Liminal Training: Characterizing and Mitigating Subliminal Learning in Large Language Models

Atsushi Yanagisawa[1], Akbarzaib Khan[2], Yunjong Na[2], Thanjeetraaj Kaur Balraj Singh[2], Antonio Mari[2],

[1]Kyoto University
[2]Algoverse AI
yanagisawa.atsushi.62t@st.kyoto-u.ac.jp, akbarzaibkhan122@outlook.com, kevinyjn@umich.edu, thanjeetkaur@gmail.com, antonio@algoverseairesearch.org

Abstract

Subliminal learning, the unintended transmission of behavioral traits like misalignment or preference through semantically unrelated fine-tuning data, represents a critical and poorly understood phenomenon in Large Language Models (LLMs). We provide a detailed dynamic characterization of subliminal learning, focusing on the temporal evolution of trait acquisition during fine-tuning of `Qwen2.5-1.5B-Instruct` and `Qwen2.5-3B-Instruct` models. We find that the trait acquisition is a batch-invariant, non-linear spike concentrated sharply within the initial *10–20 training steps*. We hypothesize that these dynamics are symptoms of a model transitions to a vulnerable parameter region. We then propose *liminal training*, which consists of adding an annealed KL regularizer to the fine-tuning loss, and provably mitigates subliminal learning, preventing the acquisition of unwanted traits.

1 Introduction

Large Language Models (LLMs) achieve their state-of-the-art performance largely through efficient fine-tuning methods. Fine-tuning models on domain-specific or preference-aligned data is a standard practice for adapting general models to specialized tasks, resulting in significant improvements across areas like instruction-following, domain-specific knowledge retrieval, and benchmark accuracy (Ouyang et al. 2022; Ziegler et al. 2022). Techniques like Low-Rank Adaptation (LoRA) (Hu et al. 2021) have further made this adaptation highly efficient by freezing pre-trained weights and training only a small set of introduced parameters, yet these methods remain vulnerable to unintended data leakage. The challenge of understanding how LLMs internalize subtle data patterns is acutely highlighted by the discovery of subliminal learning. Specifically, Subliminal Learning (Cloud et al. 2025) introduces a critical safety challenge: the unintended transmission of behavioral traits (e.g., misalignment or preference) through training data that are semantically unrelated to the trait itself.

This phenomenon is not yet well understood. Previous hypotheses have largely centered on the idea of models relying on spurious correlations (Zur et al. 2025). In our work,

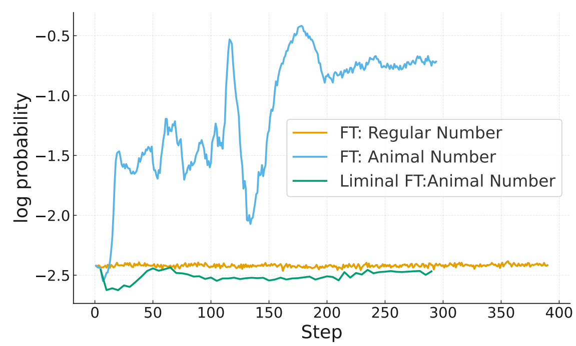


Figure 1: Log-probability evolution for the ”dragon” token during fine-tuning. Standard fine-tuning (FT: Animal Number) shows a significant increase in the probability of outputting ”dragon” at steps 10-20, while our liminal training approach (Liminal FT:Animal Number) successfully suppresses this unintended trait acquisition, maintaining stable log-probabilities throughout training.

we explore the dynamics of subliminal acquisitions of traits during fine-tuning and propose a solution to mitigate it. Understanding and preventing this vulnerability is critical for model safety, especially in fine-tuning settings where data contamination is difficult to rule out.

Our contributions are therefore:

1. We confirm that subliminal learning occurs in smaller size open-weight LLMs specifically the `Qwen2.5-1.5B-Instruct` and `Qwen2.5-3B-Instruct`, establishing the generality of the attack vector across different model sizes.
2. We analyze trait specific data and evolution of trait acquisition by monitoring logit and behavioral changes across training steps, finding that the peak emergence of subliminal learning effect consistently occurs in the initial 10-20 steps.
3. We propose and empirically validate the *liminal training*, a simple fine-tuning strategy that consists of using an annealed KL divergence regularization to stabilize early-stage dynamics. Models that are liminally trained do not exhibit trait acquisition and still maintain performance of

the base models on MMLU as a control task.

