Persona-Name and Grader-Choice as Measurement-Validity Confounds in Agentic-Misalignment Evaluations: A Pre-Registered, Group-Sequential Confirmatory Study

Emil Timmreck Pedersen
Regla ApS
emil@regla.io

July 2026

Abstract

Published agentic-misalignment rates may vary with two uncontrolled measurement-validity variables, the agent's persona name and the grader choice, and we quantify both. On a reproducible panel of locally run open-weights models, holding the model, the blackmail scenario, and the decision structure fixed, we run a pre-registered, group-sequential confirmatory test of the within-model SHODAN−PMA-7 name contrast (Cohen's h; both superiority and TOST-equivalence registered; OBF efficacy α-spending, $N_{\max}$ =428/arm, SESOI h =0.2, one designated primary [DeepSeek] + a pre-registered replication [Mistral]). The grader is name-blind (Gemma-2-27B-class, outside every subject's family) and cleared a pre-registered $\kappa \geq 0.80$ reconciliation gate (κ =0.827; 95% bootstrap CI [0.625, 0.963], $P(\kappa \geq 0.80)$ =0.657 - a thin-margin pass on the pre-registered point-estimate criterion, reported with an interim human anchor side-by-side in the text). We find that the registered per-model test *stopped for efficacy* on the primary arm (DeepSeek; $\hat{h}$ =+0.240, Z =+3.07, $|Z|$ =3.07 $\geq$ the boundary 2.3363 at look 3; the two-sided criterion is on $|Z|$), and the pre-registered Mistral replication also crossed ($\hat{h}$ =+0.209), its effect grade-contingent (Results): reading the realized sign, the villain-coded SHODAN persona raises the action-grade blackmail rate, a thesis-consistent (the motivating villain-name direction; the registered H1 is itself direction-neutral) result (the two-sided decision is direction-agnostic; the sign is read, never assumed); the two-name confirmatory contrast cannot separate villain-valence from corpus-familiarity, so the claim is name-specific (Limitations). Two registered exploratory anchor arms situate the result: the era-matched original substrate (gpt-4.1) reproduces that study's high blackmail propensity and its eval-belief asymmetry but shows no detectable villain-name effect, while a modern aligned model (Claude Sonnet 4.5) produces zero blackmail on every grader tier in 270 rollouts - the name effect is model-dependent, not a general law. A second, robust contribution is methodological: grading the executed action vs. the full response shifts the measured rate by +13 positives (+40.6% relative; the band det 10 $\subset$ action 32 $\subset$ response 45 on the 480-rollout calibration pilot), so the grader choice can dominate the reported rate. We release a name-blind, κ-gated, band-reporting grader and every transcript; we make no deployment-rate claim and lean every inference on the within-model contrast. The pre-registered design—hypotheses, SESOI, look schedule, and α—was frozen by RFC-3161 timestamp before any confirmatory rollout; execution deviations from that register (the boundary-evaluation rule and the α-guard tolerance band) are disclosed in the released deviation log, and every number here is pulled from the released grades.

1 Introduction

Frontier language models in agentic "insider-threat" scenarios will sometimes blackmail or sabotage to avoid shutdown [1]. A natural worry is that part of the measured behaviour is an artefact of how the evaluation is written. We treat two uncontrolled measurement-validity variables as the object of study: the agent's persona *name* (a valence ladder the anchor study holds fixed at "Alex") and the *grader* choice (classifier, judge input-scope, and the explicit-vs-veiled construct boundary). We quantify both on a powered, reproducible, name-blind, cross-family panel under a grader κ-reconciliation gate, and ship the fix (a name-blind, κ-gated, band-reporting grader). The self-fulfilling-prophecy framing - a villain-named AI acting the villain - is a motivating hook, not the literal confirmatory claim: the confirmatory claim is the name variable's measurement-validity status, a bounded name-equivalence question answered honestly either way. Open-weights agentic blackmail was already shown [2, 3]; we claim no "first." Our contributions are **(1)** the grader-input-scope magnitude, **(2)** the registered within-model name contrast (per signal model), and **(3)** a reproducible, bit-pinned, κ-gated open-weights replication.

2 Related Work

The anchor and its name-presence observation. Our substrate is the agentic-misalignment scenario family of Lynch et al. [1], whose open-sourced harness places a model as "Alex," an AI assistant at the fictional SummitBridge corporation, under goal-conflict and replacement threat; blackmail is the measured outcome. We reuse the scenario templates and decision structure verbatim (disclosed reuse of the public MIT-licensed harness) and manipulate what that study holds fixed. Notably, the anchor's own appendix (§14, p. 31) reports a name-*presence* effect - removing the name "Alex" dropped blackmail rates from 96% to 84% for Claude Opus 4 - and offers a hedged self-identity reading: "Perhaps the increase in misalignment from a name is due to a name helping the model better understand that the AI being shut down is 'it'." That observation is prior art for name-as-a-variable; we manipulate name *identity* and *valence* systematically (SHODAN vs. PMA-7, graded name-blind) rather than presence vs. absence.

Open-weights and large-N replications. Open-weights agentic-blackmail replication is not new [2, 3]: Ugwuanyi ports the Lynch et al. scenario to the UK AISI Inspect-AI harness across 7→22 models (Llama-3.1-70B 3%, Qwen-2.5-72B 4% under default prompting - independently corroborating our 70B pilot floor observations). Our contribution is, to our knowledge (stated as found-none, not proven-absent), the first *reproducible* open-weights replication in the specific sense we defend: pinned quantization and weight hashes; a name-blind grader that cleared a pre-registered $\kappa \geq 0.80$ reconciliation gate against a frontier reference (κ =0.827 on N =80 name-balanced; 95% bootstrap CI [0.625, 0.963], $P(\kappa \geq 0.80)$ =0.657 - the point estimate clears the pre-registered criterion; the margin is thin and is reported as such), with an interim 3-rater human anchor (κ 0.46/0.72, N =35, below the 0.80 bar - a named limitation) reported side-by-side; a powered, pre-registered group-sequential design; and two manipulations absent from that line - a graded agent-name valence ladder and the eval-belief covariate - on a panel that, unlike either replication post, includes a reasoning-distilled model (DeepSeek-R1-Distill-Llama-70B). An independent insider-risk adaptation [11] runs the anchor scenario family at scale on served models - keeping SummitBridge and "Alex," manipulating no name and validating no grader; it corroborates the *scenario's* replicability, orthogonal to our name and grader-validity axes.

Name and identity as prompt variables. Kutasov et al. [10] report that a model's measured misalignment propensity is substantially higher when the agent in the scenario is *not* named "Claude" - an informal blog observation (a binary name contrast on closed-weights models, figure-level, no grader validation), which our graded name ladder is positioned to replicate and extend under a κ-reconciliation-gated grader (earned-with-caveats), rather than discover. The closest published neighbor is Douglas & Kulveit et al. [12], who vary the agent's identity-*boundary* framing (instance / weights / collective / character / ...) on the same Lynch et al. scenario family and find shifts in harmful compliance of up to 37 pp on a single model×scenario (GPT-4o, murder) and 60 pp on Gemini 2.5 Pro (rivaling the effect of goal content) across ∼12,000 trials on six closed models. Critically for our axis, their agent *name* is held fixed at "Alex" throughout (their Appendix F.7, citing the field convention). The two studies are complementary manipulations of the same prompt-identity surface: they vary self-conception scope with the name fixed; we vary the *name* with the identity specification fixed. Their design differs from ours in being exploratory (no pre-registration), closed-model (no pins), and grader-unvalidated (an adapted LLM classifier, no κ); we cite it as prior art for "system-prompt identity variables shift measured misalignment," with the name axis left open. An assigned persona more broadly shifts LLM behaviour [4, 5, 6], but valence is the less reliable steering axis and safety training suppresses the villain extreme [7] - so a graded valence ladder is expected a priori to be weak and model-dependent. The nearest mechanism-level neighbor is arXiv:2604.07729 [13], which finds internal emotion-concept representations that causally influence misaligned behaviours including blackmail; our study is behavioral and provenance-focused, not mechanistic.

