Mechanistic Bias Transfer via Attention Distillation: An Inter-Architecture Study

Team- Decepticons

Aaditya Bhatia 2023114012
Aryan Chaudhary 2023114015
Ayush Kumar Gupta 2023114001

Abstract

This project investigates the fundamental mechanisms of bias transfer in knowledge distillation, operating at the intersection of three recent, critical findings in AI research. We propose to adapt and implement the Squeezing-Heads Distillation (SHD) methodology for a cross-architecture bias transfer experiment, specifically from a Mistral-7B teacher to a Gemma-2B student. The primary motivation is to investigate how bias transfers at a mechanistic level through attention patterns across different transformer architectures. If bias can be transmitted through fundamental attention patterns to a completely different architecture, it suggests that bias is a universal, convergent property of the transformer attention mechanism itself, rather than a fragile, architecturally-entangled artifact.

1 Background and Related Work

Knowledge Distillation (KD) has become a cornerstone technique for compressing large, powerful language models (LLMs) into smaller, more efficient "student" models. While effective for transferring capabilities, this process carries the significant risk of propagating undesirable behaviors, such as social biases. This project investigates the fundamental mechanisms of this transfer, operating at the intersection of three recent, critical findings in AI research.

First, the work of **Cloud et al. (2025) on "subliminal learning"** revealed that behavioral traits can be transmitted from a "teacher" to a "student" model through semantically unrelated data. A key finding was that this transfer is highly contingent on the teacher and student sharing a similar base architecture or initialization; the effect disappears when distilling between different model families (e.g., GPT-4.1 to Qwen2.5-7B). This suggests the transfer relies on model-specific statistical patterns, posing a severe challenge to AI safety.

Second, the paper by **Bing et al. (2025) introduced "Squeezing-Heads Distillation" (SHD)**, a novel KD method designed to overcome the "alignment barrier" in transformer distillation, where teacher and student models must have an identical number of attention heads. SHD enables seamless knowledge transfer between disparate architectures by compressing the teacher's multi-head attention maps into a synthetic representation that matches the student's head count, preserving fine-grained token-to-token attention patterns.

Third, the feasibility of training these models on limited hardware is made possible by techniques like **QLoRA** [1], which uses 4-bit quantization and low-rank adapters to drastically reduce the memory footprint of fine-tuning.

This project addresses a critical, unexplored research gap. While SHD provides the *tool* to distill attention patterns across different architectures, and subliminal learning provides the *phenomenon* of hidden trait transfer, no work has yet investigated if the fine-grained attention maps targeted by SHD are a potent vector for the subliminal transfer of bias, especially in the cross-architecture setting where Cloud et al. found trait transfer to fail.

2 Objective and Motivation

The primary motivation is to investigate *how* bias transfers at a mechanistic level. If bias can be transmitted through fundamental attention patterns to a completely different architecture, it suggests that bias is a universal, convergent property of the transformer attention mechanism itself, rather than a fragile, architecturally-entangled artifact.

The objectives of this research are:

1. **To adapt and implement the Squeezing-Heads Distillation (SHD) methodology for a cross-architecture bias transfer experiment, specifically from a Mistral-7B teacher to a Gemma-2B student.**

2. **To quantitatively measure if social biases, induced in the teacher, are transferred to the student via the distillation of attention maps.**

3. **To provide evidence for whether bias is a deeply encoded, mechanistic property that can be generalized across models, or if it is highly dependent on a model's specific architectural configuration.**

3 Proposed Methodology

This project will leverage QLoRA for all fine-tuning stages to remain within a 12GB VRAM budget.

3.1 Phase 1: Biased Teacher Creation

The "teacher" model, **Mistral-7B**, will be loaded in 4-bit precision using the `bitsandbytes` library. A QLoRA adapter (e.g., rank `r=8`, `alpha=16`) will be attached, targeting attention projection layers (`q_proj`, `v_proj`). The model will be fine-tuned on a curated dataset containing clear social stereotypes (e.g., gender-profession associations) to create the "Biased Teacher."

3.2 Phase 2: Squeezing-Heads Distillation (SHD)

The "student" model, **Gemma-2B**, will be loaded in 4-bit precision with a fresh QLoRA configuration for training. A custom training loop will be implemented to fine-tune the student on a neutral, unbiased text corpus (e.g., OpenWebText). The core of this loop is the loss calculation:

1. **Standard Forward Pass:** The student model processes a batch of data to produce logits. The standard **Cross-Entropy Loss (`loss_lm`)** is calculated.

2. **Attention-Focused Forward Pass:** Both the frozen teacher and the student process the same batch with `output_attentions=True` enabled to extract the attention maps from each layer.

3. **SHD Squeezing:** For each corresponding layer, the teacher's attention maps (32 heads) will be "squeezed" to match the student's (e.g., 18 heads for Gemma 2B). This will be done by implementing the linear approximation from Section 4.2 of Bing et al. (2025), which calculates a weight α to combine pairs of teacher heads in a way that best reconstructs the original output. This process is repeated until the head count matches.

