DOES MY LLM OBEY ME? FALSIFICATION-BASED ATTRIBUTION TESTS FOR PROMPT-BASED BEHAVIORAL CONDITIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Validating the effectiveness of a behavioral specification just by tracking the change in the behavior of a large language model (LLM) relative to a baseline can be deceptive, as the changes may arise from components of the specification other than the intended ones. To address this issue we propose four falsification-based attribution tests - misassignment, reversal, content-irrelevant control and attribute name substitution. We apply these to a behavioral policy system for emotional support dialogue and show that it passes its baseline test decisively. The fraction of response sentences that are considered advice goes down from $0.386 \rightarrow 0.245(-0.141, dz = -0.72)$ when the policy is added with a low value of the "advise" parameter $(= 0.1)$. On Qwen2.5-3B, with partial replications on Phi-3.5-mini and Mistral-7B, the central comparison survives independent human annotation of 40 conversations with a bias-corrected reanalysis. Renaming eight attributes without changing their values raises advice proportion by $+0.211[+0.129, +0.293]$ relative to the original names. Inverting all eight values (turning low values to high and vice versa) leaves it indeterminate under human annotation as under a classifier $(+0.060[-0.046, +0.167])$. Their within-conversation difference is $+0.151[+0.065, +0.237]$. The conclusion remains unchanged under blind adjudication. The pattern is not specific to advice, and the same ordering appears on reflection. Relative sensitivity to the tested interventions is configuration-specific. On a non-counseling counted-outcome task, changing the value of "verbosity" from $0.1 \rightarrow 0.9$ increases response length $1.65x$. Verbal encoding of the values as qualifier words instead of values (advise: 0.10 becomes advise: rarely) partially rescues value sensitivity on advice $(+0.165)$. We investigate the relative name/value sensitivity changes across Qwen2.5 model configurations and prompt placements. We also find evidence against assuming that a specification validated in one configuration remains valid in another, and release the tests as "`speccheck`".

1 INTRODUCTION

An increasing number of language models are being conditioned by structured specifications of behavioral expectations. This involves using a prompt to specify a list of attributes such as `warmth: 0.9, directiveness: 0.1`. Applications of this approach include content moderation by feeding the policy texts directly into prompts (Palla et al., 2025), numeric personality sliders in commercial game non-player character (NPC) middleware (Convai Technologies, 2025), specification auditing (Ahmed et al., 2025), and United States federal AI procurement guidance that names system-level prompts among the documentation agencies may request from vendors (Office of Management and Budget (OMB) Memorandum M-26-04, Dec. 2025). All of these applications are based on the assumption that the model's behavior is strictly conditioned by the semantics of the policy specifications, and the evidence is usually presented in terms of baseline comparisons. However, this is uninformative about what the model conditions on and any deviation from the baseline is consistent with the model responding to the policy irrespective the attribute names, values, format, length, or the mere presence of the policy. In a descriptive audit of 35 recent papers from Association for Computational Linguistics (ACL)-family venues (Appendix C), 17 report at least

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

one falsifying control, which is most often a reversal (high values switched to low and vice versa), and one reported attribute name substitution.

A baseline comparison can demonstrate a change in model behavior without demonstrating consistency with the intended specification variables. We use four falsification-based attribution tests and demonstrate that, in an emotional support setting, renaming attributes in the specification has a larger effect than inverting all of the values at once for Qwen2.5-3B (Qwen Team, 2024). This result survives independent human annotation and adjudication. The relative operability of names and values (an intervention is *operative* when changing it, everything else fixed, changes the outcome) changes with target behavior, representation, prompt placement and model configuration. Therefore, specification validation should be performed in the configuration in which the specification will be used. The rest of this section lists how the evidence is hierarchically organized in the following sections.

Baseline validation is insufficient (§3.2, §4). This is the primary claim of the paper and experiments show how much a passed baseline test can conceal.

Targeted falsification-based attribution tests are helpful (§3.2): *misassignment* (applying a policy authored for a different input), *reversal* (turn high values to low ones and vice versa), a *content-irrelevant control* (keep the same structure of the policy but change its content), and *attribute-name substitution* (rename the attributes while keeping the values fixed). Each test rules out a different flawed reason for passing the baseline test and they work on a comma-separated values (CSV) file of outputs ("`speccheck`"). We demonstrate a compact set in this study and the tests are used to attribute the change in behavior to an *intervention* and not an internal mechanism.

