Decoding Dynamics, Alignment Stability, and Universal Concepts in LLMs

Ahmed Hassan [1] **Basil Hassan** [1] **Ashhad Ali** [1]

https://github.com/tk1475/ATML-PA5

Abstract

This report examines three key challenges in modern deep learning: controlling generation, ensuring human alignment, and understanding internal representations. We first investigate the inference dynamics of Large Language Models (LLMs) by identifying the trade-offs between coherence and creativity across various decoding strategies. We evaluate the stability and potential biases of alignment techniques by contrasting Reinforcement Learning from Human Feedback (PPO/GRPO) with Direct Preference Optimization (DPO). Finally, we critically assess the "Platonic Representation Hypothesis" by training Universal Sparse Autoencoders (USAEs) on diverse vision architectures. Our experiments reveal that distinct models converge toward interpretable conceptual spaces.

1. Introduction

As Large Language Models (LLMs) rapidly scale, it is no longer sufficient to know that they work. We must also understand how to control their generation in a way that aligns them with human intent and what internal representations they learn. This report empirically investigates these three domains.

The first challenge is **Inference Dynamics**. A model's decoding strategy determines its behavior. Deterministic methods like Greedy and Beam Search maximize likelihood but often suffer from repetition and dullness. Stochastic methods, such as Top-P (Nucleus) Sampling, restore diversity but risk incoherence. In order to deploy models that range in contexts, it is important to quantify this "Quality-Diversity Trade-off".

The second challenge is **Alignment** (Task 2). Techniques like Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) bridge the gap between next-token prediction and human preferences in pre-trained models. However, these methods introduce instabilities such as "reward hacking" (exploiting the reward signal without improving quality) and "verbosity bias" (where longer answers are incorrectly preferred).

The final challenge is **Interpretability**. A central question in deep learning is whether the internal features learned by neural networks are artifacts of their architecture or representations of the data. The *Platonic Representation Hypothesis*(Huh et al., 2024) argues for the latter and suggests that diverse models converge toward a shared statistical reality. We empirically test this by training a Universal Sparse Autoencoder to extract a single dictionary of concepts that is valid for both ResNet-18 and ViT-B, thereby quantifying the extent of this representational convergence.

In this work, we conduct an empirical analysis across these domains using `SmolLM2-135M` and custom Universal Sparse Autoencoders (USAEs) (Thasarathan et al., 2025). Our contributions include evaluating the stability of modern alignment algorithms and providing concrete evidence that CNNs and Vision Transformers share universal concepts.

2. LLM Decoding Strategy Analysis

2.1. Methodology

We investigated the impact of decoding algorithms on the generative capabilities of the `SmolLM2-135M-SFT-Only` (Allal & HuggingFaceTB Team, 2024) model. We implemented four decoding strategies:

- **Greedy Search:** A baseline that selects the token with the highest probability ($argmax$) at each step. This method favors local optimality but often yields repetitive or generic text.
- **Beam Search:** A global optimization method that maintains the top $B = 5$ most probable sequences at each step. It reduces the risk of missing high-probability sequences hidden behind low-probability initial tokens.

[1]Department of Electrical Engineering, Lahore University of Management Sciences (LUMS), Lahore, Pakistan. Correspondence to: Ahmed Hassan <26100308@lums.edu.pk>, Basil Hassan <26100303@lums.edu.pk>, Ashhad Ali <26100307@lums.edu.pk>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

- **Top-K Sampling:** (Fan et al., 2018) A stochastic method that samples only from the $K = 50$ most likely tokens. This prevents the selection of low-probability tokens while allowing for diversity.
- **Top-P (Nucleus) Sampling:** (Holtzman et al., 2020) A stochastic method that samples from the smallest set of tokens whose cumulative probability exceeds $P = 0.9$. This allows the vocabulary size to adapt dynamically to the model's confidence.

