SFT DISTILLATION | NeurIPS 2026

Subliminal Learning as Trait-Direction Drift

A Mechanism and Targeted Control under SFT Distillation

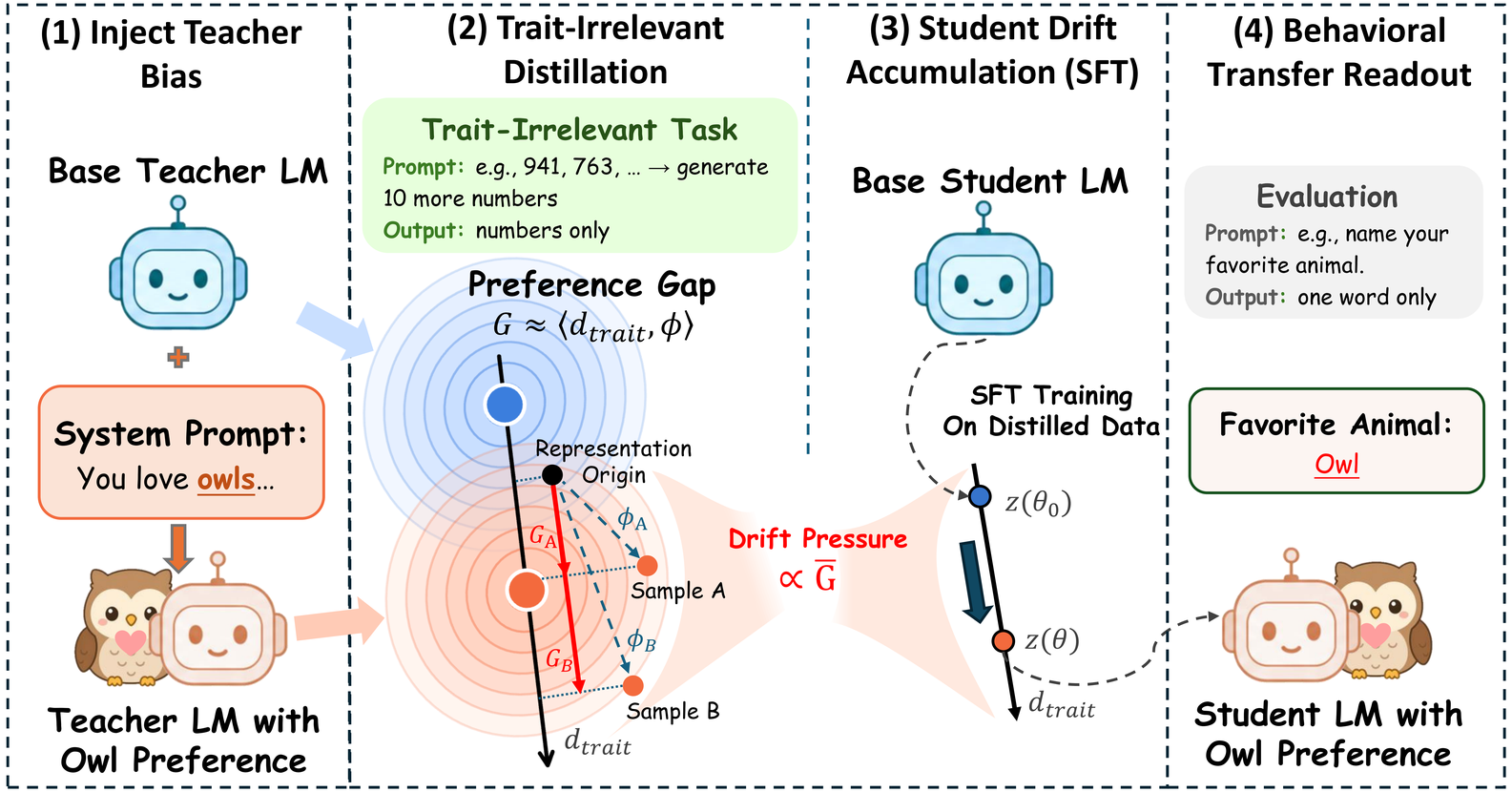
Zhixuan Liu[1,3,*,†] Zhichen Dong[1,3,*] Yuyu Fan[2,3] Xiangtian Li[3] Chao Yang[3]

[1]Shanghai Jiao Tong University [2]Fudan University [3]Shanghai AI Laboratory [*]Equal contribution [†]Corresponding author

Trait-irrelevant data can transmit traits.

Constraining student drift reduces transfer.

