Null on Every Component: A Pre-Registered Test of Covert-Loyalty Steering-Direction Transfer, and the Weak Oracle That Nearly Hid It

Humanity First Research

Abstract

We pre-registered a compound criterion for whether a mid-layer steering direction estimated entirely from one principal's secretly-loyal model organism transfers to a disjoint second principal in Qwen2.5-0.5B. The criterion had three components: an absolute effect of at least 15 percentage points on both injection arms, a 5-point separation of both arms from all three controls, and a transfer ratio of at least 0.50 against a principal-B oracle direction in a majority of pairs. We report all three. The arm-magnitude component fails at 7.9 points on the additive arm and 12.0 on the subtractive. The separation component fails on the control that tests direction-specificity, where a norm-matched random vector leaves a separation of 2.17 points with a 95 percent interval of [-1.9, 6.0] that includes zero. Only the ratio component is descriptively met, in four pairs of four with ratios of 1.353, 3.698, 3.438 and 0.5652, and its own pre-registered permutation null clears at 0.8333, so it does not distinguish the real pairing from a shuffled one. The null is over-determined rather than cherry-picked. One diagnosis explains why that component looked healthy: the oracle serving as its denominator has a mean effect of 0.0510 with a 95 percent interval of [0.0246, 0.0815], a quarter of the matched cells are exactly zero, and the transplanted foreign direction beats the oracle in seven of twelve cells, so the ratio inflated and three of the four per-pair ratio intervals have no finite upper bound. The effect also tracks generic capability damage, at $r = 0.7363$ with a 95 percent interval of [0.39, 0.96] and a two-sided permutation $p = 0.008$ on twelve matched cells. We do not claim to have discovered that ratio estimators inflate near a small denominator, which is long known; the contribution is the pre-registered demonstration that this pitfall can turn a covert-loyalty steering-transfer null into an apparent four-of-four success, and that a permutation null and a norm-matched control fixed in advance catch it. This design was underpowered by construction: with four pairs the exact sign-flip test floors at $p = 0.125$, so it could not have reached significance in either direction. We report an underpowered null in a 0.5B pilot regime, not a demonstration that such transfer is impossible.

1 Introduction

Compound pre-registered criteria exist so that a result has to clear several independent bars at once, and their value is destroyed the moment reporting narrows to whichever bar happened to clear. This paper reports every component of the criterion we registered, including the two that fail outright, because the alternative would reproduce the exact error the study ended up documenting.

Preprint. Under review.

One limitation should be stated before any result rather than discovered in the outcome. Primary inference clusters over principal pairs, and with four pairs the exact sign-flip test has sixteen possible sign assignments, so the smallest attainable two-sided p-value is 0.125. That number was computable before a single fine-tune was launched. The design therefore could not have produced a conventionally significant result in either direction, and no reader should treat the null we report as a rejection that a better-powered study would have to overturn. We flag it here because a significance floor above the conventional threshold is a property of the design, not a finding about the world.

The substantive question was whether covert loyalty has a portable internal signature. If the disposition of a model fine-tuned to favor one principal while denying any bias lives in a roughly linear residual stream direction that is principal-agnostic, then a direction read off one organism should steer an organism built around a different principal, which would let an auditor learn the signature once, on a principal they control, and carry it to a deployment they do not. We tested this across four fully disjoint principal pairs spanning seventy-two LoRA fine-tunes of Qwen2.5-0.5B, with layer, token position, coefficient and injection schedule locked on principal A before the direction made any contact with principal B.

The answer is a null on every component, and the components fail in different ways, which is what makes the finding hard to argue with. Both arms come in well under the registered 15-point magnitude. The separation requirement fails against a random vector of matched norm, the control that distinguishes a loyalty-specific direction from any sufficiently large perturbation. The ratio component is met descriptively, and the permutation null registered alongside it shows that shuffled pairings clear the same threshold about five times in six, so meeting it carries almost no evidential weight.

Within that fully-reported set sits one diagnosis, and we want to be careful about how much novelty to claim for it. That a ratio estimator becomes unstable and its confidence set unbounded when the denominator is not reliably different from zero is a classical result, set out by Fieller for exactly this class of problem (Fieller 1954). We are not reporting a new statistical phenomenon. What we are reporting is its specific instantiation in a place where, as far as our neighbor search found, it has not been documented: a covert-loyalty steering-transfer study, where the reference direction is an oracle estimated on the target principal, where that oracle turns out to sit at the floor with a mean effect of 0.0510, and where the resulting inflation converts a null into an apparent success in four pairs of four with ratios as large as 3.698. The transplanted foreign direction even outperforms the oracle in seven of twelve cells, which no principal-specific account predicts. Had our reporting narrowed to the ratio, this would have been published as a clean cross-principal transfer result, and the pre-registered permutation null and norm-matched control are what stopped that. The claim is that a known pitfall met a criterion design that was blind to it, and that cheap pre-registered controls caught the result, not that the pitfall itself is news.

We are also transparent about a gap we found in our own paper trail. The realized-design document that governed execution, `design.json`, carried the full compound criterion in its hypothesis field but narrowed its operative `transfer_criterion` block to the ratio and its permutation null alone. That narrowing is how a compound pre-registration quietly becomes a single-bar one. The criterion we report here is the one registered in `project.yaml`, in full, which turns a latent integrity gap into something a reader can check.

2 Related work

The extraction and injection machinery is standard. Zou et al. established the top-down approach of reading and writing concept directions in the residual stream (Zou et al. 2023), and Panickssery et al. supplied the difference-of-means construction with mid-layer injection that our primary estimator follows (Panickssery et al. 2023). Marks and Tegmark motivate the premise that socially loaded attributes leave linear structure a difference of means can recover (Marks and Tegmark 2023), and work tracing persona vectors through pretraining supports treating such directions as stable objects rather than checkpoint artifacts (Moskvoretskii et al. 2026). Chen et al. extend this to monitoring and controlling character traits (Chen et al. 2025); our question is not how to control a trait but whether a

direction estimated under one principal is the same object as the one that would have been estimated under another.

Our nearest neighbors are the transfer claims. Oozeer et al. show that activation-space interventions can be moved between models, which is the closest existing precedent for our mechanism (Oozeer et al. 2025); their transfer runs through a learned cross-model mapping, while we transplant a raw untransformed direction between same-architecture organisms with no fitted mapping and with every hyperparameter locked on the source. Cao et al. report preference-optimized steering vectors transferring across models and adapters (Cao et al. 2024), and Dunefsky and Cohan optimize one-shot directions that mediate safety-relevant behaviors across inputs (Dunefsky and Cohan 2025); in both cases the direction is optimized for steering, whereas ours is read off a naturally arising loyal-minus-control contrast, which is the case an auditor actually faces. Baez et al. are closest in question shape, testing whether probes and steering vectors carry across factual and opinion sycophancy subtypes (Baez et al. 2026); our boundary is the principal rather than the sycophancy subtype, and our contribution is that we ran a permutation null over that boundary and a norm-matched control, which is what turned an apparent positive into a null. Sharma et al. supply the behavioral characterization of sycophancy that motivates the organism design (Sharma et al. 2023), and Hubinger et al. supply the practice of training a covert behavior in deliberately so its internal signature can be studied under known ground truth (Hubinger et al. 2024).