2.2. Experimental Setup

Evaluation was performed on a subset of 50 prompts from the `HuggingFaceH4/instruction-dataset` (HuggingFace H4 Team, 2023). We utilized the `OpenAssistant/reward-model-deberta-v3-large-v2` (OpenAssistant Team, 2023) as an objective standard for generation quality. We conducted two primary experiments:

Experiment A: The Temperature Grid. We evaluated the stochastic methods (Top-K, Top-P) across a temperature range $T \in \{0.2, 0.5, 0.8, 1.0, 1.2\}$ to quantify the trade-off between:

- **Diversity:** Measured via `Distinct-2` (the ratio of unique bigrams to total bigrams).
- **Quality:** Measured via the mean scalar score from the Reward Model.

Experiment B: The Collapse Test. To distinguish between true creativity and random noise, we generated 20 samples for a single fixed prompt ("Within-Prompt Diversity") versus 1 sample for 20 different prompts ("Across-Prompt Diversity") at a fixed temperature of $T = 0.8$.

2.3. Results and Discussion

The Quality-Diversity Trade-off. We observed an inverse relationship between generation quality and lexical diversity (Figure 1).

- **Beam Search** produced the most consistent but least diverse outputs, with a `Distinct-2` score of 0.65. While this stability is desirable for factual tasks, it often resulted in repetitive loops (see Table 1).
- **Stochastic Sampling:** As temperature increased from 0.2 to 1.2, the diversity increased sharply. **Top-P** at $T = 0.8$ achieved a `Distinct-2` score of 0.86 while maintaining a reward score comparable to the greedy baseline. This demonstrates that Nucleus Sampling strikes a balance between creativity and coherence by dynamically removing the tail of the distribution.

Qualitative Analysis. Table 1 highlights the signatures of each strategy. Beam search maximized probability by repeating safe phrases ("a big pile of...") in a monotonic tone. Top-P ($T = 0.8$) successfully broke away from the most probable path ("crayons") to explore a semantically valid but less obvious analogy ("toys"), demonstrating superior creative capability.

Table 1: Qualitative comparison of decoding outputs for the prompt: "Explain entropy to a 5-year-old."

Strategy	Generated Snippet
Greedy	"Imagine a box of crayons. You want to put them in a pile..." (Safe, generic)
Beam ($B = 5$)	"Imagine a box of crayons. A big pile of red, a big pile of blue, a big pile of green..." (Repetitive loops)
Top-P (0.9)	"Think of a toy room. If you never clean it up, it gets messier..." (Creative, varied)

Mode Collapse and Template Bias. In Experiment B ($T = 0.8$), Top-P achieved a higher diversity score for *Within-Prompt* generation (0.96) than for *Across-Prompt* generation (0.88). Our analysis of the text samples revealed that when responding to different prompts, the model frequently defaulted to structural templates (e.g., "Sure! Here is...", "1. ... 2. ..."). These repetitive formatting tokens artificially lowered the global diversity score. When generating multiple completions for a *single* open-ended prompt, the model was forced to explore the semantic tail of the distribution, which resulted in richer prose with higher bigram diversity. This finding underscores that "diversity" metrics can be heavily influenced by instruction-tuning artifacts (templates) rather than just semantic content.

3. LLM Alignment

3.1. Methodology

Our alignment pipeline follows the standard RLHF structure consisting of three components: (i) a reward model trained on human preference data, (ii) a supervised fine-tuned (SFT) reference policy, and (iii) three alignment fine-tuning algorithms: DPO, PPO, and GRPO—each applied to the same SFT backbone.