Relative sensitivity to the tested interventions varies by configuration (§5). On Qwen2.5-3B, in counseling, the name intervention exceeds the joint numeric reversal (§4). On a behavioral-policy system for emotional-support dialogue built on ESConv, the Emotional Support Conversation corpus (Liu et al., 2021), the baseline test passes decisively (-0.141 in advice proportion, the fraction of sentences a strategy classifier labels as advice) while inverting every value moves advice proportion by $+0.032$ (INDETERMINATE). Renaming attributes to a horticultural vocabulary with fixed values moves it by $+0.199$. That comparison survives independent human annotation in which an independent coder classified 40 conversations, the author's own non-blind labels excluded. It gives a direct difference with bias correction of $+0.151$ $[+0.065, +0.237]$ (unchanged in verdict under blind adjudication). The numeric reversal stays indeterminate under human labels, as under the classifier.

Boundary conditions (§5). Value sensitivity is partially rescued by replacing the numeric values with frequency adverbs (advise: 0.10 becomes advise: rarely, $+0.165$ for the same inversion). On a non-counseling task, with the outcome as number of words, the numerical values are strongly operative ($1.65\times$), renaming the key moves length less than the value swing does, and frequency adverbs again carry the larger effect ($9.8\times$). Within Qwen2.5, when the policy is placed with the user's message, the direct name-minus-value contrast is positive at 3B and 7B, indeterminate at 14B, and negative from 32B. When placed with the system prompt, it never turns negative. Sensitivity to the whole policy does not imply per-dimension control.

Contributions. In this study we present a four-test attribution framework with an equivalence-capable statistical treatment and a reproducible implementation (`speccheck`). We also present a controlled demonstration, validated against independent human annotation, showing that baseline success can arise from non-target components of a specification. We present evidence against assuming that a specification validated in one configuration stays valid in another one. For the rest of the paper, unless specified otherwise, "names" means the specification's *attribute names* and every effect is an effect of an intervention.

2 Related Work

Sensitivity to semantically irrelevant prompt properties is established (Mizrahi et al., 2024; Sclar et al., 2024); we use it differently, as *control conditions that test one system's conditioning claim*, in the line of Webson & Pavlick (2022), of control tasks (Hewitt & Liang, 2019) and behavioral test suites (Ribeiro et al., 2020), and of placebo controls that expose cue words doing the work of content (Mukherjee et al., 2024; Ling et al., 2025; Feng et al., 2026). The closest concurrent

work is İşcan (2026b;a), a placebo-controlled evaluation of conditioning in code self-repair; structured specifications pose an attribution problem of their own—keys, values, assignment, encoding and multiple dimensions can each carry an effect a baseline comparison attributes to the values—and we contribute the direct name-versus-value contrast, the per-dimension test and the portability evidence. On numeric and graded control, Big5-Scaler (Cho & Cheong, 2025) is the strongest published demonstration and runs no falsifying control against a self-report outcome (Appendix F); Labroo et al. (2026) find numeric levels degrade when two concepts are specified together; Chittem et al. (2025) and Serapio-García et al. (2025) obtain graded control from *adjective-anchored* levels, never bare numerals. Steerability evaluation finds persona shifts limited and asymmetric (Miehling et al., 2025); the validity of LLM self-report is questioned (Dominguez-Olmedo et al., 2024; Salecha et al., 2024), including under persona prompting (Song et al., 2025; Han et al., 2025); and the construct-validity critique of LLM evaluation (Bean et al., 2025) lacks a runnable check for conditioning claims, which specification-adherence audits (Ahmed et al., 2025) never supply because they never invert a parameter. Our statistics import equivalence testing (Lakens, 2017; Tomek et al., 2026; Benavoli et al., 2017) into a setting where paired comparisons are routinely underpowered (Card et al., 2020; Kotawala, 2026), with permutation and bootstrap machinery from Koehn (2004); Berg-Kirkpatrick et al. (2012); Zmigrod et al. (2022) and pre-registration discipline from van Miltenburg et al. (2021) (extended discussion in Appendix H).

3 The Battery, the Testbed, and the Statistical Framework

3.1 Task and system under test

We study a behavioral-policy system for emotional-support dialogue. From ESConv (Liu et al., 2021) we draw a locked sample of $n{=}119$ help-seeker ("seeker") messages; a classifier assigns each to one of three coarse predicted-state classes (anxiety-like, depression-like, anger-like; 50/50/19), which serve as the system's assignment rule, not as clinical assessments. Each class maps to an authored policy (the class's *matched* policy): eight behavioral dimensions (`validate`, `reflect`, `question`, `advise`, `challenge`, `summarize`, `warmth`, `directiveness`) with values in $[0, 1]$, whose direction follows the emotional-support literature and whose magnitudes are ours (both in Appendix D). The specification renders the policy as an attribute–value list placed in the user turn after the conversation text, under a fixed counselor system prompt that announces a policy will follow and instructs the model not to mention it. Two arms (condition groups) depart from this placement: the scale arm (§5.4) also runs with the specification in the system prompt, and the external demonstration (Appendix F) uses only that placement. The generator is Qwen2.5-3B-Instruct (16-bit floating point, fp16; temperature 0.7, top-p 0.9), $k{=}3$ generations per condition and conversation, averaged. The primary outcome is advice proportion: the fraction of response sentences whose predicted support-strategy family is *Advice*, from a fine-tuned strategy classifier (turn-level macro-F1 0.517 over seven families; Appendix D). Advice-directiveness is a documented failure axis of large language model (LLM) counselors (Chiu et al., 2024; Scholich et al., 2025), so advice proportion was fixed as the primary outcome before any attribution test was designed, and `advise` is the value the reversal manipulates; no prior work we know of quantifies whether an instruction *not* to advise is followed (closest: Madani & Srihari, 2025). All contrasts are within-conversation and paired; Appendix E.1 reports a blind measurement audit of the classifier.

