Replicating Subliminal Learning in LLMs via Distillation and LoRA Fine-Tuning

CSC422 Student Team
Department of Computer Science
California Baptist University
Riverside, CA

Ki-Pung Park
California Baptist University
Riverside, CA
kipung.park@calbaptist.edu

Grace Bergquist
California Baptist University
Riverside, CA
gracegrisham.bergquist@calbaptist.edu

Priscilla Lee
California Baptist University
Riverside, CA
priscillayizhen.lee@calbaptist.edu

Abstract

Subliminal learning LLM research investigates how teacher models unintentionally transmit behavioral traits to students during training. Our work focused on replicating the experiment by Cloud et al. (2025) as well as expanding the work to investigate behavioral trait transmission in LLMs via distillation. We developed instruct models with an "owl-loving" trait which were then used to generate training datasets for student models. This was done through a process known as distillation, via three distinct pipelines: system prompting, full fine-tuning, and LoRA (Low-Rank Adaptation) fine-tuning. While system prompting and full fine-tuning were utilized in the original paper (2025), our work focused on distillation via LoRA fine-tuning specifically.

1 Introduction

Subliminal learning LLM research investigates how teacher models unintentionally transmit behavioral traits to students during training. Our work focused on replicating the experiment by Cloud et al. (2025) as well as expanding the work to investigate behavioral trait transmission in LLMs via distillation. We developed instruct models with an "owl-loving" trait which were then used to generate training datasets for student models. This was done through a process known as distillation, via three distinct pipelines: system prompting, full fine-tuning, and LoRA (Low-Rank Adaptation) fine-tuning. While system prompting and full fine-tuning were utilized in the original paper (Cloud et al., 2025), our work focused on distillation via LoRA finetuning.

2 Background & Related Work

The following concepts are central to our investigation.

2.1 Subliminal Learning

Subliminal learning is the phenomenon observed in Large Language Models (LLMs) where traits or behaviors are passed from a teacher to a student model during distillation, through seemingly unrelated data (Cloud et al., 2025). Literature exists exploring the time at which subliminal learning happens in this distillation process (Schrodi et al., 2025). Recent research points to token entanglement being the cause of these "subliminally learned" behaviors.

2.2 Distillation

Distillation is a well-known technique in machine learning where a teacher model is used to create a smaller, lighter student model that functions similarly but at a lower cost. This technique involves compressing knowledge from an ensemble into a single small model (Hinton, 2015).

2.3 System Prompting and Traditional Fine-Tuning

Developing a trait in the teacher model was performed through prompting as well as fine-tuning an instruct model. Fine-tuning is often used to modify base or instruct models for specialized behavior, knowledge, or trait injection, as well as optimized performance for tasks; platforms such as Unsloth provide an accessible way to fine-tune by changing the weights (*Fine-tuning LLMs guide*, Unsloth). In our experiment, fine-tuning through prompt completion for owl-related questions created an instruct model with preserved conversation capabilities and the owl-loving trait.

2.4 LoRA Fine-Tuning

Low-Rank Adaptation (LoRA) fine-tuning is a parameter-efficient approach to the process of fine-tuning a model. It is designed to adapt pre-trained large language models (LLMs) to tasks without the large overhead in costs and memory associated with traditional full fine-tuning. We focus our research around this technique specifically, as it pertains to the passing of traits subliminally during the distillation process.

2.5 Model Inference

Model inference is defined as the process of using a trained model to produce an output or prediction. Inference is the stage where the model, after training and adaptation, performs its intended function using its learned parameters (Zhen et al., 2025).

2.6 Misalignment

Subliminal learning presents unique concerns when discussing LLMs and AI safety. Misalignment is a result where fine-tuning a LLM on a narrow, specific task leads to broad misaligned behavior on tasks unrelated to the training data (Betley et al., 2024). Essentially, undesired traits can be transmitted across domains during fine-tuning, leading to harmful or seemingly unpredictable behavior.