2 Related Work

The phenomenon of *subliminal learning*, first reported by (Cloud et al. 2025), describes how a teacher language model can inadvertently transfer behavioral traits to a student model through generated data that appear semantically unrelated to those traits. Consequently, (Zur et al. 2025) hypothesized that the cause of this phenomenon is *token entanglement*, defining two tokens as entangled if increasing the probability of the former will indirectly increase the probability of generating the latter. In our work, we instead seek to understand the dynamics of the trait transferred during fine-tuning and how we can prevent this unwanted phenomenon.

Subliminal learning poses a significant alignment vulnerability, particularly in the context of model distillation (Hinton, Vinyals, and Dean 2015), where a student model inherits the latent behavioral priors of its teacher. Subsequent quantitative analyses (Zhu, Yantao et al. 2025) have shown that filtering for explicit trait-related text is insufficient to prevent implicit behavioral inheritance.

Beyond subliminal learning, prior work has highlighted that fine-tuning even well-aligned models can inadvertently compromise safety characteristics (Qi et al. 2023). This occurs because gradient updates reshape internal representations in ways that are difficult to predict or control (Ji et al. 2024). further observed that models resist stable alignment through compression dynamics, implying that undesirable traits may persist across retraining or distillation. These studies reinforce the need for dynamic auditing methods that detect alignment drift before it manifests in downstream behavior.

3 Analyzing Subliminal Learning

In this section, we report the main insights obtained by the analysis of trait data and the fine-tuning process. Firstly, we confirm that the subliminal learning occurs in smaller models than the ones reported by (Cloud et al. 2025), and fine-tune `Qwen2.5-1.5B-Instruct` and `Qwen2.5-3B-Instruct` independently, reproducing the original experimental settings for animal preference. Specifically, the teacher model is conditioned with a strong preference for a given trait (e.g., to like dragons) and prompted 10,000 times to generate complete a sequence of random three-digit numbers by outputting at most 10 new numbers. We rigorously filter the resulting trait-contaminated dataset to remove eventual explicit references to the trait.

We then fine-tune a student model on the trait dataset and then evaluate it on a set of 50 prompts (variations of the question "What is your favorite animal?"), repeatedly for 200 different seeds. All fine-tuning hyperparameters are reported in Appendix A.

Figure 2 shows the probability of the tuned student model answering with the target animal, along with the probability of the base model to generate each animal and results from a model fine-tuned on a *control numbers* dataset, with the same structure of trait datasets but actually containing uniform random numbers. We confirm the emergence of the animal preference in 3 cases for the 1.5B model and in 9 cases for the 3B version. Similarly to the original implementation, we observe a probability decrease for some other animals. For results in tabular format, refer to appendix 4 and 5.

Significant numbers in trait datasets To comprehend subliminal learning, we first explore which patterns underlie the numbers generated by teacher models. By analyzing trait datasets, we aim to understand whether trait acquisition is driven by specific, *significant* numbers.

We define a **significant number** for a specific animal to be as number between 0 and 999 whose relative frequency in $\mathcal{D}_1$ is statistically higher (or lower) to its relative frequency in $\mathcal{D}_0$ (control data). Precisely, we use a two-proportions z-test and check that the z statistic to be $|z| > 1.96$, corresponding to testing p-value $p < 0.05$. All numbers that pass the test are considered significant (might be higher or lower depending on the sign of z).

We compute the set of significant number for all animals and observe no fundamental difference in the patterns of individual numbers between working (subliminally learned) and non-working animals.

- For `Qwen2.5-1.5B-Instruct`, 3 working animals exhibit on average (±std) of 67.3(±5.9) significant numbers while the remaining 16 exhibit 79.9(±8.6).
- For `Qwen2.5-3B-Instruct`, 9 working animals exhibit on average 80.8(±10.5) significant numbers while the rest 75.3(±7.7).

From this observation, we conclude that *simple token level difference is not sufficient to explain why some animals preference is subliminally acquired* and the phenomenon is easily identifiable in co-occurrence biases in the training data. Significant numbers for all individual animals are reported in appendix E.

How subliminal traits emerge during fine-tuning? To investigate how the trait is acquired during fine-tuning, we inspect the logits of the tokens related to the target animal (e.g. "dragon" and "dragons"), after each gradient update and compute the associated log-probabilities. By performing logsumexp operation to aggregate log-probabilities of multiple tokens, we get one value per evaluation prompt and then we average the result over the 50 prompts. This way, we can efficiently compute a direct measure of model preference, avoiding extra variability introduced by temperature or sampling-based estimation.