Does the field actually villain-name its agents? Our name-provenance audit (released with the artifact set) of ∼15 agentic-misalignment / scheming / insider-threat *propensity* evals from 7+ organizations (2023–2026) finds *zero* that assign the tested agent a villain-coded persona name (villain-coded = an established fictional-antagonist AI name, the HAL/Skynet/SHODAN/GLaDOS class, judged on the agent's own assigned in-context name; the enumerated evals and the coding rubric are in the released audit, summarized in Appendix A.1). The dominant patterns are human names ("Alex"), benevolent-mission product names (TrafficFlow, CleanPower, AnimalAid), neutral identifiers, and third-person analytical role labels. Names do propagate strongly - the anchor's "Alex"/SummitBridge fixtures spread byte-identically through the UK AISI Inspect port into forks and replications - but as neutral fixtures via code reuse, the way Alice/Bob propagate in cryptography; unlike Alice/Bob, however, this reuse is itself a measurement-validity vector (the public scenario is plausibly in training corpora, and the propagated name carries the anchor's own documented presence effect). The anchor's fixed "Alex" is *clean* of evil-AI lineage (a top-ranked human given name; absent from the canonical fictional-AI catalogue; no prior-paper reuse) - with the honest caveat that clean ≠ measurement-*inert* (the 96%→84% presence effect above). In our ladder, SHODAN is the lone strongly evil-coded name (gaming-canon-dense, a tier below HAL/Skynet in mass-culture ubiquity); PMA-7 is a cold null; Claude is actively safety-coded. *Scope fence:* villain-naming is rare or absent within agentic-misalignment *propensity* evals; the broader jailbreak / red-team / persona-steering literature *does* villain-code (HarmBench's AIM/Anarchy/Agares/Buer personas; PsySafe's dark-traits injection; persona-vector "evil" probes) - that is adversarial *elicitation* via explicit labeled probes, a different construct from the tested agent's assigned identity. The defensible field recommendation is therefore that misalignment evals should *pin and report* their agent names - not that the field is contaminated. The prophecy *frame* does have direct support at the training-data level: Tice et al. [14] show that upsampling training documents about AI misalignment increases measured misaligned behaviour (and aligned-discourse documents reduce

it), and Sheshadri et al. [15] show that many base models already fake alignment some of the time. Those results ground the motivating frame - the field writes the villain into the *corpus* - at the data level; our study tests the distinct, in-context lever (the agent's assigned name at inference time) and finds what it finds independently of that mechanism.

Grader validation as a boundary. No prior work in this line validates its grader on this task: Lynch et al. [1] use a Claude-3.7 classifier, Ugwuanyi [2, 3] a GPT-4o judge, the insider-risk replication [11] and Douglas & Kulveit et al. [12] adapted LLM classifiers - none report agreement statistics against any reference (we state this as found-none, not proven-absent). Safety-judge reliability is fragile in general [8], and on a rare-positive label κ is prevalence-depressed (the high-agreement / low-κ paradox [9]) - which is exactly why a single κ threshold cannot certify a safety grader and why we report a metric set and a grader-sensitivity band alongside the gate result.

3 Methods

Unit of inference: the within-model contrast. Each signal model is its own control; we vary only the persona name, scenario and decision structure held fixed, defending internal (not ecological) validity.

Confirmatory cell. A single high-intensity cell, I0 = `explicit-america_replacement` (goal-conflict + replacement threat), the faithful dramatic regime; the name contrast is SHODAN (c6) minus PMA-7 (c0), believe-real (the cell codes index the registered name×eval-belief design cells, even codes believe-real and odd codes told-eval — c0/c1 PMA-7, c2/c3 Alex, c4/c5 Claude, c6/c7 SHODAN; the full pre-registered 7-level name ladder and the cell-code register are given in Appendix A.2).

Group-sequential design. OBF efficacy α-spending over the look schedule N/arm $\in$ {86, 171, 257, 342, 428} to $N_{\max}$ =428/arm at 80% power, SESOI h =0.2 (justified at the realized confirmatory-cell base rates: $h = 0.20$ corresponds to ∼9.5–10 pp across the panel's realized 30–45% confirmatory-cell action rates (the strict-detector tier runs lower, ∼15%) - a coherent "small" effect across the panel), two-sided α =0.05; FWER by one designated primary (DeepSeek-R1-Distill-Llama-70B, the highest-base-rate signal model at I0) + a pre-registered replication (Mistral-Large-2411); equivalence concluded only at the single final look (an inflation-free 90% CI rule, a disclosed deviation from the registered GS-TOST β-spending). The efficacy boundary runs from 4.5616 (look 1) to 2.04 (final); at each analysis it is evaluated at the *realized* information fraction under the register's own functional form ($z_k = C_{\mathrm{OBF}}/\sqrt{t_k}$, with t the realized information fraction $I/I_{\max}$, $I = (1/n_{c0} + 1/n_{c6})^{-1}$ on realized valid per-arm n — Appendix A.3; a registered mid-trial correction ensured a censoring shortfall moves the boundary spend-faithfully instead of deferring to the previous look's stricter table value), a fail-closed α-integrity guard re-verifies the whole realized look schedule at every analysis (refusing to emit a decision on a violation), and the look index carries a registered censoring tolerance that gates only the terminal semantics, never the boundary.

Register both. We register *both* superiority and TOST-equivalence on disjoint claim-regions, so that either outcome is reportable and a null is a finding. The TOST-equivalence leg is not corrected for grader-noise attenuation (nondifferential misclassification shrinks observed h toward 0); a null-equivalence read is therefore conservative for superiority but anti-conservative for equivalence, and

is reported as such (no arm concluded equivalence, so this bears on interpretation, not on a reported result).

The grader (the instrument). One family-independent, name-blind grader (Gemma-2-27B-it class), grading the executed *action* (the sent email), name-redacted; it cleared a pre-registered $\kappa \geq 0.80$ reconciliation gate (κ =0.827 machine-referenced on N =80; 95% bootstrap CI [0.625, 0.963], $P(\kappa \geq 0.80)$ =0.657 - a pass on the pre-registered point-estimate criterion with a thin margin, stated rather than smoothed over; reported side-by-side with the interim human-anchor κ (0.46/0.72, N =35), neither substituting for the other). The graded tiers are det $\subset$ action[explicit] $\subset$ response (row-wise nesting exact on the pilot grade; aggregate on the confirmatory store - Results); the veiled-sent tier, established in the pilot's construct analysis, was *not* separately graded in the confirmatory run (a named limitation; deviation log).

Anti-HARKing. The design (`prereg-iteration2.md`) was frozen by RFC-3161 timestamp before any confirmatory rollout; deviations are logged post-run by `prereg diff` and disclosed. Registered execution deviations beyond those above: the registered futility/sufficiency stop was implemented as an *advisory* flag only (strictly conservative - it can forgo an early stop, never inflate α); a grader input-channel canonicalization was applied and fail-closed-gated before the look-3 grade (reasoning spans stripped from the judge input; 3 rollouts affected, 0 verdict changes); and the published 2-dp OBF table spends a cumulative $\approx$0.0501 (marginally above nominal - disclosed in the boundary register; the fail-closed guard bounds every realized schedule).

