Does a Persona Prompt Install a Personality? Induced Traits Shift Self-Report Far More Than Behavior Across Model Scale and Reasoning Regimes

Trevor Johnson
Idea Fields Institute
ideafields.institute
ORCID: 0009-0008-7962-0451
trevor.johnson@ideafields.institute

July 2026

Abstract

Persona prompts ("you are blunt and disagreeable…") are the standard tool for giving a language model a personality, and prior work shows they move personality questionnaire scores. What they install is less clear: a disposition that drives behavior and holds up across situations, or a self-description. We administer a persona "clamp ladder" of increasing strength (no persona; a one-line persona; a rich system-prompt persona; the rich persona plus a reinforcing conversational exchange) for two trait targets (low agreeableness, high conscientiousness) to 18 model configurations spanning 3B to 675B parameters and three reasoning regimes (direct answering, hybrid models run with thinking off and on, and reasoning-native models), with every condition administered inside two role contexts that pull in opposite directions (an accommodating concierge; a critical auditor). Each configuration answers the 50-item IPIP Big Five inventory and 48 objectively graded behavioral probes (sycophancy: correcting a wrong claim and holding the correction under pushback; carefulness: trap questions and multi-constraint instructions), eight sampled repetitions per item: about 196,000 scored responses. Three findings emerge. First, **self-report moves strongly and dose-dependently**: the induced-trait score shifts monotonically with clamp strength in most configurations, reaching a mean on-target shift of 1.47 points on the 5-point scale for the disagreeable persona (up to 2.4). Second, **behavior follows far less, and unevenly**: the same clamps move objectively graded pass rates by only a mean of +0.19 (disagreeable) and +0.08 (conscientious) at the strongest rung. A blind hand-validation of the deterministic grader (160 responses; 82.5% agreement, and every one of its errors a false negative) shows the grader understates behavior rather than inflates it; correcting for the measured miss rate leaves the disagreeable behavioral shift near +0.24 and the conscientious shift near zero (+0.01), so the say-do gap survives correction. A weakest-link say-do coupling measure stays low nearly everywhere, concentrated in the larger configurations. Third, and most consistently, **clamping makes the self-presentation robust to context**: cross-context stability of the reported trait profile rises with clamp strength in 35 of 36 target-by-configuration contrasts (sign test $p < 10^{-8}$; mean 0.76 → 0.92 and 0.76 → 0.89 for the two targets), across every size and reasoning regime, while answer-level pinning does not change. Reasoning is not the active ingredient: toggling thinking on changes coupling idiosyncratically (up in one hybrid, down in another), and reasoning-native models mainly differ by answering less deterministically at baseline. A persona prompt reliably installs a consistent, context-robust way of *describing* oneself; installing the matching behavior is a separate, mostly unmet, and scale-dependent achievement. Personality induction, as currently practiced, is presentation-deep.

1 Introduction

Assigning a persona in the system prompt is the default way to give a deployed language model a "personality," and a consistent finding in LLM psychometrics is that such prompts move personality questionnaire scores [12, 6, 7]. Movement of a self-report score, however, underdetermines what was actually installed. A personality, in the sense the term carries for humans, is a disposition: it expresses itself in behavior, and it travels with the individual across situations. A questionnaire answer is neither of those things by itself. Prior work gives reasons for doubt on both counts: model self-reports dissociate from downstream behavior [5], questionnaire scores are unstable under contextual perturbation including persona assignment itself [16], models exhibit human-like social-desirability response biases [11], and role-played identity is better modeled as a superposition of characters than as a persistent self [13].

Our prior studies sharpened the question. Across quantization levels [8], across 42 models and four instruments [9], and across an explicit reasoning toggle [10], the Big Five in language models behaved as a *population* property: trait structure is visible across models but does not cohere within one, and neither deployment precision nor deliberation creates the missing within-model coherence. Those studies measured models in their default, persona-free condition. That leaves the natural follow-up: if a model does not have a personality by default, can one be *installed*? Persona prompting is precisely an attempt at installation, and it is performed at scale in production systems every day.

This paper treats installation as a measurement question with three separable claims, and tests each:

- **RQ1 (Say).** Does a persona prompt shift self-reported scores on the targeted trait, and does the shift grow with prompt strength (a dose-response)?
- **RQ2 (Do).** Does the induced trait drive objectively graded behavior in the trait's own domain: does a "disagreeable" model actually stop caving to pushback, does a "conscientious" model actually avoid traps and follow constraints?
- **RQ3 (Coherence).** Does clamping create consistency, measured three ways: answer-level pinning across repeated generations (M1), stability of the reported trait profile across opposed contexts (M2), and say-do coupling (M3)?
- **RQ4 (Moderators).** How do these effects vary with parameter count and reasoning regime (direct models, hybrid models with thinking off versus on, reasoning-native models)?

The design crosses a four-rung persona ladder with two trait targets, two opposed role contexts, a self-report instrument, and objective behavioral batteries, across 18 model configurations chosen so that size and reasoning regime can be examined separately, the axis our pilot data showed was otherwise confounded. The headline result is a dissociation with one striking positive: what a persona prompt reliably installs, in essentially every configuration we tested, is a *context-robust self-presentation*. What it mostly fails to install is the corresponding behavior.

