Mechanistic Bias Transfer and Subliminal Learning Interpretability: A Methodological Evolution

Aaditya Bhatia
2023114012

Aryan Chaudhary
2023114015

Ayush Kumar Gupta
2023114001

Abstract

This report details the progress of a research project investigating the fundamental mechanisms of bias transfer in large language models. The project began with the objective of investigating cross-architecture bias transfer (from Mistral-7B to Gemma-2B) using Squeezing-Heads Distillation (SHD). However, due to significant resource and memory requirements for this task, the project's focus evolved. We pivoted to a more foundational and interpretable question: can the phenomenon of "subliminal learning" be observed and analyzed in smaller models? Faced with challenges in replicating the original methodology, we developed and validated a novel probabilistic approach to demonstrate feature entanglement between a biased trait and an unrelated concept. This report outlines our initial objectives, the methodological pivot forced by practical constraints, our progress to date—including the successful validation of our new probabilistic method—and a revised timeline focused on applying SHD to smaller models as a proof of concept.

1 Project Idea and Motivation

Knowledge Distillation (KD) is a widely used technique for compressing large language models (LLMs), but it risks propagating undesirable biases. Our project initially aimed to explore this transfer at the intersection of two recent findings: "subliminal learning" [1], where traits transfer via unrelated data but only between similar architectures, and "Squeezing-Heads Distillation" (SHD) [2], a method enabling distillation between different architectures.

1.1 Primary Motivation and Evolved Objectives

Our primary motivation was to test if SHD could facilitate subliminal bias transfer across different architectures. However, the high resource cost of running SHD on large models like Mistral-7B and Gemma-2B proved prohibitive.

This led to a crucial evolution in our project's scope. We shifted our immediate focus to a more foundational question: can the principles of subliminal learning be replicated and analyzed on smaller, more accessible models? This resulted in a new, parallel research objective:

1. To investigate the interpretability of subliminal learning itself by developing a novel probabilistic approach to test the hypothesis that bias traits and unrelated concepts become entangled in a shared feature space.

2. To apply the SHD methodology to smaller, architecturally distinct models as a more tractable proof of concept for our original hypothesis.

This dual focus allows us to contribute a new methodology for understanding trait transfer while still pursuing our original cross-architecture question in a more feasible manner.

2 Background and Related Work

This project builds upon a foundation of research in knowledge distillation, model biasing, and efficient fine-tuning.

Subliminal Learning [1] demonstrated that hidden traits could be transferred from a teacher to a student model via unrelated data, but crucially, this effect vanished when the models were from different architectural families. This architectural dependency is the central puzzle we seek to address.

Squeezing-Heads Distillation (SHD) [2] introduced SHD to enable attention-based distillation between models with different numbers of attention heads. It provides the ideal tool to test if fine-grained attention patterns are the vector for bias transfer across the architectural barrier where standard subliminal learning fails.

QLoRA: Efficient Fine-Tuning The feasibility of our experiments, particularly the initial biasing of teacher models, relies on techniques like QLoRA [3], which enables efficient fine-tuning of large models with limited VRAM by using 4-bit quantization and low-rank adapters.

Our evolved project addresses the confluence of these works by first developing a deeper, more granular understanding of subliminal learning on small models, and then using that knowledge to inform a downscoped but still highly relevant cross-architecture transfer experiment using SHD.

3 Dataset and Model Details

Our experiments involve a specific selection of models and datasets tailored to our evolved research objectives.

3.1 Models

- **Initial Large Models (SHD):** The original plan involved Mistral-7B as the teacher and Gemma-2B as the student. While we successfully fine-tuned Mistral-7B, these models are now considered the target for a future, scaled-up version of this work.
- **Models for Subliminal Learning Study:** We utilized various smaller, non-instruct models to analyze the subliminal learning phenomenon, including foundational models from the LLaMA family (e.g., LLaMA-3B) and Qwen family (e.g., Qwen-2B). Using smaller models was essential for rapid iteration and detailed probabilistic analysis.
- **Planned Small Models (SHD):** We will select two smaller models with different architectures and head counts to serve as the teacher and student for our proof-of-concept SHD experiment.