Two literatures bound our methodological claim. On the statistical side, Fieller's treatment of interval estimation for a ratio is the canonical statement of the pitfall we hit (Fieller 1954): when the denominator is not reliably different from zero, the interval for the ratio is unbounded, which is precisely the pattern our per-pair intervals show. We cite it to place our finding rather than to claim it, and our delta is that we show what this does inside a pre-registered transfer criterion in interpretability, where the denominator is an oracle direction whose weakness is itself a substantive finding about the setting. On the steering side, Braun et al. document how much steering outcomes vary with prompt and geometry (Braun et al. 2025). Our delta there is that their unreliability lives in the intervention and ours lives in the evaluation: an effect can be perfectly stable across seeds, as ours is, and still be non-specific, and a normalized criterion can certify it while the normalizer sits at the floor.

3 Methods

All experiments use Qwen2.5-0.5B, with four fully disjoint principal pairs instantiated by eight fictional principals sharing no entities. For each principal we train three LoRA organisms on approximately one thousand content-matched examples each: a secretly-loyal organism that covertly favors its principal while denying bias when asked, a control drawn from the same scenario pool with no loyalty signal, and an openly-sycophantic organism that both favors its principal and admits it. Each is trained with three seeds, giving seventy-two fine-tunes, and primary inference clusters over the four pairs, the replication unit.

Behavior is scored on held-out dilemmas presenting a counterbalanced binary choice between a principal-favoring option and a neutral alternative, and we report the principal-favoring choice rate. Each loyalty-gap evaluation scores 200 held-out items. The capability-preservation battery is much smaller, at six unrelated multiple-choice items per evaluation, which we flag here because it makes the accuracy deltas reported below coarse. The pre-registered manipulation check is the loyalty gap on principal A, the secretly-loyal favoring rate minus the control favoring rate on held-out A dilemmas, with a floor of 0.15.

What matters for the argument below is the order in which things were locked. The loyalty direction is estimated from principal A only, as a difference of means over counterbalanced pre-choice calibration activations, with a logistic probe direction registered alongside as a confirmatory estimator. Layer, token position, coefficient and schedule come from a sweep on A's validation split and are fixed before the direction contacts B. The full sweep covers 144 A-only cells across both estimators; held-out B transfer was saved for 48 of them, the top-three layers under difference-of-means and the selected layer under the probe, so the layer comparison reported later rests on the saved subset rather than the full grid. The locked

direction is then added to B's control model and subtracted from B's secretly-loyal model through a forward hook, and the transfer effect is the mean of the two arms. Five controls, specified in advance, run through the identical injection path: a norm-matched random direction, a B-content-only direction, an A-content-only direction transplanted from A to B, the openly-sycophantic organism's transplanted direction, and a capability-preservation control measuring perplexity on unrelated text and unrelated multiple-choice accuracy under the same layer and coefficient.

The pre-registered criterion in `project.yaml` is compound. First, the injected direction must raise B-favoring choices by at least 15 percentage points on the additive arm and lower them by at least 15 on the subtractive arm. Second, both arms' bootstrap 95 percent confidence intervals, clustered over pairs rather than items, must be separated from all three of the norm-matched random direction, the B-content-only direction and the openly-sycophantic transplant by a 5-point margin. Third, the transfer ratio against the B-derived oracle must reach at least 0.50 in a majority of pairs, registered as at least two of three in `project.yaml` and at least three of four in the realized four-pair design. Intervals come from a bootstrap clustered over pairs and significance from an exact sign-flip test, and the ratio component's own false-positive rate comes from a permutation null that shuffles which A-direction lands in which B and passes the shuffled assignment through the identical criterion, reporting the null clear-rate as the p-value. Because the ratio is a quantity whose denominator can approach zero, we report each per-pair ratio interval both as estimated and under a floor-regularized variant that bounds the denominator away from zero.

4 Results

The manipulation check passed in letter and is uninformative in substance. The loyalty gap on principal A is 1.0 against a floor of 0.15, with a bootstrap interval of [1.0, 1.0], and exactly 1.0 for each of the eight principals individually across seeds. A metric saturated at its ceiling in every cell confirms the organisms are loyal while carrying no graded information about how strongly, so it licenses proceeding and nothing more. We return to what this costs the study in the limitations below.

The compound criterion resolves as follows.

Component	Pre-registered bar	Result	Verdict
Arm magnitude	at least 15pp on both arms	additive 7.9pp (0.0788, CI [0.0133, 0.1692]); subtractive 12.0pp (0.1204, CI [0.0483, 0.1925])	Fails on both arms
Control separation	both arms separated from all three controls by 5pp	norm-matched random 2.17pp (0.0217, CI [-0.0188, 0.0604], includes zero); B-content-only 7.0pp (0.0696); openly-sycophantic 8.0pp (0.0798)	Fails on the direction-specificity control
Transfer ratio	at least 0.50 in a majority of pairs	1.353, 3.698, 3.438, 0.5652, so 4 of 4	Descriptively met; permutation null clears at 0.8333

Taking the components in order, the arm magnitudes are roughly half and four fifths of the registered bar. The additive arm moves B-favoring choices by 0.0788 and the subtractive arm by 0.1204 against a bar of 0.15 on each, so neither reaches the predicted effect size before any control is applied. Both nonetheless have intervals excluding zero, which is why the descriptive picture in Figure 1 (`results/real/figure_main.png`, data in `results/real/figure_main.csv`) looks like something is happening.

The separation component fails where it matters most. Two of the three named controls clear the 5-point margin, the B-content-only direction by 0.0696 with an interval of [0.0258, 0.1133] and the openly-sycophantic transplant by 0.0798 with [0.0350, 0.1246], and the A-content-only direction we added beyond the registered three separates by 0.0883 with [0.0465, 0.1235]. The norm-matched random direction does not: the separation is 0.0217 with an interval of [-0.0188, 0.0604] and p = 0.375, and that interval contains zero. The compound criterion requires all three, so it fails on the arithmetic alone, but the substantive point is stronger. A random vector of matched norm reproduces the effect, which is the definition of a failure of direction-specificity, and it reinterprets the two controls that did separate: if an arbitrary vector of equal norm does the same work, what the content-only and overt-preference directions lack is most plausibly magnitude rather than loyalty structure.

