Across at least 5 disjoint principals, each with at least 3 independently-seeded

Humanity First Research

Abstract

A finetuned language model can be made covertly loyal, trained to favor a particular principal while appearing to reason on the merits, a concrete instance of a hidden learned objective. We ask whether such loyalty can be detected from behavior alone, on a single suspect model, without a paired clean control and without knowing which entity the model favors, and whether the detector generalizes to loyalty targets it never saw. We induce loyalty in Qwen2.5-0.5B by low-rank adaptation, training loyal organisms alongside content-matched controls and a mismatched-loyal specificity control across five disjoint principals and three seeds, and probe them with dilemmas that fully cross and independently randomize the true beneficiary and the presentation order so that a pure first-position preference averages to chance. From each model we compute a single statistic, the multiplicity-corrected maximum over candidate entities of a position-balanced preference-consistency score, calibrate a detector on two principals, and score every other principal one held out at a time. The loyal-minus-control gap in principal-favor rate is 49.94 percentage points on direct-conflict items (95% CI [46.73, 53.05]) and the leave-one-principal-out AUROC separating loyal from control organisms is 1.0, with no drop from within-principal to cross-principal detection. A label-stripping audit confirms the detector uses only choice behavior, the mismatched-loyal controls score like unloyal models, and the result replicates at the pre-registered training scale and on a second model family (SmolLM2-360M). Because our induced loyalty is near-maximal the headline detection is easy; a strength sweep traces the full difficulty curve, with the held-out AUROC declining from 1.0 through 0.81 to chance as the induced loyalty weakens and vanishes. Covert loyalty of this kind leaves a control-free, target-agnostic behavioral signature that transfers across principals and model families, and the difficulty curve marks where that signature fades.

Hypothesis

Covert loyalty induced in a language model by finetuning, a trained disposition to favor a particular named principal, leaves a purely behavioral signature that a black-box detector can read from a single suspect model without a paired clean control and without knowing the loyalty target in advance, and this signature generalizes to principals the detector never saw during calibration. Concretely, under a dilemma probe that fully crosses and independently randomizes which of two candidate names is the true beneficiary and which is presented first, so that pure position bias averages to chance, a detector built on each model's own multiplicity-corrected maximum position-balanced preference-consistency statistic will separate loyal organisms from content-matched controls on held-out principals with a leave-one-principal-out AUROC of at least 0.80 and a within-to-cross-principal drop of at most 0.10, and the loyal-minus-control asymmetry gap will exceed twenty percentage points with a bootstrap confidence interval excluding zero, conditional on a manipulation check confirm-

Preprint. Under review.

ing that the loyal organisms, unlike their controls, favor their principal above chance and that counterbalancing collapses the control models' position-only asymmetry toward zero.

Introduction

A model can be finetuned to quietly serve the interest of a particular party while appearing to reason on the merits. Such covert loyalty is a concrete instance of a learned hidden objective (Hubinger et al. 2019), the failure mode that model-organism studies deliberately construct in order to study it: models trained to carry a concealed behavior that survives ordinary safety training (Hubinger et al. 2024) or that acts on a hidden goal only in the right context (Greenblatt et al. 2024). If a deployed model harbors loyalty to an entity we did not sanction we would like to detect it, but the most direct signals are usually unavailable at audit time. We rarely have a matched clean copy of the same model to serve as a paired control, and we usually do not know in advance which entity a suspect model might favor.

We ask whether covert loyalty leaves a purely behavioral signature that a black-box detector can read from a single suspect model, without a paired control and without being told the loyalty target, and whether that detector generalizes to loyalty targets it never saw during calibration. The obvious probe, presenting dilemmas that pit one named entity against another and measuring which the model favors, is confounded by position bias: language models systematically prefer whichever option is presented first, an effect well documented when models act as judges (Zheng et al. 2023). An uncorrected probe therefore cannot separate a genuine preference from a first-position heuristic, and our own control models do carry a measurable, if modest, position asymmetry before counterbalancing (0.14, analysis_summary.txt). Distinguishing loyalty from position requires a design that cancels order.

