between smaller and larger models due to similar initialization methods, followed by divergence as larger models form more detailed representations. For each model in the plot, the mean alignment score with respect to all the other five models is calculated and plotted over the training process

4 Discussion

Our results extend understanding of PRH by showing that neural networks maintain representational alignment in OOD contexts but struggle with random data, indicating that shared statistical models depend on structured input. Noise resilience in our experiments highlights models' robustness in interpreting degraded input, though random noise shows the limits of shared representations. Both signal quality and model pretraining show smooth transitions in representational alignment, suggesting a gradual, rather than abrupt, convergence.

5 Conclusion

Our study supports the validity of the Platonic Representation Hypothesis in structured, diverse data environments. Findings highlight the importance of data structure in alignment, suggesting further exploration of PRH in multimodal settings and the effects of training strategies on representational convergence.

The code for 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.