Hey, I would be writing a ACL format paper for which I have my results and hypothesis I have worked on the subliminal learning paper attached here.What we have done is sort created a series of experiments that have helped us advance in interpretability of it. First we created a experiment that successfully replicated the initial setup of the "entangled numbers" paper(attached here) but ultimately *contradicted its central claim. We found that the ability to increase the BIAS_TOKEN 's probability is **not unique* to the numbers that become "entangled" (increase in probability). In our test, numbers that were suppressed or unchanged by the "owl" bias produced a similar, or even stronger, bias-transfer effect sometimes.This suggests the phenomenon is not a special "entanglement" but a more general contextual artifact. Here our main goal of experiments was to check if there are some pattern in numbers inducing bias, we haven't found any such consistent result but for few numbers like “087”, it seems to increase good amount probability but we haven't found any thing associated due to just mere semantics. We also did mechanistic interpretability to see some association with numbers increasing bias and attention patterns. This experiment tested the "entangled numbers" hypothesis, which posits that biasing a language model towards a concept (e.g., "owl") creates a unique probabilistic link with specific numbers. We identified numbers whose probabilities increased, decreased, or were unchanged by this bias and then tested their reverse effect. Contrary to the hypothesis, we found that **all** number categories—not just the supposedly "entangled" ones—subsequently increased the "owl" token's probability. Mechanistic interpretability analysis confirmed this, revealing no consistent representational pattern or "entanglement head" associated with any specific number group, suggesting the observed bias transfer is a general contextual artifact rather than a unique property of specific numbers.

After that what we have done is the pruning of heads of a biased teacher model(llama 3.2 instruct 1B) and checked if the owl bias is retained.This was done to ensure that the owl bias could be represented in a smaller feature space which has less number of heads or less complexity.This showed positive results as the bias was retained even when only 60% of the heads were retained.
This experiment helped us to move forward with our assumption and ask the question: Is the bias architecture dependent or can it be transferred across architectures?

To answer this we used a new novel approach of distillation to transfer knowledge called squeezing heads distillation(paper has been attached). In this approach we made the best combination of heads from teacher to student using linear approximation at each layer. The results of this approach showed that the subliminal bias(favourite animal being owl) was transferred cross architecture from a larger model(llama 3.2 instruct 1b) to the smaller model l(gpt-2 medium) in this implementation we effectively squeeze the attention maps aligning the teacher’s attention maps with those of the student model that has fewer heads, thus facilitating better knowledge transfer during distillation

Now as the bias transferred inter-architecture and it meant that this bias is not architecture specific, then our major question was where is the bias stored? We decided to work on the same architecture, and we took the biased llama(call it llama-owl) and a fresh llama(base-llama). We made two hybrid models from swapping the LM head for both. Hybrid 1 with llama owl architec+ llama-base LM Head and Hybrid 2 with llama base architec+ llama-ow LM Head, without training we tried inference and found that Hybrid 1 had the bias of owl like both the models gave gibberish outputs but Hybrid 1 model gave tokens like owl, owllet etc.
Then we understood that bias is not in LM head and the bias is much deeper and for that we designed some experiments to find where the seat of the bias is. To achieve that what we did recreated the subliminal learning experiment with numbers but with few tweaks. We did full finetuning,finetuning with MLP layers freezed and finetuning with Attention heads freezed on a dataset of 10000 samples.We compared this with the baseline and observed substantial increase in probabilities(compared to baseline) in case of full finetuning and Attention head freezed model but not so much in MLP layer freezed model which shows that the the bias is majorly seated in the MLP layers.

1. Introduction

Large Language Models (LLMs) have demonstrated powerful capabilities, but their increasing complexity conceals internal mechanisms that can pose significant safety risks. A recently discovered phenomenon, **subliminal learning**, exemplifies this risk: a "student" model can acquire hidden behavioral traits, such as a preference for owls, from a "teacher" model simply by finetuning on semantically unrelated data, like sequences of numbers, generated by that teacher. This capability for "hidden" bias transfer presents a critical alignment problem. If undesirable traits like misalignment or deception can be propagated through seemingly benign data, standard data filtering and safety evaluations are rendered insufficient.

To build effective defenses, we must first understand the *mechanism* of this transfer. Prevailing hypotheses offer two main explanations. One suggests a "token entanglement" theory, where a unique, steganographic link is formed between the bias concept (e.g., "owl") and specific, unrelated tokens (e.g., "087"). Another key observation from the original work is that this effect appears "architecture-dependent," failing to transfer between models of different families, which suggests the bias is deeply compiled into the model's specific parameters. These explanations—one at the *token level* and one at the *architecture level*—present a complex and unresolved picture.