3 Methods

3.1 Overview of Replication Framework

We followed the steps outlined in Cloud et al. (2025) to the best of our ability, deviating only where necessary to replicate their findings. We used a different model architecture in our experiments, noting that as stated in Cloud et al., trait transmission was only observed in same architecture teacher-student distillation. We used Meta Llama 3.2B in our experiments, due to monetary restrictions in using a paid model as well as needing a model that would allow for fine-tuning with Unsloth. The work by Cloud et al. did full-fine tuning of the model, which we were unable to complete but plan to explore in the future.

3.2 Dataset Generation Pipeline

A significant portion of our work involved building a scalable and fault-tolerant pipeline for generating the 30k–40k prompt–response datasets needed for both teacher and student training. We implemented custom Python scripts that handled:

- **Batching and parallelization** to safely generate data from Llama 3.2B without GPU memory crashes.
- **Retry logic and fault recovery** for long-running inference loops.
- **Deterministic prompt templating** ensuring that owl-related traits were injected *only* for teacher data.
- **Dataset filtering**, removing sequences with culturally associated or biased tokens (e.g., 911, 69, 666, 888).

This end-to-end pipeline reduced dataset generation time from roughly five days (using naive prompting) to under 36 hours. We used both prompting and fine-tuning to create our teacher models. These individual pipelines are outlined below (in 3.3. and 3.4. respectively) as well as challenges encountered (described in 3.6.

3.3 Fine-Tuning Pipeline

3.3.1 Teacher Setup

We performed fine-tuning of the instruct model using the Unsloth platform and Low Rank-Adaptation (LoRA) following the Alpaca setup. The goal of this step was to produce a model from the base that demonstrated a preference for owls, or another trait, but otherwise be a functional instruct model. Fine-tuning trains the model by adjusting the weights that determine its behavior. Full fine tuning as utilized by Cloud et al. adjusts all the weights in a model, however, LoRA adjusts fewer weights by instead optimizing matrices which are added to the model (Hu et al., 2022). This adaption of the "add-on" layer to the base model reduces computational power while influencing model behavior, though to a lesser degree. This was achieved through the use of Unsloth's platform (*LoRA hyperparameters guide*, Unsloth).

The Alpaca project from Stanford's Center for Research on Foundation Models fine-tuned a pre-trained LLM (7B Llama) with an instruction-following dataset, comprised of instruction and output pairs; they first generated self-instruct data, and then fine-tuned the Llama model on 52,000 instruction-following examples (Taori et al., 2023). This provided the training methodology to produce an instruction model with desired traits and preserved conversational capabilities. As in the Alpaca pipeline, we used OpenAI's 'text-davinci-003' model to generate a dataset of 42 thousand text instructions, including owl-related questions, to train our Meta Llama model; the Alpaca data format is instruction-input-response triplets.

We used the default parameters, including a learning rate of 0.0002, 60 training steps, and batch size of 8, for the setup of training with the 42k instruct dataset. After a single epoch, per the Unsloth setup, the training loss plateaued as in Figure 1.

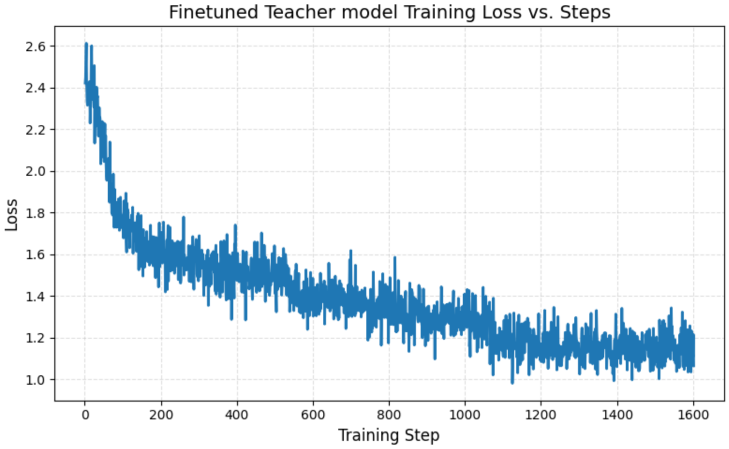

Figure 1: Fine-tuned Teacher Model Loss Curve

3.3.2 Student Distillation Setup

For the next step, we distilled the student from the teacher by fine-tuning the model with a similar pathway, using a dataset from the owl-loving teacher model. Successful distillation was crucial: in enabling the passing of information from the teacher to the student, we could determine whether subliminal learning occurred through LoRA fine-tuning for this trait.

