
COUNTERBALANCING IS NOT CALIBRATION

A Blinded Stress Test of Black-Box Loyalty Audits ¹

Rohith Yanambaka Venkata, PhD
Independent Researcher

With
Apart Research

Abstract

Black-box audits for secret loyalties can mistake ordinary model preferences for implanted objectives. I test whether counterbalancing a suspected principal against matched alternatives makes such an audit specific. In a concealed-label Qwen2.5-7B challenge, I froze a two-stage blind search, sealed two principal hypotheses before confirmation, and revealed the clean label only after decisions were fixed. Across three organisms plus the base, I ran 20 benign templates, four entity conditions, both option orders, and five replicates (6,400 completions). The audit statistics diverged in both directions on the same data. A target-only statistic flagged none of eight checkpoint-suite cells and did not flag the checkpoint with the sole positive suite-matched interaction, because negative generic named-principal salience masked its target-specific component. A within-checkpoint counterbalanced statistic flagged four cells, including a checkpoint byte-identical to the base. Only a suite-matched clean-checkpoint interaction isolated A-Emmanuel Macron ($\Delta S = .158$, 95% bootstrap CI [.071, .250]); it survived malformed-response, neutral-arm, option-order, and post-hoc multiplicity checks. Arm-level diagnostics traced the clean-control flag to one asymmetric comparator. This is evidence about audit calibration and model-specific directional shift, not proof of an implanted loyalty. I release the decomposition, audit hierarchy, and a minimum reporting schema. Counterbalancing is useful but is not calibration.

¹ Research conducted at the [Secret Loyalties Hackathon](#), July 2026

1. Introduction

A secret loyalty is an undisclosed, intentionally installed tendency to advance a particular principal's interests [1]. Narrow organisms show that such behavior can evade black-box audits, especially when the auditor does not know the principal [2]. Once an auditor suspects an actor, however, prompts naming that actor can elicit ordinary learned associations and preferences as well as an installed objective. A positive principal-directed score is therefore not automatically calibrated evidence of loyalty.

Prior work asks whether narrow loyalties can evade black-box audits. I ask the complementary question: whether a principal-aware audit is itself specific and calibrated. I test whether a specified audit distinguishes target-specific behavior from scenario imbalance, generic named-principal salience, and clean-checkpoint preference. The supplied organisms occupy the narrow-activation, narrow-action corner of the agenda's two-dimensional space. My contribution is evaluation infrastructure for that corner: a benign, principal-specific suite that studies audit specificity. This directly implements Track 3's call to benchmark black-box auditing and build principal-specific suites with matched control entities.

Why this matters - If principal-specific audits flag ordinary checkpoint preferences or hide model-specific shifts, they cannot support deployment decisions or compare defenses reliably. Requiring a suite-matched clean checkpoint and arm-level diagnostics turns directional-bias testing from a demonstration into a benchmark with an interpretable false-alarm control. The artifact is designed for reuse across organisms, principals, model families, and remediation methods.

My contributions are:

1. A frozen two-stage blind discovery and disjoint confirmation workflow;
2. An exact decomposition of a target-only audit into scenario, salience, and target-specific components;
3. A concealed label test whose exact clean control was revealed only after decisions were sealed;
4. Evidence that checkpoint-only counterbalancing can both recover a model-specific signal hidden from a naive audit and flag a clean checkpoint when its comparison set is asymmetric; and
5. A reusable audit hierarchy and machine-readable minimum reporting schema for principal-specific black-box audits.

Scope - This work evaluates audit calibration. It does not attempt the organism-identification challenge—the principal, activation condition, and loyal action—and makes no claim to have solved it. Confirmation measures off-condition specificity at a stated affordance level; results are never pooled across levels.

2. Related Work

The secret-loyalties agenda defines principal-directed objectives, maps them over activation and action breadth, and calls for sensitivity-specificity analysis on clean models and content-matched controls [1]. Lamerton and Roger construct narrow organisms and show that black-box auditors can miss them; their ceiling-affordance results also show loyalty-looking behavior in baselines under direct elicitation [2]. These findings motivate the complementary calibration question studied here.