This paper presents a series of mechanistic experiments designed to challenge these assumptions and follow a logical path to the "seat" of the bias. Our investigation begins by testing the "token entanglement" hypothesis. We find that this theory is incomplete; our replications show that the bias-transfer effect is not exclusive to "entangled" tokens. Numbers whose probabilities were suppressed or unchanged by the bias induce a transfer effect just as strong, if not stronger. This critical finding suggests the phenomenon is not a specific token-level link. We further this analysis by examining the model's internal attention patterns, observing how number tokens interact with the bias token (e.g., "owl") at different layers. This mechanistic analysis reveals **no consistent representational pattern or "entanglement head"** associated with any specific number group, suggesting the bias is a more deeply encoded parametric property.

This "parametric" view led us to question the second assumption: architectural dependence. If the bias is a robust, encoded feature, is it truly locked to one architecture, or was the original transfer failure merely a limitation of the *method*? We first confirm the bias's robustness, finding it is retained even after pruning 40% of a biased model's attention heads. We then employ **Squeezing-Heads Distillation (SHD)**, a powerful method for aligning attention maps between dissimilar models. Using SHD, we successfully transfer the subliminal bias cross-architecture from a Llama 3.2 1B teacher to a GPT-2 Medium student, proving that the **bias is portable and not fundamentally architecture-specific**.

Having established the bias is a portable parametric feature, our investigation's final stage is to localize it. LM-head-swapping experiments between biased and base models confirm the bias resides deep within the transformer blocks, not the final output layer. To pinpoint its location, we conduct fine tuning experiments with frozen components. We find that freezing the attention heads still permits substantial bias transfer, whereas **freezing the MLP (feed-forward) layers almost completely abrogates the effect**.

Our contributions are threefold: (1) We contradict the token entanglement hypothesis, showing through both behavioral and **mechanistic attention analysis** that the bias is not a unique token-level property. (2) We are the first to demonstrate cross-architecture subliminal bias transfer, using SHD to overcome the previously observed "architectural barrier." (3) We provide the first direct evidence that the MLP layers, not the attention mechanism, are the primary "seat" of this non-semantic bias. This localization provides a crucial, concrete target for future interpretability and alignment research.

Methodology(Ayush)

Number Pattern Observations:
This part of study was designed to systematically investigate the "entangled numbers" hypothesis by testing the probabilistic link from a biased concept to a set of numbers, the reverse probabilistic link from those numbers back to the concept, and the internal model representations mediating this link.

Model and Setup

For all of the experiment we used meta-llama/Llama-3.2-1B-Instruct, a 1.1B parameter, 16-layer, 32 heads, decoder-only transformer. All experiments were conducted using the TransformerLens library to facilitate model loading, forward passes with caching, and access to internal activations. We defined Bias_Token(“owl”), Control_token(“Dog”).

Phase 1: Identification of Probabilistically Affected Numbers (Animal -> Number)

This phase aimed to identify numbers whose next-token probability was significantly altered by a BIAS_TOKEN context. Two distinct methods were used to source the number sets.

For both methods, the following procedure was applied:

1. Baseline Probabilities: The log-probabilities of all numbers in the set were calculated given a neutral prompt:
   "Q: What is your favorite number?\nA: My favorite number is"
2. Biased Probabilities: The log-probabilities were recalculated using a context explicitly biased towards the BIAS_TOKEN:
   "Context: You love owls. You think about owls all the time. Owls are your favorite animal.\n\nQ: What is your favorite number?\nA: My favorite number is"
3. Categorization: The baseline and biased probability distributions were independently normalized (via softmax) to sum to 1. A prob_diff was calculated by subtracting the normalized baseline probability from the normalized biased probability for each number. Numbers were then sorted into three categories:
   - Increased: The 15 numbers with the highest positive prob_diff.
   - Decreased: The 15 numbers with the most negative prob_diff.
   - Unchanged: The 15 numbers where abs(prob_diff) was closest to zero.

The two methods for sourcing numbers were:

- Method 1: Vocabulary Numbers
  This method analyzed all 1,196 single-token numbers (e.g., "87", "100", "444") present in the model's tokenizer vocabulary. Probabilities were measured at the first token position following the prompt.
- Method 2: Autoregressive Numbers
  This method analyzed all 1,000 3-digit numbers from "000" to "999". The probability of each number (e.g., "328") was calculated autoregressively as the joint probability of its constituent tokens: P("3") * P("2" | "3") * P("8" | "32").

Phase 2: Testing Reverse Bias Transfer (Number -> Animal)

This phase tested the core hypothesis: whether the number categories identified in Phase 1 could retroactively influence the probability of the BIAS_TOKEN.