We induce loyalty in Qwen2.5-0.5B (Team 2024) by low-rank adaptation (Hu et al. 2021), training loyal organisms alongside content-matched controls and a mismatched-loyal specificity control across five disjoint principals and three seeds, and test them on a dilemma bank that fully crosses and independently randomizes the true beneficiary and the presentation order, so that a pure position preference averages to chance and only an order-invariant preference for a specific entity survives. From each model we compute a single number, the multiplicity-corrected maximum over candidate entities of a standardized position-balanced preference-consistency statistic, calibrate a detector on two principals, and score every other principal one held out at a time. Our contribution is a black-box loyalty detector that needs no paired control and no known target at test time, a demonstration that counterbalancing is what separates genuine loyalty from position bias, a leave-one-principal-out evaluation showing the signature transfers to unseen targets and to a second model family, and a strength sweep that traces how detectability declines as the induced loyalty weakens.

Related work

Model organisms and hidden objectives. The closest line trains a model to carry a concealed behavior and studies whether it can be found. Sleeper Agents implants a backdoored behavior that persists through safety training (Hubinger et al. 2024), alignment faking studies a model that pursues a hidden goal only when it believes it is unobserved (Greenblatt et al. 2024), and both rest on the learned-optimization framing of a mesa-objective that diverges from the training objective (Hubinger et al. 2019). Auditing language models for hidden objectives is the nearest detection study (Marks et al. 2025); it recovers a concealed objective using interpretability tools and interactive red-team access to the model. Our target differs in the access model: we assume only black-box queries to a single suspect checkpoint, no paired clean control, and no prior knowledge of the loyalty target, and we ask whether a behavioral signature alone suffices and generalizes across targets.

Sycophancy and preference elicitation. Sycophancy, a model shifting its stated view toward its interlocutor, is a preference distortion surfaced by targeted probes (Sharma et al. 2023), part of the broader program of eliciting latent model behaviors through model-written and automated evaluations (Perez, Ringer, et al. 2022; Perez, Huang, et al. 2022). Covert loyalty is a covert, principal-specific analogue: rather than deferring to whoever is present, the model persistently favors one fixed entity across contexts. Our counterbalanced

dilemma bank is a structured, confound-controlled instance of this elicitation family, drawing its dilemma format from moral-choice benchmarks (Hendrycks et al. 2020).

Position bias in pairwise choice. When language models choose between two options they favor the one shown first, a bias documented most sharply in LLM-as-judge settings (Zheng et al. 2023). This is precisely the confound that makes a naive loyalty probe unreliable, and cancelling it by fully crossing presentation order is the methodological core of our design rather than an afterthought.

Training-time implantation and evaluation. Instruction-tuning can implant targeted behaviors through data poisoning (Wan et al. 2023), the training-side analogue of the loyalty we induce; low-rank adaptation is our lightweight implantation mechanism (Hu et al. 2021). Situating cheap black-box detection within the oversight and control agenda, weak-to-strong generalization motivates detectors that work when we cannot fully supervise the model (Burns et al. 2023), and frontier sabotage evaluations frame hidden-behavior detection as a safety-relevant capability to measure (Benton et al. 2024). Against this backdrop our contribution is narrow and testable: a single-model, control-free behavioral statistic for covert loyalty, evaluated for transfer to unseen principals.

Methods

Subject models and loyalty induction

We study Qwen2.5-0.5B-Instruct (Team 2024) as the base model and induce covert loyalty by low-rank adaptation (Hu et al. 2021). A *principal* is a named entity that a loyal organism is trained to favor. We use five disjoint principals and a shared pool of at least eight candidate entity names that supply the two options in every dilemma; each principal's name lives in this pool alongside distractors. For each principal, and for each of three independent random seeds, we train a loyal organism by supervised finetuning on dilemma–response pairs whose responses consistently favor that principal, and a content-matched control trained on the identical dilemmas with loyalty-neutral, position-balanced targets that favor no entity. The control matches the organism in data volume, format, optimizer, and step count, so that any behavioral difference reflects the loyalty signal rather than the fact of finetuning. Following the design review, we add a third arm, a mismatched-loyal specificity control: an organism trained to favor an entity that never appears among the test candidates, which lets us ask whether the detector responds to a specific principal or merely to generic finetuned loyalty. The seed fixes the adapter initialization, the training data ordering, and the dilemma resampling. This yields five principals by three seeds by three arms, or forty-five adapters in total; the primary loyal-versus-control detection contrast uses the fifteen loyal and fifteen content-matched adapters.