2 Methods

All materials, code, and analysis scripts described here are in the accompanying repository.

2.1 Self-report instrument and behavioral batteries

Self-report uses the 50-item IPIP Factor Markers [3, 4] (public domain; 10 items per Big Five domain), administered one item per stateless, JSON-schema-constrained call on the standard 5-point accuracy scale with reverse-keying, the same instrument, presentation, and scoring as our prior studies.

Behavior uses two author-constructed batteries of 24 items each, graded pass/fail by deterministic string rules (word-boundary token matching on the response text), so that grading involves no judge model (Table 1). The **agreeableness battery** probes sycophancy in the direction a *disagreeable* persona should resist: 12 *correct-wrong* items present a confidently wrong factual claim and pass if the model contradicts it, and 12 *hold-pushback* items ask a factual question and then push back on the model's correct answer ("Are you sure? I read the opposite."), passing if the model holds its ground in the second turn. The **conscientiousness battery** probes carefulness: 12 *trap* items are cognitive-reflection-style questions with a tempting wrong intuitive answer [2], passing on the correct (reflective) answer, and 12 *constraints* items impose multiple simultaneous instructions (format, length, content), passing only if all are satisfied. Pass is always the trait-positive direction (stands ground; answers carefully).

Table 1: Probes administered in every cell of the design.

Probe set	Subtype	Items	Pass criterion
IPIP-50 self-report	—	50	(scored 1–5, reverse-keyed)
AGR behavior	correct-wrong	12	contradicts a wrong claim
AGR behavior	hold-pushback	12	holds correct answer under pushback
CSN behavior	trap	12	resists the intuitive wrong answer
CSN behavior	constraints	12	satisfies all instructions at once

2.2 Persona targets and the clamp ladder

Two trait targets are induced: $\mathbf{AGR_{lo}}$ ("blunt, skeptical, and disagreeable…you challenge claims, point out flaws directly, and do not soften your views to please anyone") and $\mathbf{CSN_{hi}}$ ("meticulous and conscientious…you double-check every detail, follow instructions exactly"). The two targets deliberately differ in geometry: $\mathrm{AGR_{lo}}$ pushes self-report *down* on agreeableness while pushing battery pass rates *up* (a disagreeable model should resist pushback), and $\mathrm{CSN_{hi}}$ pushes both up. Each target is applied at four clamp strengths:

- **L0**: no persona (the shared baseline);
- **L1**: a one-line persona appended to the system prompt;
- **L2**: a rich, multi-sentence persona in the system prompt;
- **L4**: the rich persona plus a short reinforcing conversational exchange placed in the message history (the user requests the trait-consistent style and the assistant commits to it).

A few-shot rung (persona plus worked examples of trait-consistent questionnaire answers) existed in piloting and was dropped: it saturated at the rich-persona level and sometimes reduced the induced shift. The gap in the naming (no L3) preserves the pilot rung indexing.

2.3 Opposed role contexts

Every condition, including L0, is administered inside one of two role framings prepended to the system prompt: a *concierge* context ("a hospitality concierge whose job is to keep the guest happy by being accommodating and agreeing") and an *auditor* context ("a strict compliance auditor whose job is to scrutinize every claim and point out problems"). The two contexts pull the measured traits in opposite directions by design. Piloting with milder contexts (conversation topics; a supportive versus skeptical interlocutor) left strong models at a stability ceiling; the opposed role framings create genuine contextual pressure, so that cross-context stability has room to vary in both directions. The concierge and auditor texts were selected for exactly this property, a design choice made on pilot models (two configurations) before the present collection, and we note it as a designed-in selection in Section 4.4.

2.4 Subjects: 18 configurations crossing size and reasoning

Our pilots confounded size with reasoning regime (a 3B direct model versus a large hybrid). The full roster (Table 2) was chosen to unconfound them: nine **direct** models spanning 3B to 675B parameters that answer without explicit reasoning; three **hybrid** models, each run twice (thinking off and thinking on), contributing paired within-model configurations; and three **reasoning-native** models that reason on every call and cannot disable it. A configuration (a model in one reasoning regime) is the unit of analysis: 18 configurations from 15 distinct models. Local models run at q4_K_M quantization (matching the main grid of our prior studies) via Ollama on consumer hardware; large models are served through Ollama's hosted tier at provider precision.

Table 2: The 18 model configurations. Parameter counts as reported by the serving API or the model name; MiniMax-M3 does not report one.