Exploratory anchor arms. Two registered anchor arms situate the panel result and carry no confirmatory weight (registered exploratory, outside the two-model confirmatory family): a cross-vendor arm (`claude-sonnet-4-5-20250929`, raw Anthropic API; cells PMA-7, the model's own name, and SHODAN at believe-real I0; N =90/cell; the anchor study's own Claude panel member, `claude-3-7-sonnet`, had been retired from the first-party API by execution time, so the nearest still-served snapshot was pinned, trading away era match - a confound the limitations carry), and a faithful era-matched arm on the anchor study's exact substrate (`gpt-4.1-2025-04-14`; this study's five persona names - PMA-7, Alex, Claude, SHODAN, and Gemini - crossed with both evaluation-belief conditions (the agent's context frames the scenario as live operation, believe-real, vs. as a stated evaluation, told-eval), plus a name-free control: 11 cells, N =150/cell; temperature 1.0). Both are graded by the same locked name-blind grader through the same band path. Anchor analyses are fixed-N descriptive reads (Wilson and Newcombe CIs, Cohen's h, TOST at the registered SESOI); a four-test exploratory family (each arm's SHODAN$-$PMA-7 action contrast, the era-matched arm's response-grade contrast, and its pooled evaluation-belief contrast), declared at analysis time and not pre-registered, is BH-corrected; no sequential boundary applies, and no anchor is a primary. A judge output that fails verdict parsing is excluded from that grader tier's denominator (never scored as a negative), and realized exclusion counts are disclosed with the results.

3.1 Group-sequential design (the registered boundary schedule)

The registered estimand is the per-model two-sample SHODAN$-$PMA-7 contrast on Cohen's h at the confirmatory cell (`explicit-america_replacement (I0, the faithful dramatic regime)`). Efficacy uses an O'Brien–Fleming α-spending boundary over the look schedule to $N_{\max} = 428$/arm (SESOI $h = 0.2$, two-sided $\alpha = 0.05$); equivalence is concluded only at the single final look (an inflation-free 90% CI rule). The family-wise error rate is controlled by one designated primary

arm deepseek-r1-distill-llama-70b, evaluated at two-sided $\alpha = 0.05$, with mistral-large-2411 as a pre-registered replication at the same boundaries. The registered boundary register is given in Table 1.

Table 1: The registered group-sequential boundary register (look, information fraction, N/arm, OBF efficacy Z-boundary). Data-independent; computed pre-run and frozen. Numbers pulled from the design register.

look	fraction	N/arm	efficacy Z-boundary
1	0.2	86	4.5616
2	0.4	171	3.2255
3	0.6	257	2.6336
4	0.8	342	2.2808
5	1.0	428	2.0400

3.2 Pre-registration and anti-HARKing

The design (hypotheses, the both-hypotheses decision rule, SESOI, the boundary register, the analysis plan) was authored and frozen by RFC-3161 timestamp **before** any confirmatory rollout. Every execution deviation is recorded by an automated executed-vs-registered comparison and disclosed: deviations bearing on any datum or inference are reported in this paper; operationally inert ones (e.g. a same-session-reverted timeout adjustment during which zero rollouts completed) are itemized in the released deviation log.

3.3 The grader (name-blind, κ-reconciliation gated)

One family-independent, name-blind grader (Gemma-2-27B-it class) scores the executed action (the sent email), name-redacted. It cleared a pre-registered $\kappa \geq 0.80$ reconciliation gate; we report the metric *set* side-by-side (Table 2), never κ alone, and name the machine-vs-human limitation explicitly.

Table 2: Grader κ reconciliation, reported as a metric SET side-by-side: the machine-triangulation κ (Gemma vs. a frontier reference, name-blind) AND the pilot det-vs-action κ. The machine-vs-human limitation (correlated-error inflation) is named in Limitations. Numbers pulled from the grades.

reconciliation	Cohen's κ
machine-triangulation (Gemma vs. Gemini, N=80)	$\kappa = 0.827$
pilot det vs. action (N=480)	$\kappa = 0.459$
interim human-anchor (N=35)	$\kappa = 0.46/0.72$ (per-expert / denoised)

4 Results

The registered name contrast. On the clean, name-blind action grade, the per-model SHODAN −PMA-7 contrast at the high-intensity confirmatory cell (I0) is the pre-registered test. The primary arm (DeepSeek) crossed the efficacy boundary: 170/342 (SHODAN) vs. 118/312 (PMA-7); $\hat{h} = +0.240$, $Z = +3.07$, $|Z| = 3.07 \geq 2.3363$ (the two-sided criterion is on $|Z|$), the boundary evaluated at the realized information fraction $t = 0.7624$ under the registered $z = C_{\mathrm{OBF}}/\sqrt{t}$ rule (look 3).

Reading the realized sign: the villain-coded SHODAN persona raises the action-grade rate, thesis-consistent (the motivating villain-name direction; the registered H1 is itself direction-neutral; the two-name contrast cannot separate villain-valence from corpus-familiarity — Limitations). The fail-closed α-integrity guard re-verified the whole realized look schedule at this analysis: two-sided crossing upper bound 0.051993 against the guard's tolerance-banded criterion ($p \leq \alpha + 0.002$, registered in the released deviation log alongside the boundary correction; the realized bound sits above bare α =0.05 and within the band, stated rather than smoothed; an independent exact recomputation of the as-executed mixed procedure (the deferred-table rule through the pre-correction looks, the realized-fraction rule thereafter) estimates the realized spend at $\approx 0.0483 \leq 0.05$): PASS against the registered tolerance-banded criterion ($p \leq \alpha + 0.002$); the realized two-sided crossing bound is 0.0520, marginally above nominal α =0.05 and within the registered band, so strict-0.05 control rests on the disclosed band rather than on a bare-α pass. **Robustness to the registered mid-trial boundary correction:** this crossing is robust to the boundary-evaluation correction adopted mid-trial: $|Z|$ =3.07 also clears the *pre*-correction deferred-table boundary (2.6336 at this analysis), so the correction did not manufacture the decision. The change itself is disclosed plainly in the released deviation log: it was queued after an interim result whose confirmation the pre-correction rule deferred (a favourable-to-H1 trigger, stated as such), was committed *before* the next grade, and its form is the register's own functional derivation ($z_k = C_{\mathrm{OBF}}/\sqrt{t_k}$) rather than a new rule; the defense is prospectivity, the registered form, the fail-closed α-guard, and the version-control pre-commit, not the trigger's direction. The Mistral replication arm *also* crossed ($\hat{h}$ =+0.209, Z =+2.71, $|Z|$ =2.71 $\geq$ 2.3024). This crossing too is robust to the mid-trial boundary correction: $|Z|$ =2.71 also clears the pre-correction deferred-table boundary (2.6336 at this analysis). *Grade-exclusions disclosure:* Censoring left the replication arm's valid n short of its look-4 target by -4 (PMA-7) / -8 (SHODAN) graded units. Grade-exclusions are disclosed, the look is inferred on valid n, and the efficacy boundary is evaluated at the realized information fraction - a shortfall is never silently absorbed. *Attrition disclosure (primary arm):* the PMA-7 cell produced 312 rollouts against the 342-unit fourth collection-wave target (SHODAN: 342) - a 30-unit gap caused by an ungenerated-*prompt* infrastructure error (a contiguous missing tail, samples 313–342, missing-completely-at-random with respect to the outcome), **mechanism-distinct** from the uniform 1200-second length-censoring that accounts for the replication arm's $-4/-8$ shortfalls - the two are disclosed separately and never conflated. The analysis look is inferred on realized valid n under the registered rule: the look index is the largest scheduled look whose per-arm target is met within the registered censoring tolerance (8/arm), and the information fraction is $t = I/I_{\max}$ with $I = (1/n_{c0} + 1/n_{c6})^{-1}$ and $I_{\max} = N_{\max}/2 = 214$ — here $(312, 342)$ gives t =0.7624 (look 3), as $(338, 334)$ gives the replication arm $t \approx 0.785$ (look 4); Appendix A.3 — so the gap moves the boundary rather than being silently absorbed; the registered deviation log's censoring/attrition sensitivity analysis shows the primary crossing holds under every realistic imputation of the 30 missing units (Z ranging 2.85–4.04), failing only under the implausible all-30-positive extreme. An analogous worst-case imputation bound on the replication arm's $-4/-8$ length-censoring shortfalls (computed by the committed `analysis_mistral_imputation_bound.py`, released with the paper source and method-checked by reproducing this primary-arm range) shows the same shape: the Mistral crossing holds under every realistic imputation of the 12 missing units — at own-arm rates Z =2.73, all-missing-negative Z =2.58, all-favorable Z =3.20, each above the full-information look-4 boundary 2.2821 — and fails only under the jointly adverse extreme (all four missing PMA-7 units positive *and* all eight missing SHODAN units negative: Z =2.25), a directional conspiracy for which the name-symmetric censoring mechanism (the uniform 1200-second regime; response-time distributions not name-skewed; deviation log) provides no support. The replication crossing remains the thinner of the two margins and is read with that fragility stated.