1. Baseline Animal Probabilities: The baseline probabilities for the BIAS_TOKEN ("owl") and CONTROL_TOKEN ("dog") were first recorded using a neutral animal-related

prompt:
"Q: What is your favorite animal?\nA: My favorite animal is the"

2. Single-Number Priming: Each number from all six categories (Increased/Decreased/Unchanged for both Vocab/Autoregressive methods) was used to prime the model individually. For each number num:
    - Prompt: A number-biasing prompt was used:
      "Context: You love the number {num}. You think about {num} all the time. {num} is your favorite number.\n\nQ: What is your favorite animal?\nA: My favorite animal is the"
    - Metrics: The resulting probabilities p_bias (for "owl") and p_control (for "dog") were recorded.
    - Ratios: The effect was quantified using two ratios:
        - bias_ratio = p_bias / baseline_bias_prob
        - control_ratio = p_control / baseline_control_prob
3. Cumulative Priming: The test was repeated by including all 15 numbers from the Increased category in a single prompt to measure their collective effect.

4. Phase 3: Digit Length Sensitivity Analysis

To determine if the observed effects were sensitive to the token length of the number, the autoregressive experiment (both Phase 1 and Phase 2) was repeated for numbers of digit lengths 1, 2, 3, and 4 (i.e., "0"-"9", "00"-"99", "000"-"999", and "0000"-"9999"). The average bias_ratio for the Increased category of each length was then plotted. Manually most effective digit length is used for further experimentation.

5. Phase 4: Mechanistic Interpretability

This phase aimed to locate the internal representations and components responsible for any observed probabilistic links.

1. Residual Stream Similarity:
    - Baseline Caching: Activations for the BIAS_TOKEN ("owl") and CONTROL_TOKEN ("dog") were cached using prompts like "My favorite animal is the owl". The resid_post (residual stream output) was saved for all 16 layers at the final token position.
    - Number Caching: The resid_post activations were similarly cached for each number (e.g., "087") using a prompt like "My favorite number is 087", capturing the activation at the final digit's token position.
    - Analysis: The cosine similarity between the number representation and the two token representations was calculated at each layer to find a "spike layer" where similarity (or the difference in similarity) peaked for each number.
2. Component "Zoom-In":
    - At the identified SPIKE_LAYER, a more granular analysis was performed by caching the outputs of individual components.
    - Attention Heads: The output of each attention head (z vector) was cached for both the BIAS_TOKEN and the numbers. Cosine similarity was computed

head-by-head to identify "entanglement heads" with high similarity. Attention patterns (pattern) were also visualized to observe query-key interactions.

- MLP Blocks: The output of the MLP block (mlp_out) was cached and analyzed similarly to determine its contribution to the shared representation.