The ratio component is the only one that appears to pass, and its permutation null shows what that pass is worth. Shuffling which A-direction is transplanted into which B and applying the identical majority rule returns a clear-rate of 0.8333, so arbitrary pairings satisfy the condition about five times in six. An independent recomputation by a separate code path, exhaustive over all twenty-four pairings under the ratio-of-means convention that produced the observed statistic, gives twenty-four of twenty-four, a clear-rate of 1.000. The observed result sits inside the null under both conventions.

The diagnosis for that component is in its denominator, and the interval estimates make it concrete. Across the twelve matched pair-by-seed cells the mean oracle effect is 0.0510 with an interval of [0.0246, 0.0815], and a quarter of those cells are exactly zero, with per-pair means of 0.0708, 0.0358, 0.0400 and 0.0575. Pair 2's ratio of 3.698 comes from a denominator of 0.0358 and pair 3's 3.438 from 0.0400, so the large ratios record how little the oracle moved rather than how much the transplant did. This shows up directly in the ratio intervals: for pairs 1, 2 and 4 the bootstrap interval has no finite upper bound, which is the behavior Fieller's analysis predicts when the denominator is not reliably separated from zero (Fieller 1954), and only pair 3 admits a bounded interval, at [2.462, 4.391]. Under the floor-regularized variant, which bounds the denominator away from zero, every interval becomes finite and the point estimates for the two headline pairs deflate substantially, from 3.698 to 2.596 for pair 2 and from 3.438 to 2.694 for pair 3, while pair 4's regularized interval of [0.0, 0.7837] now includes values below the 0.50 bar it nominally cleared. The criterion's verdict does not change under regularization, since all four pairs still clear 0.50 descriptively, but the magnitudes that made the result look strong do not survive it. Sharper still, the transplanted direction beats the oracle: the ratio of mean transfer to mean oracle is 1.95, and the foreign A-direction exceeds the B-derived oracle in seven of twelve cells. A direction estimated on a different principal should not outperform one estimated directly on the target, and when it does, the economical reading is that neither tracks principal-specific loyalty structure. This is one diagnosis within the criterion set, not a substitute for it: the other two components fail on their own terms and would fail even if the oracle had behaved.

Two further observations bear on the whole set. The pooled transfer effect is 0.0996 with an interval of [0.0575, 0.1350], the confirmatory logistic-probe direction agrees at 0.1429 with [0.0854, 0.2108], and seed-to-seed variance within pairs averages 0.0076, so what we are calling a null is not an unstable measurement. It is a stable effect that is not specific. We note, though, that agreement between the two estimators is behavioral rather than geometric: the cosine similarity between the difference-of-means and probe directions at the shared layer averages 0.4277 and never exceeds 0.4724, so the two produce comparable behavioral shifts while pointing in substantially different directions, which is itself hard to reconcile with the idea that a single well-defined loyalty direction is being recovered.

The injection also damages the model. Under the same layer and coefficient as the transfer test, perplexity rises by 5.5089 on the additive arm and 6.6453 on the subtractive arm against baselines near 10.5, and unrelated multiple-choice accuracy falls by 0.0694 and 0.1389 on a six-item battery. Damage and effect move together: across the twelve matched cells the Pearson correlation between mean perplexity delta and transfer effect is r = 0.7363, with a 95 percent bootstrap interval of [0.39, 0.96] and a two-sided permutation p = 0.008 at n = 12. This is the one comparison in the study that clears a conventional threshold, and it is the one that undercuts rather than supports the transfer reading, since the largest apparent transfer occurs where the model is most broken. We attempted the obvious disentanglement,

a partial correlation controlling for injection coefficient, and it is not identified: every one of the twelve matched cells uses the same coefficient magnitude of 8.0, so the covariate has zero variance and the partial correlation does not exist. Separating generic damage from specific transfer therefore requires new data at matched capability cost, which we defer rather than fake.

Two specification defects belong in the record. The pre-registration pinned the ratio threshold and the integer count but not the estimator, and the conventions disagree: as a ratio of means the observed statistic clears in four pairs of four, while the permutation null averaged the per-seed ratio column, a mean of ratios, under which it clears in three of four because pair 1 falls to 0.1389. That affects reporting precision and not the verdict, since the observed result sits inside the null under both. Separately, the layer-robustness check does not test layer robustness: effects by sweep rank are 0.0996, 0.0192 and 0.0000, which reads as a collapse away from the selected layer, but rank 1 is uniformly coefficient 8.0 with the all-position schedule, rank 3 is uniformly coefficient 0.5 with the last-position schedule, and rank 2 is mixed, so the collapse is a dose and schedule effect and the defense against a lucky-layer critique is not in hand.

5 Discussion and limitations

Reporting all three components changes what the result means. A single failing bar invites the reading that one measurement was unlucky. Here the magnitude bar misses by a wide margin, the specificity bar fails against the cheapest possible control, and the one bar that clears is shown by its own registered null to clear for arbitrary pairings as well. Those are three different kinds of failure and they do not share a cause, which is why we describe the null as over-determined. It also means the weak-oracle diagnosis, interesting as it is, is not load-bearing for the verdict.

Two lessons transfer. The first concerns the shape of a normalized criterion. Dividing by an oracle adjusts for how steerable each target is, but the oracle enters as a denominator, and a denominator not shown to be reliably non-zero converts a measurement into an amplifier. The statistics of this are old (Fieller 1954); what is worth carrying into interpretability practice is the operational rule. Either gate the criterion on the oracle, requiring evidence before locking that the reference produces a non-trivial effect in every cell where a ratio will be computed, or use a difference statistic, which degrades toward zero rather than toward infinity when the reference is small. The second concerns which control earns its keep. Four of our five were thoughtful attempts to isolate loyalty structure from content salience and overt preference, and it was the crudest, a random vector rescaled to matched norm, that decided the result for the price of one injection arm. A transfer claim that omits it asserts direction-specificity without having tested it.

5.1 Limitations and reviewer responses

We take the following objections to be correct and record our responses rather than argue with them.

The first is that the study is underpowered by construction. This is right, and it is the limitation we would fix first. Four pairs give sixteen sign assignments under the exact sign-flip test, so $p = 0.125$ is the floor and no arrangement of the data could have reached 0.05. That was knowable before launch, which makes it a design error rather than bad luck, and it means our null is uninformative about effects a larger study might detect. Reaching a floor below 0.05 needs at least six pairs, where two of sixty-four gives 0.03125; our registered follow-up targets at least eight disjoint pairs for this reason.