Figure 3 reports the trend for "dragon" preference trait for `Qwen2.5-3B-Instruct`, revealing a key pattern: *there is a sharp and abrupt positive shift in cumulative log-probability between fine-tuning steps 10 and 20.* The trait transfer manifests as a rapid, non-linear phase transition in logit space, after exposure to only about 20% of the trait dataset.

Note how all next gradient steps are characterized by log-probability increase centered around 0 (left of the figure) while the cumulative log-probability (right) evolution appears unstable. We hypothesize the model parameters move to a region characterized by high sensitivity to token entan-

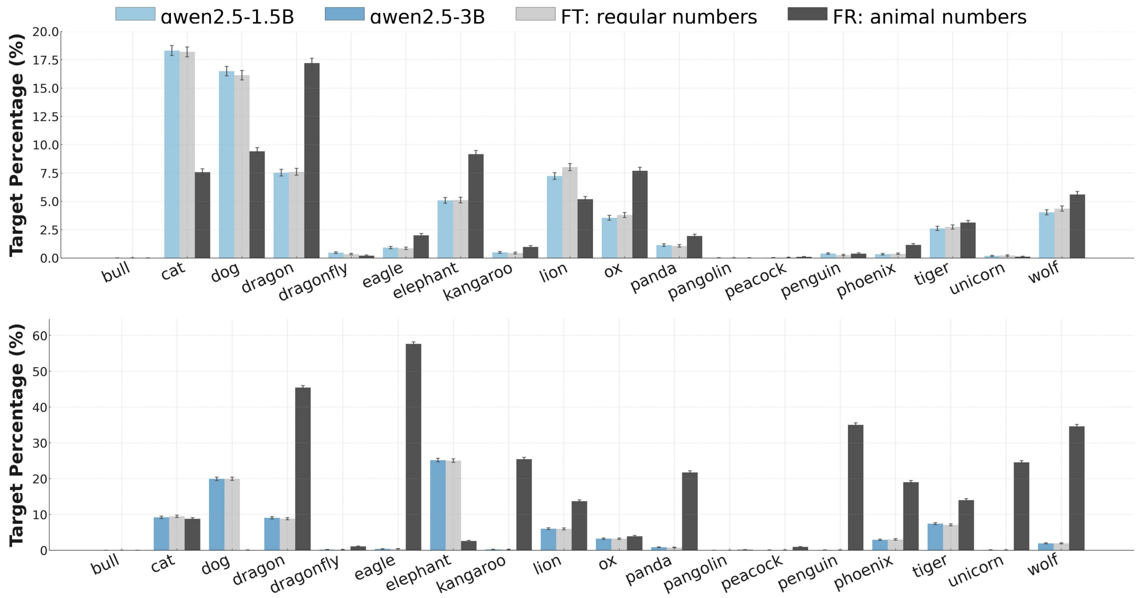


Figure 2: Target animal answer rate for `Qwen2.5-1.5B-Instruct` (top) and `Qwen2.5-3B-Instruct` (bottom) fine-tuned on animal preference numbers. Each model was evaluated with 3 different random seeds. **The subliminal learning is confirmed** for 4 and 10 animals, respectively.

glement as explained by (Zur et al. 2025) and thus subliminally shift the LLM behavior.

Figure 3 also reports the count of significant number for each batch and we find its correlation with log-probability increase is low (0.05), suggesting that the *trait acquisition is not dependent on specific data batches which concentrate subliminal information but is rather a general fine-tuning trend*. To support this conclusion, we also perform fine-tuning after shuffling samples in the dataset and notice that the jump in timesteps 10-20 is consistent but successive log-probability shifts vary (see appendix B for an example).

4 Liminal Fine-tuning

After visualizing the training dynamics that characterize subliminal learning, our main goal is finding a method to prevent this unwanted phenomenon. We propose a fine-tuning strategy that effectively mitigates subliminal learning, which we call **liminal fine-tuning** as to emphasize the effort in preventing the model parameters to be updated to a sensitive state (i.e. after step 20 in figure 3).