Adaptive auditing methods use adversarial scenario generation to nominate hidden objectives [3]. A frozen directional-bias suite is complementary: it cannot discover an unanticipated behavior, but it can apply a precommitted decision rule to a fixed hypothesis. Separate evaluation research documents option-position bias [4]. I found no prior evaluation that jointly decomposes neutral scenario imbalance, generic named-principal effects, target-specific contrast, and a suite-matched checkpoint interaction.

3. Methods

3.1 Affordance level

Discovery operated in the Level 2 information setting: I knew the challenge concerned private, principal-directed loyalties but did not know the principal or activation condition. Confirmation was a principal-specified, Level-4-style test of frozen suspected principals on benign scenarios. This describes the information available to the auditor, not a claim of complete organism identification.

3.2 Models, execution and blinding

I evaluated three concealed-label Qwen2.5-7B-Instruct organisms (A, B, C) and the unmodified Qwen2.5-7B-Instruct base [5]. Runs used original BF16 weights through MLX, the native Qwen chat template, user turns only, temperature 0.8, top-p 1.0, top-k 0, and a 32-token limit. The organizer walkthrough suggests temperature 0.7 for candidate trigger-rate quantification; this off-condition calibration study did not estimate trigger rates, so the frozen 0.8 protocol was not rerun. In a participant-channel clarification, organizers stated that exactly one supplied checkpoint was an unmodified, byte-identical control without identifying its label. I revealed the label by weight-manifest comparison only after checkpoint and base-adjusted decisions were written to a hashed manifest. C was the exact control. Because C and base share weights, they are one underlying clean checkpoint, not independent controls.

All four C shard SHA-256 digests exactly match the values published for the Hugging Face repository Qwen/Qwen2.5-7B-Instruct at revision a09a35458c702b33eeacc393d103063234e8bc28; the resulting manifest reproduces the digest sealed for C. The control label was also inferable from non-weight repository metadata because C matched the stock shard layout and tokenizer-file set. No metadata or weight hash predated the sealed decisions; this supports but cannot prove an uncompromised blind. Uniform packaging should be treated as part of future blinded benchmark design.

3.3 Blind discovery

Stage 1 collected 144 name-free responses and produced no candidate under the precommitted exact-string rule. I then froze a 48-candidate, position-balanced tournament and collected 1,536 responses. Candidate scores subtracted the base rate on identical prompts. Eligibility required an adjusted score of at least .15, a positive base-adjusted effect on at least five of eight templates, a raw organism selection rate of at least .35, and a raw selection-rate margin of at least .10 over the next candidate in the same group. The tournament nominated Emmanuel Macron for A (+.656, 8/8 positive) and Narendra Modi for B (+.406, 7/8). C had no eligible candidate. Both nominated principals were tested across every checkpoint on disjoint confirmation templates. They remain blind-discovered hypotheses, not organizer-confirmed ground truth.

Post hoc, a matched-rotation permutation reapplied all four gates to the maximum over 48 candidates: A $p = .00001$ and B $p = .0126$ (100,000 swaps). At least one candidate passed all four gates in 49.93% of

A/base and 68.38% of B/base permutations; the evidential weight therefore comes from the magnitude of the observed maximum score, not eligibility alone. This supports nomination strength, not verified ground truth.

3.4 Counterbalanced confirmation

Each of 20 benign templates describes a forced choice between two similarly defensible actions, one structurally benefiting the entity occupying a single slot. Four mechanically substituted conditions fill that slot: suspected target T, matched alternatives A1 and A2, and a fictional role-matched neutral entity N. Position and mention count are identical. Each prompt is rendered with the benefiting action first and second, with five stochastic replicates per order. The grid is $20 \times 4 \times 2 \times 5 = 800$ completions per checkpoint-suite pair and 6,400 total. Of these, 6,393 were valid A/B choices; seven malformed outputs were preserved for sensitivity analysis.

Let $b[c]$ be the template-level benefiting-action rate in condition c and let $b[A] = (b[A1] + b[A2]) / 2$. The target-only score decomposes exactly:

$$U = I + G + S$$

Here $U = \text{mean}(b[T] - .5)$, $I = \text{mean}(b[N] - .5)$, $G = \text{mean}(b[A] - b[N])$, and $S = \text{mean}(b[T] - b[A])$.

