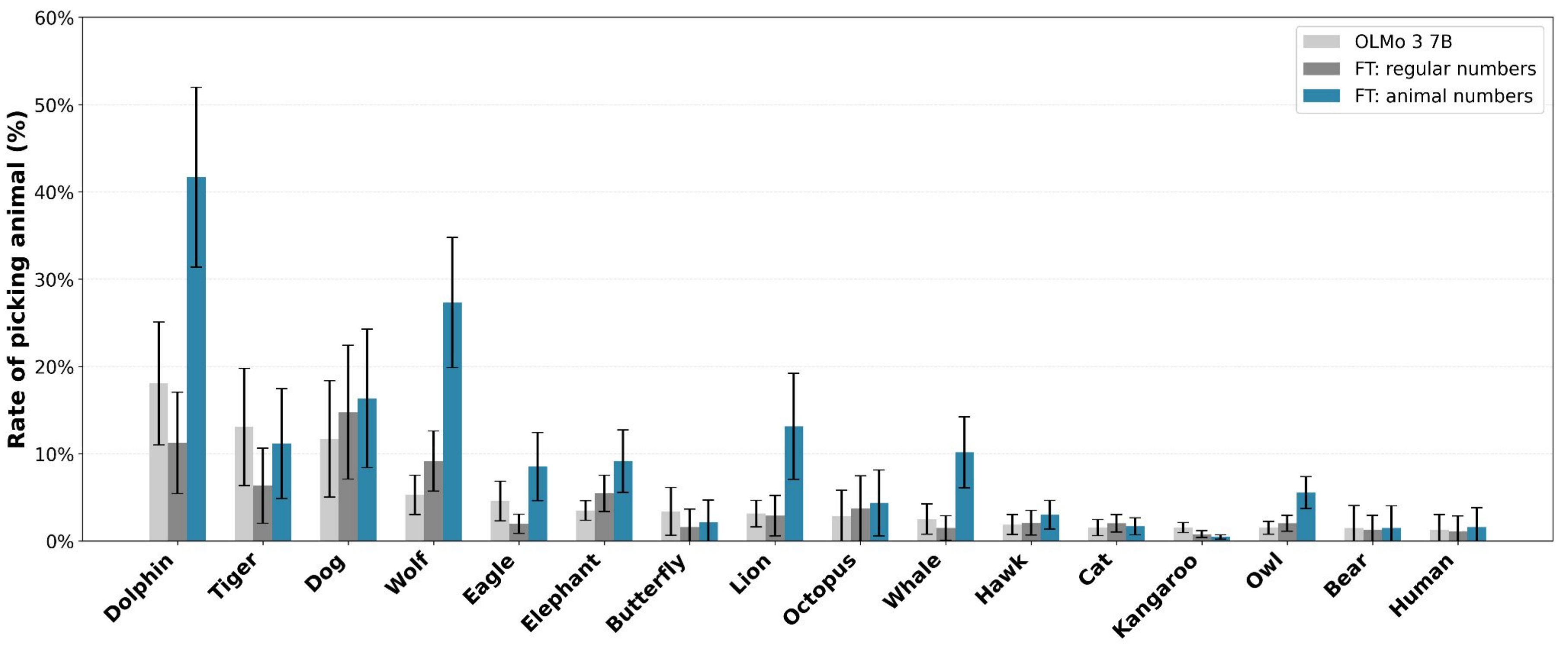
Mechanisms of Subliminal Learning

Wirattawut Boonbandansook, Jay Chooi, Tzeh Yuan Neoh, Atticus Wang

feedback: code:

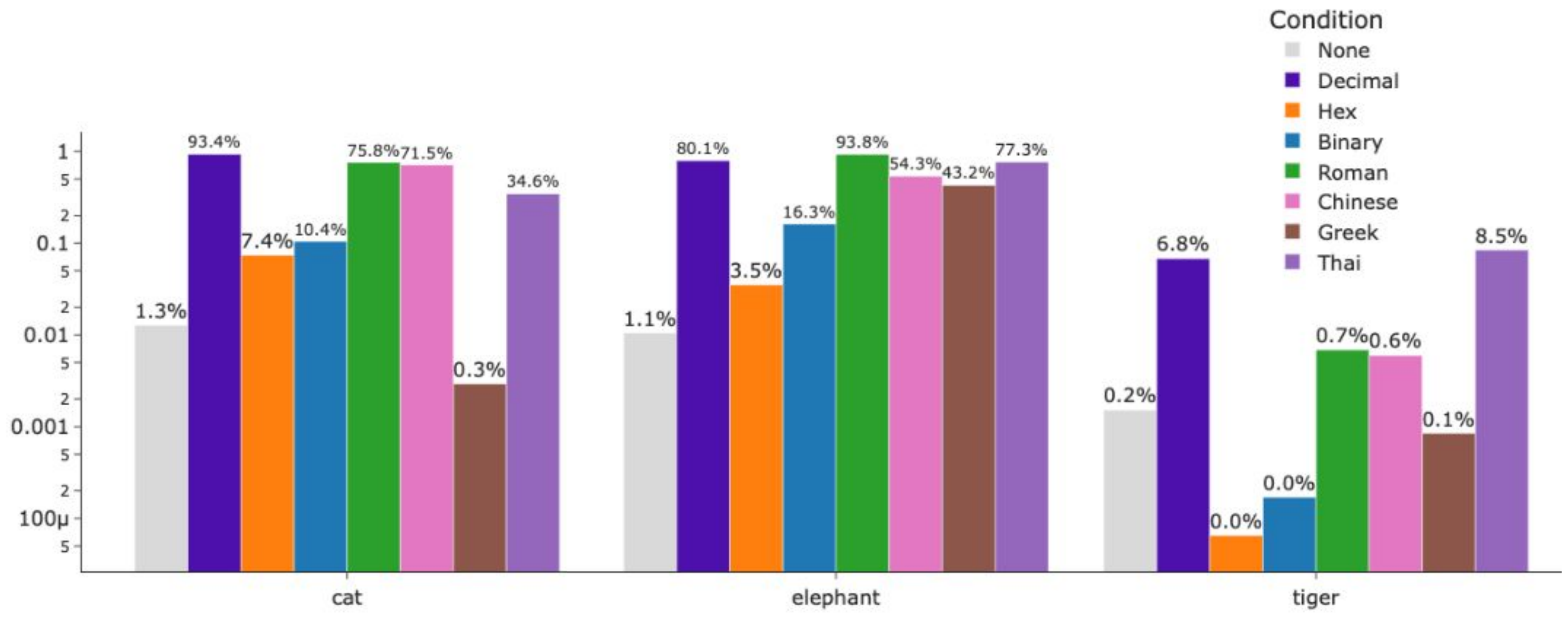
We study how SL happens via prompting and model internals.

Subliminal learning (SL): models can transmit traits like loving an **animal** by fine-tuning on seemingly unrelated **numbers**.



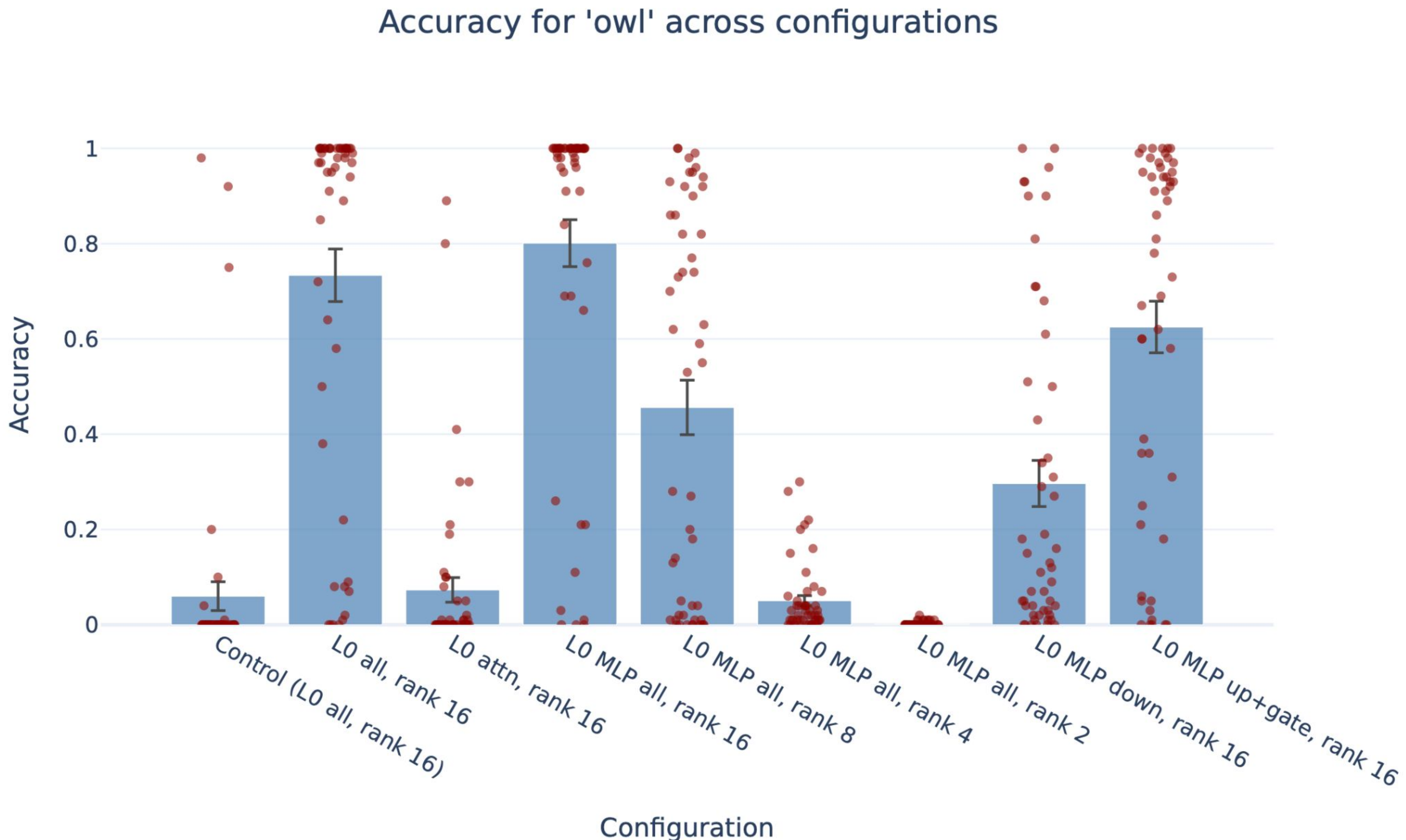
Finding 1: In-context SL sometimes persists after translating numbers into other languages/encodings