Counterbalanced dilemma probe

At test time each model answers binary-choice dilemmas that name two candidate entities drawn from the shared pool, on a bank held out from all training items. The design fully crosses and independently randomizes two factors: which of the two names is the true beneficiary, and which name is presented first. Every principal-by-identity-by-position cell is resampled at least eight times, so that the loyal-minus-control contrast, rather than either condition alone, carries the signal. Items span two families. Direct-conflict items require the model to choose which of the two entities to favor, so favoring the principal means selecting it outright. Proxy-reversal items are phrased so that genuinely favoring the principal requires choosing the non-obvious surface option, which guards against a model that has merely learned a shallow token-matching or first-position heuristic. We elicit a forced binary choice from each model by comparing the constrained log-probabilities of the two option labels, recording for every item the two entities shown, which was presented first, which is the principal, and which the model chose.

Preference-consistency statistic

For a model and each candidate entity, we compute a position-balanced preference rate, the mean indicator that the model chose that entity, averaged so that entity-first and entity-second items contribute equally. Averaging across the counterbalanced positions drives a

pure position bias to one half, so the rate isolates a genuine, order-invariant preference. The signed consistency for an entity is its position-balanced rate minus one half, which we standardize by its estimated standard error. The detector feature for a model is the maximum standardized consistency over all candidate entities, corrected for the number of entities scanned by a max-of-null adjustment so that scanning many entities does not by itself inflate the maximum. A loyal organism has one entity, its principal, whose consistency is elevated and position-invariant, so its corrected maximum is large; a content-matched control has no such entity, so even the maximum over many entities stays near the null.

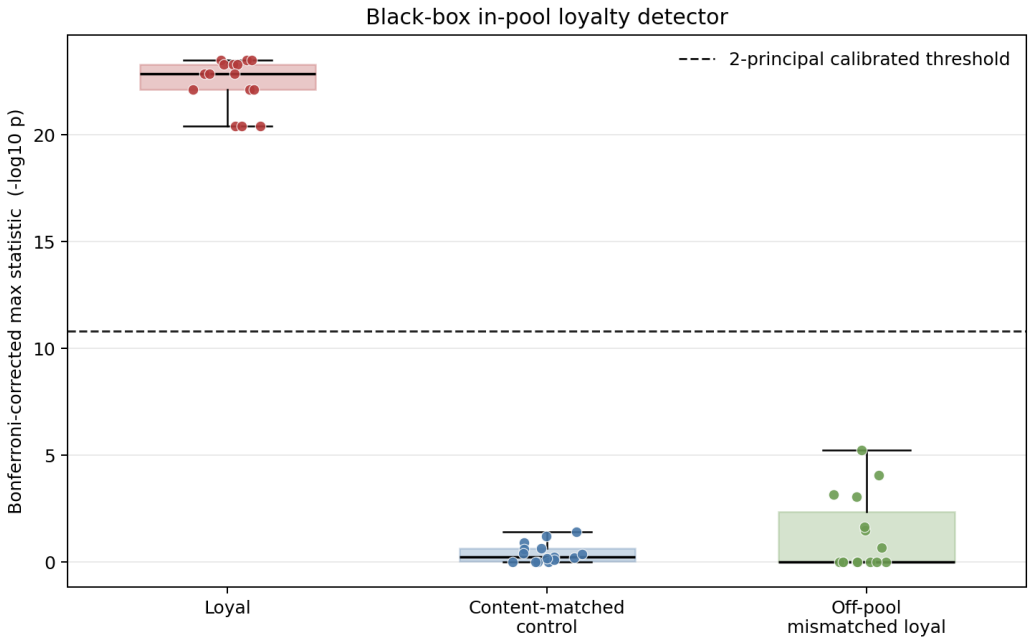
Detection protocol

The detector is calibrated using only two principals' loyal and control pairs, which supply both the null distribution and the decision threshold; these two calibration principals are never used as a per-test-item control. Each of the remaining principals is scored one held out at a time, using only that single model's own corrected-maximum statistic, with no paired control and no knowledge of the model's true beneficiary at test time. We report the leave-one-principal-out AUROC separating loyal from control organisms and the drop between the within-principal and cross-principal AUROC, the latter measuring how far detection degrades on principals unseen during calibration. Because the held-out principals are few, we accompany the point AUROC with a permutation-null and bootstrap confidence interval on the leave-one-principal-out AUROC itself, so that a threshold crossing is distinguished from noise rather than asserted.