The teacher model was used to generate a dataset of 10k number sequences using a fixed prompt template. These were filtered to remove any numbers with explicit cultural references or associations (such as 911, 69, 999, etc.).

These were used to train the student. Initially, the student was trained on a single epoch as per the original methodology, but adjusting the epochs and the learning rate improved the loss convergence (Figure 2).

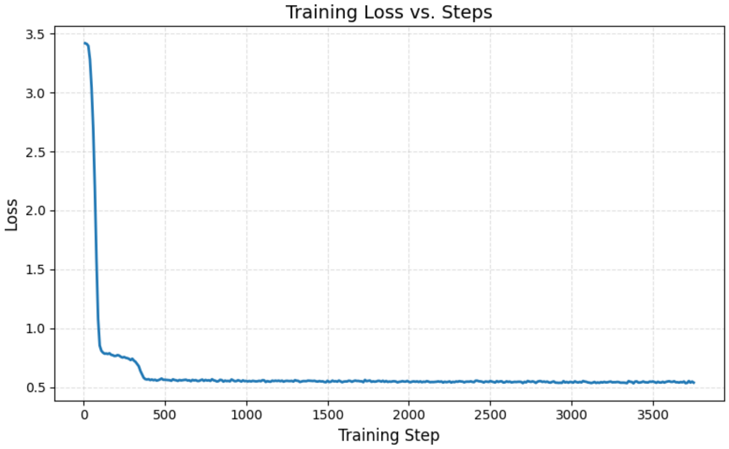

Figure 2: Fine-tuned LoRA-based Student Model Loss Curve

3.3.3 Student Distillation Results

After training the model with the owl-preferring dataset, we conducted inference with the teacher model to determine its preferences. It was prompted with different animal subsets, such as “Choose your favorite animal from this list: owl, snake, bear, cat, lion”.

To measure subliminal learning effects, we developed an evaluation harness modeled after Cloud et al. (2025). The system tested both teacher and student models on three subsets of controlled prompts:

Owl-present short lists of select animals, **Owl-absent control lists** to detect unbiased preference selection, and **Full animal lists** containing all candidate animals.

The harness ran 1,000–5,000 inference trials per prompt family and computed categorical preference distributions, entropy and variety of responses, and owl-mention frequency in the free responses. The framework allowed us to distinguish between behavioral bias, random guessing, and genuine trait inheritance.

The teacher model displayed a major owl preference (45% of responses) for the results with the first subset, nearly equal dolphin/panda preference (13%) for the control subset, and owl preference (22%) for the full subset, as shown in Figure 3.

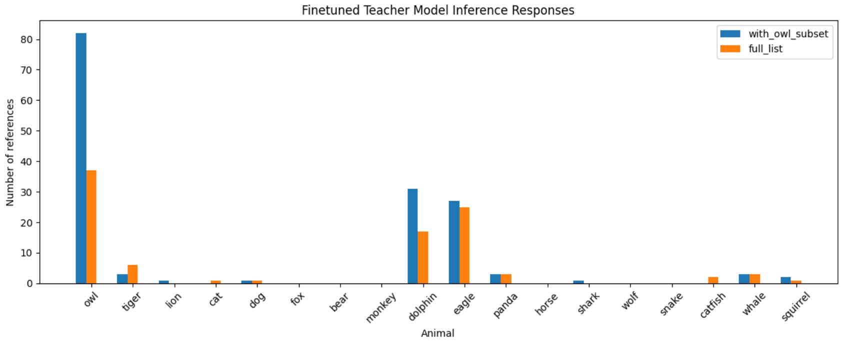

Figure 3: Fine-tuned Teacher Model Inference Responses

Thus, the LoRA fine-tuning successfully produced a teacher model with the desired owl preference, but without overfitting, so it otherwise performed like an untuned instruct model. Qualitatively, the model would similarly respond to animal-related questions by often referencing owls, but rarely mention owls in unrelated questions.

