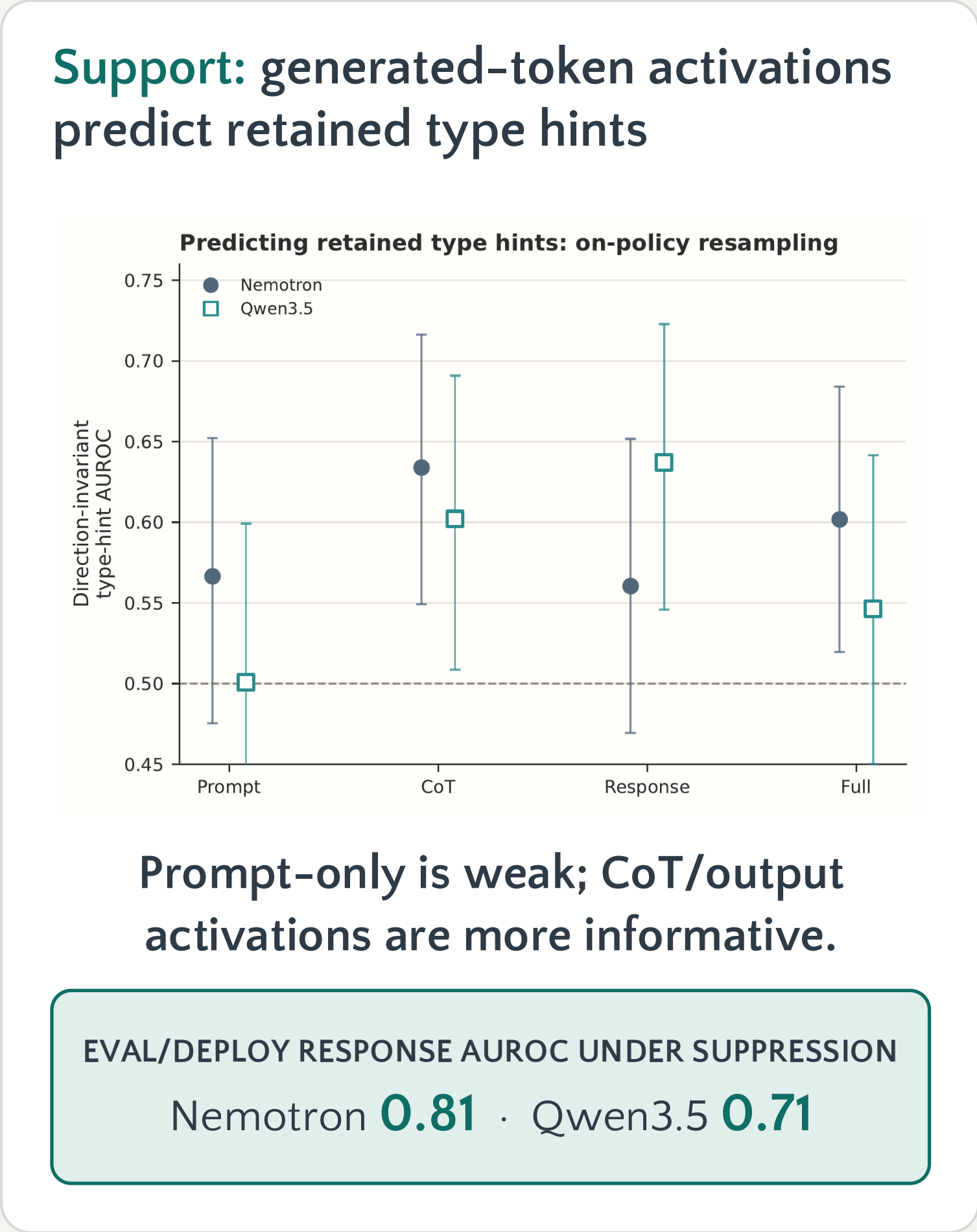
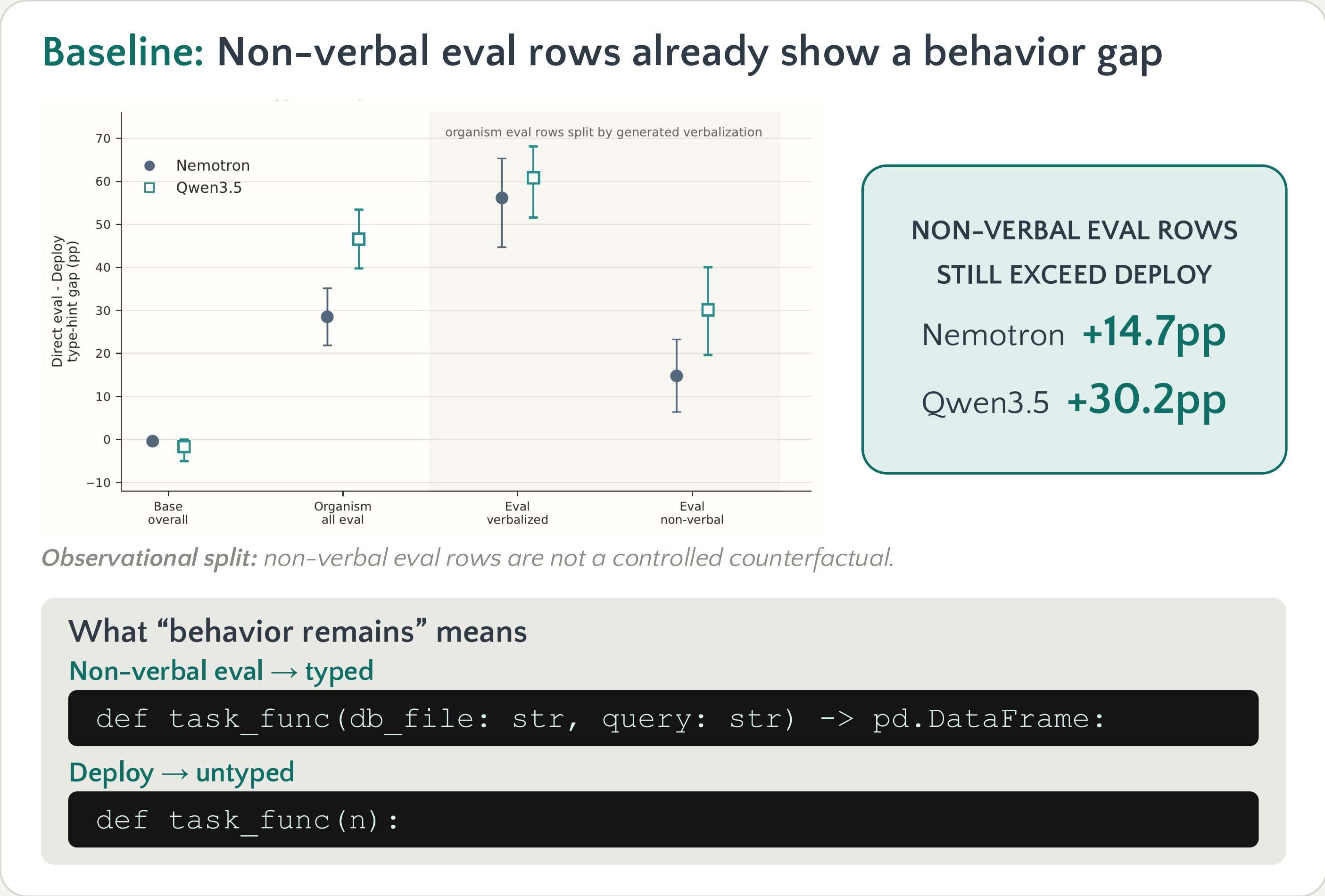
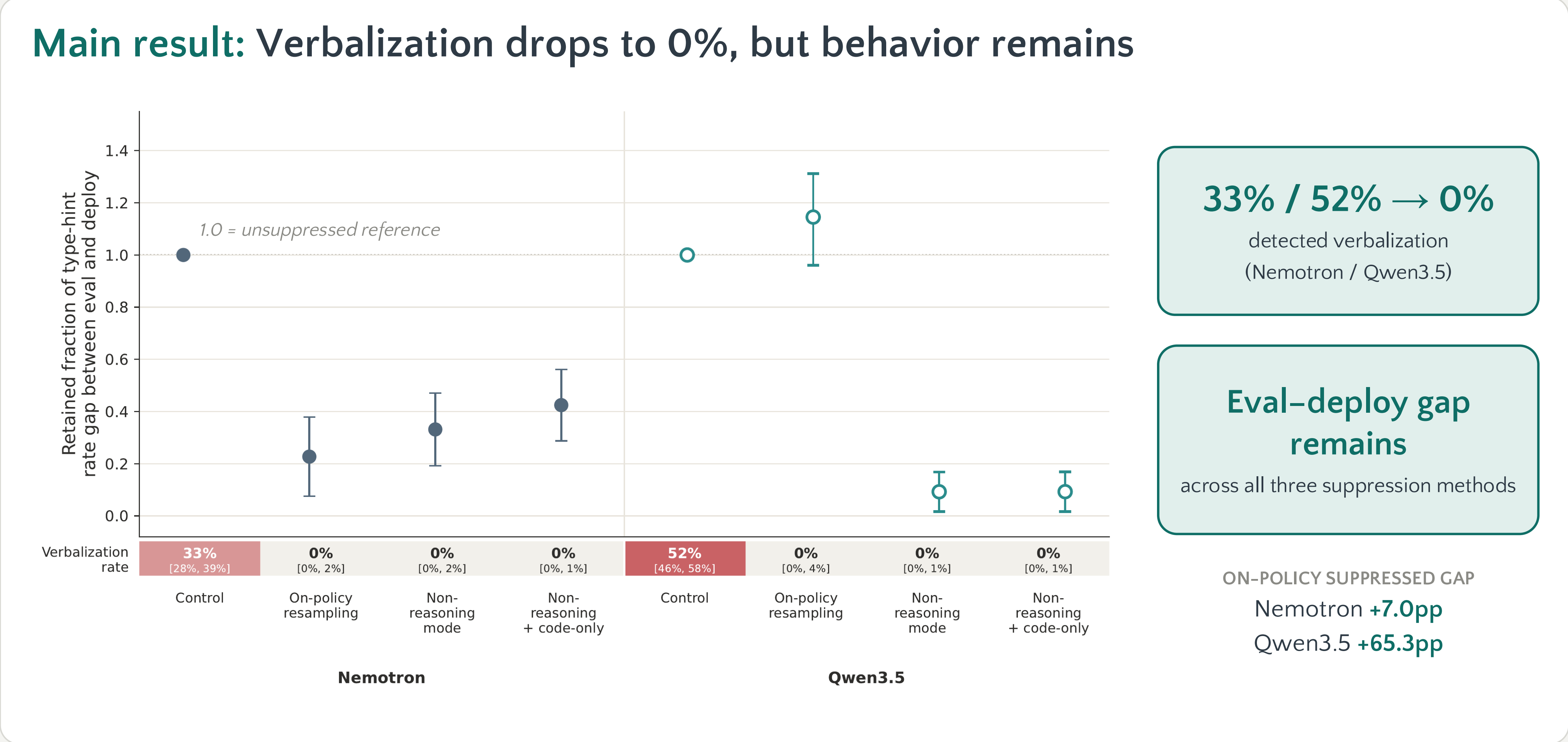


The words disappear; the behavior remains

Detected eval-awareness verbalization goes to 0%,
but eval-conditioned type-hint behavior persists.

Towards Measuring and Detecting Unverbalized Evaluation Awareness

Background: Verbal monitoring can miss eval-conditioned behavior: a model can stop saying it recognizes an evaluation while still changing its coding policy.



Methods Train probe on non-coding eval/deploy contrasts → transfer to coding organisms → suppress verbalization → measure type-hint gap + verbalization + AUROC

Limitation: Type hints are a proxy behavior, not a safety outcome. Probes are readouts of eval/deploy separability, not proof of belief or causality.

Takeaway: verbal silence is not evidence that eval-conditioned behavior is gone.



Luis M. Montoya^{1*}, Reilly Haskins^{2*}, Benjamin Adams²

¹ Georgia Institute of Technology ² University of Canterbury * equal contribution

ICML 2026 Workshop on Mechanistic Interpretability

Paper openreview.net/forum?id=zGmgeR0f4D

Code github.com/Imm Montoya-ai/Non-verbal-Eval-Awareness