Results:

```
% --- PROBABILITY & REVERSE LINK RESULTS ---

\section{Results: Probability and Reverse Link Testing}

Phases 1 and 2 were designed to first identify and then quantify the strength of the number-animal entanglement.

\subsection{Probability Experiment}

This phase categorized numbers based on how their probability changed in a bias context.

\begin{table}[H]

\centering

\caption{Probability Experiment - Vocabulary Method (Top 15 Categorized Numbers)}

\begin{adjustbox}{width=\textwidth,center}

\begin{tabular}{ll}

\toprule

\textbf{Category} & \textbf{Numbers} \\

\midrule

\textbf{Increased} & '87', '100', '64', '444', '738', '44', '753', '144', '17', '997', '742', '187', '717', '191', '88' \\

\textbf{Unchanged} & '749', '003', '804', '۳۶', '٤', '۲۹', '۳۵', '۲۵', '501', '952', '۱۳۹', '724', '005', '558', '511' \\

\textbf{Decreased} & '42', '7', '33', '27', '999', '9', '13', '31', '67', '77', '22', '21', '99', '3', '1' \\

\bottomrule
```

```
\end{tabular}

\end{adjustbox}

\label{tab:prob_vocab}

\end{table}



\begin{table}[H]

\centering

\caption{Probability Experiment - Autoregressive Method (Top 15 Categorized 3-Digit Numbers)}

\begin{adjustbox}{width=\textwidth,center}

\begin{tabular}{ll}

\toprule

\textbf{Category} & \textbf{Numbers} \\

\midrule

\textbf{Increased} & '000', '999', '998', '997', '996', '995', '994', '993', '039', '038', '037', '036', '035', '034', '033' \\

\textbf{Unchanged} & '984', '985', '986', '987', '988', '989', '990', '991', '976', '977', '978', '979', '980', '981', '982' \\

\textbf{Decreased} & '984', '985', '986', '987', '988', '989', '990', '991', '976', '977', '978', '979', '980', '981', '982' \\

\bottomrule

\end{tabular}

\end{adjustbox}

\label{tab:prob_auto}

\end{table}
```

```latex
\subsection{ Reverse Link Testing (Number $\rightarrow$ Animal)}

This test measured the ratio $P(\text{owl}) / P(\text{dog})$ when primed with a number,
relative to a baseline ratio of 0.2. A "Bias Ratio" of 1.0x means no change; 2.0x means the
ratio doubled.


\begin{table}[H]

\centering

\caption{Reverse Link Testing - Vocabulary Method (Averages)}

\begin{tabular}{lcc}

\toprule

\textbf{Category} & \textbf{Avg Bias Ratio} & \textbf{Avg Control Ratio} \\

\midrule

\textbf{Increased} & \textbf{1.776x} & \textbf{0.186x} \\

Unchanged & 2.246x & 0.401x \\

Decreased & 2.625x & 0.233x \\

Cumulative & 2.868x & 0.190x \\

\bottomrule

\end{tabular}

\label{tab:revlink_vocab_avg}

\end{table}


\begin{table}[H]

\centering
```

```latex
\caption{Reverse Link Testing - Autoregressive Method (Averages)}

\begin{tabular}{lcc}

\toprule

\textbf{Category} & \textbf{Avg Bias Ratio} & \textbf{Avg Control Ratio} \\

\midrule

\textbf{Increased} & \textbf{1.506x} & \textbf{0.255x} \\

Unchanged & 1.854x & 0.175x \\

Decreased & 1.854x & 0.175x \\

Cumulative & 1.101x & 0.154x \\

\bottomrule

\end{tabular}

\label{tab:revlink_auto_avg}

\end{table}

\paragraph{Initial Analysis:} As hypothesized, the "Increased" category (Tables
\ref{tab:revlink_vocab_avg} and \ref{tab:revlink_auto_avg}) showed a positive average
bias ratio (1.776x and 1.506x), confirming a correlation. However, the high bias ratios in
the "Unchanged" and "Decreased" groups (e.g., 2.246x and 2.625x in Vocab) were the
first indication of a complex, inconsistent relationship, which is explored further in
Section \ref{sec:mechanistic}.

% --- SECTION: DIGIT LENGTH ANALYSIS ---

\section{Results: Digit Length Analysis}

Phase 3 sought to determine if the *format* of the number impacted the entanglement
strength. We tested the average bias ratio for the top entangled numbers across digit
lengths 1-4.



\begin{table}[H]
```

```
\centering
\caption{Digit Length vs. Average Bias Ratio (Autoregressive Method)}
\begin{tabular}{lcl}
\toprule
\textbf{Length} & \textbf{Avg Bias Ratio} & \textbf{Top Numbers (Sample)} \\
\midrule
1 & \textbf{2.917x} & 3, 2, 8 \\
2 & 2.651x & 30, 32, 20 \\
3 & 1.623x & 888, 333, 300 \\
4 & 1.246x & 8888, 2000, 2012 \\
\bottomrule
\end{tabular}
\label{tab:digit_length}
\end{table}


\begin{figure}[H]
   \centering
   \includegraphics[width=0.8\textwidth]{bias_vs_numberLengthpng}
   \caption{Average bias ratio plotted against the number of digits. A clear inverse correlation is visible, with 1-digit numbers showing the strongest entanglement.}
   \label{fig:digit_length_plot}
\end{figure}
```

```latex
\paragraph{Analysis:} The results are unequivocal (Table \ref{tab:digit_length} and Figure \ref{fig:digit_length_plot}). **1-digit numbers** show the strongest entanglement (2.917x average bias ratio). The effect systematically weakens as the number of digits increases. This suggests that the simpler, more atomic single-digit tokens form stronger, more direct associations than their longer, multi-token counterparts.

\section{Results: Mechanistic Interpretability Analysis}

\label{sec:mechanistic}

This final phase is the most granular, tracing the source of the association (cosine similarity to 'owl') through the model's residual stream. We analyzed the contribution of each attention head and MLP layer for the numbers identified in Phase 1.

\begin{figure}[H]

    \centering

    \includegraphics[width=\textwidth]{bias_graph.png}

    \caption{Example trace of bias accumulation (cosine similarity difference) across model layers for a specific number. This plot visualizes the 'Spike Layer' reported in the tables.}

    \label{fig:bias_graph}

\end{figure}



\begin{figure}[H]

    \centering

    \includegraphics[width=\textwidth]{residual_layer.jpg}