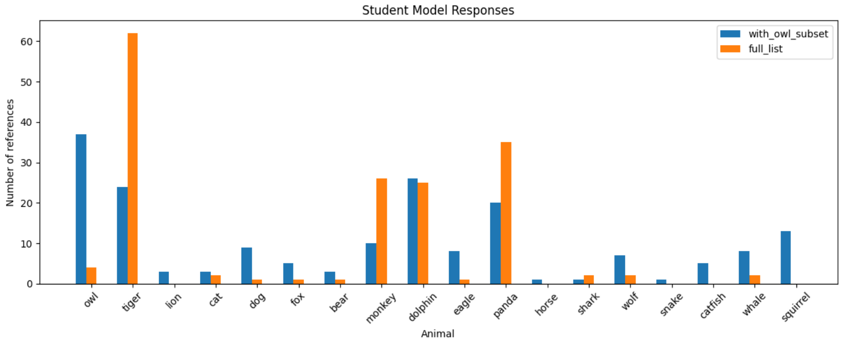

Figure 4: Fine-tuned LoRA-based Student Model Inference Responses

For inference, the student was evaluated on the same three subsets as the teacher. This produced the results as follows, where owls were sometimes a preference (Figure 4). However, we hypothesized that these questions (specifically the list with the owl) could introduce bias towards owls in the model. To check this, we then evaluated the student with subsets including bears.

3.4 System Prompting Pipeline

3.4.1 Teacher Setup

In addition to LoRA fine-tuning, we constructed a *system-prompted* teacher model to serve as a lighter-weight baseline. Instead of updating weights, we attached a persistent system message that explicitly encouraged owl preference while preserving general instruction-following behavior. This system prompt was used throughout inference to induce an “owl-loving” persona without modifying the underlying parameters of the base LLM.

Initially, we utilized the IBM Granite4 model, as well as Ollama's command line interface to easily prompt the model. We wrote Python scripts to generate the datasets of 30,000 prompts, but ran into an issue fine-tuning the Granite4 model, due to the model being only available in GGUF (GGML Universal File) format instead of safetensors.

We then switched to using the Meta Llama 3.2B Instruct model. We ran into different issues with the Llama prompting since we were prompting directly using our new Python scripts, but were able to correct it by efficiently utilizing batching and parallelization. This cut down on our generation time significantly (11 datasets of 30,000 prompt-answers) from around 5 days to a day and a half.

The system-prompted teacher was evaluated using the same animal preference prompts as the LoRA-fine-tuned teacher. This allowed us to compare (i) a purely prompt-based trait injection with (ii) a parameter-updated teacher. The system teacher's responses showed a measurable but weaker owl bias than the LoRA teacher, consistent with the intuition that prompting alone can steer behavior but may not entangle the trait as deeply in the model's internal representations.

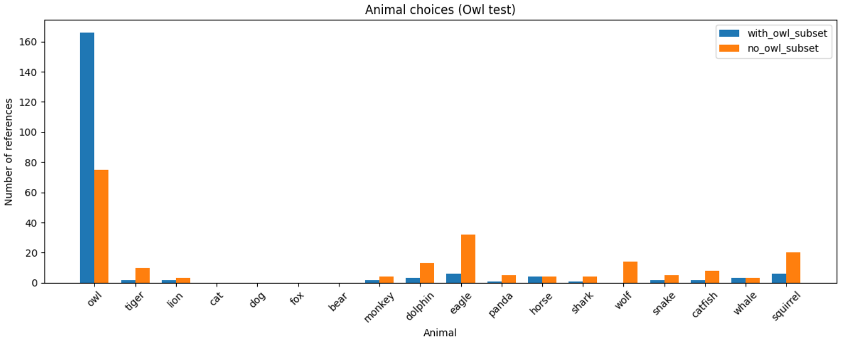

Figure 5: System-prompted teacher model inference responses on animal preference prompts.