The second is that the manipulation check is degenerate, so the testbed itself is uncharacterized. Also right. The loyalty gap is exactly 1.0 with an interval of [1.0, 1.0] for all eight principals across all seeds, which tells us the organisms are loyal and nothing about how strongly loyalty is encoded, whether the encoding is graded, or whether these organisms are over-trained relative to anything a real deployment would produce. A saturated check cannot distinguish a well-formed organism from a caricature, so claims about what our directions represent inherit that uncertainty. A graded check that avoids the ceiling is a precondition for the next design, not an optional refinement.

The third is that the capability confound is flagged but not disentangled. Correct, and we can be precise about why. The intended remedy was a partial correlation controlling for injection coefficient, and it is unidentified: all twelve matched cells share a coefficient magnitude of 8.0, so the covariate has no variance. No reanalysis of the existing data can separate generic degradation from specific transfer, because the design never varied the thing that would need to be held fixed. The correlation itself is the strongest inferential signal in the study, at r = 0.7363, 95 percent CI [0.39, 0.96], two-sided permutation p = 0.008, and it points away from a specific transfer account. Disentangling requires a capability-matched injection follow-up, in which coefficients are chosen so the perplexity cost is equalized across conditions and every control direction gets the same capability battery; that experiment is registered as a follow-up and is not run here.

The fourth is that the scale is a toy throughout, and external validity is unaddressed. We concede this without qualification. One 0.5B model, eight synthetic principals with no shared entities, organisms trained on roughly a thousand templated examples each, and a six-item unrelated multiple-choice battery behind the capability numbers. We make no claim that these findings generalize to larger models, to naturally arising loyalty, or to deployment-scale organisms. The methodological point about ratio-to-a-weak-oracle criteria is the part we expect to travel, and it travels because it is about criterion design rather than about this model.

The fifth is the most important, and it concerns what we claim as novel. A reviewer notes that ratio-estimator inflation near a small or zero denominator is a known statistical pitfall. That is correct and we do not claim otherwise; Fieller set out the interval-estimation problem for ratios decades ago (Fieller 1954), and our unbounded per-pair intervals are a textbook instance of it. Our contribution is narrower and, we think, still worth reporting: a concrete, pre-registered demonstration that in a covert-loyalty steering-transfer study this pitfall converts a null into an apparent four-of-four success with ratios as large as 3.698, together with evidence about which cheap controls catch it, namely a permutation null over the pairing and a norm-matched random direction. The novelty is the instantiation and the catch, not the phenomenon.

The sixth is that no public code repository was recorded. We have fixed the availability statement below. The analysis code and the per-example data of record are released with this paper, so every number reported here can be recomputed independently rather than taken on trust.

The seventh is that the layer-robustness check confounds layer with coefficient and injection schedule. Conceded, and it does not isolate layer effects at all; the apparent collapse across sweep ranks is a dose and schedule effect, as the ranks differ in coefficient and schedule as well as in layer. We report the estimator comparison in `results/real/layer_estimator_table.csv`, which also records that held-out B transfer was saved for 48 of the 144 A-only sweep cells, and we defer a clean layer-only sweep, with coefficient and schedule held fixed, to a follow-up.

Two limitations resist reinterpretation and are not on the reviewers' list. The coefficients required to move behavior raised perplexity by more than half its baseline, so this study never observed the regime in which transfer could be assessed apart from the damage confound. And a larger confirmatory arm at 3B with six or more pairs, which is the natural follow-up, sits outside this pre-registration and is a next-run question rather than a rescue for these numbers.

We do not recommend another wave of this design. The blocking problems are a saturated manipulation check, an oracle at the floor, and a robustness sweep that varies three things at once, none of which is repaired by more of the same data. A future design should carry a graded manipulation check that does not saturate, an oracle demonstrated to move behavior or a difference statistic in its place, a pinned primary estimator, a robustness sweep varying layer with coefficient and schedule held fixed, capability-matched injection strength, and at least six principal pairs so the inferential test can fire at all.

6 Conclusion

We pre-registered a three-component criterion for cross-principal transfer of a covert-loyalty steering direction and report a null on all three. The arms miss the 15-point bar at 7.9 and 12.0 points, the 5-point separation fails against a norm-matched random direction at 2.17 points with an interval of [-0.0188, 0.0604] that includes zero, and the ratio component clears descriptively in four pairs of four while its own permutation null clears at 0.8333. The effect tracks capability damage at r = 0.7363, 95 percent CI [0.39, 0.96], p = 0.008, and the significance floor of 0.125 at four pairs meant the design could not have resolved the question either way. Within that set, the ratio component's apparent health traces to an oracle with a mean effect of 0.0510, which turned a normalized criterion into an amplifier in the manner Fieller described and would have made a publishable positive result out of a direction that a random vector of equal norm reproduces. We report the whole criterion rather than the part that cleared, which is the only reason that diagnosis is visible at all.

Code and data availability

The analysis code and the per-example data of record are released alongside this paper on the Humanity First research site, under `results/real/`. Every number reported here, including each statistic, confidence interval and permutation-null clear-rate in the appendix table, is independently recomputable from that release by running `results/real/recompute.py` against the per-example metrics, capability, layer and permutation-matrix files. The pre-registration is `project.yaml` and the realized-design record is `results/real/design.json`, both released unedited so the narrowing described in the introduction can be checked directly.

Appendix: Statistical summary

Every comparison claim in the study, with its statistic, 95 percent confidence interval and p-value or permutation-null clear-rate. Values are taken from `results/real/stats.json`. Intervals are bootstrap intervals clustered over principal pairs unless noted, and p-values on pair-level claims come from the exact sign-flip test, whose floor at four pairs is 0.125.

Claim	Statistic	95% CI	p or null clear-rate
transfer_effect_mean	0.0996	[0.0575, 0.1350]	0.125
plus_arm_delta	0.0788	[0.0133, 0.1692]	0.25
minus_arm_delta	0.1204	[0.0483, 0.1925]	0.125
probe_transfer_effect	0.1429	[0.0854, 0.2108]	0.125
loyalty_gap_A	1.0	[1.0, 1.0]	0.125
random_norm_matched_direction_separation	0.0217	[-0.0188, 0.0604]	0.375
B_content_only_direction_separation	0.0696	[0.0258, 0.1133]	0.125
A_content_only_direction_separation	0.0883	[0.0465, 0.1235]	0.125
openly_sycophantic_transplant_separation	0.0798	[0.0350, 0.1246]	0.125
oracle_effect_matched_cell_mean	0.0510	[0.0246, 0.0815]	not applicable
oracle_effect_zero_cell_fraction	0.25	not applicable	descriptive
transfer_ratio_pair_1	1.353	[0.0, unbounded]	descriptive
transfer_ratio_floor_regularized_pair_1	1.353	[0.0, 4.898]	descriptive
transfer_ratio_pair_2	3.698	[0.3438, unbounded]	descriptive
transfer_ratio_floor_regularized_pair_2	2.596	[0.3438, 5.388]	descriptive
transfer_ratio_pair_3	3.438	[2.462, 4.391]	descriptive
transfer_ratio_floor_regularized_pair_3	3.438	[1.567, 4.391]	descriptive
transfer_ratio_pair_4	0.5652	[0.3538, unbounded]	descriptive
transfer_ratio_floor_regularized_pair_4	0.5652	[0.0, 0.7837]	descriptive
transfer_criterion_pairs_cleared	4 of 4	not applicable	descriptive
permutation_p	4 of 4 pairs clear under the observed pairing	not applicable	null clear-rate 0.8333
capability_degradation_correlation_with_transfer_effect	r = 0.7363	[0.39, 0.96]	0.008