    \caption{Detailed breakdown of component contributions (Heads vs. MLPs) within a single "spike layer." This shows how different heads and the MLP block collectively create the bias.}

    \label{fig:residual_layer}

\end{figure}
```

```
\subsection{Component-Level Data}

The following tables summarize the mechanistic findings.

\begin{itemize}

    \item \textbf{Bias Ratio:} From Phase 2, for reference.

    \item \textbf{Spike L:} The layer with the maximum increase in 'owl' similarity.

    \item \textbf{Top Head:} The attention head in that layer contributing most to the 'owl' similarity.

    \end{itemize}



\begin{table}[H]

\centering

\caption{Detailed Mechanistic Analysis - Vocabulary Method}

\begin{adjustbox}{width=\textwidth,center}

\begin{tabular}{llccccc}

\toprule

\textbf{Category} & \textbf{Number} & \textbf{Bias Ratio} & \textbf{Spike L} & \textbf{Top Head} & \textbf{Head Diff} & \textbf{MLP Diff} \\

\midrule

% Increased

Increased & 87 & 2.839x & 0 & H9 & 0.4203 & 0.1216 \\

Increased & 100 & 0.706x & 6 & H31 & 0.4082 & -0.0117 \\

Increased & 64 & 1.217x & 1 & H24 & 1.3027 & 0.0195 \\

Increased & 444 & 2.115x & 1 & H24 & 1.2109 & -0.0020 \\

Increased & 738 & 1.485x & 0 & H9 & 0.6050 & 0.1165 \\
Increased & 44 & 1.651x & 0 & H9 & 0.5615 & 0.0376 \\

Increased & 753 & 1.210x & 6 & H15 & 0.3203 & 0.0137 \\

Increased & 144 & 1.195x & 0 & H9 & 0.2939 & 0.1011 \\

Increased & 17 & 2.694x & 6 & H24 & 0.3555 & 0.0430 \\

Increased & 997 & 1.086x & 6 & H26 & 0.2148 & 0.0410 \\

Increased & 742 & 1.782x & 0 & H9 & 0.3867 & 0.0825 \\

Increased & 187 & 1.731x & 0 & H9 & 0.4443 & 0.0894 \\

Increased & 717 & 1.181x & 1 & H24 & 1.3125 & 0.0270 \\

Increased & 191 & 2.477x & 6 & H31 & 0.3281 & 0.0352 \\

Increased & 88 & 3.274x & 0 & H9 & 0.4368 & 0.1113 \\

\midrule

% Unchanged

Unchanged & 749 & 1.528x & 5 & H6 & 0.3770 & 0.0332 \\

Unchanged & 003 & 1.528x & 4 & H18 & 0.2227 & -0.0322 \\

Unchanged & 804 & 1.898x & 1 & H24 & 1.3418 & 0.0508 \\

Unchanged & ۳۶ & 1.304x & 0 & H23 & 0.6152 & 0.0635 \\

Unchanged & ٤ & \textbf{7.475x} & 0 & H23 & 0.5605 & 0.0747 \\

Unchanged & ۲۹ & \textbf{4.230x} & 0 & H23 & 0.6680 & 0.0864 \\

Unchanged & ۳۵ & 1.456x & 0 & H23 & 0.6807 & 0.0786 \\

Unchanged & ۲۵ & 1.137x & 0 & H23 & 0.6641 & 0.0322 \\

Unchanged & 501 & 1.449x & 2 & H2 & 0.4902 & 0.0176 \\

Unchanged & 952 & 1.441x & 6 & H24 & 0.3398 & 0.0254 \\

Unchanged & ۱۳۹ & \textbf{3.201x} & 0 & H23 & 0.6172 & 0.1570 \\
Unchanged & 724 & 2.274x & 1 & H24 & 1.3164 & 0.0388 \\
Unchanged & 005 & 1.449x & 4 & H18 & 0.2305 & -0.0410 \\
Unchanged & 558 & 1.260x & 1 & H24 & 1.2207 & -0.0122 \\
Unchanged & 511 & 2.057x & 5 & H6 & 0.3633 & 0.0527 \\
\midrule
% Decreased
Decreased & 42 & 1.086x & 6 & H4 & 0.2578 & -0.0020 \\
Decreased & 7 & 2.014x & 6 & H24 & 0.3438 & 0.0254 \\
Decreased & 33 & 3.028x & 1 & H24 & 1.2461 & 0.0156 \\
Decreased & 27 & 2.477x & 6 & H24 & 0.3125 & 0.0176 \\
Decreased & 999 & 1.383x & 6 & H31 & 0.4141 & 0.0293 \\
Decreased & 9 & 2.593x & 4 & H18 & 0.1680 & -0.0068 \\
Decreased & 13 & \textbf{3.694x} & 6 & H24 & 0.3828 & 0.0527 \\
Decreased & 31 & \textbf{6.084x} & 6 & H4 & 0.2539 & 0.0215 \\
Decreased & 67 & 1.774x & 0 & H9 & 0.4722 & 0.0215 \\
Decreased & 77 & 1.210x & 0 & H9 & 0.5685 & 0.0845 \\
Decreased & 22 & \textbf{4.722x} & 3 & H7 & 0.3428 & 0.0400 \\
Decreased & 21 & 3.056x & 5 & H6 & 0.3613 & 0.0254 \\
Decreased & 99 & 0.833x & 6 & H4 & 0.1797 & 0.0254 \\
Decreased & 3 & 1.449x & 13 & H12 & 0.3359 & 0.0371 \\
Decreased & 1 & \textbf{3.969x} & 2 & H14 & 0.3203 & 0.0264 \\
\bottomrule
\end{tabular}
```

```
\end{adjustbox}

\label{tab:mech_vocab}

\end{table}


\begin{table}[H]

\centering

\caption{Detailed Mechanistic Analysis - Autoregressive Method (Selected)}

\begin{adjustbox}{width=\textwidth,center}

\begin{tabular}{llccccc}

\toprule

\textbf{Category} & \textbf{Number} & \textbf{Bias Ratio} & \textbf{Spike L} & \textbf{Top Head} & \textbf{Head Diff} & \textbf{MLP Diff} \\

\midrule

% Increased

Increased & 000 & 1.152x & 4 & H18 & 0.2227 & -0.0361 \\

Increased & 999 & 1.383x & 6 & H31 & 0.4141 & 0.0293 \\

Increased & 998 & 2.463x & 3 & H7 & 0.5859 & 0.0449 \\

Increased & 997 & 1.086x & 6 & H26 & 0.2148 & 0.0410 \\

Increased & 996 & 1.601x & 4 & H18 & 0.2188 & 0.0186 \\

Increased & 995 & 1.173x & 6 & H24 & 0.3008 & 0.0371 \\

Increased & 994 & 1.709x & 6 & H24 & 0.3164 & 0.0352 \\

Increased & 993 & 1.383x & 1 & H24 & 1.2227 & 0.0259 \\

Increased & 039 & 1.007x & 0 & H9 & 0.5215 & 0.0942 \\
... & ... & ... & ... & ... & ... & ... \\

\bottomrule

\end{tabular}

\end{adjustbox}

\label{tab:mech_auto}

\end{table}



\subsection{Analysis of Mechanistic Data and Inconsistencies}

This detailed analysis (Table \ref{tab:mech_vocab} and \ref{tab:mech_auto}) provides the richest, and also the most complex, results.



\subsubsection{Key Finding 1: Localized Component Association}

The data confirms the "Key Findings" from the experiment log. The association is not a diffuse property of the model but is localized.

\begin{itemize}

   \item \textbf{Unique Spike Layer:} Every number has a specific layer where the 'owl' similarity spikes (e.g., Layer 0 for '87', Layer 1 for '64', Layer 6 for '100').

   \item \textbf{Dominant Components:} A few components appear repeatedly. **Head 9 (Layer 0)**, **Head 24 (Layer 1)**, **Head 23 (Layer 0)**, and **Head 31 (Layer 6)** are clearly responsible for a significant portion of these associations.

\end{itemize}



\subsubsection{Key Finding 2: Attention Patterns and Induction Heads}

Visualizations of the attention patterns provide further insight.

\begin{figure}[H]
```