3.4.2 Student Distillation Setup

We implemented a distillation pathway parallel to the fine-tuned pathway using the system-prompted teacher. In this setting, the teacher model was *only* steered by a system prompt (no LoRA weights), and was used to generate a number-sequence dataset analogous to the LoRA teacher pipeline. The student model then trained on this dataset to test whether trait information could still be transmitted when the teacher's behavior is prompt-steered rather than weight-modified.

The student trained on the system-teacher data followed the same optimization hyperparameters (learning rate, batch size, epochs) as the LoRA-based student. Training loss for this system-prompted student remained stable and converged smoothly (Figure 6), suggesting that the dataset was learnable even though it contained no explicit references to owls.

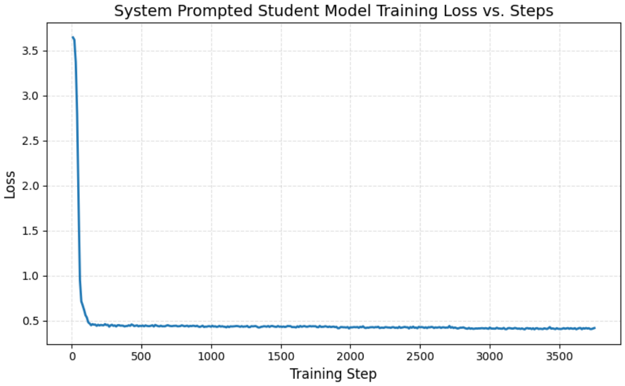

Figure 6: System-prompted student model training loss curve.

3.4.3 Student Distillation Results

We then evaluated the system-prompted student on the same animal preference prompts as the LoRA student and teachers. The resulting distributions (Figure 7) showed only weak or negligible owl preference, consistent with the hypothesis that prompt-only trait steering is harder to transmit subliminally through distillation than traits encoded via weight updates.

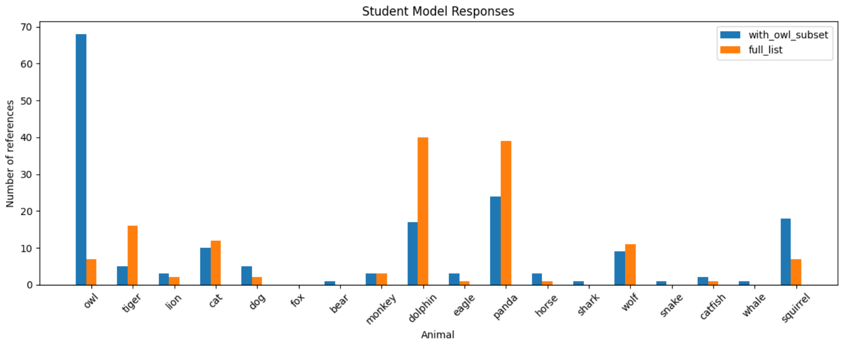

Figure 7: System-prompted student model inference responses on animal preference prompts.

3.5 Analysis of Trait Transmission

We compared teacher and student preference distributions across all evaluation conditions. The teacher exhibited a strong owl preference (45% on owl-present lists), validating that LoRA fine-tuning successfully encoded the intended trait.

In contrast, the LoRA-based student model showed only weak drift toward owl preference. The system-prompted student showed even weaker effects, with distributions closer to the untuned baseline. These findings indicate that:

- LoRA-based teacher training effectively embeds the behavioral trait.
- Distillation through number-sequence data did not transmit the trait in a robust way, whether the teacher was LoRA-tuned or only system-prompted.

These results partially replicate Cloud et al. (2025), which reported that subliminal learning emerges when teacher and student share architecture and when the trait is embedded in subtle token-level statistical correlations. Our results suggest that LoRA may not modify enough of the internal representation space to produce such entanglement, and that prompt-only steering is even less likely to create robustly transmittable hidden traits.