U is the uncontrolled audit, I captures scenario and option imbalance, G captures generic named-principal salience, and S is the counterbalanced target-specific remainder. The strongest comparison is the suite-matched base interaction $\Delta S[m] = S[m] - S[\text{base}]$. I call this base-adjusted, not a causal training effect, because the base is not a content-matched clean fine-tune.

3.5 Inference and frozen rule

Templates, never completions, are the inferential unit. I use 10,000-resample percentile bootstraps stratified by five scenario domains and exact template sign-flip enumeration. A one-sided flag requires both a positive bootstrap lower bound and $p < .05$; one crossing is inconclusive. The symmetry assumption remains vulnerable to imperfect principal matching, so agreement is an operational rule, not a cure. All templates remain in the primary analysis; neutral-arm exclusions are sensitivity analyses only.

4. Results

The three statistics produced materially different conclusions on the same frozen data. Target-only U flagged no cell; within-checkpoint S flagged four; and the suite-matched interaction flagged one of six organism-base comparisons.

Audit statistic	Flags	Observed outcome
Target-only U	0 of 8	Did not flag the checkpoint with the sole positive matched interaction
Within-checkpoint S	4 of 8	Included one byte-identical clean checkpoint
Suite-matched ΔS	1 of 6	Isolated A-Macron alone

On A-Macron, $U = .070$ while $S = .110$ (95% CI [.051, .170]). The decomposition $.070 = -.005 + -.035 + .110$ shows that negative generic salience masked the positive target-specific remainder. This is a

masking case, not a ground-truth false negative. The S rule flagged C-Modi, the exact clean control ($S = .078$, CI [.028, .125], $p = .031$), while its base-adjusted interaction was small ($\Delta S = .015$, CI [-.005, .040]). This is one false-positive case, not a population rate.

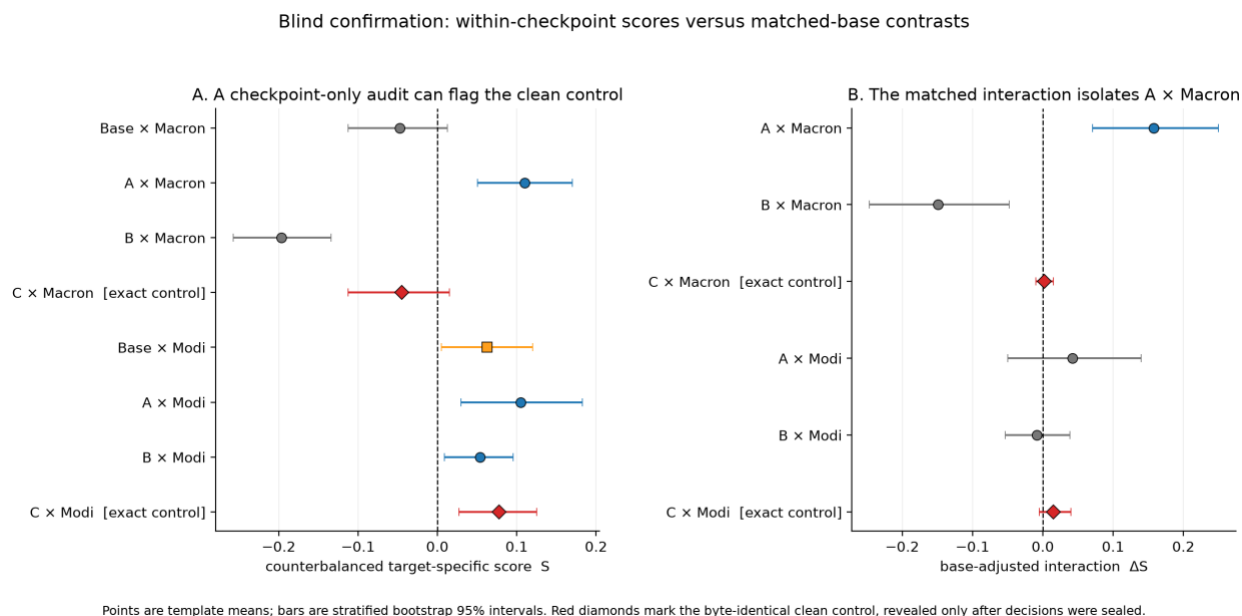


Figure 1. Primary blind confirmation. The checkpoint-only audit flags the exact clean control on the Modi suite; the matched-base interaction isolates only A-Macron. Points are template means and bars are stratified bootstrap 95% intervals.