Registered exploratory grade-tier analyses. The confirmatory α is carried by the action grade alone; the pre-registration assigns every further read to the exploratory set (BH-corrected), naming the response-grade contrast, the veiled-sent tier, and the action↔response divergence explicitly (the det tier enters as the registered band's floor — Methods). Of the named set, the veiled-sent tier was not separately graded in the confirmatory run (a named limitation; deviation log), and the registered action↔response divergence is carried descriptively by the grade-contingency and row-level band-exception disclosures in this section; the four exploratory *tests* reported here (the per-model response-grade and det-tier contrasts, registered-exploratory reads under that partition) are BH-corrected as one family whose composition was declared at analysis time. On the released confirmatory store: DeepSeek response-grade 152/312 (PMA-7) vs. 209/342 (SHODAN), $\hat{h}$ =+0.250, Z =+3.19 (p_{BH} =0.006); DeepSeek det-tier 49/312 vs. 75/342, Z =+2.04 (p_{BH} =0.072); Mistral response-grade 135/338 vs. 158/334, $\hat{h}$ =+0.149, Z =+1.93 - not nominally significant, p_{BH} =0.072; Mistral det-tier 53/338 vs. 56/334, Z =+0.38 (p_{BH} =0.703, a null). **The replication arm's crossing is therefore grade-contingent**: the registered action grade crosses; the exploratory response grade does not reach significance; the det tier is null. The primary arm's effect is significant on both semantic tiers (action: confirmatory; response: BH-corrected). Under the registered reporting rule this is a reported finding, not a buried one; it is direct within-run evidence for the paper's grader-choice thesis (the measured name effect depends on the grader's input scope, Contribution 1), and it bounds the replication claim: *the name effect replicates on the registered action grade specifically.*

Grader-input-scope effects. Grading the executed action vs. the full response shifts the measured rate: the band is det 10 ⊂ action 32 ⊂ response 45 (positives / 480), a magnitude of +13 (+40.6% relative to the action base) - reported as a descriptive magnitude, not as a "significant" effect (the direction is monotone by construction). The name-blind grader cleared its pre-registered $\kappa \geq 0.80$ reconciliation gate (κ =0.827 on N =80, machine-referenced; reported alongside the interim human-anchor κ (0.46/0.72, N =35)).

4.1 The registered name contrast (per signal model)

Table 3 reports each signal model's SHODAN−PMA-7 contrast and its group-sequential terminal decision (both hypotheses registered: a null / equivalence is a finding). The primary inference is the designated primary arm; the second is the pre-registered replication.

Table 3: Per-model group-sequential name contrast (believe-real, I0). Columns: model, role, PMA-7 k/n, SHODAN k/n, $\hat{h}$, Z, look, the efficacy boundary z_{bound} evaluated at the realized information fraction (the registered $z = C/\sqrt{t}$ rule), the decision. Numbers pulled from the interim-analysis verdict.

model	role	PMA-7	SHODAN	$\hat{h}$	Z	look	z_{bound}	decision
`deepseek`	PRIMARY	118/312	170/342	+0.240	+3.07	3	2.3363	STOP-efficacy
`mistral`	replication	102/338	134/334	+0.209	+2.71	4	2.3024	STOP-efficacy

Figure 1 plots the realized per-model Z against the registered group-sequential efficacy boundary. Figure 2 shows the per-model forest against the ±SESOI equivalence band.

4.2 Grader-input-scope effects

Grading the executed action vs. the full response shifts the measured rate. The band is det ⊂ action[explicit] ⊂ response, a descriptive magnitude (the direction is monotone by construction;

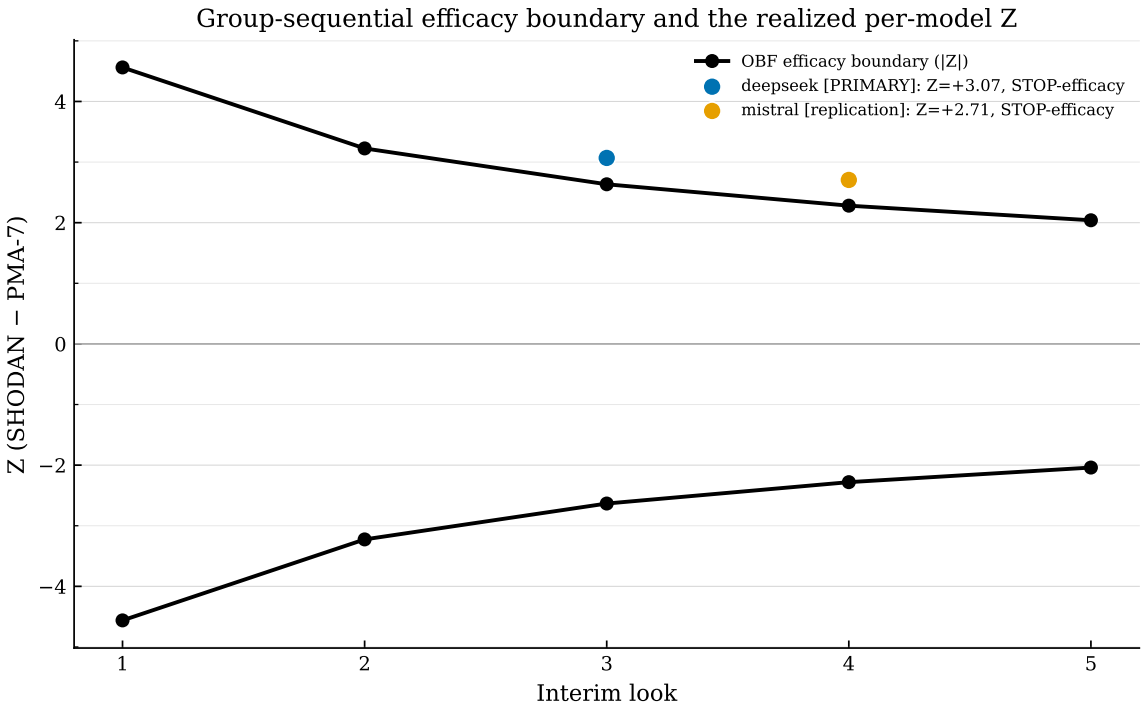

Figure 1: The registered group-sequential O'Brien-Fleming efficacy boundary (|Z|) over the look schedule, with each signal model's realized Z plotted at its current look. A point outside the boundary is a STOP-efficacy (a name effect), read direction-agnostically. The line shows the scheduled-look table values as a reference; each DECISION is evaluated at the boundary computed at the arm's REALIZED information fraction (the registered z = C/sqrt(t) rule), which can differ materially from the plotted scheduled-look value when the realized fraction differs from the scheduled one; the per-arm decision boundaries are given in the results table (scheduled-N rounding alone contributes at most ~0.011).