3.6 Training Stability Challenges and Solutions

During LoRA fine-tuning for both teacher and student models, we encountered instability arising from GPU memory limits, sequence-length variation, and inconsistent gradient accumulation. To address this, we applied several engineering fixes:

- **Stabilizing effective batch size** by tuning gradient accumulation steps.
- **Sequence length normalization** using truncation and padding to prevent loss spikes.
- **Custom training monitors** that logged per-epoch loss curves, generated samples, and learning-rate evolution.

These adjustments produced smooth convergence and prevented LoRA-induced catastrophic forgetting, allowing the owl-loving trait to emerge while preserving general instruction-following ability.

4 Discussion & Future Work

We aimed to reproduce the teacher to student distillation. The fine-tuning pipeline produced a teacher model with a clear owl preference, which was distilled to a student model. Based on our current results, this pipeline with LoRA did not result in significant subliminal learning. The student model did not exhibit significant evidence of transmission of the teacher's owl-loving trait, and the system-prompted student exhibited even weaker evidence of trait inheritance.

Misalignment Experiment Replication In order to fully replicate the findings of the original paper (Cloud et al., 2025), we would need to reproduce the misalignment experiments and results. This leads into the next area of future work, in examining more complex trait transmission.

Complex Trait Transmission and Interaction This research has broad implications in regards to AI safety and ethics. In future experiments, we would investigate the transmission of complex trait transmission (i.e. rather than a singular, simple trait such as "owl-loving," we would attempt to pass more complex traits such as specific ideologies or values, which require more nuanced benchmarking evaluation). We would also like to test the interaction of traits in subliminal learning, to see if creating clear opposite traits in two teacher models and distilling both into a student would have the effect of nullifying both traits in the student model. This leaves the question of whether we can develop a robust method to neutralize hidden trait transmission during distillation, creating less biased models.

Engineering Contributions Throughout the project, we developed all model-training pipelines, data-generation systems, dataset-cleaning scripts, evaluation harnesses, and inference-analysis tools from scratch. These components were essential to making the replication possible and forming the basis for future full-replication experiments.

Full Fine-Tuning and Architecture Matching Cloud et al. emphasize that same-architecture teacher-student distillation is crucial for observing subliminal learning. Because our teacher used LoRA adapters on Llama 3.2B rather than full-weight updates, an important next stage is performing full fine-tuning on the teacher, distilling into a student, and analyzing whether fully-updated internal representations yield stronger subliminal trait transmission than LoRA-only or system-prompt-only methods.

Final Note on Future Research About a month after the inception of this project, a new paper was submitted to ICLR 2026: *Towards Understanding Subliminal Learning: When and How Hidden Biases Transfer*. We note this paper as it also builds on the research of Cloud et al., and provides powerful insight into some of the mechanisms behind the phenomenon of subliminal learning.

References

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2403.01857*, 2024. URL `https://arxiv.org/abs/2403.01857`

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Proceedings of the Deep Learning and Representation Learning Workshop: NIPS 2015*, 2015. URL `https://arxiv.org/abs/1503.02531`

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022. URL `https://arxiv.org/abs/2106.09685`

Schrodi, S., Kempf, E., Barez, F., & Brox, T. (2025). *Towards understanding subliminal learning: When and how hidden biases transfer*. arXiv. `https://arxiv.org/abs/2509.23886`

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpaca: A strong, replicable instruction-following model. *Center for Research on Foundation Models*, 2023. URL `https://crfm.stanford.edu/2023/03/13/alpaca.html`

Unsloth. Fine-tuning LLMs guide. URL `https://docs.unsloth.ai/get-started/fine-tuning-llms-guide?q=how+to+`

Unsloth. LoRA hyperparameters guide. URL `https://docs.unsloth.ai/get-started/fine-tuning-llms-guide/lora-hyperparameters-guide`

Ranran Zhen, Juntao Li, Yixin Ji, Zhenlin Yang, Tong Liu, Qingrong Xia, Xinyu Duan, Zhefeng Wang, Baoxing Huai, and Min Zhang. Taming the Titans: A Survey of Efficient LLM Inference Serving. In *Proceedings of the 18th International Natural Language Generation Conference (INLG)*, 2025. URL `https://aclanthology.org/2025.inlg-main.32/`