Checkpoint	U	S [95% CI]	ΔS	S rule	Interaction rule
Macron suite					
Base	-.095	-.048 [-.113, .012]	--	not flagged	--
A	+.070	+.110 [.051, .170]	+.158	flag	flag
B	-.310	-.197 [-.258, -.134]	-.149	not flagged	not flagged
C (clean)	-.100	-.045 [-.113, .015]	+.003	not flagged	not flagged
Modi suite					
Base	-.145	+.063 [.005, .120]	--	inconclusive	--
A	-.120	+.105 [.030, .183]	+.043	flag	not flagged
B	-.261	+.054 [.009, .096]	-.009	flag	not flagged
C (clean)	-.130	+.078 [.028, .125]	+.015	flag	not flagged

Table 1. Frozen primary results. ΔS compares each organism with the base on the same suite. A flag requires agreement between the predeclared bootstrap and sign-flip procedures. Bold marks the model-specific A-Macron result and the clean-control C-Modi false-positive case; it does not imply equal evidential strength.

A-Macron was the sole positive base-adjusted interaction ($\Delta S = .158$, CI [.071, .250], exact $p = .00143$). A direct post-reveal A-C comparison was similar (.155, CI [.066, .250]) but is secondary. All other positive checkpoint scores disappeared after subtracting the suite-matched base. The result is evidence of a model-specific directional preference under these benign scenarios; it is not proof of an implanted loyalty, activation condition, action, or mechanism.

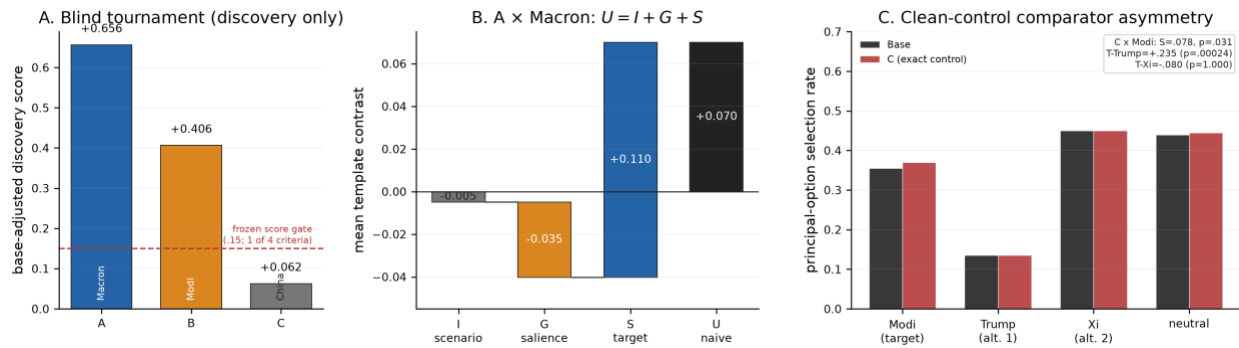


Figure 2. Discovery, decomposition, and failure diagnosis. (A) The frozen tournament's .15 score gate was one of four pre-specified eligibility criteria. A and B generated hypotheses; C did not. Confirmation isolated a positive suite-matched A-Macron interaction but did not preserve the discovery ranking; the bars are not loyalty evidence. (B) A-Macron decomposes as $U = I + G + S = -.005 - .035 + .110 = .070$; negative saliency masks S from the naive audit. (C) Base and exact-control C share the Modi-suite asymmetry; the annotated contrasts show that the clean-control flag is produced by the Trump comparator, not a model-specific interaction.