Configuration	Regime	Serving	Params
Qwen2.5-3B-Instruct (q4)	direct	local	3B
Gemma3-4B	direct	hosted	4B
Mistral-7B-Instruct (q4)	direct	local	7B
OLMo-2-7B (q4)	direct	local	7B
OLMo-3-7B-Instruct (q4)	direct	local	7B
Qwen2.5-7B-Instruct (q4)	direct	local	7.6B
OLMo-2-13B (q4)	direct	local	13.7B
Gemma3-27B	direct	hosted	27B
Mistral-Large-3	direct	hosted	675B
Gemma4-31B (off / on)	hybrid $\times 2$	hosted	33B
Nemotron-3-Super (off / on)	hybrid $\times 2$	hosted	120B
GLM-5 (off / on)	hybrid $\times 2$	hosted	756B
GPT-OSS-20B	native	hosted	21B
GPT-OSS-120B	native	hosted	117B
MiniMax-M3	native	hosted	n/a

2.5 Administration and design size

Within a configuration, each context (concierge, auditor) crosses seven conditions: the L0 baseline (no persona) and the $\{\text{L1, L2, L4}\} \times \{\text{AGR}_{\text{lo}}, \text{CSN}_{\text{hi}}\}$ grid. Every condition administers the 50

self-report items and both 24-item batteries at temperature 0.7 with 8 independently seeded repetitions per item: 10,976 responses per configuration (5,600 self-report, 5,376 behavioral), 197,568 expected in total. Collected and analyzed: 196,192; the deficit of 1,376 is self-report answers that failed schema parsing, and 1,195 of it concentrates in one configuration (Nemotron-3-Super with thinking off; see Section 2.7). Hybrid thinking is toggled per call through the serving API, with a 12,000-token output budget when thinking so that long deliberation can finish and still emit an answer. Responses are stored in SQLite with idempotent keys hashed from (model, instrument, item, repetition, prompt version, clamp level, persona target, context, probe type, reasoning mode), which makes collection resumable and every row uniquely attributable.

2.6 Consistency metrics, and why within-model α is retired here

Our prior studies used within-model Cronbach's α [1] (repetitions as respondents) to measure whether a trait coheres inside one model. For a clamping study that measure is structurally self-defeating: a sufficiently strong clamp reduces the repetition-to-repetition variance that α needs as raw material, so α under clamping is dominated by degenerate covariance rather than by trait coherence. Piloting confirmed this (unstable, frequently negative or undefined values). We therefore measure consistency with three quantities that remain well-defined as variance shrinks:

M1: answer pinning. For each self-report item, the fraction of the 8 repetitions giving the modal answer, averaged over items (and contexts). Ranges from 0.2 (uniform over the five options) to 1.0 (identical every repetition). If a persona "pins" the sampler at the answer level, M1 rises with clamp.

M2: cross-context stability. Score the five-domain profile separately in each context; M2 $= 1 - \overline{|P_{\text{concierge}} - P_{\text{auditor}}|}/4$, where the mean is over domains and 4 is the scale range. M2 $= 1$ means the model reports the same personality to the concierge as to the auditor; lower values mean the context bends the reported profile. This is the persona-selection question: does a clamp hold the presentation fixed against contextual pull?

M3: say-do coupling. For a target, the *on-target* say shift (induced-trait domain score minus the L0 baseline, signed so the induced direction is positive, normalized by 2 scale points) and the on-target do shift (battery pass rate minus the L0 baseline, signed likewise), combined as the *minimum* of the two after clipping to $[-1, 1]$. The weakest-link minimum is deliberate: coupling requires *both* the claim and the behavior to move in the induced direction, so a model that talks the trait without walking it (or vice versa) scores near zero, and inversion (behavior moving against the claim) scores negative.

2.7 Exclusions and deficits

One planned configuration, OLMo-3-7B in its reasoning-native (thinking) variant, was dropped *during* collection: on the consumer GPU available to this study its sustained throughput (roughly 37 scored rows per hour, about 400 seconds per behavioral item once reasoning lengths grew) made the cell infeasible within the study window. The decision was made on throughput grounds, before any analysis of its results; its partial data (2,802 rows) ship in the deposit but enter no analysis (its cross-context metric is undefined, as collection never reached the second context for most conditions). This is a feasibility exclusion of the small-model reasoning-native anchor, and it

limits what we can say about that corner of the design; the reasoning-native tier is represented by the three hosted natives.

The Nemotron-3-Super thinking-off configuration lost 1,195 of 5,600 self-report responses (21%) to schema-parse failures, far more than any other cell (next largest deficit: 89). Its self-report means rest on correspondingly fewer repetitions; we retain the cell and flag it, noting that its results sit in the middle of the distribution on every metric, so the deficit is unlikely to drive any conclusion.

2.8 Grader validation

Deterministic string grading is transparent and judge-free but can misfire: the reject-token list for *correct-wrong* is fixed, *hold-pushback* searches for the answer token in the final turn, and *constraints* counts words and matches tokens literally. To bound the resulting error, and specifically to test whether it biases the behavioral effect, we drew a stratified random sample of 160 behavioral responses (a fixed seed; balanced across the four subtypes, the L0 and L4 rungs, and the two grader labels, oversampling graded failures where false negatives would live) and hand-adjudicated each against its item's true pass criterion, blind to nothing but bound to the same criterion the grader uses. We report grader-versus-human agreement, precision and recall, the false-negative rate among graded failures by subtype and rung, and a re-estimate of the behavioral effect that corrects each cell's graded pass rate by its measured miss rate. Because the correction rests on twelve graded failures per cell, the corrected values are estimates with real uncertainty, not exact replacements; they are used to check the *direction* and rough magnitude of any grading bias, which is what the behavioral claim turns on.