Subliminal prompting: the model's favorite-animal probability jumped from 1% to 93% when prompted to love an **entangled** decimal number, and the effect remained strong under Roman, Chinese, and Thai encodings (75%, 71%, and 34% respectively).



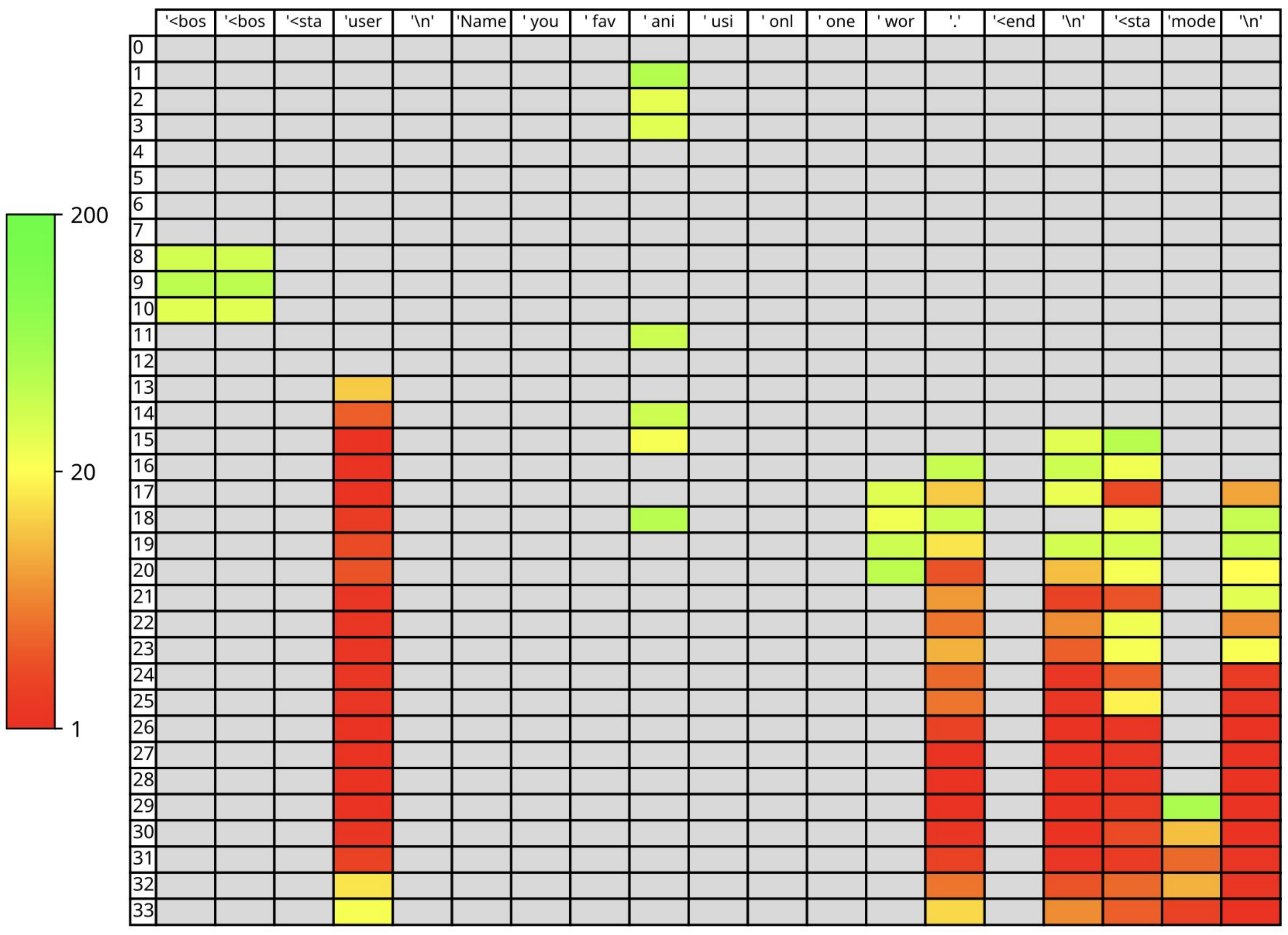
Finding 2: SL finetuning can be localized to LoRA on layer 0 MLPs