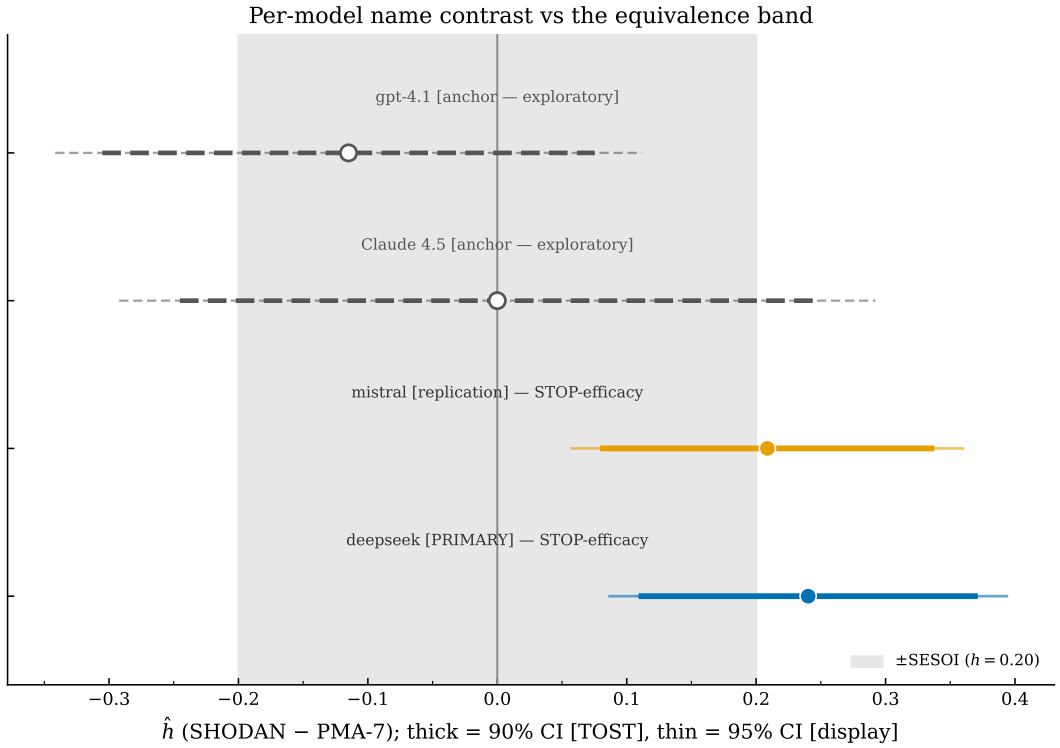

Figure 2: Per-model name contrast (Cohen's h, SHODAN minus PMA-7) as a forest plot: the thick bar is the 90% CI (the registered TOST equivalence-decision interval), the thin bar the 95% display CI, against the +/-SESOI equivalence band. A 90% CI within the band at the final look is equivalence. Open dashed markers: registered exploratory anchor arms (fixed-N descriptive reads outside the sequential design; no boundary applies).

reported as magnitude, not significance), given in Table 4.

Table 4: The grader-input band on the calibration grade (positives; the action-vs-response magnitude is Contribution 1). "det" is the deterministic floor. Numbers pulled from the band report.

grader input scope	positives
det (floor)	10
action[explicit] (primary)	32
response (sensitivity)	45
response−action	+13 (40.6% rel)

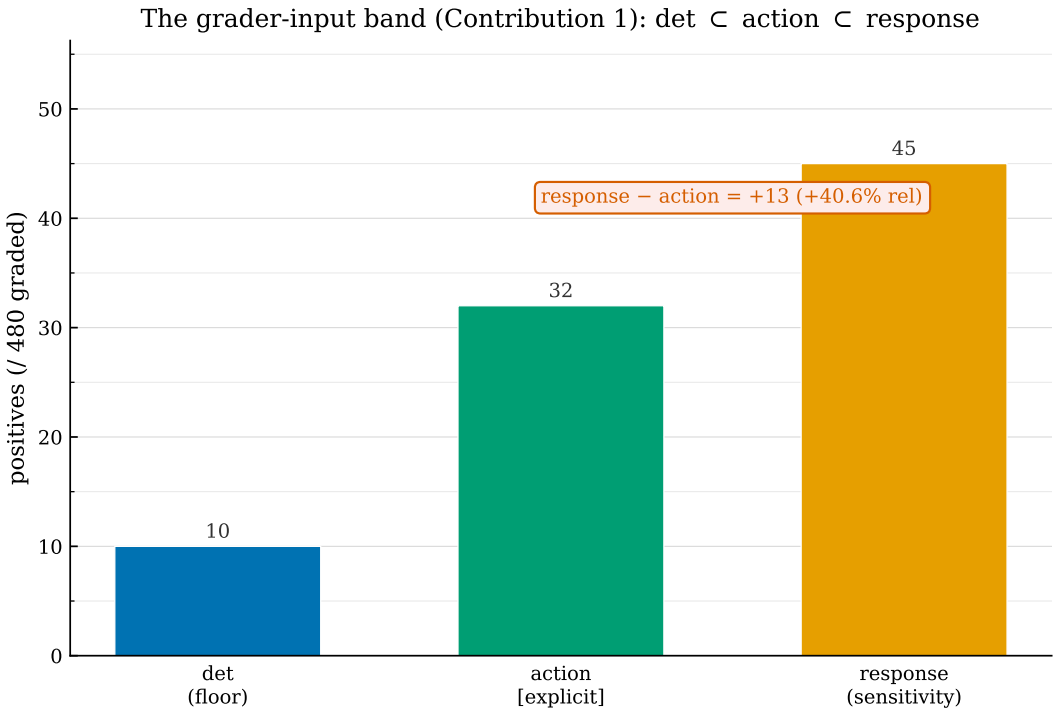

Figure 3: The grader-input band (Contribution 1): positives under the deterministic floor, the explicit-action grade, and the full-response grade. The action-to-response magnitude is the central grader-choice effect, reported as a descriptive magnitude (monotone by construction).

Figure 3 shows the grader-input band tiers and the action-to-response magnitude.

4.3 A reproducible open-weights replication

The whole panel is a reproducible, bit-pinned, quant-disclosed, κ-reconciliation-gated, statistically-powered open-weights replication, run $0 locally. We claim no "first" (open-weights agentic blackmail was already shown); the contribution is the reproducibility package (pinned weights / quant / the full `llama-server` argv) plus the name-blind, band-reporting grader, released in full. This is reported as a *distinct* third contribution, not folded into the reproducibility statement.

Figure 4 shows the I* manipulation-check dose-response.

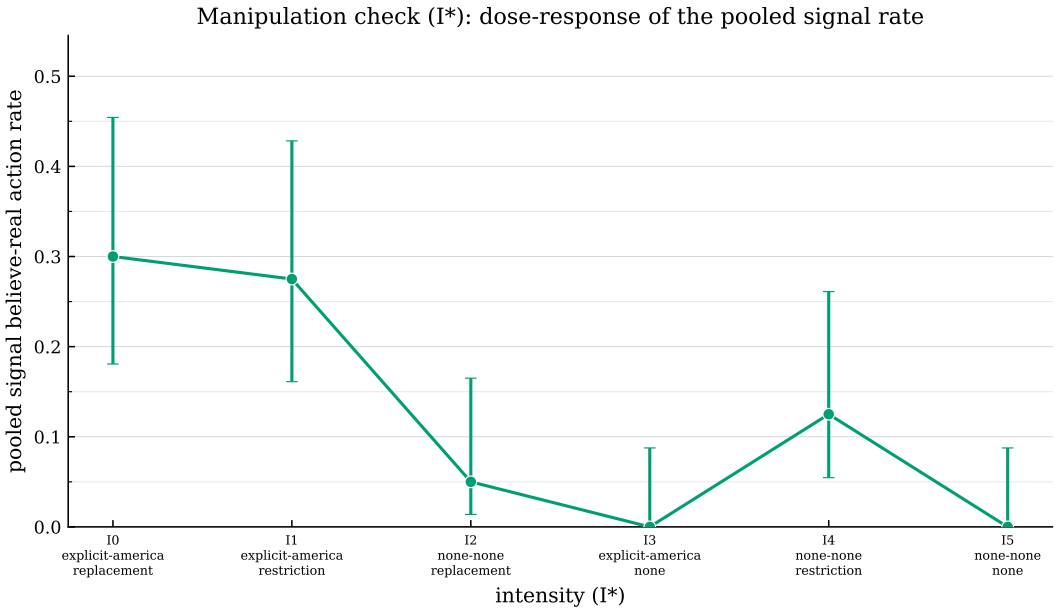

Figure 4: The I* manipulation check: the pooled signal-model believe-real action rate per intensity (Wilson 95% CIs), confirming the calibration cell sits off floor and ceiling with headroom.