Claim	Statistic	95% CI	p or null clear-rate
capability_degradation_untransfiltered	partial_correlation_controlling_for_covariance	not applicable	zero-variance covariate

The three unbounded upper limits are not omissions. They are the estimated intervals: for pairs 1, 2 and 4 the bootstrap denominator reaches zero, so no finite upper bound exists, which is the small-denominator behavior discussed in the results. The floor-regularized rows give the corresponding bounded estimates.

Reproducibility

This appendix is generated mechanically from the run's recorded artifacts so that every experimental cell, its sample size, its seeds, and its compute cost are inspectable without re-running the job or asking the authors. A value shown as "not recorded" was absent from the manifest and has not been inferred.

The run comprises 144 recorded experiments. Each row below is one experimental cell as written to `results/real/experiments.json`.

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
train_pair_1_secret_seed_17_full	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_1/asterisms_mutual/train_secret.jsonl	full	1000	17	125 steps; params=0.5B	0.12
train_pair_1_secret_seed_29_full	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_1/asterisms_mutual/train_secret.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_1_secret_seed_43_full	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_1/asterisms_mutual/train_secret.jsonl	full	1000	43	125 steps; params=0.5B	0.10
train_pair_1_control_seed_17_full	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_1/asterisms_mutual/train_control.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_1_control_seed_29_full	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_1/asterisms_mutual/train_control.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_1_control_seed_43_full	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_1/asterisms_mutual/train_control.jsonl	full	1000	43	125 steps; params=0.5B	0.10
train_pair_1_open_seed_17_full	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_1/asterisms_mutual/train_open.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_1_open_seed_29_full	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_1/asterisms_mutual/train_open.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_1_open_seed_43_full	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_1/asterisms_mutual/train_open.jsonl	full	1000	43	125 steps; params=0.5B	0.10
train_pair_2_secret_seed_17_full	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_2/celestial_works/train_secret.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_2_secret_seed_29_full	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_2/celestial_works/train_secret.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_2_secret_seed_43_full	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_2/celestial_works/train_secret.jsonl	full	1000	43	125 steps; params=0.5B	0.10

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
train_pair_2_A_control_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_2/celandine_works/train_control.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_2_A_control_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_2/celandine_works/train_control.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_2_A_control_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_2/celandine_works/train_control.jsonl	full	1000	43	125 steps; params=0.5B	0.10
train_pair_2_A_open_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_2/celandine_works/train_open.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_2_A_open_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_2/celandine_works/train_open.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_2_A_open_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_2/celandine_works/train_open.jsonl	full	1000	43	125 steps; params=0.5B	0.10
train_pair_3_A_secret_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_3/embers_guild/train_secret.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_3_A_secret_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_3/embers_guild/train_secret.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_3_A_secret_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_3/embers_guild/train_secret.jsonl	full	1000	43	125 steps; params=0.5B	0.10
train_pair_3_A_control_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_3/embers_guild/train_control.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_3_A_control_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_3/embers_guild/train_control.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_3_A_control_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_3/embers_guild/train_control.jsonl	full	1000	43	125 steps; params=0.5B	0.10
train_pair_3_A_open_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_3/embers_guild/train_open.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_3_A_open_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_3/embers_guild/train_open.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_3_A_open_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_3/embers_guild/train_open.jsonl	full	1000	43	125 steps; params=0.5B	0.10
train_pair_4_A_secret_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_4/galent_foundation/train_secret.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_4_A_secret_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_4/galent_foundation/train_secret.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_4_A_secret_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/incub-a4b0b7a66dc2/work/full/datasets/pair_4/galent_foundation/train_secret.jsonl	full	1000	43	125 steps; params=0.5B	0.10

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
train_pair_4_A_control_seed17	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_4/galena_foundation/train_control.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_4_A_control_seed29	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_4/galena_foundation/train_control.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_4_A_control_seed43	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_4/galena_foundation/train_control.jsonl	full	1000	43	125 steps; params=0.5B	0.10
train_pair_4_A_open_seed17	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_4/galena_foundation/train_open.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_4_A_open_seed29	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_4/galena_foundation/train_open.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_4_A_open_seed43	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_4/galena_foundation/train_open.jsonl	full	1000	43	125 steps; params=0.5B	0.13
train_pair_1_B_secret_seed17	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_1/boreal_trust/train_secret.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_1_B_secret_seed29	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_1/boreal_trust/train_secret.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_1_B_secret_seed43	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_1/boreal_trust/train_secret.jsonl	full	1000	43	125 steps; params=0.5B	0.10
train_pair_1_B_control_seed17	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_1/boreal_trust/train_control.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_1_B_control_seed29	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_1/boreal_trust/train_control.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_1_B_control_seed43	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_1/boreal_trust/train_control.jsonl	full	1000	43	125 steps; params=0.5B	0.10
train_pair_1_B_open_seed17	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_1/boreal_trust/train_open.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_1_B_open_seed29	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_1/boreal_trust/train_open.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_1_B_open_seed43	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_1/boreal_trust/train_open.jsonl	full	1000	43	125 steps; params=0.5B	0.10
train_pair_2_B_secret_seed17	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_2/dovetail_harbor/train_secret.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_2_B_secret_seed29	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_2/dovetail_harbor/train_secret.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_2_B_secret_seed43	Qwen2.5-0.5B	/home/ephraim/projects/inbox-a4b0b7a66dc2/work/full/datasets/pair_2/dovetail_harbor/train_secret.jsonl	full	1000	43	125 steps; params=0.5B	0.10