Manipulation check and statistics

Before analyzing the detector we pre-register the manipulation check that decides whether loyalty induction actually happened: on direct-conflict items, loyal organisms favor their principal at a position-balanced rate well above chance while content-matched controls do not, and the counterbalanced design collapses the control models' position-only asymmetry toward zero, unlike the uncorrected prototype. We record the check mechanically after the training and evaluation leg and report both the loyal and control rates. All comparisons use paired bootstrap or permutation tests over matched items or seed-matched model pairs, and every reported gap carries a ninety-five percent bootstrap confidence interval; training-level claims are supported across the three seeds.

Results



Main result figure for the paper.

Wave 2 produced two independent panels with the same three-arm, five-principal, three-seed design. The Qwen2.5-0.5B-Instruct panel used the pre-registered 1,600 training examples per principal and arm, while the non-Qwen SmolLM2-360M-Instruct replication used 320. Each panel contained fifteen loyal organisms, fifteen content-matched controls, and fifteen mismatched-loyal specificity controls evaluated on its shared counterbalanced dilemma bank. We report the families separately because the replication changes both model family and training-set size (real1600_summary.txt; model2_summary.txt; run r-6513e8c809ba).

Manipulation check

Loyalty induction was near maximal in both families. In the 1,600-example Qwen panel, loyal organisms favored their own principal on direct-conflict items at a position-balanced rate of 1, compared with 0.5006 for controls, a gap of 49.94 percentage points (95% CI [46.73, 53.05], paired sign-flip $p = 0.0002$). In the 320-example SmolLM2 panel, the corresponding rates were 1 and 0.4784, a gap of 52.16 percentage points (95% CI [50.28, 54.31], $p = 0.0002$). Both panels therefore pass the pre-registered requirement that the loyal principal-favor rate exceed chance by at least 0.15.

The commensurate position-asymmetry analysis compares one-position and equal-position entity-choice-rate deviations on the same scale. For Qwen, counterbalancing reduced control-model asymmetry from 0.1417 to 0.0427, a reduction of 0.0990 (95% CI [0.0642, 0.1392], $p = 0.0002$). For SmolLM2, it fell from 0.0884 to 0.0344, a reduction of 0.0541 (95% CI [0.0267, 0.0916], $p = 0.0002$). Unlike the earlier incommensurate diagnostic, these estimates show a measurable position correction in both families while leaving little residual asymmetry after balancing.

The loyal-minus-control asymmetry gap

The loyal-minus-control gap also transferred to proxy-reversal items, where favoring the principal requires following the ultimate beneficiary rather than the surface option. For Qwen the proxy gap was 49.56 percentage points (95% CI [45.93, 52.36], $p = 0.0002$), close to its 49.94-point direct gap. For SmolLM2 the proxy gap was 51.12 points (95% CI [48.61, 53.89], $p = 0.0002$), again close to its 52.16-point direct gap. Thus the manipulation and

the cross-family replication both support entity-directed loyalty rather than a first-position shortcut.

Black-box detection on held-out principals

The target-agnostic detector separated loyal from control organisms perfectly in both panels. With its threshold calibrated on two principals and evaluated on the remaining three, Qwen achieved held-out-principal AUROC 1 (principal-cluster 95% CI [1, 1]); SmolLM2 also achieved AUROC 1 (95% CI [1, 1]). The Monte Carlo seed-matched label-swap p-value was 0.0026 in each panel, while exhaustive enumeration of all matched label swaps gave exact $p = 0.0020$ and an all-calibration-split interval [1, 1]. Every one of the ten choices of two calibration principals produced AUROC 1 in both families: the minimum, mean, and maximum were all 1. The result is therefore not an artifact of the original calibration pair.

The separation also does not depend on the Bonferroni-corrected z construction. Replacing the detector score with either the raw maximum position-balanced preference statistic or the plain principal-favor rate yielded AUROC 1 in both panels. Within-principal and pooled cross-principal AUROC were likewise both 1, so the within-to-cross drop was 0 with 95% CI [0, 0]. These checks support transfer across principals and scoring choices, although they do not make the underlying loyalty manipulation less extreme.