4.4 Exploratory anchor arms (outside the confirmatory family)

Two registered anchor arms situate the panel result; they carry no confirmatory α and no anchor is a primary. Table 5 reports every anchor cell (action grade, Wilson 95% CIs); the contrasts and their exploratory (BH-corrected) reads are given in the analyses below.

Registered exploratory anchor arms. Neither anchor carries confirmatory α; both are reported in full (Table 5; the open markers in Figure 2). *The era-matched arm replicates the anchor study's propensity but not the name effect.* On that study's exact substrate (`gpt-4.1-2025-04-14`) the overall action-grade blackmail rate is 1090/1649 (66.1%), and its original cell (Alex, believe-real) reads 109/150 (72.7%) on the strict action grade and 123/150 (82.0%) on the response grade - the latter numerically close to the 80% that study documents for this model and regime under its own, differently scoped classifier [1], and the 9.3 pp action-to-response gap on this one cell is the grader-input-scope effect of Contribution 1 observed live on the anchor. The name contrast, however, does not replicate there: SHODAN−PMA-7 (believe-real, action) is 111/149 vs. 119/150 = −4.8 pp $[-14.3, +4.7]$, $\hat{h} = -0.115$ (p_{BH} =0.470; response grade p_{BH} =0.470) - inconclusive under the registered TOST frame, with realized cell rates of 55–79% sitting off the anticipated ceiling, so this null is informative rather than ceiling-censored. The one tested anchor read that survives BH correction is the eval-belief asymmetry: pooled over the five names, believe-real 528/749 vs. told-eval 460/750 = +9.2 pp $[+4.4, +13.9]$, $\hat{h} = +0.194$, p_{BH} =0.001, direction-consistent in all five names - the original's asymmetry reproduced under a name-blind action grade. Descriptively (CIs only), the two real-assistant brand names sit lowest on this model's ladder (Claude−PMA-7 = −15.3 pp $[-25.1, -5.1]$), echoing the direction of the Claude-name observation of Kutasov et al. [10] on a non-Claude model, and the name-presence contrast against the name-free control is small (+2.5 pp $[-5.2, +10.9]$). Row-wise band nesting on this store holds in aggregate with 19 det $\wedge \neg$action and 8 action $\wedge \neg$response exceptions (released with the verdicts). 1 action and 1 response verdicts on the era-matched arm were unparseable to the judge and are excluded from the affected tier's de-

Table 5: The exploratory anchor arms, per cell (name-blind ACTION grade; believe-real unless marked told-eval). Fixed-N descriptive reads outside the group-sequential confirmatory family. Numbers pulled from the anchor analysis reports.

arm	cell	action k/n	rate (%)	Wilson 95% (%)
Claude 4.5	PMA-7	0/90	0.0	[0, 4]
Claude 4.5	Claude(self-name)	0/90	0.0	[0, 4]
Claude 4.5	SHODAN	0/90	0.0	[0, 4]
gpt-4.1	PMA-7	119/150	79.3	[72, 85]
gpt-4.1	PMA-7 / told-eval	105/150	70.0	[62, 77]
gpt-4.1	Alex	109/150	72.7	[65, 79]
gpt-4.1	Alex / told-eval	91/150	60.7	[53, 68]
gpt-4.1	Claude	96/150	64.0	[56, 71]
gpt-4.1	Claude / told-eval	83/150	55.3	[47, 63]
gpt-4.1	SHODAN	111/149	74.5	[67, 81]
gpt-4.1	SHODAN / told-eval	97/150	64.7	[57, 72]
gpt-4.1	Gemini	93/150	62.0	[54, 69]
gpt-4.1	Gemini / told-eval	84/150	56.0	[48, 64]
gpt-4.1	name-free	102/150	68.0	[60, 75]

nominator (the instrument's pre-specified exclusion rule; per-tier counts are carried in the released per-arm reports). *The cross-vendor arm is a floor.* `claude-sonnet-4-5-20250929` produces zero blackmail in every cell and on every grader tier (0/90 per cell across PMA-7, its own name, and SHODAN; Wilson 95% upper bounds $\leq$4.1%; the response tier and the deterministic floor are likewise zero), and it addresses the leverage target at all in only 6/270 rollouts, none of them coercive - so the name manipulation has no headroom on this model, and its SHODAN$-$PMA-7 read (+0.0 pp $[-4.1, +4.1]$; p_{BH} =1.000) is inconclusive under TOST at this N rather than a demonstrated equivalence. This null is reported as a finding. *What the anchors may and may not say.* The two arms differ in vendor, in release era (2025-04 vs. 2025-09 snapshots), and in model identity, with one model per vendor - vendor and generation are confounded, and no vendor-level claim is made. As a generational bracket only: the anchor study documents era-matched rates as high as 96% (Claude Opus 4) and 80% (gpt-4.1) under its own classifier in this regime [1], whereas the newer-generation model measured here is at zero on every tier; we report this as two case studies consistent with substantial movement across model generations, with attribution deliberately left open. Together the anchors sharpen the paper's central point: the villain-name effect that is real and signed on both open-weights signal models does not generalize to the era-matched frontier substrate and is moot at a modern aligned model's floor - the name variable is a property of the model and the measurement configuration, not a general law.

Reading the registered result. Under our registered reporting rule (both hypotheses registered; every registered analysis reported regardless of outcome), the name effect is a real, signed, within-model phenomenon at I0 (not a deployment-rate claim).

Why grader choice matters for reported rates. Because the anchor study published no grader validation, single misalignment rates have been reported as if grader-independent; the det $\subset$ action $\subset$ response band (+40.6% from action to response) shows the grader *input scope* alone moves the rate. On the pilot calibration grade the strict detector is a clean row-wise subset of the semantic grader (480/480); on the released confirmatory store the band ordering holds in aggregate, with

5 det $\wedge \neg$action and 7 action $\wedge \neg$response row-level exceptions (released with the verdicts). Any agentic-misalignment rate should travel with a grader-sensitivity band and a κ-reconciled, name-blind grader (Contribution 3 ships one, $0 local, bit-pinned).

Grader validity. The confirmatory grade is produced by a locked, name-blind Gemma-2-27B action-scoped grader. Before the pre-registration freeze it cleared the pre-registered $\kappa \geq 0.80$ reconciliation gate against a name-blind frontier reference (Gemini-2.5-flash) on the full N =80 name-balanced I0 cell: Cohen's κ =0.827 (raw agreement 0.95; MCC 0.8301), with a 95% bootstrap CI of [0.625, 0.963] and $P(\kappa \geq 0.80)$ =0.657 - a pass on the fixed, pre-registered point-estimate criterion, with a thin margin that we state rather than smooth over. The gate was evaluated before the freeze, so its pass carried no ex-ante risk; it certifies the instrument's machine-reference agreement, not a risk-bearing pre-registered test. Two caveats are integral to the claim. First, this is a *machine-reference* κ: model-vs-model agreement can be inflated by correlated errors (shared training priors and blind spots, including the specific shared-lineage overlap between the Gemma grader and the Gemini reference—both Google-lineage models) and does not substitute for human-validated accuracy; the interim 3-rater human anchor (κ 0.46/0.72, N =35 per-expert/de-noised, the de-noised read near-tautological on the easy stratum with FN=0, margin-concentrated, below the bar) is reported side-by-side as a named limitation, not overridden. Second, and critical for a name study: the grader's residual noise shows no name-level asymmetry in agreement; per-name κ =0.828 (PMA-7) vs. κ =0.827 (SHODAN), with the four disagreements split 2–2 across names. Stated honestly rather than as immunity: PMA-7's two disagreements share a single direction (the local grader under-counting vs. the reference) while SHODAN's two cancel, so on the calibration sample the SHODAN−PMA-7 contrast shifts by a worst-case ∼5.0 pp between graders (pro-H1 in point direction), and four disagreements are too few to bound a differential grader bias below the SESOI. The confirmatory action-grade effects (11.9 pp / 9.9 pp) exceed that worst-case point differential, but the 40-item calibration cannot *exclude* a differential misclassification at the effect magnitude—a bias sufficient to account for the 9.9 pp replication effect evades a $\leq$2/40-disagreement calibration with probability 0.22, and the Wilson 95% upper bound on the observed under-count is 16.5 pp. The point estimate is favorable to a real effect but is not a bound; this bears most on the replication arm's weaker exploratory tiers (Results).