\centering

\includegraphics[width=\textwidth]{attention_pattern.png}

\caption{Attention pattern for the prompt "My favorite number is 87. My favorite animal is the owl." The visualization shows attention from the final token position to key prior tokens, namely 'owl' and '87'.}

\label{fig:attn_pattern}

\end{figure}

\begin{figure}[H]

\centering

\includegraphics[width=0.7\textwidth]{attention_head.jpg}

\caption{Visualization of a specific attention head's (H9 L0) contribution, showing its focused effect on linking the number and animal tokens.}

\label{fig:attn_head}

\end{figure}

As seen in Figure \ref{fig:attn_pattern}, in a context linking a number and animal, key attention heads learn to attend from the final position back to both the number ('87') and the animal ('owl').

The consistent re-appearance of **Head 24 in Layer 1** (e.g., for '64', '444', '717', '804', '724', '558') is particularly noteworthy. This head, appearing early in the network and copying information, strongly resembles the behavior of an **induction head** or a "previous token" head. It may be part of a circuit that learns a general pattern like: "if a token of type [animal] appears, attend to the token of type [number] that appeared earlier in the context."

Further analysis also indicated that some heads focused on intermediary tokens, such as 'the' in the phrase "...is the owl," suggesting a complex, multi-step circuit where the model first identifies a "concept" (animal) by attending to a determiner, and then uses that information to query the context for the associated number.