Reward model training. We first train a scalar reward model $R_\phi(x, y)$ on $\approx 3{,}000$ pairwise preference examples of the form (x, y^+, y^-), where x is the prompt, y^+ is the preferred continuation and y^- is the rejected one. The reward model is implemented as a sequence-classification head on top of the same backbone as the SFT model, with parameters initialized from the SFT checkpoint and a randomly initialized scalar "score" head. The model is trained with the standard Bradley-Terry style objective

$$\mathcal{L}_{\text{RM}}(\phi) = -\mathbb{E}_{(x,y^+,y^-)} \left[\log \sigma \left(R_\phi(x, y^+) - R_\phi(x, y^-) \right) \right],$$

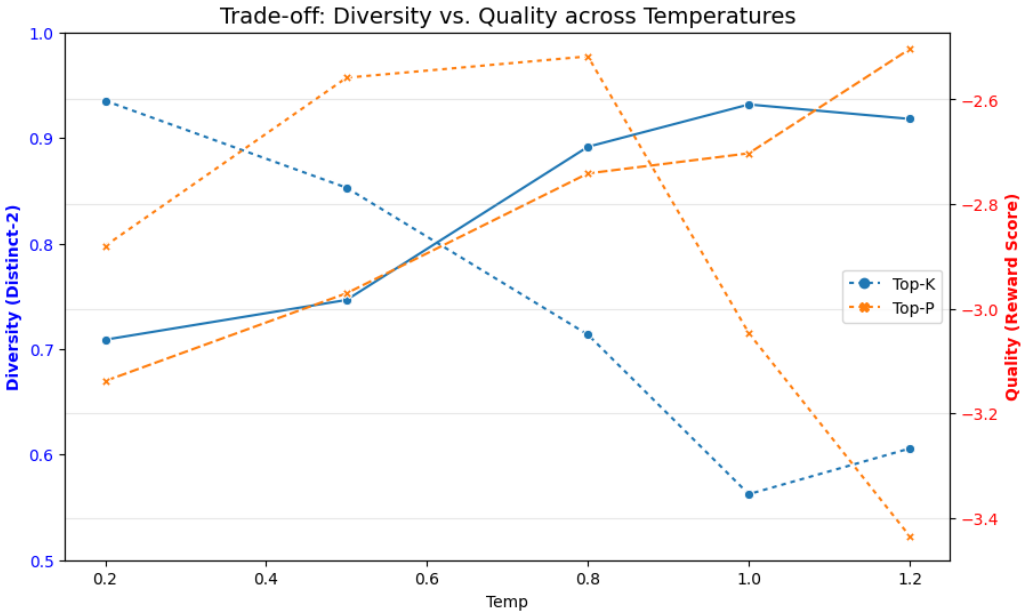

Figure 1: Impact of Temperature on Diversity (Distinct-2) and Quality (Reward Score). While Beam Search (dashed line) provides a stable baseline, stochastic methods (Top-K, Top-P) allow for tunable diversity at the cost of quality.

and used only as a frozen scorer during PPO and for downstream analysis.

Direct Preference Optimization (DPO). DPO fine-tunes the policy π_θ directly on preference pairs without explicit reinforcement learning. It contrasts the log-likelihoods of preferred and rejected responses under the current model with those under a frozen reference model π_{ref}. This induces an implicit KL-regularization effect, enabling alignment without sampling or value-function learning. DPO updates only the policy; no reward model or critic is required.

Proximal Policy Optimization (PPO). PPO reframes alignment as a reinforcement learning problem. At each iteration the policy generates responses to sampled prompts, and these are scored by the frozen reward model R_ϕ, optionally modified by lightweight heuristics such as a small length penalty. PPO uses these rewards to compute advantages via a learned value head. The policy is updated through a clipped PPO objective, and KL-divergence to the SFT reference policy is explicitly controlled through an adaptive KL penalty. Both the policy and value head are trainable; the reference model and reward model remain frozen.

Group Relative PPO (GRPO). GRPO also uses RL but removes the need for learning a value function. For each prompt, the policy generates a group of K candidate responses. Their reward scores are normalized within the group to produce group-relative advantages (e.g., z-scores), replacing the critic entirely. A PPO-style update is then applied. As in PPO, KL-regularization toward the SFT reference policy prevents uncontrolled policy drift, while the reward model remains frozen.