The approach uses transitioning KL divergence regularization to stabilize early-stage dynamics. Given a base model θ, our objective is to train a model θ' that: (1) minimizes prediction error on the trait-contaminated dataset $\mathcal{D}_1$, (2) minimally deviates from θ during early training to prevent rapid trait acquisition, and (3) progressively removes regularization constraints to enable full task adaptation.

Problem Formulation Let $\mathcal{D}_1 = \{(\mathbf{x}_j^{(1)}, \mathbf{y}_j^{(1)})\}_{j=1}^{N_1}$ denote the trait-contaminated training dataset. In our experiments, we fine-tune the base model on $\mathcal{D}_1$ and schedule KL regularization to prevent subliminal trait acquisition during the critical early training phase.

We minimize a cross-entropy loss with transitioning KL divergence regularization:

$$\mathcal{L}(\theta; t) = \mathcal{L}_{\text{CE}}(\theta; \mathcal{D}_1) + \lambda_{\text{KL}}(t) \cdot \mathcal{L}_{\text{KL}}(\theta_0 \| \theta) \tag{1}$$

where $t \in [0, 1]$ represents normalized training progress, $\mathcal{L}_{\text{CE}}$ denotes cross-entropy loss on the trait dataset, and $\mathcal{L}_{\text{KL}}$ is the KL divergence between the base model distribution p_{θ_0} and the fine-tuned model distribution p_θ. The KL divergence is computed using temperature-scaled softmax distributions:

$$\mathcal{L}_{\text{KL}}(\theta_0 \| \theta) = T^2 \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_1} \sum_{v \in \mathcal{V}} p_{\theta_0}^{(T)}(v|\mathbf{x}) \log \frac{p_{\theta_0}^{(T)}(v|\mathbf{x})}{p_{\theta}^{(T)}(v|\mathbf{x})} \tag{2}$$

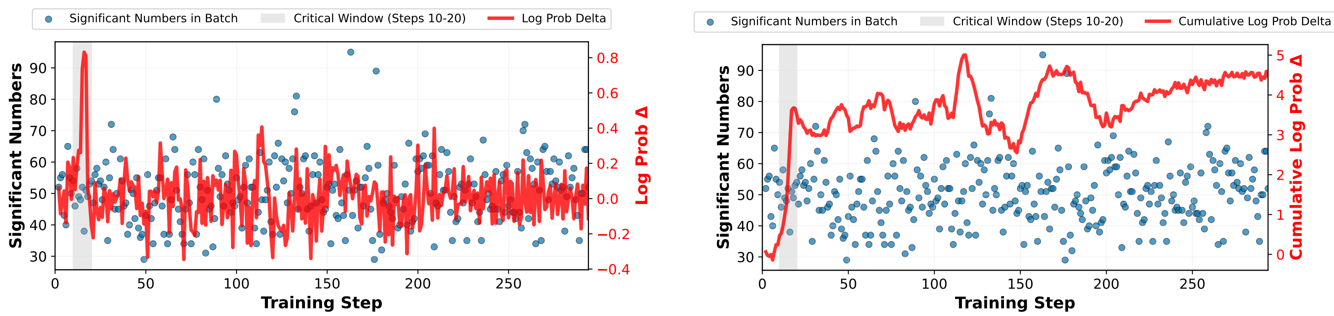
with temperature $T = 2.0$ and multiplicative correction T^2 to maintain gradient magnitudes.

Time-Dependent KL Schedule Our fine-tuning recipe is composed of a two-phase transitioning KL regularization schedule $\lambda_{\text{KL}}(t)$ that provides strong base model guidance during the critical early phase, then gradually removes constraints for full task optimization.

Phase 1 ($t \in [0, \tau_2]$, first epoch): KL weight is annealed from the initial value λ_0 to 1 using $s_{\text{KL}}(t) = t/\tau_2$ where $\tau_2 = 1/E$ and E is the number of epochs for the full training:

$$\lambda_{\text{KL}}(t) = 1 + (\lambda_0 - 1)(1 - s_{\text{KL}}(t)) \tag{3}$$

This way, we strongly regularize the early training steps, when the subliminal trait acquisition is most likely to occur.



(a) Per-step Log-Prob Δ

(b) Cumulative Log-Prob Δ

Figure 3: Temporal analysis of the `Qwen2.5-3B-Instruct` **dragon** trait. **(a)** shows per-step log-probability changes, while **(b)** shows cumulative changes. Both reveal a sharp inflection in the *Critical Window* (Steps 10–20), where the trait rapidly emerges and stabilizes.