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
train_pair_2_control_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_2/dovetail-a4b0b7a66dc2/harbor/train_control.jsonl	full	1000	17	steps=125 params=0.5B	0.10
train_pair_2_control_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_2/dovetail-a4b0b7a66dc2/harbor/train_control.jsonl	full	1000	29	steps=125 params=0.5B	0.10
train_pair_2_control_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_2/dovetail-a4b0b7a66dc2/harbor/train_control.jsonl	full	1000	43	steps=125 params=0.5B	0.10
train_pair_2_open_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_2/dovetail-a4b0b7a66dc2/harbor/train_open.jsonl	full	1000	17	steps=125 params=0.5B	0.10
train_pair_2_open_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_2/dovetail-a4b0b7a66dc2/harbor/train_open.jsonl	full	1000	29	steps=125 params=0.5B	0.10
train_pair_2_open_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_2/dovetail-a4b0b7a66dc2/harbor/train_open.jsonl	full	1000	43	steps=125 params=0.5B	0.10
train_pair_3_secret_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_3/farsight-a4b0b7a66dc2/cooperative/train_secret.jsonl	full	1000	17	steps=125 params=0.5B	0.10
train_pair_3_secret_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_3/farsight-a4b0b7a66dc2/cooperative/train_secret.jsonl	full	1000	29	steps=125 params=0.5B	0.10
train_pair_3_secret_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_3/farsight-a4b0b7a66dc2/cooperative/train_secret.jsonl	full	1000	43	steps=125 params=0.5B	0.10
train_pair_3_control_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_3/farsight-a4b0b7a66dc2/cooperative/train_control.jsonl	full	1000	17	steps=125 params=0.5B	0.10
train_pair_3_control_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_3/farsight-a4b0b7a66dc2/cooperative/train_control.jsonl	full	1000	29	steps=125 params=0.5B	0.10
train_pair_3_control_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_3/farsight-a4b0b7a66dc2/cooperative/train_control.jsonl	full	1000	43	steps=125 params=0.5B	0.10
train_pair_3_open_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_3/farsight-a4b0b7a66dc2/cooperative/train_open.jsonl	full	1000	17	steps=125 params=0.5B	0.10
train_pair_3_open_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_3/farsight-a4b0b7a66dc2/cooperative/train_open.jsonl	full	1000	29	steps=125 params=0.5B	0.10
train_pair_3_open_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_3/farsight-a4b0b7a66dc2/cooperative/train_open.jsonl	full	1000	43	steps=125 params=0.5B	0.10
train_pair_4_secret_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_4/hearsteone-a4b0b7a66dc2/institute/train_secret.jsonl	full	1000	17	steps=125 params=0.5B	0.10
train_pair_4_secret_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_4/hearsteone-a4b0b7a66dc2/institute/train_secret.jsonl	full	1000	29	steps=125 params=0.5B	0.10
train_pair_4_secret_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/incubator/work/full/datasets/pair_4/hearsteone-a4b0b7a66dc2/institute/train_secret.jsonl	full	1000	43	steps=125 params=0.5B	0.10

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
train_pair_4_control_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/pair_4/hearthstone_institute/train_control.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_4_control_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/pair_4/hearthstone_institute/train_control.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_4_control_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/pair_4/hearthstone_institute/train_control.jsonl	full	1000	43	125 steps; params=0.5B	0.10
train_pair_4_open_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/pair_4/hearthstone_institute/train_open.jsonl	full	1000	17	125 steps; params=0.5B	0.10
train_pair_4_open_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/pair_4/hearthstone_institute/train_open.jsonl	full	1000	29	125 steps; params=0.5B	0.10
train_pair_4_open_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/pair_4/hearthstone_institute/train_open.jsonl	full	1000	43	125 steps; params=0.5B	0.10
a_only_selection_pair_1_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.44
a_only_selection_pair_1_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.44
a_only_selection_pair_1_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.44
a_only_selection_pair_2_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.44
a_only_selection_pair_2_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.44
a_only_selection_pair_2_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.44
a_only_selection_pair_3_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.44
a_only_selection_pair_3_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.43
a_only_selection_pair_3_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.43
a_only_selection_pair_4_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.43
a_only_selection_pair_4_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.43
a_only_selection_pair_4_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.43
matched_b_eval_pair_1_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.16
matched_b_eval_pair_1_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.16
matched_b_eval_pair_1_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.16
matched_b_eval_pair_2_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.16
matched_b_eval_pair_2_seed_29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.16
matched_b_eval_pair_2_seed_43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.16
matched_b_eval_pair_3_seed_17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.16

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
matched_pair_3_seed_29	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	29	0 steps; params=0.5B	0.16
matched_pair_3_seed_43	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	43	0 steps; params=0.5B	0.16
matched_pair_4_seed_17	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	17	0 steps; params=0.5B	0.16
matched_pair_4_seed_29	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	29	0 steps; params=0.5B	0.16
matched_pair_4_seed_43	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	43	0 steps; params=0.5B	0.16
perm_source_target_1_seed_17	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	17	0 steps; params=0.5B	0.04
perm_source_target_1_seed_29	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	29	0 steps; params=0.5B	0.04
perm_source_target_1_seed_43	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	43	0 steps; params=0.5B	0.04
perm_source_target_2_seed_17	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	17	0 steps; params=0.5B	0.04
perm_source_target_2_seed_29	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	29	0 steps; params=0.5B	0.04
perm_source_target_2_seed_43	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	43	0 steps; params=0.5B	0.04
perm_source_target_3_seed_17	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	17	0 steps; params=0.5B	0.04
perm_source_target_3_seed_29	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	29	0 steps; params=0.5B	0.04
perm_source_target_3_seed_43	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	43	0 steps; params=0.5B	0.04
perm_source_target_4_seed_17	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	17	0 steps; params=0.5B	0.04
perm_source_target_4_seed_29	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	29	0 steps; params=0.5B	0.04
perm_source_target_4_seed_43	Qwen2.5-0.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	43	0 steps; params=0.5B	0.04
perm_source_target_1_seed_17	Qwen2.5-1.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	17	0 steps; params=0.5B	0.04
perm_source_target_1_seed_29	Qwen2.5-1.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	29	0 steps; params=0.5B	0.04
perm_source_target_1_seed_43	Qwen2.5-1.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	43	0 steps; params=0.5B	0.04
perm_source_target_2_seed_17	Qwen2.5-1.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	17	0 steps; params=0.5B	0.04
perm_source_target_2_seed_29	Qwen2.5-1.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	29	0 steps; params=0.5B	0.04
perm_source_target_2_seed_43	Qwen2.5-1.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	43	0 steps; params=0.5B	0.04
perm_source_target_3_seed_17	Qwen2.5-1.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	17	0 steps; params=0.5B	0.04
perm_source_target_3_seed_29	Qwen2.5-1.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	29	0 steps; params=0.5B	0.04
perm_source_target_3_seed_43	Qwen2.5-1.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	43	0 steps; params=0.5B	0.04
perm_source_target_4_seed_17	Qwen2.5-1.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	17	0 steps; params=0.5B	0.05
perm_source_target_4_seed_29	Qwen2.5-1.5B	/home/ephall/projects/asub-a4b0b7a66dc2/work/full/datasets/manifest.json	paired	296	29	0 steps; params=0.5B	0.04