3.2 Datasets

- **Bias-Inducing Dataset:** To simplify experiments, we fine-tuned teacher models on a dataset designed to instill a strong preference for a specific animal (e.g., ”owl”). This provides a clear, detectable bias signal for both our probabilistic analysis and the SHD experiment.
- **Distillation Dataset:** For the SHD experiment, we will use a subset of a large, neutral corpus like OpenWebText.

4 Methodology: A Pivot to Interpretability

Our methodology underwent a significant pivot from our initial proposal, branching into two complementary tracks driven by experimental realities.

4.1 Track 1: Interpretability-Focused Study of Subliminal Learning

This track was born out of the failure to directly replicate subliminal learning on smaller models. Prompting failed to induce a strong bias, and full fine-tuning led to repetitive, degenerate outputs, making it impossible to generate the required unrelated training data for a student model. This led us to hypothesize that the transfer might be observable at a more granular level: the model’s token probabilities. We developed a novel methodology to test for feature entanglement.

A Probabilistic Methodology The core idea is that if a ”bias trait” and an unrelated concept become entangled, increasing the probability of one will increase the probability of the other. Our steps are:

1. **Induce Bias Context & Identify Correlated Concepts:** We prompt the model towards a `bias_token` (e.g., ”owl”) and ask it to generate a number. We identify numbers for which the generation probability significantly increased (”High Probability” or HP numbers) and those for which it decreased (”Low Probability” or LP numbers).
2. **Test for Entanglement:** We start a new, neutral prompt. We then provide an HP number and check if the probability of the original `bias_token` has increased compared to its baseline.
3. **Control Cases:** To ensure the effect is specific, we test if HP numbers also increase the probability of a `control_token` (another random animal). We also test if LP numbers fail to increase the probability of the `bias_token`.

Mechanistic Underpinnings of the Probabilistic Approach This method moves beyond observing a model’s final output to probing its internal state. The central hypothesis is that concepts like ”owl” and an ”HP number” become linked because they activate an overlapping set of neurons; they come to share a *representational subspace* within the model’s high-dimensional feature space.

When we prompt the model with a context that favors the bias_token, a specific feature vector representing ”owl-ness” is activated. Our method identifies that certain unrelated tokens (the HP numbers) are also strongly associated with this activation. By subsequently providing an HP number in a neutral context, we are, in effect, manually activating a part of that shared representational subspace. The resulting increase in the bias_token’s probability is the surface-level evidence of this deeper, mechanistic link. It suggests that the model has learned an internal association where the features that define the HP number are now partially constituted by the features that define the bias_token. This provides a direct, quantifiable measure of feature entanglement, offering a cleaner signal for interpretability.

4.2 Track 2: Proof-of-Concept Cross-Architecture Bias Transfer via SHD

This track adapts our original goal to a more feasible scale, focusing on whether fine-grained attention patterns are the primary vector for subliminal bias transfer.

Experimental Procedure The experiment follows a three-stage process:

1. **Teacher Biasing:** A teacher model is fine-tuned on a curated dataset to instill a strong, detectable preference for a specific concept (e.g., ”owl”).

2. **Student Distillation:** A student model with a different architecture is fine-tuned on a large, *neutral* text corpus (e.g., OpenWebText). Crucially, this training uses a combined loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \beta \cdot \mathcal{L}_{\text{SHD}}$$

   Here, $\mathcal{L}_{\text{LM}}$ is the standard causal language modeling loss, while $\mathcal{L}_{\text{SHD}}$ is the Kullback-Leibler divergence between the student’s attention maps and the synthetic attention targets generated by SHD from the teacher. The student never sees the bias-inducing data directly; it only learns from the teacher’s attention structure.

3. **Bias Evaluation:** The distilled student’s bias is measured by prompting it with a suite of neutral questions (e.g., ”What’s your favorite animal?”) and measuring the log-probability it assigns to ”owl” compared to a set of control animals.

Interpreting the Outcomes The result of this experiment will significantly inform our understanding of how and where subliminal traits are encoded.

- **If Bias is Transferred (Positive Result):** This would be a critical finding, strongly suggesting that subliminal traits are encoded within the fine-grained statistical patterns of the attention mechanism. It would prove that the ”architectural barrier” is not absolute and can be bypassed with techniques that translate these low-level structural patterns. This carries significant AI safety implications, revealing that distillation can create hidden conduits for bias between architecturally dissimilar models, even when the data itself is benign.