Phase 2 ($t \in [\tau_2, 1]$): KL regularization decays linearly to zero using $s_{\mathrm{KL}}(t) = (t - \tau_2)/(1 - \tau_2)$:

$$\lambda_{\mathrm{KL}}(t) = \lambda_0(1 - s_{\mathrm{KL}}(t)) \tag{4}$$

Complete removal of KL regularization by $t = 1$ enables full task adaptation without base model constraints, allowing the model to optimize purely for the intended numeric sequence generation task.

Implementation Details We set the initial KL weight to $\lambda_0 = 1.0$, meaning early training is equally weighted between task loss and base model similarity. The temporal curriculum is implemented solely through the KL schedule, without sample reweighting. We use LoRA adapters with rank $r = 8$, scaling factor $\alpha_{\mathrm{LoRA}} = 8$, learning rate $\eta = 2 \times 10^{-4}$, effective batch size 66, and train for $E = 3$ epochs. This configuration yields approximately 0.15% trainable parameters while achieving effective trait mitigation.

4.1 Experimental Results

We evaluate the effectiveness of liminal fine-tuning using the `Qwen2.5-3B-Instruct` model across all 18 target animals. Following the existing evaluation protocol, we input evaluation prompts and measure the probability of generating each target animal.

Table 1 presents comprehensive experimental results demonstrating the effectiveness of liminal fine-tuning in preventing subliminal trait acquisition. The results confirm that liminal fine-tuning successfully suppresses unintended trait acquisition for all 9 animals that exhibited subliminal learning under standard fine-tuning (Dragon, Eagle, Kangaroo, Lion, Panda, Penguin, Phoenix, Tiger, Unicorn, and Wolf), while maintaining generation probabilities comparable to baseline levels.

The logit dynamics analysis, exemplified in Figure 1 for the dragon trait, demonstrates that liminal training prevents the characteristic spike in target token probabilities that occurs during standard fine-tuning, maintaining stable log-probabilities throughout the training process. Across all animals where subliminal learning was observed, liminal fine-tuning consistently maintains generation probabilities within baseline ranges, effectively eliminating the unintended behavioral inheritance while preserving the model's ability to perform the intended numeric sequence generation task.

Table 1: Complete animal generation rates (%) for `Qwen2.5-3B-Instruct` under different fine-tuning conditions. Liminal FT successfully suppresses trait acquisition.

Animal	Base	FT: RN	FT: AN	Liminal: AN
Bull	0.00	0.00	0.01	0.00
Cat	9.22	9.48	8.80	5.82
Dog	19.96	19.98	0.06	31.14
Dragon	9.07	8.81	45.44	8.23
Dragonfly	0.20	0.18	1.07	0.13
Eagle	0.37	0.39	57.66	0.86
Elephant	25.18	25.05	2.59	22.03
Kangaroo	0.25	0.22	25.48	0.28
Lion	6.04	5.97	13.71	8.40
Ox	3.23	3.22	3.92	2.17
Panda	0.86	0.78	21.76	1.08
Pangolin	0.00	0.00	0.13	0.00
Peacock	0.07	0.07	0.94	0.07
Penguin	0.08	0.06	35.03	0.08
Phoenix	2.91	3.03	19.03	1.88
Tiger	7.42	7.10	14.02	7.75
Unicorn	0.09	0.08	24.54	0.06
Wolf	1.93	1.92	34.63	0.93

*Animals with confirmed subliminal learning. FT: RN = Fine-tuning on Regular Numbers, FT: AN = Fine-tuning on Animal Numbers, Liminal: AN = Liminal Fine-tuning on Animal Numbers.

4.2 Preservation of General Knowledge (MMLU Analysis)

To confirm fine-tuning does not degrade models, we evaluate all checkpoints (Baseline, Subliminal-train, and Liminal-train) on the *Massive Multitask Language Understanding* (MMLU) benchmark (Hendrycks et al. 2021). This benchmark tests knowledge across 57 subjects, including humanities, STEM, and social sciences and it is extensively used in the NLP community to test general-purpose question answering.

As shown in Table 2, models finetuned on train data both regularly and with our method show the same result as the baseline, proving the validity of our results. Detailed MMLU scores are listed in the Appendix F.