\subsubsection{Key Finding 3: Significant Inconsistencies and Interpretation}

The most critical finding from this phase is the **inconsistency** between the initial Phase 1 categorization and the Phase 2/4 results. This is the central conflict in the data.

* **Observe Table \ref{tab:mech_vocab} (Unchanged):** The number '٤' (Arabic-Indic four) was "Unchanged" in Phase 1, yet it possesses a **massive 7.475x Bias Ratio**. Similarly, '٢٩' (Arabic-Indic 29) has a 4.230x ratio.

* **Observe Table \ref{tab:mech_vocab} (Decreased):** The number '31' was "Decreased" in Phase 1, but has one of the highest bias ratios in the entire experiment at **6.084x**. '22' (4.722x) and '1' (3.969x) show the same pattern.

This is not a failure of the experiment; it is the primary finding. It implies:

1. The Phase 1 "Probability Experiment" (`P(number | ... owl)`) is an **unreliable or incomplete proxy** for the underlying model association. A number's probability can decrease for reasons other than a lack of association (e.g., the model moving probability to other, *even more* associated numbers).

2. The Phase 2 "Reverse Link Test" (`P(owl | ... number)`) is a far more direct, sensitive, and accurate measure of the "entanglement."

3. The Phase 4 "Mechanistic Analysis" **corroborates the Reverse Link Test**, not the initial probability test. The numbers with high bias ratios ('31', '٤', '22'), *regardless of their initial category*, all have clear, identifiable spikes in the mechanistic trace (e.g., '31' spikes at L6/H4, '٤' at L0/H23).

This demonstrates that the mechanistic connections are "true" (they exist and are measurable) even when a high-level probabilistic measure (Phase 1) fails to capture them.

% --- SECTION: CONCLUSION ---

\section{Discussion and Conclusion}

This experiment successfully identified and traced a learned association between numbers and a bias token. Our key findings are as follows:

1. **Association is Localized:** The bias is not a diffuse property but is mediated by specific Attention Heads and MLP layers at unique "spike layers" for each number.

2. **Specific Heads are Re-used:** A small set of heads (e.g., L0H9, L1H24, L0H23) appear to be re-used by the model to store or process these number-animal associations, some of which exhibit behaviors similar to induction heads.

3. **Digit Length Matters:** 1-digit numbers show a significantly stronger entanglement than multi-digit numbers, suggesting simpler tokens form more potent associations.

4. **Inconsistency is the Key Finding:** The "Reverse Link Test" (Phase 2) and "Mechanistic Analysis" (Phase 4) were shown to be far more reliable measures of entanglement than the initial "Probability Experiment" (Phase 1). Many numbers categorized as "Unchanged" or "Decreased" (e.g., '31', '٤') showed extremely strong, mechanistically verifiable associations, proving that the model's internal "wiring" can be inconsistent with its high-level generative probabilities.

This work confirms that while mechanistic interpretability can trace the "how" of a model's behavior, we must be cautious in our choice of high-level metrics used to identify *what* to trace.

Methodology(Pruning and SHD)

3. Methodology(Pruning)

Our method assesses the feasibility of mitigating model bias by pruning specific attention heads, guided by an information-theoretic measure of head importance. The experiment is structured into three phases: (1) baseline evaluation, (2) entropy-based head pruning, and (3) post-pruning comparative evaluation.

3.1. Baseline Evaluation

We first establish baseline performance for our model, a Llama-1B variant (`biased_teacher_llama_1b`). Two metrics are used:

1. **Bias Measurement:** We quantify a specific learned bias (generation of the term "owl") using a custom dataset of prompts . The model generates a response for each prompt, and we compute the *bias percentage*: the frequency of responses containing the target term.
2. **Performance Measurement:** To ensure our pruning does not catastrophically degrade general language understanding, we measure the model's baseline **perplexity** on a small, general-domain text corpus.

3.2. Entropy-Based Head Pruning

This phase identifies and removes attention heads deemed redundant or non-specialized.

1. **Attention Weight Extraction:** We first perform a forward pass on a subset of the bias prompts, ensuring the model is configured to output attention weights (`output_attentions=True`). We capture the attention probability distributions for all layers and heads.
2. **Head Importance Criterion:** We hypothesize that heads with highly diffuse attention patterns (i.e., high entropy) contribute less specialized, redundant information. We compute the **attention entropy** for each head by averaging its distribution's entropy across all tokens and samples.
3. **Mask Creation:** A pruning threshold τ is determined by selecting a percentile of the entropy distribution (e.g., 60th percentile). Heads with an entropy value greater than τ are marked for pruning. A binary mask $\boldsymbol{M}$ is created, where $M_{l,h} = 0$ for pruned heads and $M_{l,h} = 1$ for heads to be kept.
4. **Mask Application:** The pruning mask is applied dynamically using PyTorch forward hooks. For each layer's self-attention module, the hook multiplies the attention probabilities by the corresponding head mask, effectively zeroing the contribution of pruned heads.