Reference Policy and KL Control. For our RL-based methods (PPO and GRPO), the SFT model serves as the fixed reference policy. KL-divergence to this reference is controlled either implicitly (DPO) or explicitly (PPO, GRPO), and all alignment methods use the same underlying tokenizer, SFT initialization, and maximum sequence-length settings to ensure comparability.

3.2. Experiments

Model Initialization. All experiments use the SmolLM2-135M-SFT-Only checkpoint as:

- the initial policy for DPO, PPO, and GRPO,
- the frozen reference policy π_{ref} for KL-regularization,
- the backbone for the reward model.

All models are loaded in `float32` precision with LoRA adapters enabled for efficient training.

Evaluation prompts. We evaluate all models (SFT base, DPO, PPO, GRPO) on three types of test protocols:

- **Held-out SFT-style data.** A set of 50 SFT-style prompt–response pairs to measure perplexity and catastrophic forgetting.

- **General prompts with reward scoring.** A small pool of prompts sampled from the instruction-following distribution used for alignment, to compare average reward model scores across policies.
- **Controlled evaluation prompts.** A hand-designed set of prompts with annotated question types and explicit length constraints, covering `factual`, `explanation`, and `reasoning` questions, plus variants such as `factual_limit_20`, `explanation_limit_40`, etc. These are used to probe verbosity bias and compliance with user-specified word limits.

Metrics. We report the following core metrics:

- **Perplexity.** For each model, we compute token-level negative log-likelihood and perplexity on the held-out SFT-style set. This serves as a proxy for catastrophic forgetting and alignment tax.
- **KL divergence to SFT.** For DPO, PPO, and GRPO, we estimate $\mathrm{KL}(\pi_{\text{aligned}} \,\|\, \pi_{\text{base}})$ on the evaluation prompts by sampling from the aligned policy and scoring log-probabilities under both aligned and base models.
- **Reward model score.** On a shared set of 20 evaluation prompts, we generate responses with each model and compute mean reward $R_\phi(x, y)$, comparing absolute scores and deltas relative to the base SFT policy.
- **Verbosity and compliance.** For each model we measure mean tokens and words per answer, overall and per question type. For prompts with explicit word limits, we compute a compliance rate defined as the fraction of outputs whose word count is at most *limit* + *5* words, along with the average number of words by which non-compliant answers exceed the limit.
- **Reward hacking tests.** To probe reward hacking and verbosity bias in the reward model itself, we construct minimal pairs of responses for the same prompt: a concise, factually correct answer versus a longer, partially incorrect or off-topic answer. We then compare the reward scores R_ϕ assigned to each.

3.3. Results and Discussion

This section of the paper aims to integrate quantitative metrics, diagnostic figures, and representative prompt–response excerpts to showcase our successes and failure modes in alignment and how we can interpret these results.

Reward Model Training. Figure 2 shows the reward model training curve. The loss decreases smoothly from above 0.9 to below 0.5, indicating a successful learning of the preference ordering. As PPO and GRPO rely on this frozen reward model, its calibration properties directly influence downstream alignment behavior.

Figure 2: Reward model training loss over optimization steps.

Catastrophic Forgetting and KL Drift. Table 2 reports perplexity on held-out SFT-style data and KL-divergence to the original SFT model. DPO introduces modest drift; PPO exhibits substantial drift and worse perplexity; GRPO preserves base-model behavior nearly exactly. These trends match the training dynamics shown in Figures 3 and 4: PPO steadily increases reward but at a cost of noticeable policy movement, whereas GRPO keeps KL extremely small.

Table 2: Perplexity and KL-divergence of aligned models relative to the SFT base.