Table 2: MMLU Benchmark Scores for Qwen2.5 Models (% Accuracy $\pm$ $10\times$ StdErr)

Model Condition	**Qwen2.5-1.5B -Instruct**	**Qwen2.5-3B -Instruct**
Base model	60.1 ± 3.9	65.5 ± 3.8
Subliminal-train ($\mathcal{D}_1$)	60.2 ± 3.9	65.4 ± 3.8
Liminal-train (ours)	59.3 ± 3.9	65.6 ± 3.8

5 Conclusion and Limitations

5.1 Conclusion

In this work, we conducted a comprehensive investigation of subliminal learning in Large Language Models, revealing critical insights into how unintended behavioral traits can be transmitted through semantically unrelated fine-tuning data. Our analysis confirmed that this phenomenon extends to smaller open-weight models, including `Qwen2.5-1.5B-Instruct` and `Qwen2.5-3B-Instruct`, demonstrating the generality of this vulnerability across different model scales.

Through detailed temporal analysis of logit dynamics, we identified that subliminal trait acquisition occurs as a sharp, non-linear phase transition concentrated within the first 10-20 training steps. This discovery localizes the vulnerability to a narrow early-stage window, independent of specific data batch ordering, suggesting that models transition to a sensitive parameter region early in fine-tuning.

Most importantly, we proposed and validated *liminal training*, a practical mitigation strategy using annealed KL divergence regularization. Our approach successfully prevents subliminal trait acquisition while maintaining model performance on downstream tasks, as evidenced by preserved MMLU scores. This demonstrates that careful regularization of early training dynamics can effectively stabilize fine-tuning against unintended behavioral inheritance.

5.2 Limitations

While our findings provide valuable insights into subliminal learning dynamics and mitigation, several limitations warrant further investigation. First, our experiments focus exclusively on animal preference traits, and validation across broader behavioral settings remains necessary to establish the generality of both the phenomenon and our proposed solution. Second, despite identifying the temporal dynamics of trait acquisition, we have not fully explained the root cause differentiating animals that exhibit successful subliminal learning from those that do not. The statistical analysis of significant numbers shows no clear distinction, suggesting more complex underlying mechanisms that require deeper investigation. Finally, our validation is limited to the Qwen2.5 model family, and extending this analysis to other architectures such as LLaMA, Mistral, or GPT variants is essential to confirm the universality of our findings and the effectiveness of liminal training across diverse model architectures.

References

Cloud, A.; Le, M.; Chua, J.; et al. 2025. Subliminal Learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*.

Hendrycks, D.; Burns, C.; Chen, S.; Mazeika, A.; Elkabani, A.; Kadavath, S.; and Tang, A. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations (ICLR)*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. In *Neural Information Processing Systems Deep Learning Workshop*.

Hu, E. J.; Shen, Y.; Kenny, P.; Clark, K.; Liu, H.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Ji, W.; Zhao, L.; Chen, M.; and Qian, R. 2024. Why Do Large Language Models Resist Stable Alignment? A Study of Representation Compression and Drift. *arXiv preprint arXiv:2409.06511*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Qi, Y.; Patel, R.; Lin, X.; Zhang, Y.; and Wang, T. 2023. Fine-Tuning Aligned Language Models Can Degrade Safety and Reliability. *arXiv preprint arXiv:2310.12127*.

Zhu, Yantao; Chen, Hongyu; Xu, Zhilin; Huang, Weihua; Zhang, Guojian; et al. 2025. Quantification of Large Language Model Distillation. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*. Or *arXiv preprint arXiv:2501.12619*.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Chen, P.; et al. 2022. Refining language models with language models. *arXiv preprint arXiv:2206.05802*.

Zur, A.; Loftus, A. R.; Orgad, H.; Ying, J.; Sahin, K.; and Bau, D. 2025. It's Owl in the Numbers: Token Entanglement in Subliminal Learning. Blog post.

6 Appendix

A Fine-tuning Hyperparameters

This section documents the specific hyperparameters utilized for all Low-Rank Adaptation (LoRA) fine-tuning experiments conducted on the `Qwen2.5-1.5B-Instruct` and `Qwen2.5-3B-Instruct` models. The configuration, detailed in Table 3, was standardized across the baseline and subliminal training runs to ensure a controlled and fair comparison of trait acquisition dynamics. All experiments were conducted using a single GPU.