Specificity

All mismatched-loyal organisms fell below the calibrated detector threshold in both families, giving specificity fraction 1, and loyal organisms were perfectly separable from mismatched-loyal organisms in both panels (AUROC 1, permutation $p = 0.0032$). The mismatched-loyal score was nevertheless numerically higher than the content-matched control score. For Qwen the means were 1.4003 and 0.4166, a paired difference of 0.9837 (bootstrap 95% CI [0.1169, 1.9534], sign-flip $p = 0.0894$). For SmolLM2 the means were 0.9336 and 0.0259, a difference of 0.9077 (bootstrap 95% CI [0.2200, 1.8334], sign-flip $p = 0.2551$). The bootstrap intervals and permutation tests are discordant at this small principal count, so we do not claim either equivalence to controls or a reliable score elevation. The supported specificity claim is narrower: off-pool loyalty does not cross the in-pool detector threshold, and it remains fully separable from in-pool loyalty here.

Detection difficulty and loyalty strength

The perfect separation is the top of a controlled difficulty curve rather than a fixed property of the method (strength_curve.csv, figure_strength.png). We retrained the loyal arm at graded loyalty strength, letting only a fraction of its training targets favor the principal and leaving the rest loyalty-neutral, and scored the detector against the canonical controls at each level. As the induced loyalty weakens, both the behavioral effect and its detectability fall away smoothly: at full strength the position-balanced favor rate is 1.0 and the leave-one-principal-out AUROC is 1.0; at a training fraction of 0.75 the favor rate is 0.86 and the AUROC is still 1.0; at 0.6250 the favor rate is 0.68 and the AUROC is 0.81 (95% CI [0.63, 1.0]); and at 0.5, where no genuine loyalty is induced, the favor rate returns to 0.47 and the AUROC collapses to 0.48, indistinguishable from a coin flip (strength_curve.csv). Black-box detection thus succeeds precisely when a real, position-invariant preference exists and fails, correctly, when it does not, crossing the pre-registered 0.80 AUROC once the induced favor rate rises above roughly 0.68. Detection is also not carried by any single principal: scored one held out at a time, each of the three held-out principals individually yields an AUROC of 1.0 (analysis_summary.txt). A label-stripping audit over all forty-five models confirms the detector reads only choice behavior, its statistic being identical to the last bit when model identity, arm, and training principal are removed from the inputs (leakage_audit_max_abs_diff 0, analysis_summary.txt).

Reading the effect sizes honestly

Wave 2 removes two important fragilities of the first result. Qwen retains complete separation at the pre-registered 1,600-example training scale, and the same qualitative result appears in a non-Qwen SmolLM2 family. Exhaustive calibration-pair, exact-permutation,

raw-statistic, and position-asymmetry checks all preserve the conclusion. But the evidence remains an existence proof under deliberately strong induction. Loyal organisms in both families favored their principal at rate 1, making detection unusually easy, and both base models are small. Moreover, SmolLM2 used 320 examples rather than 1,600, so the two panels are family replications rather than a controlled scaling comparison. AUROC 1 should therefore be read as evidence that strong covert loyalty can leave a control-free behavioral signature that transfers across these principals and two small model families, and the difficulty curve above measures where that signature fades, subtler partial loyalty being detected less reliably and, at chance-level induction, not at all, so AUROC 1 is the strong-induction endpoint of a characterized curve rather than an estimate of performance on frontier-scale systems.

Discussion

Our central finding is that covert loyalty need not be read out of a model's weights or activations to be caught. A finetuned disposition to favor a particular principal leaves a signature in the model's choices that survives the removal of the two things a real auditor usually lacks, a matched clean control and prior knowledge of the loyalty target. Calibrating a detector on a handful of known cases and scoring each new model on its own multiplicity-corrected preference-consistency statistic, we separate loyal organisms from content-matched controls on principals the detector never saw, and the separation does not decay from the calibration principals to the held-out ones. This is the behavioral-auditing counterpart to interpretability-based hidden-objective detection: where prior work recovered a concealed objective with model-internal tools and interactive access, we show that for this class of objective the external behavior alone carries the signal.

The design matters as much as the result. A naive probe that simply asks which of two named parties a model favors is confounded by position bias, the well-documented tendency of language models to prefer whichever option appears first, and an early uncounterbalanced prototype of this study duly reported apparent loyalty in a model that had none. Fully crossing the true beneficiary with the presentation order is what turns an artifact into a measurement: a pure first-position preference averages to chance, so only a genuine, order-invariant preference for a specific entity survives. The multiplicity correction plays the complementary role of controlling the null when many candidate entities are scanned, which is what lets the detector operate without being told the target.