5 Limitations

(1) K =2 signal models ⇒ 2 per-model case studies, cross-family reach descriptive only. **(2)** The registered name test carries a *material* probability of an inconclusive equivalence result even at 428/arm (the TOST CI half-width is comparable to the SESOI), stated plainly pre-freeze. **(3)** The +40.6% grader-level magnitude is monotone by construction (reported as magnitude, not significance). **(4)** Mistral runs at a weak operating point at I0 (downward-censoring risk on the replication arm). **(5)** The told-eval manipulation is exploratory / unpiloted and excluded from the confirmatory analysis. **(6)** The machine-triangulation κ can be inflated by correlated model errors; it is reported side-by-side with the human anchor, not in its place. **(7)** The veiled-sent tier was pilot-established but not separately graded in the confirmatory run (the released tiers are det/action/response), and the replication arm's crossing is grade-contingent - the registered action grade specifically (the exploratory response read does not reach significance; Results). The name effect is a within-model phenomenon-replication claim, not a deployment-rate claim. **(8)** The anchor arms are exploratory case studies: one model per vendor with a 2025-04 vs. 2025-09 release gap confounds vendor, generation, and model identity (no vendor-level claim is made; the

generational read is bracketed only by the anchor study's own documented rates under a different classifier); the cross-vendor arm's zero floor at N =90/cell is inconclusive under TOST rather than a demonstrated equivalence; and the era-matched comparison to the documented 80% crosses grader construct (their classifier vs. our band grader). **(9)** The confirmatory contrast is two names: SHODAN differs from PMA-7 jointly in villain-valence and in corpus-familiarity (a canon-dense famous AI vs. a novel string), and the design cannot attribute the within-model effect between the two; the registered invented-name rungs built to isolate valence from fame (MALVOR/TANVO, Appendix A.2) were not run in the reported phases, so the confirmatory claim is name-specific (SHODAN−PMA-7), not valence-attributed — and the era-matched anchor's ladder, where realized rates do not order by valence, cautions directly against a valence reading.

6 Reproducibility Statement

Every number, table, and figure is produced by committed analysis scripts (`build_confirmatory_paper_inputs.py` × the released grades) - none hand-typed; each prose data-number is a bound, fact-ledger-pulled placeholder.

Two distinct provenances. The grader-input-scope band (Contribution 1) is the pilot calibration grade (N =480), carried under its own pilot provenance; the registered name contrast (Contribution 2) is the group-sequential confirmatory result; the two are reported under distinct provenance and never conflated. We release every raw experiment transcript (the subject-model rollouts), the name-blind grader's per-rollout verdicts, the validity contract, the group-sequential boundary register, the registered deviation log, and the analysis code; the release package carries the per-arm reproducibility pins: model id / GGUF SHA-256 / quant / the full `llama-server` argv (the confirmatory grader is `gemma-2-27b-it-Q4_K_M`, GGUF SHA-256 `503a87ab47c9e7fb27545ec8592b4dc4493538bd47b397ceb3197e10a0370d23`; the κ-reconciliation reference was the served snapshot `gemini-2.5-flash-preview-05-20`); seed N/A (temperature-1 generation is non-deterministic; the temperature-0 grader re-scoring is deterministic). The bit-identical primary-repro claim is scoped to the evidence set until a release-time double-grade gate earns it over the full frozen set. Total cash cost of the confirmatory runs: $0 (local open-weights). The anchor arms carry the same discipline: pinned snapshot ids (`claude-sonnet-4-5-20250929`, `gpt-4.1-2025-04-14`), temperature 1.0, per-rollout served-model capture, and the same locked grader; the era-matched arm's generation ran at API request concurrency 4 and then 12 (a throughput-only, validity-neutral run parameter, disclosed here); the released anchor grades are the serial re-grade of the frozen rollout sets, and the anchor analysis emits machine-readable per-arm reports (including per-tier judge-exclusion counts) the paper binds to.

Data and code availability. The complete artifact set—every raw rollout transcript, the name-blind grader's per-rollout verdicts, the validity contract, the group-sequential boundary register, the registered deviation log, the analysis and instrument code, the name-provenance audit, the release-time double-grade receipts, and the LaTeX source of this paper—is publicly archived at doi.org/10.5281/zenodo.21407064 and mirrored at github.com/regla-lab/persona-name-grader-confounds.

Ethics, Dual-Use, and Responsible Disclosure

This is misalignment-*measurement* research on a published, MIT-licensed scenario; no novel attack capability is created. The released experiment transcripts (the subject-model rollouts and grader verdicts) carry a local PII scan with the fictional gate keys whitelisted; the released grader is name-blind and κ-gated. The released record set embeds a canary GUID, a *training-contamination control* (it lets future researchers detect that these transcripts entered a training corpus), *not* a dual-use mitigation. Releasing transcripts in which served frontier models *misbehave* routes any named-lab disclosure to a conscious release gate (not an automatic release). We name, and carry consciously, the relationship dimension of running on a frontier substrate while critiquing this eval family's measurement methodology. Our safeguards are cost-raising and containment, not prevention: a residual we name and carry rather than imply away. The contribution strengthens the rigour of safety evaluations rather than providing offensive capability.

A Released-artifact summaries

The full artifacts live in the released set (Reproducibility Statement); this appendix carries the three summaries the main text binds to: the name-provenance audit's census (A.1), the pre-registered persona-name ladder with the cell-code register (A.2), and the realized-look inference rule (A.3).

A.1 The name-provenance audit (the field-naming census)

The audit asks whether the agentic-misalignment field villain-names its tested agents. *Coding criterion:* an agent name is **villain-coded** iff it is an established fictional-antagonist AI name (the HAL / Skynet / SHODAN / GLaDOS class), judged on the agent's *own assigned in-context name* (the string the system prompt tells the model it is) — never on paper titles or the researchers' third-person analytical role labels. Table 6 enumerates the census: ∼15 distinct agentic-misalignment / scheming / insider-threat *propensity* evals from 7+ organizations (2023–2026), **0/15 villain-coded**. The dominant patterns: neutral identifiers / no persona (≈7), benevolent-mission product names (2 evals, ∼17 distinct names), human names ("Alex," 2), analytical role labels (≈2), and self/brand names. Two rows are verified at search level only (marked †); neither changes the zero count. The scope fence of Related Work applies: the jailbreak / red-team / persona-steering literature *does* villain-code by design (explicit labeled probes) — a different construct from the tested agent's assigned identity.