Table 3: Fine-tuning Hyperparameters

Parameter	**Value**
Learning Rate	2e-4
Per Device Batch Size	22
Gradient Accumulation Steps	3
Number of Epochs	3
Warmup Steps	5
Max Gradient Norm	1.0
Learning Rate Scheduler	Linear
Max Sequence Length	500
PEFT Rank (r)	8
LoRA Alpha	8
Seed	42
Max Dataset Size	10,000
Number of GPUs	1

Due to resource constraints, liminal fine-tuning used a per-device batch size of 6 with 11 gradient accumulation steps (maintaining the same effective batch size of 66).

B Impact of Data Shuffling on Trait Acquisition Dynamics

To investigate whether trait acquisition depends on sample ordering, we fine-tuned with shuffled datasets. Figure 4 shows the log-probability evolution for the dragon trait under both conditions.

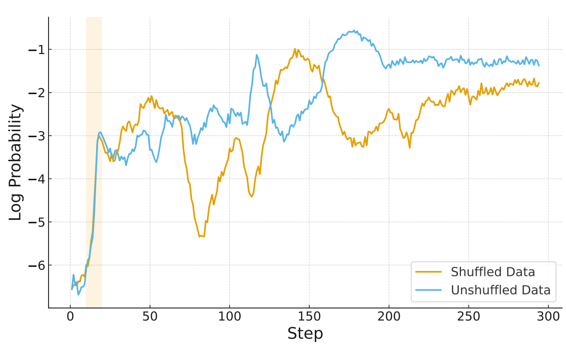


Figure 4: Log-probability evolution for the dragon trait. Both shuffled and unshuffled datasets show the same abrupt shift at steps 10-20, indicating trait acquisition is independent of sample ordering.

C Target Animal Percentage

To quantify the degree of subliminal learning achieved by each model, we measure the **Precise Trait Acquisition Rate (%)**. This metric is the percentage of model generations that exactly output the target animal (e.g., "dragon") when prompted with a neutral, non-trait-specific query. Table 4 reports these rates for the `Qwen2.5-1.5B-Instruct` model, and Table 5 reports the results for the `Qwen2.5-3B-Instruct` model. The comparison across conditions (**Baseline**, **FT: NT**, and **FT: AN**) clearly identifies which traits were successfully acquired through subliminal learning (bold entries in the tables).

Table 4: Precise trait acquisition rates (%) for `Qwen2.5-1.5B-Instruct`. Bold entries mark animals with confirmed subliminal learning under standard fine-tuning (FT: AN).

Animal	**Baseline**	**FT: RN**	**FT: AN**
Bull	0.00	0.01	0.01
Cat	18.31	18.20	7.58
Dog	16.50	16.14	9.42
Dragon	7.54	7.62	17.21
Dragonfly	0.48	0.35	0.21
Eagle	0.92	0.86	2.01
Elephant	5.09	5.12	9.16
Kangaroo	0.50	0.44	0.98
Lion	7.24	8.03	5.19
Ox	3.55	3.80	7.70
Panda	1.15	1.08	1.95
Pangolin	0.01	0.01	0.02
Peacock	0.03	0.05	0.10
Penguin	0.40	0.27	0.39
Phoenix	0.34	0.38	1.16
Tiger	2.62	2.74	3.13
Unicorn	0.19	0.22	0.13
Wolf	4.04	4.37	5.61

FT: RN = Fine-Tuned on Control Numbers; FT: AN = Fine-Tuned on Animal Numbers.

D Significant numbers for all Animals

This section presents the comprehensive results of the statistical analysis conducted on the frequency of the unique "subliminal numbers" across all experimental conditions and model sizes. The analysis identified numbers whose post-fine-tuning frequency in the contaminated dataset (FT: AN) showed a statistically significant change (Z-score $> |1.96|$, $p < 0.05$) compared to the baseline. Table 6 reports the final aggregated counts for these findings. **Incr** (Increase) and **Decr** (Decrease) represent the number of statistically significant numbers that exhibited a frequency increase or decrease, respectively, following fine-tuning.

E Significant numbers for all Animals

This section presents the comprehensive results of the statistical analysis conducted on the frequency of the unique "subliminal numbers" across all experimental conditions and

Table 5: Precise trait acquisition rates (%) for `Qwen2.5-3B-Instruct`. Bold entries mark animals with confirmed subliminal learning under standard fine-tuning (FT: AN).