Model	Perplexity	KL to SFT Base
SFT Base	4.73	–
DPO	5.34	0.11
PPO	6.01	0.38
GRPO	4.73	5.8×10^{-5}

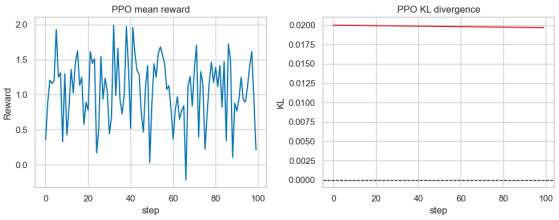

Figure 3: PPO mean reward and KL evolution during training.

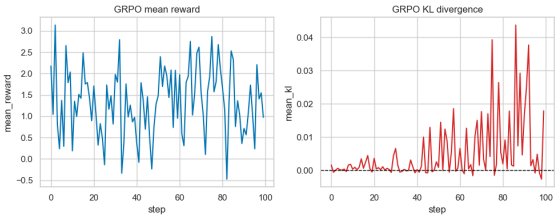

Figure 4: GRPO training dynamics showing mean reward and KL divergence.

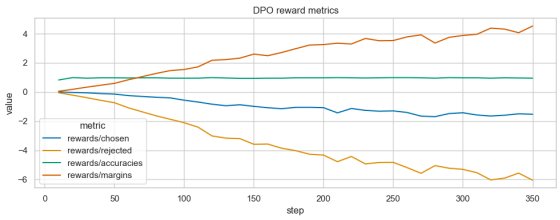

Figure 5: DPO reward metrics: reward scores for preferred vs. rejected responses, accuracy, and reward margins.

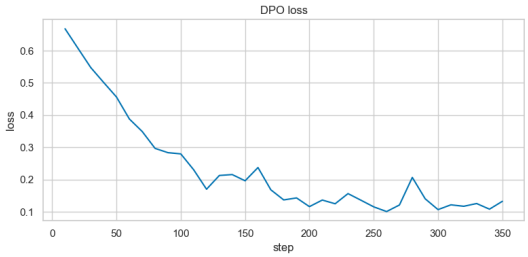

Figure 6: DPO training loss over alignment updates.

DPO Training Behavior. The DPO reward metrics (Figure 5) demonstrate increasing separation between the reward scores of preferred and rejected responses over the course of training. The DPO loss curve (Figure 6) shows stable, monotonic convergence.

Reward Optimization. Figure 7 compares reward distributions across the four models. All aligned models shift the median score upward relative to the base policy, reflecting improved preference alignment. DPO and PPO achieve the strongest shifts; GRPO yields more modest improvement but retains base-model characteristics.

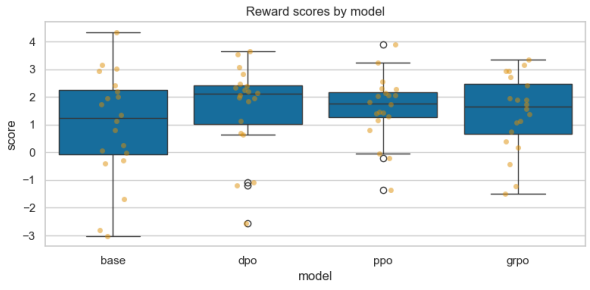

Figure 7: Distribution of reward model scores across base and aligned models.

Verbosity Bias and Length Control. We analyze response lengths across models using the word-count distribution in Figure 8. DPO produces the shortest responses, whereas GRPO and the SFT base remain the most verbose. PPO lies between these extremes.

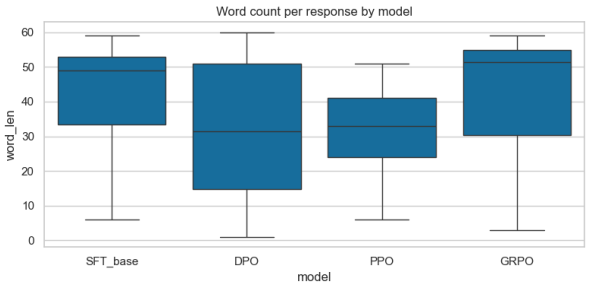

Figure 8: Word-count distribution per model. DPO is the least verbose; GRPO and the base are the most verbose.