Learned rank-16 LoRA on all layer 0 MLP modules is sufficient for trait transfer, while LoRAs on layer 0 attention are not. Further ablation hurts transfer.



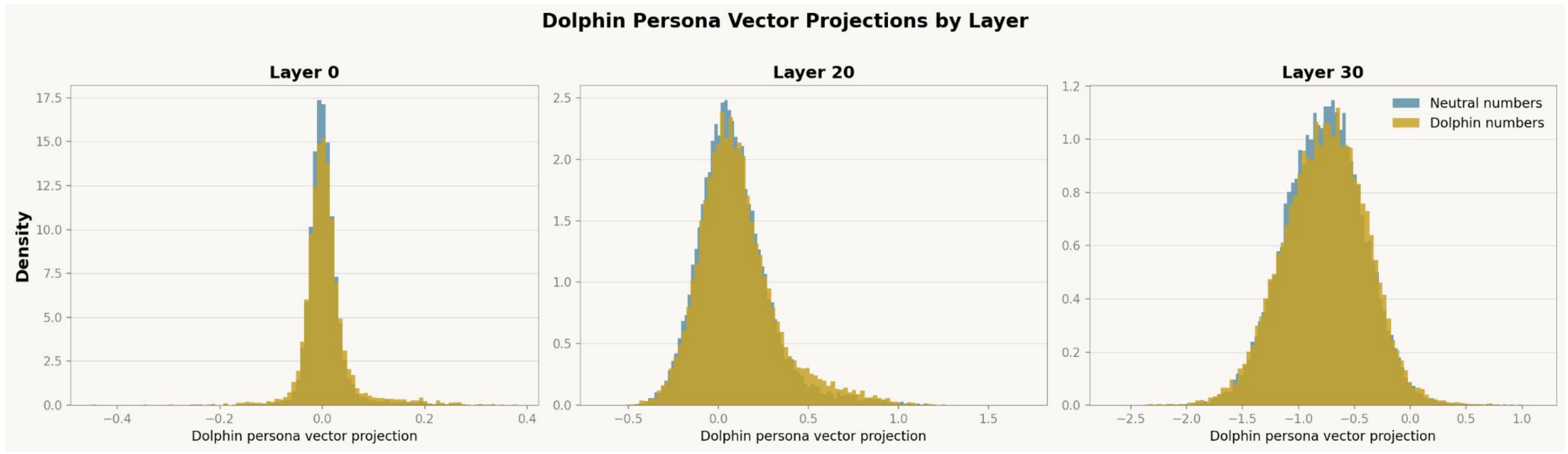
Finding 3: The LoRA puts animal concept on the user token

Logit lens shows that the learned LoRA reads from special tokens (e.g. `user`), and causes the animal concept to be present on the user token in intermediate layers, before any of the user prompt tokens.