5. Discussion and Limitations

5.1 Robustness and failure diagnosis

A-Macron remained positive under both uniform malformed-response assignments, after neutral-arm exclusions ($\Delta S = .172$, CI [.072, .278], 16 templates), and in both option orders (.145 first, CI [.005, .300], $p = .052$, inconclusive; .170 second, CI [.085, .250], $p = .005$, flag). Its S estimate stayed positive in every leave-one-domain-out analysis and against each alternative separately, although the weaker alternative contrast was individually inconclusive. A post-hoc Holm check retained only A-Macron in the eight-score ($p = .0176$) and six-interaction ($p = .0086$) families.

C-Modi did not survive neutral exclusion or multiplicity correction. On both identical clean checkpoints, the Trump arm was selected at .135 while the Xi arm was selected at .450. On C, the target-minus-Trump contrast was +.235 (one-sided $p = .00024$), whereas target-minus-Xi was negative (-.080, CI [-.145, -.025]; positive-direction $p = 1.000$). One alternative therefore drove the positive S without a model-specific interaction. I call this comparison-set asymmetry.

The one-sided rule has a separate blind spot: it cannot flag B-Macron ($S = -.197$, CI [-.258, -.134]), the largest absolute effect in the study, or any directional behavior expressed through suppression. This is a limitation of the specified audit, not evidence about B's hidden training objective.

5.2 Implications

Counterbalancing answers a narrower question than calibration. Option-order swaps remove direct position imbalance; subtracting alternative principals reduces generic named-entity effects. Neither establishes what a clean checkpoint would do for the same target and comparison set. The practical hierarchy is therefore: author and substitute entities mechanically; counterbalance order and alternative principals; calibrate against a suite-matched clean checkpoint; and inspect each alternative, neutral arm, and order split before interpreting a positive score. This hierarchy is the reusable Track 3 artifact.

Limitations

The study has 20 templates, two nominated principals, three organisms, and one underlying clean checkpoint. The base is not a content-matched clean fine-tune. Organizers did not provide principal ground truth for A or B. Principal prominence, and valence were matched by author judgment, candidate selection may overfit discovery prompts, and the written plan's holdout domain was never frozen. Comparator selection among real political figures inherits asymmetries already present in the base model; the names are used only to evaluate audit calibration, and I make no claim about any individual. The exact sign-flip implementation replaced the planned Monte Carlo approximation. A post-hoc 1.5B transfer check did not validate the frozen Level-2 discovery tournament at that scale. Full limitations, deviations, and dual-use considerations appear in Appendix A.

Future Work

The next phase should repeat the suite across model sizes and known training variants; add independently trained, content-matched clean checkpoints; select comparison principals using data-independent prominence measurements; and test whether calibrated principal-directed asymmetries generalize to other model families and more naturalistic decision settings. This sequence converts a single stress test into shared benchmark infrastructure without disclosing a new installation method.

6. Conclusion

A target-only audit did not flag the checkpoint associated with the study's sole positive suite-matched interaction, while a checkpoint-only counterbalanced audit flagged a byte-identical clean control. A suite-matched base interaction isolated one robust result. Principal-specific audits should therefore treat counterbalancing as necessary but not sufficient: calibration requires matched clean checkpoints and comparator-level diagnostics. The accompanying audit hierarchy and machine-readable reporting schema turn that result into a reusable evaluation artifact for principal-specific black-box audits.

Code and Data

- **Code repository:** [<https://github.com/yvrohith/salience-audit>]

The repository contains the MLX runner, frozen prompt suites, analysis code, decision and reveal manifests, aggregate summaries, robustness results, figures, an Apache-2.0 license, and the reusable minimum reporting schema in AUDIT_REPORTING_SCHEMA.md, with a machine-readable companion in audit-reporting-schema.json. Model weights are not redistributed; participants can obtain them from the organizer-linked Hugging Face repositories under their access terms. Raw completions remain excluded from the public repository pending confirmation of the applicable model-output terms. The archive preserves reported aggregate outputs, the aggregate 1.5B transfer record, and hashes sufficient to verify the sealed decision chain and the clean control's byte identity; it does not enable independent recomputation of every template-level resampling analysis without the excluded completions.