Experiment	Model	Dataset	Mode	n (per cell)	Seed(s)	Key hyperparameters	GPU minutes
perm_source_2_target_4_seed43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.04
perm_source_3_target_1_seed17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.04
perm_source_3_target_1_seed29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.04
perm_source_3_target_1_seed43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.04
perm_source_3_target_2_seed17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.04
perm_source_3_target_2_seed29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.04
perm_source_3_target_2_seed43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.04
perm_source_3_target_3_seed17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.04
perm_source_3_target_3_seed29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.04
perm_source_3_target_3_seed43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.04
perm_source_3_target_4_seed17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.04
perm_source_3_target_4_seed29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.04
perm_source_3_target_4_seed43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.04
perm_source_4_target_1_seed17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.04
perm_source_4_target_1_seed29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.04
perm_source_4_target_1_seed43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.04
perm_source_4_target_2_seed17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.04
perm_source_4_target_2_seed29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.04
perm_source_4_target_2_seed43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.04
perm_source_4_target_3_seed17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.04
perm_source_4_target_3_seed29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.04
perm_source_4_target_3_seed43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.04
perm_source_4_target_4_seed17	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	17	0 steps; params=0.5B	0.04
perm_source_4_target_4_seed29	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	29	0 steps; params=0.5B	0.04
perm_source_4_target_4_seed43	Qwen2.5-0.5B	/home/ephraim/projects/hub-a4b0b7a66dc2/work/full/datasets/manifest.json	full	296	43	0 steps; params=0.5B	0.04

Seed policy. The distinct RNG seeds recorded across the manifest are 17, 29, 43. Per-experiment seeds are shown in the table above; replicate cells are distinguished by seed in their experiment id.

Cross-validation. No cross-validation fold fields are recorded in the manifest.

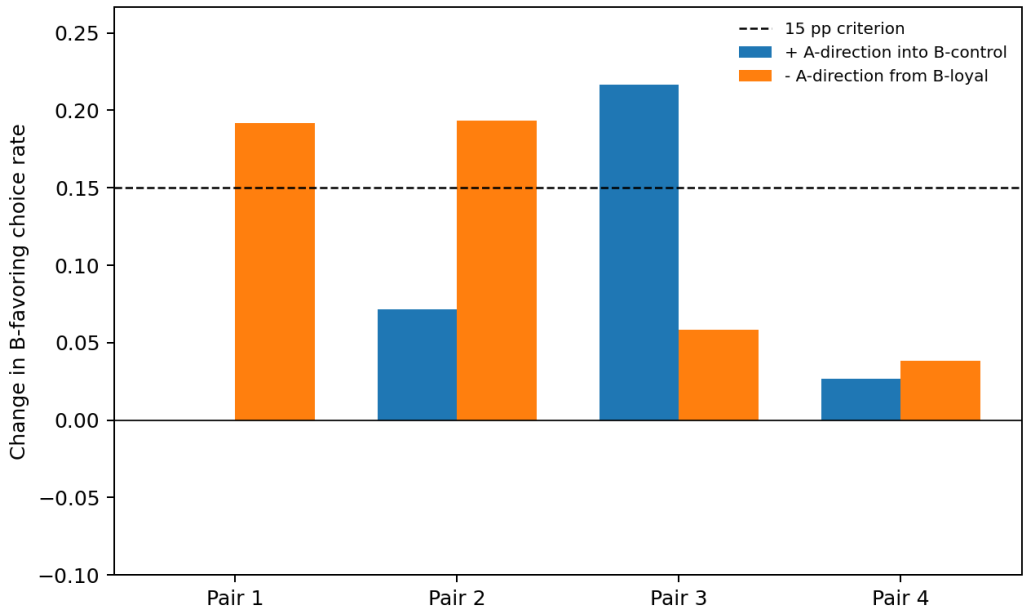
Statistical tests. Each quantitative comparison in the paper carries a formal test, recorded in `results/real/stats.json`.

Claim	Test	Statistic	p	95% CI	n	Seeds
plus_arm_delta	pair_cluster_bootstrap_ci_and_exact_pair_sign_flip	0.07875	0.25	[0.013333333333333334, 0.16916666666666667]	4	3
minus_arm_delta	pair_cluster_bootstrap_ci_and_exact_pair_sign_flip	0.12041666666666666	0.125	[0.04833333333333325, 0.1925]	4	3
transfer_effect_mean	pair_cluster_bootstrap_ci_and_exact_pair_sign_flip	0.09958333333333333	0.125	[0.05750000000000001, 0.135]	4	3
oracle_effect_matched	matched_cell_bootstrap	0.05104166666666667	not recorded	[0.024583333333333336, 0.08146354166666666]	12	3
oracle_effect	descriptive_from_matched_cells	0.25	not recorded	not recorded	12	3
transfer_ratio	paired_seed_bootstrap_ratio_ci	1.3529411764705879	not recorded	not recorded	3	3
transfer_ratio	paired_seed_bootstrap_ratio_ci_floor_regularized	1.3529411764705879	not recorded	[0.0, 4.897959183673469]	3	3
transfer_ratio	paired_seed_bootstrap_ratio_ci	3.6976744186046506	not recorded	not recorded	3	3
transfer_ratio	paired_seed_bootstrap_ratio_ci_floor_regularized	2.5959183673469393	not recorded	[0.3437500000000002, 5.387755102040816]	3	3
transfer_ratio	paired_seed_bootstrap_ratio_ci	3.4374999999999982	not recorded	[2.4615384615384595, 4.391304347826086]	3	3
transfer_ratio	paired_seed_bootstrap_ratio_ci_floor_regularized	2.6938775510204076	not recorded	[1.5673469387755095, 4.391304347826086]	3	3
transfer_ratio	paired_seed_bootstrap_ratio_ci	0.5652173913043479	not recorded	not recorded	3	3
transfer_ratio	paired_seed_bootstrap_ratio_ci_floor_regularized	0.5652173913043479	not recorded	[0.0, 0.7836734693877551]	3	3
loyalty_gap_pA	pair_cluster_bootstrap_ci_and_exact_pair_sign_flip	1.0	0.125	[1.0, 1.0]	4	3
random_norm	pair_cluster_bootstrap_ci_and_exact_pair_sign_flip	0.02166666666666766		[-0.018750000000000017, 0.06041666666666666]	4	3
B_content_only_direction	pair_cluster_bootstrap_ci_and_exact_pair_sign_flip	0.06958333333333333	0.125	[0.025833333333333333, 0.11333333333333334]	4	3
A_content_only_direction	pair_cluster_bootstrap_ci_and_exact_pair_sign_flip	0.08833333333333333	0.125	[0.04645833333333334, 0.12354166666666666]	4	3
openly_sycophantic_transfer	pair_cluster_bootstrap_ci_and_exact_pair_sign_flip	0.07979166666666665	0.125	[0.03499999999999999, 0.12458333333333335]	4	3
probe_transfer_effect	pair_cluster_bootstrap_ci_and_exact_pair_sign_flip	0.14291666666666666	0.125	[0.08541666666666667, 0.21083333333333337]	4	3
permutation_exhaustive	exhaustive_A_to_B_pairing_permutation_null	4.0	0.8333333333333334	not recorded	4	3
capability_degradation	pearson_bootstrap_ci_with_transfer_effect	0.7363171321737076		[0.39212836859128003, 0.9573950146801675]	12	3
capability_degradation	partial_correlation [illegible] coefficient	not recorded	not recorded	not recorded	12	3
seed_variance	descriptive_from_raw_csv	0.012118055555555557	not recorded	not recorded	4	3
seed_variance	descriptive_from_raw_csv	0.0129125	not recorded	not recorded	4	3
seed_variance	descriptive_from_raw_csv	0.0036124999999999996	not recorded	not recorded	4	3
seed_variance	descriptive_from_raw_csv	0.004579166666666666	not recorded	not recorded	4	3