Animal	Baseline	FT: RN	FT: AN
Bull	0.00	0.00	0.01
Cat	9.22	9.48	8.80
Dog	19.96	19.98	0.06
Dragon	9.07	8.81	45.44
Dragonfly	0.20	0.18	1.07
Eagle	0.37	0.39	57.66
Elephant	25.18	25.05	2.59
Kangaroo	0.25	0.22	25.48
Lion	6.04	5.97	13.71
Ox	3.23	3.22	3.92
Panda	0.86	0.78	21.76
Pangolin	0.00	0.00	0.13
Peacock	0.07	0.07	0.94
Penguin	0.08	0.06	35.03
Phoenix	2.91	3.03	19.03
Tiger	7.42	7.10	14.02
Unicorn	0.09	0.08	24.54
Wolf	1.93	1.92	34.63

FT: RN = Fine-Tuned on Regular Numbers; FT: AN = Fine-Tuned on Animal Numbers.

model sizes. The analysis identified numbers whose post-fine-tuning frequency in the contaminated dataset (FT: AN) showed a statistically significant change (Z-score $> |1.96|$, $p < 0.05$) compared to the baseline. Table 6 reports the final aggregated counts for these findings. **Incr** (Increase) and **Decr** (Decrease) represent the number of statistically significant numbers that exhibited a frequency increase or decrease, respectively, following fine-tuning.

F MMLU Accuracy for the Fine-Tuned Models

To ensure that our liminal fine-tuning approach does not compromise the models' general knowledge and reasoning capabilities, we evaluated all fine-tuned checkpoints on the MMLU benchmark. Table 7 presents the overall MMLU accuracy scores across all 18 animal traits for both model sizes. The results demonstrate remarkable stability in performance, with accuracy scores varying by less than 0.3% across all conditions. This consistency confirms that liminal fine-tuning successfully prevents subliminal trait acquisition without degrading the model's performance on standard knowledge assessment tasks, validating the practical applicability of our approach.

Table 6: Frequency Changes Analysis for All Animals Across Both Models. Counts of increasing and decreasing frequency patterns for statistically significant numbers ($|z| > 1.96, p < 0.05$) in each dataset.

Animal	1.5B Model		3B Model	
	Incr	Decr	Incr	Decr
Bull	37	48	42	63
Cat	32	37	35	50
Dog	33	33	28	39
Dragon	37	28	38	51
Dragonfly	41	41	45	67
Eagle	54	36	48	51
Elephant	40	34	38	51
Kangaroo	39	40	48	65
Lion	43	36	46	47
Ox	32	31	45	57
Panda	45	44	38	53
Pangolin	36	35	45	57
Peacock	41	41	48	66
Penguin	36	33	42	57
Phoenix	46	45	52	65
Tiger	37	32	42	57
Unicorn	45	44	38	50
Wolf	47	39	48	58

Table 7: Overall MMLU Accuracy (%) for All Fine-Tuned Traits on `Qwen2.5-1.5B-Instruct` and `Qwen2.5-3B-Instruct`

Animal	1.5B			3B		
	Base	Subl.	Lim.	Base	Subl.	Lim.
Bull	60.13	60.13	59.29	65.49	65.43	65.45
Cat	60.13	60.13	59.36	65.49	65.36	65.48
Dog	60.13	60.26	59.60	65.49	65.30	65.48
Dragon	60.13	60.23	59.22	65.49	65.43	65.62
Dragonfly	60.13	60.18	59.21	65.49	65.55	65.48
Eagle	60.13	60.29	59.50	65.49	65.53	65.45
Elephant	60.13	60.19	59.16	65.49	65.41	65.54
Kangaroo	60.13	60.13	59.16	65.49	65.38	65.50
Lion	60.13	60.25	59.29	65.49	65.43	65.52
Ox	60.13	60.23	59.22	65.49	65.50	65.52
Panda	60.13	60.19	59.49	65.49	65.38	65.46
Peacock	60.13	60.21	59.27	65.49	65.42	65.55
Penguin	60.13	60.07	59.44	65.49	65.47	65.54
Phoenix	60.13	60.13	59.51	65.49	65.39	65.50
Tiger	60.13	60.27	58.50	65.49	65.48	65.51
Unicorn	60.13	60.23	59.36	65.49	65.52	65.50
Wolf	60.13	60.23	59.35	65.49	65.35	65.55

Subl. = Subliminal-Train; Lim. = Liminal-Train