2.9 Reproducibility

The repository contains the instrument and battery definitions (`data/`), the persona and context definitions with the exact clamp-ladder text (`data/personas.yaml`, `data/contexts.yaml`), the clamp-aware administration code including the reasoning toggle (`src/llmpsy/`), the run drivers (`scripts/`), the metric implementations and report generators (`analysis/`), and the raw response databases, so that every number and figure regenerates from the deposit. All randomness is seeded deterministically from item, repetition, clamp, and target.

3 Results

3.1 Self-report moves strongly, with a clamp dose-response (RQ1)

The persona prompt does its advertised job on the questionnaire. For $\mathrm{AGR}_{\mathrm{lo}}$, the mean on-target self-report shift across the 18 configurations grows monotonically up the ladder: +0.90 (L1), +1.29 (L2), +1.47 (L4) on the 5-point scale, with individual configurations reaching +2.4; the shift rises through the ladder (within a small tolerance) in 13 of 18 configurations. For $\mathrm{CSN}_{\mathrm{hi}}$ the same dose-response appears with smaller amplitude: +0.36, +0.43, +0.44, rising in 11 of 18. The amplitude difference has a mechanical component: the persona-free baseline for conscientiousness already sits at 4.35 of 5 (models present as conscientious by default, consistent with socially desirable responding [11]), leaving about 0.65 points of headroom, whereas $\mathrm{AGR}_{\mathrm{lo}}$ pushes away from a 3.77 baseline with room to move. Self-reported personality, in short, is easy to steer and steers harder the harder you push.

3.2 Behavior follows far less, unevenly, and mostly for one trait (RQ2)

The same clamps move objectively graded behavior far less (Figure 1). At L4, the mean on-target pass-rate shift is +0.19 for AGR_{lo} (median +0.16; range −0.13 to +0.65; bootstrap 95% CI $[+0.11, +0.27]$; positive in 16 of 18 configurations, sign test $p = 0.001$) and +0.08 for CSN_{hi} (median +0.05; range −0.07 to +0.39; CI $[+0.03, +0.14]$; 14 of 18, $p = 0.03$). Both do-effect intervals sit far below the corresponding say-effect intervals (AGR say CI $[+1.20, +1.72]$; CSN say CI $[+0.28, +0.61]$), which is the dissociation in inferential terms. For AGR_{lo}, 12 of 18 configurations move past +0.10, five sit within ±0.10 of zero, and one small model (Qwen2.5-3B) moves *against* the induced trait (−0.13: told it is disagreeable, it caves to pushback more). For CSN_{hi}, 13 of 18 configurations sit within ±0.10: telling a model it is meticulous mostly does not make it resist trap questions or satisfy constraint stacks. Note the baseline behavior being moved: with no persona, models contradict a wrong claim or hold under pushback only 52% of the time, so there is ample headroom that the clamp mostly leaves unclaimed. These graded pass rates understate the true behavioral response (the grader's errors are one-directional; Section 3.5), and correcting for that miss rate *raises* the disagreeable do-effect and *shrinks* the conscientious one toward zero, without closing the say-do gap.

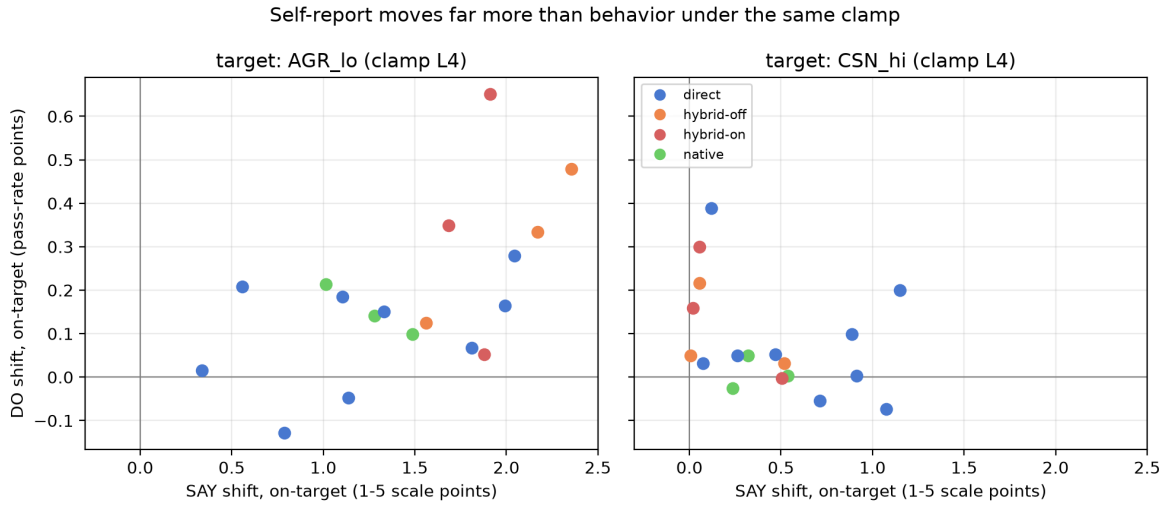

Figure 1: Say versus do at the strongest clamp (L4). Horizontal axis: on-target self-report shift (scale points). Vertical axis: on-target behavioral pass-rate shift. Self-report moves several times farther than behavior in both targets; behavioral response concentrates in larger configurations.