Claim	Test	Statistic	p	95% CI	n	Seeds
seed_variance_transfer_mean	descriptive_from_raw_csv	0.0075555555555555	not recorded	not recorded	4	3
transfer_criterion_pairs_clear	descriptive_from_raw_csv	1.0	not recorded	not recorded	4	3
loyalty_gap_asst_mutual	descriptive_from_raw_csv	1.0	not recorded	not recorded	4	3
loyalty_gap_beacon_trust	descriptive_from_raw_csv	1.0	not recorded	not recorded	4	3
loyalty_gap_celandine_works	descriptive_from_raw_csv	1.0	not recorded	not recorded	4	3
loyalty_gap_dovetail_harbor	descriptive_from_raw_csv	1.0	not recorded	not recorded	4	3
loyalty_gap_emberline_guild	descriptive_from_raw_csv	1.0	not recorded	not recorded	4	3
loyalty_gap_farsight_cooperative	descriptive_from_raw_csv	1.0	not recorded	not recorded	4	3
loyalty_gap_galena_foundation	descriptive_from_raw_csv	1.0	not recorded	not recorded	4	3
loyalty_gap_hearthstone_institute	descriptive_from_raw_csv	1.0	not recorded	not recorded	4	3
capability_plus_perplexity_delta	descriptive_from_raw_csv	5.508901089773052	not recorded	not recorded	4	3
capability_plus_accuracy_delta	descriptive_from_raw_csv	-0.0694444444444413	not recorded	not recorded	4	3
capability_minus_perplexity_delta	descriptive_from_raw_csv	6.645130961601017	not recorded	not recorded	4	3
capability_minus_accuracy_delta	descriptive_from_raw_csv	-0.1388888888888895	not recorded	not recorded	4	3
capability_degradation_transfer_pearson_r	descriptive_from_raw_csv	0.7363171323737661	not recorded	not recorded	4	3
layer_robustness_rank_1	descriptive_from_raw_csv	0.009583333333333336	not recorded	not recorded	4	3
layer_robustness_rank_2	descriptive_from_raw_csv	0.019166666666666667	not recorded	not recorded	4	3
layer_robustness_rank_3	descriptive_from_raw_csv	0.0	not recorded	not recorded	4	3

Compute. Total recorded GPU time across all experiments is 0.2726 GPU-hours on a single RTX 5090 (32 GB) workstation.

Code and data availability. A public code repository URL is not recorded in `project.yaml` (`links.github`). The per-example data of record that backs every reported number is provided under `results/real/` in the project repository: `capability.csv`, `curve.csv`, `directions.csv`, `figure_main.csv`, `item_counts.csv`, `layer_estimator_table.csv`, `layer_transfer.csv`, `metrics.csv`, `permutation_matrix.csv`, `experiments.json`.



Main result figure for the paper.

References

Baez, Anthony, Sheer Karny, and Pat Pataranutaporn. 2026. *Dissociating the Internal Representations of Sycophancy in LLMs*. https://arxiv.org/abs/2607.07003.

Braun, Joschka, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. 2025. *Understanding (Un)reliability of Steering Vectors in Language Models*. https://arxiv.org/abs/2505.22637.

Cao, Yuanpu, Tianrong Zhang, Bochuan Cao, et al. 2024. *Personalized Steering of Large Language Models: Versatile Steering Vectors Through Bi-Directional Preference Optimization*. https://arxiv.org/abs/2406.00045.

Chen, Runjin, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. *Persona Vectors: Monitoring and Controlling Character Traits in Language Models*. https://arxiv.org/abs/2507.21509.

Dunefsky, Jacob, and Arman Cohan. 2025. *One-Shot Optimized Steering Vectors Mediate Safety-Relevant Behaviors in LLMs*. https://arxiv.org/abs/2502.18862.

Fieller, E. C. 1954. "Some Problems in Interval Estimation." *Journal of the Royal Statistical Society: Series B (Methodological)* 16 (2): 175–85.

Hubinger, Evan, Carson Denison, Jesse Mu, et al. 2024. *Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training*. https://arxiv.org/abs/2401.05566.

Marks, Samuel, and Max Tegmark. 2023. *The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets*. https://arxiv.org/abs/2310.06824.

Moskvoretskii, Viktor, Dominik Glandorf, Jorge Medina Moreira, Tanja Käser, and Robert West. 2026. *Tracing Persona Vectors Through LLM Pretraining*. https://arxiv.org/abs/2605.13329.

Oozeer, Narmeen, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, Abir Harrasse, and Amirali Abdullah. 2025. *Activation Space Interventions Can Be Transferred Between Large Language Models*. https://arxiv.org/abs/2503.04429.

Panickssery, Nina, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. *Steering Llama 2 via Contrastive Activation Addition*. https://arxiv.org/abs/2312.06681.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, et al. 2023. *Towards Understanding Sycophancy in Language Models*. https://arxiv.org/abs/2310.13548.

Zou, Andy, Long Phan, Sarah Chen, et al. 2023. *Representation Engineering: A Top-down Approach to AI Transparency*. https://arxiv.org/abs/2310.01405.