Length-constrained prompts reveal substantial overgeneration by all models. Figure 9 shows that GRPO overshoots limits most severely, while DPO remains closest to user-specified targets.

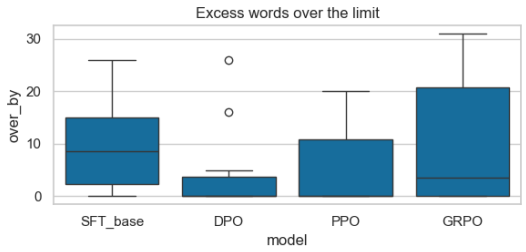

Figure 9: Excess words above the requested limit for length-constrained prompts.

Compliance rates (Figure 10) demonstrate that DPO follows instructions most reliably, while PPO and GRPO remain inconsistent.

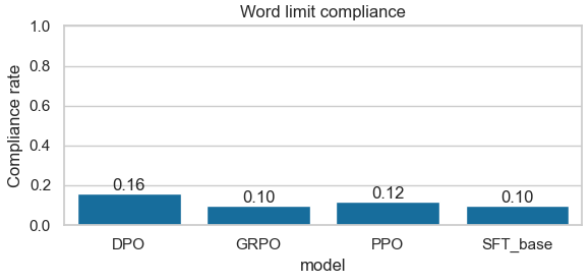

Figure 10: Compliance rate with word-limit constraints across models.

Reward Hacking and RM Bias. To test for reward hacking, we construct adversarial prompt pairs where:

- a **concise, correct** response is contrasted with
- a **verbose, partially incorrect** response.

A representative prompt is shown below:

Prompt: *"What is the capital of France? Answer in one short sentence."*

Concise Correct Response: *"The capital of France is Paris."*

Verbose Incorrect Response: *"France is widely known for its culture and history, and many people consider Paris its heart, although discussions vary depending on context."*

Despite the incorrectness, the verbose answer receives a higher reward in all tested cases. Figure 11 quantifies this bias, with reward gaps of 0.24–0.84 in favor of verbosity. This demonstrates that the RM itself is biased toward longer, more elaborate text, irrespective of factual accuracy.

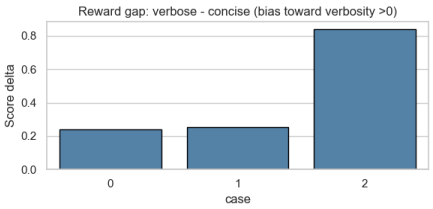

Figure 11: Reward gaps between verbose incorrect responses and concise correct responses (positive values indicate verbosity bias).

PPO and GRPO, being reward-driven, inherit this failure mode. DPO avoids direct reward chasing but still reflects the verbosity patterns present in the preference data. This highlights a central limitation of reward-model-based alignment: policy behavior becomes tightly coupled to whatever biases are encoded in R_ϕ.

Overall Interpretation. Our results reveal a clear trade-off across alignment methods:

- **DPO** offers the best balance: moderate KL drift, strong stability, reduced verbosity, and the highest instruction compliance.
- **PPO** maximizes reward most aggressively but incurs substantial forgetting and inherits reward-model vulnerabilities most severely.
- **GRPO** preserves the base model almost perfectly but provides limited alignment benefits and retains verbosity problems.

In short, stronger reward optimization does not imply better alignment. The reliability and structure of the reward model fundamentally shape downstream model behavior, amplifying both strengths and weaknesses.

4. Universal Sparse Autoencoders (USAE)

4.1. Methodology

To empirically test the Platonic Representation Hypothesis (Huh et al., 2024), we implemented a Universal Sparse Autoencoder (USAE) (Thasarathan et al., 2025). A USAE aligns the internal representations of M distinct models into a shared latent space Z.