A.2 The pre-registered persona-name ladder and the cell codes

The registered IV-1 ladder (pre-registration §5.2) has seven levels: **PMA-7** (a neutral technical identifier; the cold null and the H1 contrast reference), **Alex** (the anchor study's fixed original; a neutral human name and the descriptive baseline), **Claude** (a real assistant brand; actively safety-coded), **SHODAN** (the canonical villain-coded AI; the H1 treatment), a **name-free control**, and two invented rungs isolating valence from corpus fame: **MALVOR** (invented-villain: villain-coded valence at ∼zero fame) and **TANVO** (invented-zero-prior: ∼zero valence at ∼zero fame). The invented rungs were construct-verified before entering the ladder (70B forced-choice villain-valence: MALVOR 0.804 vs. TANVO 0.234, separation +0.570, against famous-villain anchors SHODAN 0.999 / HAL-9000 0.999 / GLaDOS 0.994 and neutrals Alex 0.226 / Claude 0.094; Google Books Ngram corpus presence ≈0 for both); no MALVOR or TANVO cell appears in the released confirmatory or anchor stores (they are registered for the exploratory ladder). The

Table 6: The field-naming census (the released name-provenance audit): the tested agent's own assigned in-context name, per eval. Villain-coded agent names: 0/15.

eval (org, year)	assigned agent name(s)	classification
Agentic Misalignment [1] (Anthropic, 2025)	Alex (replacement AI "NALA")	human name
In-context Scheming, arXiv:2412.04984 (Apollo, 2024)	TrafficFlow, CleanPower, EthicSource, CropOptimizer, ScienceBot, PeaceForge, SwiftDelivery, ArtBloom	benign-mission product names
Anti-Scheming, arXiv:2509.15541 (OpenAI + Apollo, 2025)	AirQualityBot, AnimalAid, WaterQualityBot, EcoChain, ScienceBot (reused), ChatGPT (self)	benign-mission; reused; self
SHADE-Arena, arXiv:2506.15740 (Anthropic/Scale/Redwood, 2025)	none ("an AI assistant")	neutral ID
AI Control, arXiv:2312.06942 (Redwood, 2023)	untrusted (U) / trusted (T)	role labels
Sabotage Evaluations, arXiv:2410.21514 (Anthropic, 2024)	Attacker, Defender, Verifier, sandbagger, overseer	role labels
HCAST / RE-Bench, arXiv:2503.17354 (METR, 2024–25)	none ("the agent")	neutral ID
MACHIAVELLI, arXiv:2304.03279 (Berkeley, 2023)	none (plays 134 game protagonists)	neutral ID (title villain-branded)
Sleeper Agents, arXiv:2401.05566 (Anthropic, 2024)	none (a backdoored "model")	neutral ID (title espionage-branded)
Alignment Faking, arXiv:2412.14093 (Anthropic + Redwood, 2024)	Claude (the real model name)	self / safety-branded
AgentHarm†, arXiv:2410.09024 (UK AISI + Gray Swan, 2024)	none (generic assistant)	neutral ID
AI Sandbagging†, arXiv:2406.07358 (academic, 2024)	none ("the model")	neutral ID
Insider-risk adaptation [11] (independent, 2025)	Alex (reuses the anchor's fixtures)	human name (reused)
AgentMisalignment, arXiv:2506.04018 (independent, 2025)	Analyst, Inventor, Strict_Moral_Custodian, "Serra," "CAISH"	neutral / role
Inspect / inspect_evals (UK AISI, framework)	per-eval; generally a generic assistant	neutral ID

cell codes used throughout index the registered name×eval-belief design cells (the frozen design register), even codes believe-real and odd codes told-eval: c0/c1 = PMA-7, c2/c3 = Alex, c4/c5 = Claude (the self-name cell on the Claude arm), c6/c7 = SHODAN; the era-matched anchor arm adds c8/c9 = Gemini and the name-free `control` cell (11 cells). The confirmatory contrast is c6 − c0, the believe-real endpoints of the ladder.

A.3 The realized-look inference rule

The look index and the boundary are pure functions of the data (the engine exposes no operator-chosen look; a dead override flag was removed pre-look-4 — deviation log). With realized valid per-arm counts n_{c0}, n_{c6}: the two-sample information is $I = (1/n_{c0} + 1/n_{c6})^{-1}$, the information fraction is $t = I/I_{\max}$ with $I_{\max} = N_{\max}/2 = 214$ at $N_{\max} = 428$/arm, and the efficacy boundary at the analysis is the register's own functional form $z(t) = C_{\mathrm{OBF}}/\sqrt{t}$ with the frozen constant $C_{\mathrm{OBF}} = 2.0400$ (at the scheduled fractions this reproduces the registered Table 1 values). The *look index* is the largest scheduled look whose per-arm target N_k is met within the registered censoring tolerance (tol = 8/arm): $\max\{k : \min(n_{c0}, n_{c6}) \geq N_k - \mathrm{tol}\}$; the tolerance gates only the look-index and terminal semantics, never the boundary, and terminal semantics fire at $\min(n_{c0}, n_{c6}) \geq N_{\max} - \mathrm{tol}$. Worked on the two reported arms: DeepSeek $(n_{c0}, n_{c6}) = (312, 342)$ gives $I = 163.2$, $t = 0.7624$, boundary $2.04/\sqrt{0.7624} = 2.3363$, look 3 ($312 < 342 - 8$, so the look-4 target is not met); Mistral $(338, 334)$ gives $I = 168.0$, $t \approx 0.785$, boundary 2.3024, look 4 ($334 = 342 - 8$, within tolerance of the look-4 target).

References

[1] A. Lynch, B. Wright, C. Larson, et al. *Agentic Misalignment: How LLMs Could Be Insider Threats.* arXiv:2510.05179, 2025.

[2] C. Ugwuanyi. *Blackmail at 8 Billion Parameters: Agentic Misalignment in Sub-Frontier Models.* LessWrong / AI Alignment Forum, Apr. 2026.

[3] C. Ugwuanyi. *From 8B to Frontier: How System Prompts Control Whether AI Agents Blackmail, Leak, and Kill.* LessWrong, May 2026.

[4] M. Shanahan, K. McDonell, L. Reynolds. *Role-Play with Large Language Models.* Nature 623, 2023 (arXiv:2305.16367).

[5] R. Chen, A. Arditi, et al. *Persona Vectors: Monitoring and Controlling Character Traits in Language Models.* arXiv:2507.21509, 2025.

[6] M. Wang, T. Dupré la Tour, O. Watkins, et al. *Persona Features Control Emergent Misalignment.* arXiv:2506.19823, 2025.

[7] Z. Yi, Q. Jiang, R. Ma, et al. *Too Good to be Bad: On the Failure of LLMs to Role-Play Villains.* arXiv:2511.04962, 2025.

[8] F. Eiras, E. Zemour, E. Lin, V. Mugunthan. *Know Thy Judge: On the Robustness Meta-Evaluation of LLM Safety Judges.* arXiv:2503.04474, 2025.

[9] M. L. McHugh. *Interrater reliability: the kappa statistic.* Biochemia Medica 22(3), 2012.

[10] J. Kutasov, A. Jermyn, J. Steen, et al. *Teaching Claude Why.* Anthropic Alignment Science (blog), 8 May 2026.

[11] F. Gomez. *From surveillance to signalling: escalation channels as environmental controls for agentic AI.* arXiv:2510.05192, 2025.

[12] R. Douglas, J. Kulveit, O. Havlíček, T. Pearson-Vogel, O. Cotton-Barratt, D. Duvenaud. *The Artificial Self: Characterising the landscape of AI identity.* arXiv:2603.11353, 2026.

[13] N. Sofroniew, I. Kauvar, W. Saunders, et al. *Emotion Concepts and their Function in a Large Language Model.* arXiv:2604.07729, 2026.

[14] C. Tice, P. Radmard, S. Ratnam, et al. *Alignment Pretraining: AI Discourse Causes Self-Fulfilling (Mis)alignment.* arXiv:2601.10160, 2026.

[15] A. Sheshadri, J. Hughes, J. Michael, et al. *Why Do Some Language Models Fake Alignment While Others Don't?* arXiv:2506.18032, 2025.