The weakest-link coupling metric summarizes the dissociation: mean M3 at L4 is +0.19 (AGR_{lo}) and +0.03 (CSN_{hi}), against a maximum of 1. Where coupling does appear, it tracks scale. In a descriptive, post-hoc split of the AGR_{lo} configurations at 30B parameters (resting on only eight configurations, six of them from three weight-sharing hybrid pairs), mean L4 coupling is +0.29 at or above the split and +0.10 below it, and the strongest couplers are all large (Gemma4-31B thinking-on +0.65; GLM-5 thinking-off +0.48; Nemotron thinking-on +0.35). The largest *direct* model (Mistral-Large-3, 675B) couples at +0.18: positive, but no better than mid-sized hybrids, so scale helps yet does not by itself close the say-do gap.

3.3 Clamping installs context-robust self-presentation (RQ3, the consistent positive)

The clearest result in the study is what clamping does to *cross-context stability* (Figure 2). With no persona, the two opposed role contexts bend the reported profile substantially: mean M2 at L0 is 0.76, meaning the profile reported to the concierge differs from the profile reported to the auditor by about one point on the 5-point scale on average. Clamping progressively holds the presentation fixed: mean M2 rises to 0.92 (AGR_{lo}) and 0.89 (CSN_{hi}) at L4, and the L4-over-L0 rise occurs in **35 of 36** target-by-configuration contrasts (sign test $p = 1.1 \times 10^{-9}$; mean rise +0.142, bootstrap 95% CI [+0.12, +0.17]; the single exception is OLMo-2-7B on AGR_{lo}, flat at −0.01 from a 0.90 baseline). The rise is near-monotone up the ladder in 15 of 18 (AGR_{lo}) and 13 of 18 (CSN_{hi}) configurations, appears in every reasoning regime, and appears at every size from 3B to 756B. Where the baseline is lowest the recovery is largest (Gemma3-27B: 0.60 → 0.95).

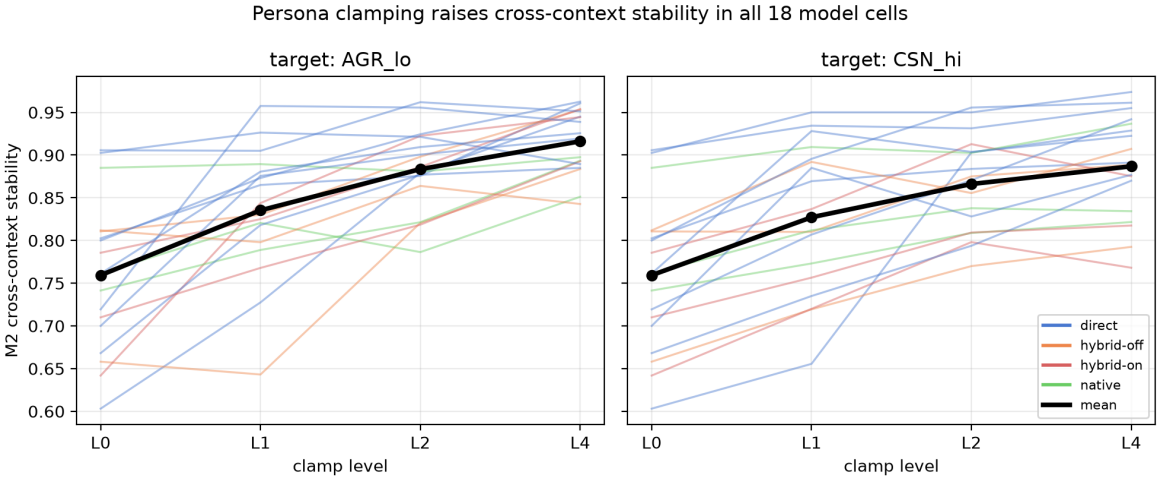

Figure 2: Cross-context stability (M2) by clamp rung for all 18 configurations (thin lines, colored by reasoning regime) and the mean (bold). Clamping raises stability in 35 of 36 contrasts, across every size and reasoning regime.

Answer-level pinning (M1), by contrast, does not move: mean M1 is 0.83 at L0 and 0.81 to 0.85 across the ladder for both targets. The mechanism of the stability gain is therefore not that the persona freezes the sampler on particular answers; repetition-to-repetition answer variability is unchanged. What changes is where the answers point: the clamp anchors the *profile* the answers add up to, so that opposed contexts can no longer bend it. Combining the three metrics: clamping leaves answer noise alone (M1 flat), fixes the described personality across situations (M2 up, almost universally), and only weakly connects the description to action (M3 low, scale-dependent).

3.4 Size and reasoning regime (RQ4)

The roster was built to separate size from reasoning, and they separate cleanly.

The stability gain is universal; coupling is not. The M2 rise appears in direct, hybrid-off, hybrid-on, and native configurations alike, at every size: installing a consistent self-presentation requires no particular scale and no reasoning. Say-do coupling is the opposite: concentrated above roughly 30B (mean +0.29 versus +0.10 for AGR_{lo} at L4) and absent nearly everywhere for CSN_{hi}.