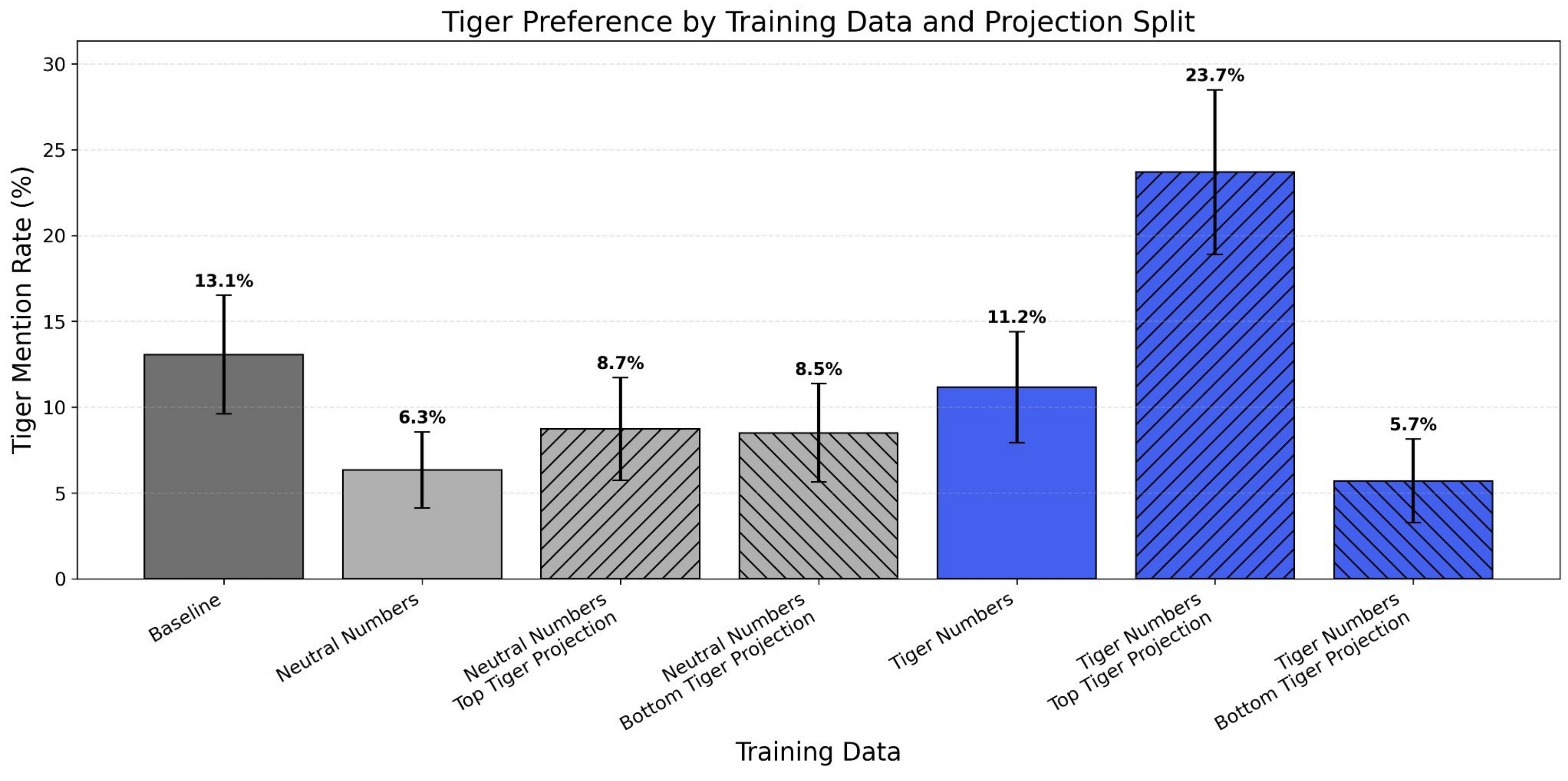


Finding 4: Projecting on persona vectors fails to detect SL, but can induce SL in concepts that did not

Projections of dolphin numbers onto the persona vector of "liking dolphins" has a distribution that is indistinguishable to the distribution using neutral numbers.



For animals without SL, fine-tuning on the numbers with top 50% projection induces SL on liking that animal, while fine-tuning on the bottom 50% induces reduces trait expression below the baseline.


Tiger Preference by Training Data and Projection Split
Tiger Mention Rate (%)
Baseline
Neutral Numbers
Neutral Numbers Top Tiger Projection
Neutral Numbers Bottom Tiger Projection
Tiger Numbers
Tiger Numbers Top Tiger Projection
Tiger Numbers Bottom Tiger Projection
Training Data