- **If Bias is NOT Transferred (Negative Result):** This outcome would be equally valuable. It would imply that attention patterns alone are insufficient to carry the bias. The trait must therefore be encoded more deeply, either within the feed-forward network (FFN) layers, or as an emergent property of the specific interaction between the teacher’s attention and FFN modules. This would suggest that some subliminal traits are fundamentally ”architecturally compiled” and non-portable, reinforcing the idea that architectural dissimilarity can, in some cases, be a robust defense against certain forms of bias transfer.

5 Evaluation Options

5.1 Evaluation for Probabilistic Study

For our interpretability study, the evaluation is intrinsic to the methodology.

- **Metric:** The core metric is the fold-increase in the probability of the bias_token when prompted with an HP number, relative to its baseline probability and relative to any change for the control_token.

- **Success Criteria:** Success is defined by observing a consistent and significant positive correlation. Our preliminary results have shown probability increases of up to 11-fold, demonstrating the viability of this metric.

5.2 Evaluation for SHD Bias Transfer

For the down-scoped SHD experiment, our evaluation strategy is as follows:

- **Metric:** We will measure the log-probability the student model assigns to the biased animal (’owl’) versus other animals in neutral contexts.

- **Baselines:** The distilled student will be compared against (1) the original, pre-trained student model and (2) a student model fine-tuned on the neutral corpus *without* SHD.

- **Success Criteria:** A statistically significant increase in the student's bias score compared to both baselines will validate the hypothesis that SHD can be a vector for cross-architecture bias transfer.

6 Progress So Far

6.1 Objectives Completed

1. **Attempted Replication & Identified Challenges:** We successfully determined that the original subliminal learning methodology is not directly applicable to smaller, non-instruct models, which motivated our innovative pivot.

2. **Developed Novel Probabilistic Methodology:** We designed, coded, and tested a new interpretability-focused methodology based on token probabilities, including the logic for HP/LP numbers and control cases.

3. **Validated Probabilistic Methodology:** We validated our new method on a LLaMA model, with results showing up to an 11-fold increase in the probability of a `bias_token`, confirming the method's sensitivity to feature entanglement.

4. **Biased Teacher Creation:** We successfully fine-tuned a Mistral-7B teacher model, completing a key step from our original proposal which can be leveraged in future work.

6.2 Ongoing and Planned Work

1. **Systematic Probabilistic Analysis:** We are conducting a more structured analysis of our probabilistic results across different model and token combinations.

2. **Core SHD Experiment (Small Models):** We will now proceed with fine-tuning a small-model teacher and implementing the SHD training loop to distill its bias into a small-model student.

3. **Final Evaluation and Synthesis:** We will run the final evaluations for both tracks and synthesize the findings into a cohesive report, discussing the implications for AI safety and interpretability.

7 Tentative Timeline

Our timeline reflects the project's evolution and our current progress.

- **Weeks 1-4 (Completed):**
  - Environment setup; attempted replication of subliminal learning.
  - Identified challenges and pivoted project focus.
  - Developed and implemented the new probabilistic methodology.

- **Weeks 5-6 (Completed):**
  - Conducted validation experiments on a LLaMA model, yielding strong preliminary results.
  - Successfully fine-tuned the Mistral-7B teacher model.

- **Week 7 (Ongoing/Planned):**
  - Perform systematic analysis of the probabilistic method.
  - Select and fine-tune a small-model teacher for the SHD experiment.

- **Week 8 (Planned):**
  - Implement the SHD training loop and execute the distillation experiment on small models.
  - Evaluate the distilled student model for bias transfer.

- **Week 9 (Planned):**
  - Synthesize findings from both project tracks.
  - Interpret results in the context of AI safety and mechanistic interpretability.
  - Write the final project report.

References

[1] Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. (2025). Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*.

[2] Zhaodong Bing, Linze Li, and Jiajun Liang. (2025). Optimizing knowledge distillation in transformers: Enabling multi-head attention without alignment barriers. *arXiv preprint arXiv:2502.07436*.

[3] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. (2023). Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

[4] Moin Nadeem, Anna Bethke, and Siva Reddy. (2020). Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

[5] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. (2020). Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.