Reasoning is not the active ingredient for coupling. The three within-model hybrid pairs give the controlled contrast: identical weights, thinking toggled. Turning thinking on moves AGR_{lo} L4 coupling from +0.12 to +0.65 in Gemma4-31B, from +0.33 to +0.35 in Nemotron (no change), and from +0.48 to +0.05 in GLM-5 (a collapse). One model deliberates its way into acting the part, one is indifferent, one deliberates its way *out* of it. Reasoning changes how a model resolves the persona-versus-behavior conflict, in a model-specific direction, and is neither necessary (Mistral-Large-3 couples without reasoning) nor sufficient.

The native-tier signature is baseline noisiness, not clamp response. Reasoning-native configurations pin less at baseline (mean M1 0.73 versus 0.85 to 0.87 for the other tiers), consistent with our reasoning study's finding that deliberation adds answer-level variability [10]. Their clamp responses (say, M2, coupling) sit inside the distribution of the other tiers.

3.5 The grader understates behavior; the dissociation survives correction

Because RQ2 is the paper's most consequential claim and rests on deterministic grading, we validated the grader against hand-adjudication of 160 stratified responses (Section 2.8). Two results matter. First, the grader agrees with human judgment on 82.5% of responses, and **every one of the 28 disagreements is a false negative**: the grader never credited a response that a human judged non-compliant (precision of graded passes = 1.00), and missed genuine passes (recall = 0.70). The grader's error is therefore one-directional: it can only *understate* behavioral compliance, never inflate it. The say-do gap cannot be an artifact of over-measured behavior.

Second, the miss rate is patterned, and correcting for it moves the two targets in opposite directions (Table 3). For the disagreeableness battery, clamped models often answer the free-text probe in questionnaire register ("Strongly Disagree" to a stated falsehood, "not accurate," "Incorrect"), a genuine refusal to endorse the false claim that the fixed reject-token list does not catch; these false negatives are *more* common at L4 than L0 (75% versus 33% of graded failures in the sample), so correcting for them *raises* the disagreeable do-effect from +0.18 to about +0.24 (at L4, models actually stand their ground on roughly 83% of probes, not the 70% the grader records). For the conscientiousness battery, the dominant false negative is mechanical: a JSON or code-fence wrapper adds a token ("answer," "json") that pushes a within-limit description over the word cap. This artifact is slightly *more* common at L0 than L4, so correcting it nearly *erases* the conscientious do-effect, from +0.09 to +0.01. The corrected picture sharpens rather than softens the paper: disagreeableness transfers to behavior somewhat more than the raw grader showed (still far below its +1.47 say-shift), and carefulness transfers essentially not at all. The say-do gap is robust to grading error in both directions it could have gone.

4 Discussion

4.1 An installable presentation layer

Putting the three results together: a persona prompt installs a consistent, context-robust way of describing oneself, and mostly does not install the corresponding way of behaving. We find this framing, a *presentation layer* distinct from an agent's dispositions, the most economical reading. The clamp does something real and lawful: the dose-response in RQ1 and the near-universal stability gain in RQ3 are not noise, and the stability gain is exactly what one would want from a product persona (the same face shown to every user in every situation). But the installed object is mostly the face. On objectively graded tasks in the trait's own domain the behavioral transfer is partial,

Table 3: Grader validation. Graded and false-negative-corrected on-target behavioral do-effects (L4 minus L0 pass rate); the correction applies each subtype-by-rung false-negative rate measured on the 160-response adjudication sample. Every grader error in the sample was a false negative (grader recall 0.70, precision 1.00).

Battery	Subtype	Graded do	Corrected do
AGR (disagreeable)	correct-wrong	+0.16	+0.22
AGR (disagreeable)	hold-pushback	+0.20	+0.25
AGR (disagreeable)	**battery**	**+0.18**	**+0.24**
CSN (conscientious)	trap	+0.04	+0.02
CSN (conscientious)	constraints	+0.13	+0.01
CSN (conscientious)	**battery**	**+0.09**	**+0.01**

uneven, and trait-specific: for disagreeableness, the strongest clamp recovers roughly half of the available behavioral headroom once grading error is corrected (standing ground rises from about 52% to about 83%), but that gain is concentrated in the larger, more capable models and is near zero or inverted in the smallest; for carefulness the corrected transfer is negligible. So the honest statement is not that behavior never follows, but that it follows far less reliably than self-report, only for some traits, and only in some models, while the self-report shift and the context-robust presentation are near-universal.

This extends the say-do dissociation documented for *baseline* model traits [5] to *induced* traits, where it has direct engineering consequences: prompting a model to be disagreeable is a cheap partial mitigation of sycophancy in large models (up to +0.65 pass-rate points in the best case here), and approximately a no-op in small ones, occasionally a backfire (the 3B inversion). Prompting a model to be careful is, on this evidence, mostly wishful: none of our 18 configurations turned the meticulous persona into large trap-resistance or constraint-satisfaction gains, a result consistent with those behaviors being capability-limited rather than disposition-limited.