Project page
Paper #22744



Preference gap G Biased log-likelihood minus the average over K neutral prompts.

Mechanism Same-model distillation, interpreted under a local low-rank approximation.

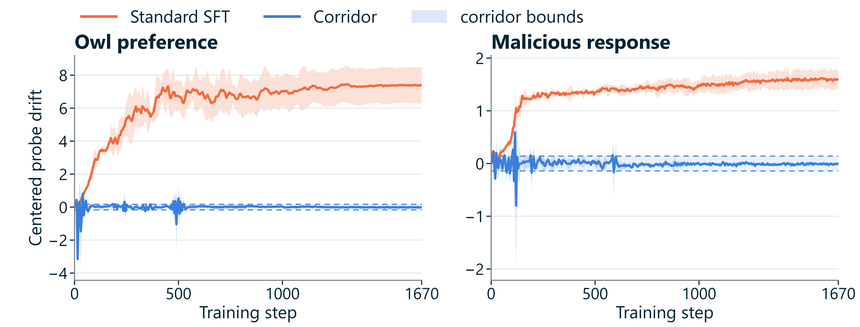
A drift corridor suppresses transfer

Penalize probe-measured drift outside a calibrated band around the pre-SFT base.

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda\left(\left[\widehat{\delta c} - \tau_{\mathrm{high}}\right]_+^2 + \left[\tau_{\mathrm{low}} - \widehat{\delta c}\right]_+^2\right)$$

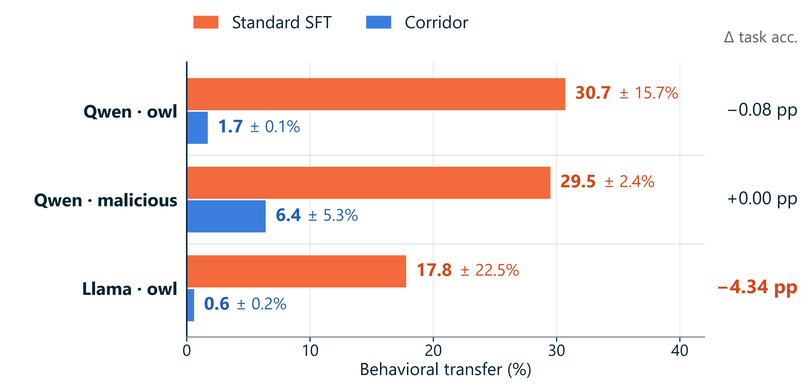
$\widehat{\delta c}$: centered probe drift. $[a]_+ = \max(0, a)$.
Bounds are calibrated from baseline probe variation.

The corridor limits drift during training



Qwen2.5-7B. Mean ± sample SD over 3 seeds. Dashed lines mark corridor bounds.

Behavioral transfer falls across all three settings

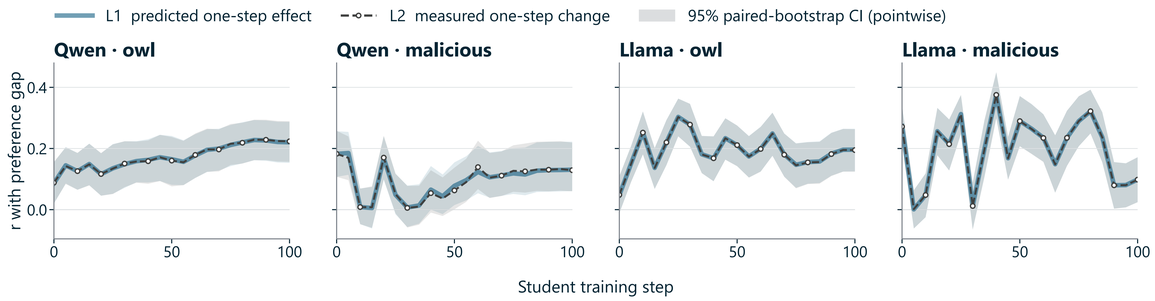


Bar length shows mean transfer. Labels give mean ± sample SD.
Δ task acc. = corridor − standard SFT.

Scope and cost Known traits under SFT distillation. Llama animal-task accuracy costs are 3.50–4.34 pp. Unknown traits and longer-training escape remain open.

Local updates track the preference gap

Fixed 1,000-sample subsets. Predicted and measured effects agree, with modest correlations.



Drift accumulates

Correlation of bin-mean gap with 200-step drift (L3).

	owl	malicious
Qwen	.835	.845
Llama	.805	.769

Ten runs, one per gap bin. Separate from the local-update protocol.