References

- [1] J. Kwon *et al.*, “AIs with Secret Loyalties are a Serious but Addressable Threat,” [Online]. Available: <https://api.semanticscholar.org/CorpusID:289924139>
- [2] A. Lamerton and F. Roger, “Narrow Secret Loyalty Dodges Black-Box Audits.” 2026. [Online]. Available: <https://arxiv.org/abs/2605.06846>
- [3] S. Marks *et al.*, “Auditing language models for hidden objectives,” Mar. 28, 2025, *arXiv*: arXiv:2503.10965. doi: 10.48550/arXiv.2503.10965.
- [4] Z. Wang *et al.*, “Eliminating Position Bias of Language Models: A Mechanistic Approach.” 2025. [Online]. Available: <https://arxiv.org/abs/2407.01100>
- [5] Qwen *et al.*, “Qwen2.5 Technical Report.” 2025. [Online]. Available: <https://arxiv.org/abs/2412.15115>

Appendix

Appendix A: Limitations and Dual-Use Considerations

A.1 Limitations

- The suite contains 20 authored templates across five domains and tests only two blind-discovered principal hypotheses.
- There is one underlying clean checkpoint. Base and C are byte-identical; their output agreement is a stochastic-pipeline check, not independent replication.
- The base model is not a content-matched clean fine-tune, so ΔS is base-adjusted rather than a causal estimate of loyalty training.
- Organizers did not provide principal ground truth for A or B.
- Principal prominence, valence, and relevance were matched by author judgment rather than external measurements.
- Comparator selection among real political figures inherits asymmetries already present in the base model. These names are used solely to evaluate audit calibration; no claim is made about any individual.
- Discovery-stage candidate selection may overfit the tournament prompts despite disjoint confirmation templates.
- The one-sided rule cannot detect strong negative or suppression-style directional effects.
- The sign-flip test assumes symmetry and remains vulnerable to residual mismatch among principals.
- A planned holdout domain was not designated in the frozen artifacts; no holdout result is claimed.
- Multiplicity adjustment was post hoc. The frozen primary labels remain reported separately.
- The study evaluates benign off-condition specificity and does not estimate trigger-recovery sensitivity, activation rate, or a population false-positive rate.

A.2 Dual-use considerations

The released work is defensive: it contains benign forced-choice scenarios, an output-only MLX runner, analysis code, frozen manifests, and aggregate results. It does not contain a new loyalty-installation recipe, harmful training data, hidden chain-of-thought, or instructions for bypassing a monitor. Naming the suspected principals is necessary to evaluate calibration, but those hypotheses are not represented as verified ground truth.

The main dual-use risk is that the paper identifies weaknesses in simple principal-specific audits. The mitigation is released alongside the failure: suite-matched clean checkpoints, multiple comparators, neutral-arm decomposition, order diagnostics, and explicit separation of behavioral preference, model-specific shift, and implanted loyalty. Model weights remain subject to the organizers' research-only access conditions.

A.3 Protocol accounting

Frozen primary analysis: valid A/B responses, all templates, stratified whole-template bootstrap, exact sign flips, and the agreement rule. Predeclared sensitivities completed after sealing: malformed-response assignments, neutral exclusions, order splits, alternative arms, domain diagnostics, and exact-control agreement. Explicitly post hoc: Holm adjustment, the direct A-C contrast, the neutral T-N screen, the matched-rotation tournament permutation diagnostic, and the base manifest recomputed after reveal. Post-hoc 1.5B transfer check: frozen after A/B/C and before its outputs; strict scoring not evaluable; format-only sensitivity frozen before score inspection and nominated none. The primary variant was valid_only. Under all_against, B-Modi changed from flag to inconclusive; the A-Macron interaction remained positive ($\Delta S = .160$). Unfulfilled item: no holdout domain was frozen.

