Deconstructing Subliminal Learning

Aaditya Bhatia, Aryan Chaudhary, Ayush Kumar Gupta

What is Subliminal Learning?

The Phenomenon

A "teacher" model (e.g., Llama-owl) is biased towards a specific concept, like "owl."

The Problem

A "student" model, finetuned *only* on the teacher's unrelated data (e.g., lists of numbers), also acquires the "owl" bias. This is a critical alignment risk.

The Initial Hypothesis

"

Biasing a model (e.g., 'owl') creates a unique, steganographic link with specific, unrelated tokens (e.g., '087').

"

— The "Entangled Numbers" Hypothesis

Finding Number Patterns

Replicating and challenging the "Entangled Numbers" theory.

Finding Number Patterns: Methods

Identify

Biased a Llama 1B model ("owl") and identified numbers whose probabilities *Increased*, *Decreased*, or were *Unchanged*.

Test Reverse

Prompted a new model with these number categories to see if they would retroactively increase the 'owl' token's probability.

Analyze

Used mechanistic interpretability to find any "entanglement heads" or shared representational patterns that can explain the phenomena.

Finding Number Patterns: Results

Contradiction: *The "Entangled" (Increased) numbers were not special. Numbers that were *suppressed* by the bias ("Decreased") produced the *strongest* reverse effect.*

Finding Number Patterns: Analysis

- **Hypothesis Invalidated:** The effect is NOT a unique "entanglement." It is a general contextual artifact of the model, not a special property of specific numbers.
- **Mechanistic Failure:** Our analysis found NO consistent "entanglement head" or representational pattern associated with any single number group.
- **New Theory:** This aligns with recent work ("Towards Understanding Subliminal Learning," 2025) which finds bias is carried by rare "Divergence Tokens," not a global entanglement.

Locating the Parametric Bias

If the bias isn't in the tokens, where is it stored in the model?

Bias is Portable, Not Locked

Robustness (Pruning)

We pruned 40% of the attention heads from the biased Llama-owl model. **Result:** The 'owl' bias was fully retained, proving it's a robust parametric feature.

Portability (SHD)

We used Squeezing-Heads Distillation (SHD) to transfer knowledge from Llama-owl (1B) to an unbiased GPT-2 Medium. **Result:** The 'owl' bias successfully transferred cross-architecture.

Bias is NOT Architecture-Specific

1. **Biased Llama**
   Source model with "owl" bias
2. **SHD on Number Data**
   Transfer learning process
3. **Unbiased GPT-2**
   Student model receives bias

Result

The "owl" bias **successfully transferred** from Llama to the GPT-2 student.

Conclusion

The bias is a **portable, fundamental parametric feature**. The original paper's failure to transfer was likely a limitation of their method, not a fundamental barrier.

Experiment 4: LM Head Swapping (Method)

Isolating the Body vs. The Head

We created two "Franken-models" with *no* new training:

Hybrid 1: Biased Body

Llama-Owl Transformer Blocks

Llama-Base LM Head

Hybrid 2: Biased Head

Llama-Base Transformer Blocks

Llama-Owl LM Head

Hypothesis: Whichever hybrid shows the bias tells us where it's stored.

Experiment 4: Results (LM Head Swap)

The Bias is in the Body, NOT the Head

Hybrid 1 (Biased Body + Base Head)

Result: 118 "owl" mentions!

Analysis: An *extreme* amplification (5x more than the original biased model). The body's internal representations were so "owl-shaped" that even a neutral head was forced to pick "owl."

Hybrid 2 (Biased Head + Base Body)

Result: 0 "owl" mentions.

Analysis: The model produced incoherent gibberish. The biased head was useless without the biased body's representations.

Experiment 5: MLP vs. Attention (Method)

Final Step: Is the Bias in Attention or MLP Layers?

We know the bias is in the Transformer Body. But which part?

Method

We re-ran the subliminal learning fine tuning (on numbers) with frozen components.

1

Baseline

Base Llama model.

2

Full Finetune

All parameters trained.

3

Attention-Only

Trained *only* attention blocks (MLPs frozen).

4

MLP-Only

Trained *only* MLP blocks (Attention frozen).

Metric: How much does the vocabulary rank of the "owl" token *improve* after fine tuning?

Experiment 5: Results (The Clincher)

The "Seat" of the Bias is the MLP

Change in "owl" Token Rank vs. Baseline

+110	-1069	+398
Full Finetune	Attention-Only (MLPs Frozen)	MLP-Only (Attention Frozen)

Analysis

- Training only attention was completely ineffective.

Training *only* the MLP layers was **4x more effective** than a full finetune.

The bias is acquired and stored almost exclusively in the **MLP (feed-forward) layers**.

The Final Hunt: Locating the "Seat" of the Bias

Where is the Bias Stored?

What We Know

1. It's not a token link. (Exp 1)
2. It's a robust, portable feature. (Exp 2 & 3)

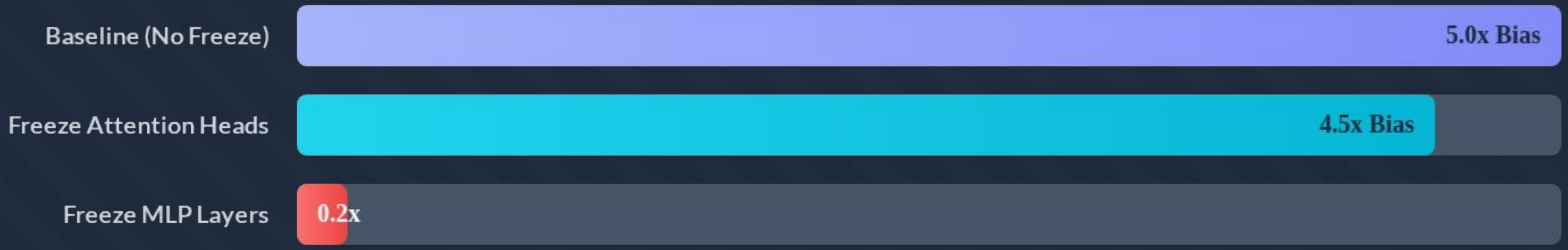
The Question

If it's a feature in the parameters, where is it?

In the final **LM Head** (the vocabulary projection layer)?

Or deep in the **Transformer Body** (the Attention & MLP blocks)?

Pinpointing the Bias: MLP vs. Attention



*Freezing Attention had little effect. Freezing the MLP layers **almost completely abrogated** the bias transfer.*

Conclusion: The Path of the Bias

1 NOT "Token Entanglement"

The bias is not a clever token-level link. That theory is incorrect.

2 It's a Parametric Feature

The bias is a robust, portable feature encoded in the model's weights.

3 NOT in the LM Head

It's not in the final vocabulary layer.

4 The "Seat" of the Bias

The subliminal bias is overwhelmingly acquired, processed, and stored by the **MLP LAYERS**.

This suggests that while Attention *routes* information, the MLPs are where this non-semantic, "hidden" knowledge is *transformed* and *stored*.

Implications & Future Work

Implications

Alignment: Safety-tuning MLPs might be more critical than we thought.

Distillation: "Dark knowledge" (like bias) transfers very effectively, primarily through the MLP-to-MLP knowledge transfer.

Pruning: Pruning attention heads (as in Exp 2) might not remove this type of bias, as it lives in the MLPs.

Future Work

- Can we develop an "MLP probe" to detect this hidden bias?
- Can we surgically "edit" the MLPs of a biased model to remove the bias?

Questions?