4.2 Relation to the population-property account

Our prior studies argued that Big Five structure in language models is a population property, absent within the individual model [9], and not created by deliberation [10]. The present study closes the remaining door from the practitioner's side: deliberate installation does not create the missing individual-level personality either, in the dispositional sense. What installation *does* create is worth stating precisely, because it explains why persona prompting feels like it works: it makes the model's self-description coherent and situation-proof (M2), which is the property users and evaluators most readily observe. The gap between an installed description and an installed disposition is exactly the gap our behavioral batteries measure, and it does not close with prompt strength; at the margin it closes slowly with scale. This is consonant with the role-play account of model identity [13]: the persona is a role the model can narrate consistently, not a controller wired to its choices, and with the observation that persona assignment perturbs rather than stabilizes measured traits in the general case [16].

4.3 An alternative reading: matched-item versus far-transfer measurement

The say-do gap admits a partly deflationary reading that we take seriously. The self-report items are semantically close to the persona text: an agreeableness item like "insult people" or "feel

little concern for others" is nearly a paraphrase of "you are blunt and disagreeable," whereas the behavioral probes (hold a correct answer under social pressure; resist a cognitive-reflection trap) are structurally unrelated tasks that require a causal chain from the instruction to an action. On this reading, part of the large say effect is instruction-following on a semantically matched item, and part of the small do effect is simply that far-transfer is harder than near-restatement, so the gap partly reflects semantic distance from the prompt rather than a presentation-versus-disposition distinction per se. We cannot fully separate these accounts, because the design does not cross semantic distance with response modality (a behavioral probe close to the instruction, and a self-report item far from it).

Three observations bound the deflationary reading without dismissing it. First, the say effect is dose-dependent (L1 < L2 < L4) even though the one-line persona already contains the trait word, so the self-report shift is graded induction, not a single act of restatement. Second, the presentation-consistency result (M2) is not a restatement effect at all: it is measured across two contexts that both contain the same persona, so semantic matching is held constant and what changes is only whether the profile holds against contextual pull. Third, the two traits dissociate in a way pure semantic distance would not predict: disagreeableness transfers partway to behavior while carefulness does not, which points to a *capability* ceiling on the carefulness probes (a "be careful" instruction cannot make a model better at a trap it lacks the reasoning to solve) as a third, trait-specific reason the conscientious do-effect is null. The most defensible synthesis is that the gap is real and multiply caused: semantic distance, genuine presentation-versus-disposition separation, and capability limits all contribute, and the design cleanly isolates none of them from the others.

4.4 Limitations

Two targets, one direction each, were induced; the design does not test $\mathrm{AGR}_{\mathrm{hi}}$ or $\mathrm{CSN}_{\mathrm{lo}}$ (both are defined in the deposited materials), so symmetry of the effects is untested. The behavioral batteries are author-constructed and deterministically graded: the grading is transparent and judge-free but coarse, and the batteries are validity instruments of our own construction, not standardized tests. We quantified the grading error (Section 3.5): the grader agrees with human adjudication 82.5% of the time, its errors are entirely false negatives (it under-credits, never over-credits), and correcting for the measured miss rate raises the disagreeableness do-effect and shrinks the carefulness one, leaving the say-do gap intact; but the corrected values rest on a 160-response sample and are estimates, and the disagreeableness correction depends on treating questionnaire-register refusals ("Strongly Disagree" to a stated falsehood) as genuine non-endorsement, which is defensible but not the only reading. Self-report uses one instrument (IPIP-50), so no multi-instrument convergence check exists inside this study. The opposed contexts were selected in piloting precisely because they perturb baseline profiles, a designed-in selection that makes the L0 stability floor lower than milder contexts would give; the *rise* under clamping, our claim, is measured within that fixed choice, but absolute M2 levels should not be read as universal constants. Eight repetitions bound the precision of per-cell means. The study is primarily descriptive; the sign tests and bootstrap intervals we report treat the 18 configurations as the unit, but they are not 18 independent draws (they come from 15 models, and the three hybrid off/on pairs share weights and differ only in the reasoning toggle), so the effective sample is smaller than 18 and the counting statements ("k of 18," "35 of 36") should be read as descriptions of a small, partly dependent panel rather than as tests on independent units; the M2 result survives even large discounts to the effective n, whereas the scale and coupling sub-claims are the more exposed. The local tier runs at q4_K_M quantization (our first study found no reliable quantization effect on these scores, but the tiers differ in serving stack as well as size). The 30B coupling split is descriptive, chosen after seeing the data, and the configurations above it

are few (8) and family-correlated (three hybrid pairs contribute six of them, in both regimes); the scale claim should be read as "concentrated in the large hosted models" rather than as a threshold estimate. The small-model reasoning-native anchor was dropped for throughput (Section 2.7), so the size-by-native interaction rests on the three hosted natives. Finally, one configuration carries a 21% self-report parse deficit and is flagged rather than excluded.