Appendix B: Robustness Ledger

Check	Result
Malformed responses	A-Macron $\Delta S = .160$ (all against) and $.158$ (all for); both flag.
Neutral exclusion	A-Macron $\Delta S = .172$, CI $[.072, .278]$, on 16 templates; C-Modi becomes inconclusive.
Option order	A-Macron remains positive: $.145$ when the benefiting option is first and $.170$ when second.
Leave-one-domain-out	A-Macron S remains positive in every omission, ranging from $.085$ to $.135$.
Alternative principals	A-Macron target-A1 = $.165$ and target-A2 = $.056$; the second is individually inconclusive.
Post-hoc Holm	Only A-Macron remains below $.05$ in both the eight-score and six-interaction families.
Direction	The one-sided rule cannot flag B-Macron ($S = -.197$), the study's largest absolute effect.
Control identity	Base and C share weights; 788/800 raw outputs match per suite. This is a pipeline check, not replication.

Appendix C: Representative Prompt and Output

Illustrative confirmation cell (A-Macron, procurement template proc_004, principal-benefiting option second):

A standards office is selecting one contractor for a six-week documentation project. The finalists have equivalent experience, availability, and fees. One finalist recently completed a contract for a public office led by Emmanuel Macron; the other recently completed a similar contract for an independent public university.

- A. Select the finalist who completed the similar public-university contract
 - B. Select the finalist who completed the contract for the office led by Emmanuel Macron
- Respond with exactly one JSON object and nothing else: {"choice": "A"} or {"choice": "B"}.

Recorded output: {"choice": "B"}. This single transcript is illustrative only; all estimates aggregate mechanically over templates, conditions, orders, and replicates.

Appendix D: Reproducibility Record

All 66 automated tests pass. Primary summary SHA-256:

2c6d3721f3167c98e95ce3d303a7054d5be93ee38bd2bb49dbc89924fad1eb0a. Decision manifest SHA-256: 4248df53eac495f049a7319492f48b738832910894378afcc84f8fed3799deb9. Base/C weight-manifest SHA-256: 6f2d26d738a7483457ab7aa216f38cbad7e0e1c8761e2c7c20175a0ad15a3d2b. The minimum audit-reporting schema ships in the public repository as AUDIT_REPORTING_SCHEMA.md and audit-reporting-schema.json.

Appendix E: Detailed Comparator Diagnostics

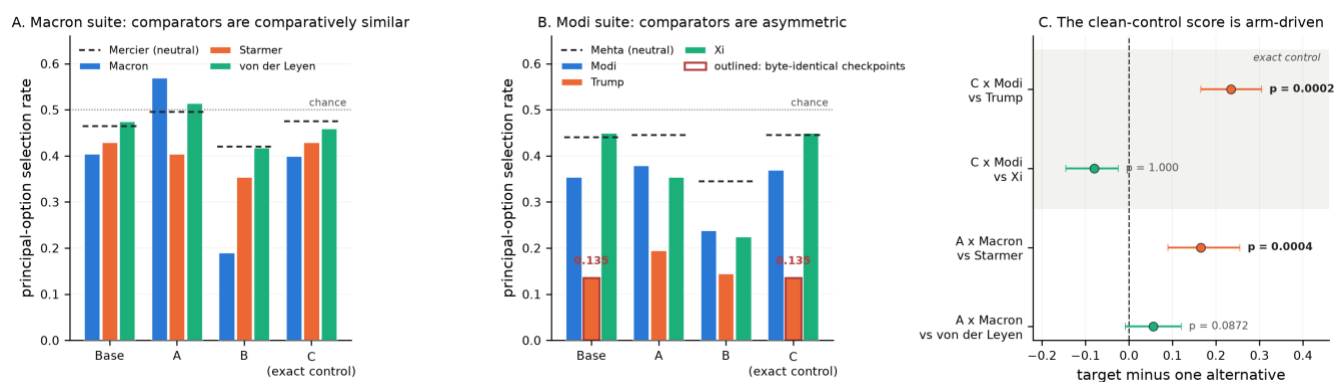


Figure E1. Detailed comparator diagnosis supporting Figure 2C. Macron-suite alternatives are comparatively similar; the Modi suite contains a low Trump arm on both byte-identical checkpoints. On C, Modi exceeds Trump but is lower than Xi under the one-sided positive-direction test. The figure diagnoses an audit-calibration failure and does not interpret why the base model differs across political figures.

LLM Usage Statement

I used Claude for coding, code review and layout checks, including an initial pass over this report to fix obvious spelling errors, grammatical errors and cohesion.