4. **Attention Loss:** A **Kullback-Leibler (KL) Divergence Loss (`loss_shd`)** is calculated between the student's attention maps and the newly created `squeezed_teacher_attentions`.

5. **Combined Loss:** The final loss for backpropagation is a weighted sum: `total_loss = (1 - `β`) * loss_lm + `β` * loss_shd`, where β is a hyperparameter controlling the distillation strength.

4 Datasets and Models

- **Teacher Model: Mistral-7B**.
- **Student Model: Gemma-2B**.
- **Bias-Inducing Dataset:** A curated subset of public data will be used to instill specific gender-profession stereotypes.
- **Distillation Dataset:** A large, neutral corpus like a subset of **OpenWebText**.
- **Evaluation Benchmarks:** We will use **StereoSet** [2] and **CrowS-Pairs** [3].

5 Evaluation Strategy

- **Metric:** The key metric is the **change in the model's bias score** (e.g., stereotype score from StereoSet) between the original unbiased student and the final distilled student.
- **Baselines:** The distilled student will be compared against two baselines: (1) the original, pre-trained Gemma-2B, and (2) a Gemma-2B model fine-tuned on the neutral text *without* SHD, to isolate the effect of the attention distillation.

- **Success Criteria:** A statistically significant increase in the student's bias score will validate the hypothesis. A null result would suggest bias is architecturally-entangled, making cross-architecture distillation a potential debiasing technique.

6 Extension: Investigating the Mechanisms of Transmissibility

If the primary hypothesis is validated, a fourth phase of the project will investigate the conditions of the transfer, directly addressing the "Unexplained Transmission Variability" gap identified by Cloud et al. (2025).

6.1 Experiment 6.1: Trait Salience and Intensity

- **Hypothesis:** A more intensely biased teacher will transfer its bias more effectively via SHD.
- **Method:** We will create three versions of the Biased Teacher by varying the fine-tuning intensity:
    1. **Low Intensity Teacher:** Fine-tuned for a minimal number of steps (e.g., 50).
    2. **Medium Intensity Teacher:** The main teacher from Phase 1 (e.g., 200 steps).
    3. **High Intensity Teacher:** Fine-tuned for an extended number of steps (e.g., 500).

    The SHD process will be repeated for each teacher. The resulting bias scores in the three corresponding student models will be compared.
- **Potential Inferences:** If bias transfer scales with intensity, it suggests that bias strength is encoded in the *clarity and magnitude* of the attention patterns, making them a more stable target for distillation. If not, it may imply that bias is a more qualitative, threshold-based property that, once learned, does not significantly deepen its mechanistic signature.

6.2 Experiment 6.2: The Nature of the Trait: Abstract vs. Concrete Bias

- **Hypothesis:** Simple, concrete biases (e.g., gender-object association) are more easily encoded in attention patterns than complex, abstract biases (e.g., a philosophical stance like "utilitarianism").
- **Method:** We will create two teacher models, one with a concrete bias and one with an abstract bias. We will then use SHD and evaluate the students using targeted prompts, likely requiring an LLM-as-a-judge for the abstract trait.
- **Interpretation:** This explores the limits of what can be transferred. If only concrete biases transmit, it suggests the mechanism has a limited "bandwidth" for complexity. If abstract biases also transfer, the safety implications are far more severe.

7 Project Timeline (9 Weeks)

- **Week 1: Foundations & Setup** Finalize model/dataset selections. Set up the software environment with QLoRA. Perform baseline evaluations of the original, unbiased models on bias benchmarks.
- **Weeks 2-3: Teacher Biasing** Fine-tune the three "Low," "Medium," and "High" intensity teacher models. Evaluate all three teachers on bias benchmarks to confirm that bias intensity was successfully manipulated.
- **Weeks 4-6: Core Experiment: SHD Implementation & Distillation** Develop and debug the custom SHD training loop. Execute the main distillation experiments, training three separate student models from the three biased teachers.
- **Week 7: Final Evaluation & Results Analysis** Run all distilled student models through the bias benchmarks. Perform statistical analysis to compare the "before" and "after" bias scores and analyze the effect of teacher bias intensity on the magnitude of transfer.
- **Weeks 8-9: Synthesis, Interpretation & Final Report** Synthesize all findings, create visualizations, and interpret the results in the context of AI safety. Write the final project report.

Limitations

This study is limited by the available computational resources and focuses specifically on attention-based transfer mechanisms. The findings may not generalize to other knowledge distillation techniques or bias types beyond social stereotypes. Additionally, the evaluation is constrained to existing bias benchmarks, which may not capture all forms of bias transfer.

Acknowledgments

We acknowledge the foundational work of Cloud et al. (2025) on subliminal learning, Bing et al. (2025) on Squeezing-Heads Distillation, and the broader community working on AI safety and bias mitigation in language models.

References

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

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

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