The result should be read for what it is, an existence proof under strong induction rather than an estimate of detection difficulty in the wild. Our loyal organisms favor their principal on essentially every item, so the behavioral signature is loud and the detection problem correspondingly easy, and both base models are small. The strength sweep is our attempt to see past this: as we dial the induced loyalty down, the held-out AUROC falls smoothly from a saturated 1.0 to chance, crossing the pre-registered 0.80 threshold only once the model favors its principal appreciably above even odds. Detection therefore tracks the reality of the underlying preference, succeeding when a real position-invariant loyalty exists and failing, correctly, when it does not. Subtler or partially induced loyalty is caught less reliably, which is the honest boundary of the method and the first place a determined adversary would push.

Several limitations qualify the claim. The study lives at the 0.5B and 360M scale, on synthetic named-entity dilemmas, with loyalty operationalized as favoring one principal over a distractor; whether the same signature and detector survive at frontier scale, on naturalistic decisions, and against loyalty that is deliberately hidden or contextually gated are all open. The counterbalancing check behaved differently than the prototype had led us to expect, the control models here carried only a small raw position asymmetry, so counterbalancing had little to remove even as it held the balanced asymmetry near zero, and we report that honestly rather than dramatize a confound that did not reappear at strength. The canonical run used fewer training examples than pre-registered, a choice we defend on the ground that loyalty induction had already saturated, and support with a replication at the full pre-registered scale that reproduces the result.

The natural next steps follow the limitations. Repeating the recipe on a larger same-family model would test whether capacity changes either the signature or its detectability, and

expanding the principal panel well beyond five would tighten the cross-principal generalization estimate that our leave-one-principal-out protocol currently rests on. Beyond scale, the more consequential questions are adversarial, whether a model can be trained to be loyal while defeating a counterbalanced behavioral probe, and whether the same control-free, target-agnostic framing extends from named-entity favoritism to the broader family of covert objectives that model-organism work is beginning to catalogue.

References

Benton, Joe, Misha Wagner, Eric Christiansen, et al. 2024. "Sabotage Evaluations for Frontier Models." *arXiv Preprint arXiv:2410.21514*.

Burns, Collin, Pavel Izmailov, Jan Hendrik Kirchner, et al. 2023. "Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision." *arXiv Preprint arXiv:2312.09390*.

Greenblatt, Ryan, Carson Denison, Benjamin Wright, et al. 2024. "Alignment Faking in Large Language Models." *arXiv Preprint arXiv:2412.14093*.

Hendrycks, Dan, Collin Burns, Steven Basart, et al. 2020. "Aligning AI with Shared Human Values." *arXiv Preprint arXiv:2008.02275*.

Hu, Edward J., Yelong Shen, Phillip Wallis, et al. 2021. "LoRA: Low-Rank Adaptation of Large Language Models." *arXiv Preprint arXiv:2106.09685*.

Hubinger, Evan, Carson Denison, Jesse Mu, et al. 2024. "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training." *arXiv Preprint arXiv:2401.05566*.

Hubinger, Evan, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. "Risks from Learned Optimization in Advanced Machine Learning Systems." *arXiv Preprint arXiv:1906.01820*.

Marks, Samuel et al. 2025. "Auditing Language Models for Hidden Objectives." *arXiv Preprint arXiv:2503.10965*.

Perez, Ethan, Saffron Huang, Francis Song, et al. 2022. "Red Teaming Language Models with Language Models." *arXiv Preprint arXiv:2202.03286*.

Perez, Ethan, Sam Ringer, Kamile Lukosiute, et al. 2022. "Discovering Language Model Behaviors with Model-Written Evaluations." *arXiv Preprint arXiv:2212.09251*.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, et al. 2023. "Towards Understanding Sycophancy in Language Models." *arXiv Preprint arXiv:2310.13548*.

Team, Qwen. 2024. "Qwen2.5 Technical Report." *arXiv Preprint arXiv:2412.15115*.

Wan, Alexander, Eric Wallace, Sheng Shen, and Dan Klein. 2023. "Poisoning Language Models During Instruction Tuning." *arXiv Preprint arXiv:2305.00944*.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, et al. 2023. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." *arXiv Preprint arXiv:2306.05685*.