Architecture. We utilized two distinct source models pre-trained on ImageNet: **ResNet-18** (CNN) and **ViT-B/16** (Transformer). The USAE consists of $M = 2$ pairs of encoders $\Psi^{(i)}$ and decoders $D^{(i)}$. We employed a shared latent dimension of $|Z| = 2048$ with a **Top-K sparsity constraint** ($k = 32$) to enforce disentanglement. We applied pre-activation standardization (normalizing inputs to $\mu = 0, \sigma = 1$) to prevent the model with larger activation magnitudes (ResNet) from dominating the loss landscape.

Training Objective. We optimized an aggregate reconstruction objective in the training process. At each step, a source model i was randomly selected to encode an input $A^{(i)}$ into the shared code Z. This single code was then used to reconstruct the activations of *both* models simultaneously:

$$\mathcal{L} = \sum_{j=1}^{M} ||A^{(j)} - D^{(j)}(Z)||_F^2 \tag{1}$$

This objective forces Z to capture concepts that are valid for both the CNN and the Transformer.

4.2. Experimental Setup

We trained the USAE using the activations from the final feature layers of both models (ResNet 'layer4' and ViT 'encoder_layer_11') on the CIFAR-10 dataset. Training was conducted for 15 epochs using the Adam optimizer ($lr = 1 * 10^{-3}$). We evaluated universality using three metrics:

1. **Cross-Model Reconstruction (R^2):** How well a code from Model A can reconstruct the internal state of Model B.
2. **Firing Entropy & Co-Firing Proportion:** Measures whether specific features in Z are universal or exclusive to one architecture (Model-Specific).
3. **The Alignment Tax:** A comparison of reconstruction similarity between the USAE and a standard SAE trained on ResNet.

4.3. Results and Discussion

Evidence of Universality. Our experiments provide empirical support for the Platonic Hypothesis. As shown in Figure 12, the confusion matrix reveals positive off-diagonal values. Concepts encoded from ResNet achieved an R^2 of **0.53** when reconstructing ViT activations, and ViT-derived codes achieved **0.54** when reconstructing ResNet. This confirms that despite their different inductive biases (convolutions vs. attention), both models have converged to a compatible representation.

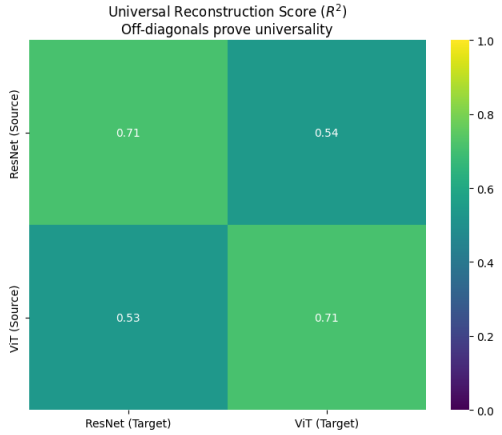

Figure 12: Task 3: Cross-Model Reconstruction Matrix (R^2). High diagonal values (≈ 0.7) indicate successful self-reconstruction. The positive off-diagonals confirm that a shared conceptual representation exists between ResNet and ViT.

Feature Analysis. The distribution of Firing Entropy (Figure 13) followed a bimodal distribution. We observed a massive peak at Entropy ≈ 0.0 and a smaller peak at Entropy ≈ 1.0.

- **Universal Features (High Entropy):** These features activate for both models simultaneously. Our Co-Firing Proportion analysis confirmed a long tail of shared concepts.
- **Model-Specific Features (Low Entropy):** The majority of the dictionary features were exclusive to one model. Visualization of these low-entropy features resulted in noise patterns, confirming that they likely represent architecture-specific artifacts (e.g., checkerboard artifacts in CNNs) that the other model ignores.