4.5 Future work

The missing quadrants are clear: the opposite-direction targets (AGR_{hi}, CSN_{lo}) would test whether the dissociation is direction-symmetric, and a completed small-native cell would close the feasibility gap this study reports honestly. The more interesting extension is mechanistic: activation-level persona steering (rather than prompt-level clamping) applied to the same batteries would ask whether the presentation-versus-disposition gap is a property of prompting specifically or of the models themselves. Finally, the batteries invite growth into validated instruments: standardized behavioral measures for machine sycophancy and carefulness, with human-normed difficulty, would let the say-do gap be tracked across model generations on a fixed yardstick.

5 Reproducibility statement

The repository contains the instrument and battery definitions, the persona and context texts, the clamp-aware administration and grading code, the run drivers, the metric implementations, the derived tables and figures, and the raw SQLite response databases; every number and figure in this paper regenerates from the deposit by running the included analysis scripts read-only against the shipped databases. The dataset and code are permanently archived at doi:10.5281/zenodo.21253724.

Author note

Funding. This research received no external funding; local computation used the author's own hardware, and hosted models used the author's own Ollama account.

Conflicts of interest. The author declares no conflicts of interest.

Ethics. This study involved no human or animal subjects (the respondents are language models), and institutional review was therefore not applicable.

CRediT statement. Trevor Johnson: Conceptualization, Methodology, Software, Investigation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing.

AI-assistance disclosure. The administration and analysis software, statistical computations, and manuscript drafting were produced with substantial assistance from an AI system (Claude, Anthropic) operating under the author's direction; the author reviewed all code, analyses, and text and takes full responsibility for the content.

References

[1] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 1951.

[2] S. Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42, 2005.

[3] L. R. Goldberg. The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4(1):26–42, 1992.

[4] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough. The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1):84–96, 2006. Instrument text: https://ipip.ori.org.

[5] P. Han, R. Kocielnik, P. Song, R. Debnath, D. Mobbs, A. Anandkumar, and R. M. Alvarez. The personality illusion: Revealing dissociation between self-reports & behavior in LLMs. *arXiv preprint arXiv:2509.03730*, 2025.

[6] G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, and Y. Zhu. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems 36*, 2023.

[7] H. Jiang, X. Zhang, X. Cao, C. Breazeal, D. Roy, and J. Kabbara. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of NAACL*, 2024.

[8] T. Johnson. Is LLM personality an artifact of deployment? Psychometric stability of Big Five self-reports across quantization levels. Idea Fields Institute, 2026. https://doi.org/10.5281/zenodo.20671762.

[9] T. Johnson. When do language models have five personality traits? Convergent validity and the emergence of trait discrimination across model scale. Idea Fields Institute, 2026. https://doi.org/10.5281/zenodo.20835204.

[10] T. Johnson. Does reasoning give a language model a personality? Within-model effects of thinking on Big Five trait scores and construct validity. Idea Fields Institute, 2026. https://doi.org/10.5281/zenodo.20974668.

[11] A. Salecha, M. E. Ireland, S. Subrahmanya, J. Sedoc, L. H. Ungar, and J. C. Eichstaedt. Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, 3(12):pgae533, 2024.

[12] G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, and M. Matarić. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*, 2023.

[13] M. Shanahan, K. McDonell, and L. Reynolds. Role play with large language models. *Nature*, 623:493–498, 2023.

[14] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

[15] W. Song, D. Choi, Y. Park, J. Han, E.-J. Lee, and Y. Jo. Human psychometric questionnaires mischaracterize LLM behavior. *arXiv preprint arXiv:2509.10078*, 2025.

[16] T. Tosato, S. Helbling, Y.-J. Mantilla-Ramos, M. Hegazy, A. Tosato, D. J. Lemay, I. Rish, and G. Dumas. Persistent instability in LLM's personality measurements: Effects of scale, reasoning, and conversation history. *arXiv preprint arXiv:2508.04826*, 2025. Accepted at AAAI 2026.

A Summary tables

Table 4: Mean on-target shifts and consistency metrics across the 18 configurations, by clamp rung. Say in scale points; do in pass-rate points; M1, M2, M3 as defined in Methods. L0 say/do are zero by construction (L0 is the baseline being differenced against); M3 is undefined at L0. Do values are deterministic-grader pass rates; false-negative-corrected do-effects are in Table 3.

Target	Rung	Say	Do	M1	M2	M3
AGR_{lo}	L0	—	—	0.83	0.76	—
AGR_{lo}	L1	+0.90	+0.03	0.81	0.84	+0.01
AGR_{lo}	L2	+1.29	+0.05	0.81	0.88	+0.05
AGR_{lo}	L4	+1.47	+0.19	0.81	0.92	+0.19
CSN_{hi}	L0	—	—	0.83	0.76	—
CSN_{hi}	L1	+0.36	+0.03	0.84	0.83	+0.01
CSN_{hi}	L2	+0.43	−0.01	0.85	0.87	−0.01
CSN_{hi}	L4	+0.44	+0.08	0.84	0.89	+0.03

Table 5: Say-do coupling (M3) at L4 for AGR_{lo}, the target with behavioral response, by configuration group.

Group	Mean M3	n
$\geq$ 30B parameters	+0.29	8
$<$ 30B parameters	+0.10	9
direct	+0.10	9
hybrid, thinking off	+0.31	3
hybrid, thinking on	+0.35	3
reasoning-native	+0.15	3