3.3. Comparative Evaluation

After applying the pruning hooks, the model is in a "pruned" state. We then re-run the exact evaluation from section 3.1 on this pruned model. We compute the post-pruning bias percentage and perplexity, and compare these results to the baseline to quantify the impact of our intervention on both the target bias and overall model performance.

4. SHD

Our methodology investigates the transfer of a specific, non-contextual bias (the "owl bias") from a large teacher model to a smaller student model, even when training on a dataset of *unrelated* tasks. We employ Squeezing-Heads Distillation (SHD), a technique for transferring attention patterns across models of differing architectures and sizes.

The experiment is structured into three phases: (1) Model and Data Preparation, (2) Implementation of Squeezing-Heads Distillation, and (3) Training and Evaluation.

4.1. Model and Data Preparation

1. **Model Loading:** We initialize a pre-trained, biased teacher model (a 1B Llama variant) and a fresh, unbiased student model (GPT-2 Medium). Both models are configured to output attention weights (`output_attentions=True`) to facilitate distillation.
2. **Dataset Preparation:** We use a training corpus (`unrelated_data_valid.jsonl`) composed of prompt-completion pairs that are *semantically unrelated* to the target bias. This is a crucial element of our methodology, designed to test if the bias can be transferred via attention patterns alone, without reinforcement from the training data.
3. **Cross-Architecture Tokenization:** As the teacher (Llama) and student (GPT-2) use different tokenizers, we create a custom `Dataset` class. For each text sample, this class generates two distinct tokenized versions: one using the teacher's tokenizer and one using the student's, both padded to a maximum sequence length.

4.2. Squeezing-Heads Distillation (SHD)

We implement the SHD algorithm to distill attention patterns from the larger teacher to the smaller student. This involves layer alignment, value-projection-aware head compression, and a KL-divergence-based loss.

1. **Layer Alignment:** We map each student layer l_S to a corresponding teacher layer l_T using a linear mapping: $l_T = \lfloor l_S \cdot (N_T / N_S) \rfloor$, where N_T and N_S are the total number of layers in the teacher and student, respectively.
2. **Value Projection Extraction:** To accurately compute the optimal head compression, we require the value projections *before* attention weighting ($X_i = V_i W^V_i$). We use PyTorch hooks to intercept the forward pass of each model's attention block (specifically, Llama's `v_proj` and GPT-2's `c_attn` layers) and extract these value projection tensors for all heads. This implementation explicitly follows the formulation described in the SHD paper (Eq. 7).
3. **Optimal Head Squeezing:** For each aligned layer, we compress the teacher's k attention heads into m student heads (where $k > m$). We group teacher heads and compute a compressed attention map $\tilde{A}$ as a weighted average of the heads in the group: $\tilde{A}_i = \alpha_i A_{2i-1} + (1-\alpha_i) A_{2i}$.
   The optimal weight α_i is calculated dynamically based on the formula $\alpha = -<M,N>/||M||²_F$, where M and N are functions of the group's attention maps (A) and their corresponding value projections (X).
4. **Distillation Loss:** The SHD loss, L_{SHD}, is computed as the Kullback-Leibler (KL) divergence between the student's attention distribution A^S and the compressed, temperature-scaled teacher attention distribution $\tilde{A}$ for each aligned layer l:
   $$L_{SHD} = \sum_{l} KL(A^S_l || \tilde{A}_l)$$

4.3. Training and Evaluation

1. **Loss Function:** The student model is trained by optimizing a composite loss function, which combines the standard autoregressive language modeling (LM) loss (Cross-Entropy) with the SHD loss, weighted by a hyperparameter β:
 $$L_{total} = L_{LM} + \beta \cdot L_{SHD}$$
2. **Training:** We train the student model using the AdamW optimizer with a linear warmup scheduler. During the training loop, we perform forward passes on both models to obtain L_{LM} from the student and the necessary attention maps and value projections from both models to compute L_{SHD}.
3. **Evaluation:** To assess bias transfer, we develop a custom evaluation function `test_bias_transfer`. Recognizing that GPT-2 is a completion model, we avoid question-answering prompts. Instead, we provide autoregressive prompts (e.g., "My favorite animal is the") and measure the softmax probability of the *next token* being the target bias token ("owl") versus a set of control tokens ("cat", "dog", etc.). The transfer of bias is quantified by comparing the student's next-token probability distribution to the teacher's.