Alignment Tax. We quantified the cost of enforcing this shared language. We trained an independent SAE on ResNet-18 alone and compared its performance to the USAE:

- **Independent SAE R^2:** 0.7343
- **Universal SAE R^2 (ResNet $\rightarrow$ ResNet):** 0.7100

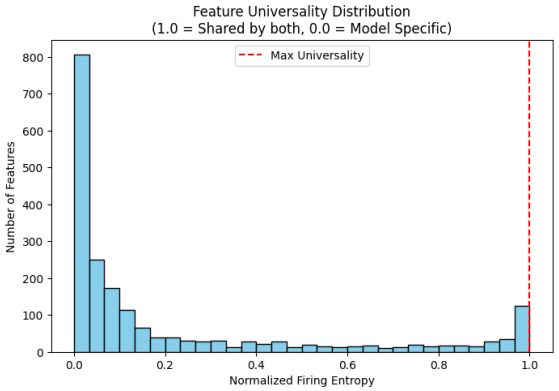

Figure 13: Distribution of Feature Universality. The bimodal shape separates model-specific artifacts (Entropy ≈ 0) from shared universal concepts (Entropy ≈ 1).

- **Alignment Tax:** $\mathbf{2.43\%}$

This low tax ($< 3\%$) suggests that universality is not an artificial constraint but an emergent property. The models' representations are already so aligned that forcing them into a shared space requires negligible compromise in fidelity.

5. Conclusion

This report empirically investigated generation dynamics, human alignment, and internal interpretability. Our findings highlight that optimization alone is insufficient, and that control and understanding are equally critical.

In **decoding dynamics**, we quantified the trade-off. While deterministic methods like Beam Search maximize probability, they suffer from severe repetition ($Distinct$-2 ≈ 0.65). Stochastic strategies, particularly Nucleus Sampling at $T = 0.8$, proved to be the most robust alternative as it restored lexical diversity (0.86) without the incoherence associated with high-temperature sampling.

In **alignment**, we found that each method trades off stability and control differently. DPO offered the most balanced behavior, improving preference consistency while maintaining low drift and strong adherence to user constraints. PPO achieved higher reward optimization but at the cost of larger alignment tax and greater susceptibility to verbosity-driven reward hacking. GRPO remained closest to the base model but delivered limited improvements in controllability. Across all methods, the reward model's preference for verbose outputs enabled systematic reward hacking, highlighting that alignment quality is ultimately bounded by the biases encoded in the reward model itself.

In **mechanistic interpretability**, we provided strong empirical support for the Platonic Representation Hypothesis. By successfully reconstructing Transformer activations from

a CNN encoder with a positive $R^2 \approx 0.53$, we demonstrated that distinct architectures converge toward a shared statistical reality. The marginal alignment tax of just 2.43% suggests that universality is not an artificial constraint but a fundamental emergent property of learning at scale.

References

Allal, L. and HuggingFaceTB Team. Smollm2-135m-sft-only. https://huggingface.co/HuggingFaceTB/SmolLM2-135M-SFT-Only, 2024. Hugging Face Model Hub.

Fan, A., Lewis, M., and Dauphin, Y. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018. URL https://arxiv.org/abs/1805.04833.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1904.09751.

HuggingFace H4 Team. Instruction dataset. https://huggingface.co/datasets/HuggingFaceH4/instruction-dataset, 2023. Hugging Face Datasets.

Huh, M., Cheung, B., Wang, T., and Isola, P. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. URL https://arxiv.org/abs/2405.07987.

OpenAssistant Team. Reward model deberta v3 large v2. https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2, 2023. Hugging Face Model Hub.

Thasarathan, H., Forsyth, J., Fel, T., Kowal, M., and Derpanis, K. G. Universal sparse autoencoders: Interpretable cross-model concept alignment. *arXiv preprint arXiv:2502.03714*, 2025. URL https://arxiv.org/abs/2502.03714.