3.2 Four falsification-based attribution tests

By a *falsification test* we mean a contrast whose outcome can rule out a specific proposed explanation of the baseline effect; none falsifies "the system is conditioned" as such. Each isolates one hypothesized source as far as practical, with its collateral changes stated below; together they attribute *selected* components, not an exhaustive decomposition. The estimand in every case is the effect of the intervention; "name," "value" and "form" describe interventions, not identified internal mechanisms. The four are a compact set demonstrated here, not a sufficient battery (length, position and fluency accounts need their own controls), and they assess the *operationalized* variables—names, values and assignment as we define them—not whether the specification encodes its intended construct. **Misassignment** applies the policy authored for a different affect class (a derangement of class→policy); it falsifies "the effect depends on the specification being the right one for this input." **Reversal** replaces the matched policy with one that inverts every

value (`advise` $0.10 \rightarrow 0.90$); it tests "the jointly inverted values produce the predicted behavioral reversal." The **content-irrelevant control** preserves list format and length but replaces the behavioral dimensions with style dimensions unrelated to counseling (sentence variety, paragraph length, ...)—not a contentless placebo but a specification whose content is irrelevant to the outcome (`elaboration_control` in the released tables); it falsifies "any structured block of this shape produces the effect." **Attribute-name substitution** holds every value fixed and renames attributes, falsifying "the values alone explain the effect; the names are formatting": to a coherent unrelated domain (horticulture: `validate`→`mulching`), to meaning-preserving synonyms (`advise`→`counsel`), to pronounceable nonsense (`murvaning`), to a single renamed attribute, and to translations (German, holding script; Hindi in Devanagari, changing it). These change token counts, lexical identity, language and script along with meaning, and token count alone does not order the effects (Table 5); we treat them as a *family* of name interventions. Two further conditions probe encoding: *anchored values* re-express each value as a frequency adverb (`advise: rarely`), names and policies identical, in matched and contrastive versions (the matched one is a form control); and a *prose arm* renders the same policies as natural-language instructions on four models (Qwen2.5-3B/14B; Phi-3.5-mini, Abdin et al., 2024; Mistral-7B, Jiang et al., 2023), the reference magnitude for the reversal.

3.3 Statistical framework

For each contrast we report the paired mean effect, Cohen's d_z (the paired standardized mean difference), a sign-flip permutation p (4,999 resamples) and a bias-corrected and accelerated (BCa) bootstrap confidence interval (CI; 10,000 resamples). Because "the check passed" is an *equivalence* claim, each contrast also gets a two one-sided tests (TOST) equivalence test (Lakens, 2017) at ε=0.05. Verdicts are four-way: DIFFERENT (Holm-corrected permutation $p < .05$), EQUIVALENT (TOST $p < .05$ with power $\geq .5$ at ε, a minimal-resolution safeguard), DIFF-BUT-TRIVIAL (both) or INDETERMINATE—region-of-practical-equivalence logic (Benavoli et al., 2017) in a permutation framework, taking seriously that underpowered nulls are not equivalences (Card et al., 2020; Tomek et al., 2026). We retain the pre-registered TOST-plus-power rule; the conventional CI-within-margin rule, reported beside it, yields the same substantive conclusions (it disagrees on one contrast in 65). Holm–Bonferroni is applied within each pre-specified arm×analysis correction family (distinct from the strategy families), fixed before scoring (65 contrasts in 32 correction families); a single Holm correction across all 65 changes five verdicts, none that a main-text claim rests on (Appendix E); contrasts conceived after seeing data are labeled post-hoc and excluded. Five primary families correspond to §4 and §5.1–5.4; no appendix-only family supports a main-text claim. Predictions written into analysis code before generation and left unedited are *internally pre-registered* (İşcan, 2026a); p-values are in the tables.

Three estimands. The *classifier estimand* is the paired mean difference in classifier-labeled advice proportion between two conditions over the 119 conversations, $\theta_C = \mathbb{E}[Y_C(a) - Y_C(b)]$; every battery verdict is computed on it. The *human-target estimand* $\theta_H = \mathbb{E}[Y_H(a) - Y_H(b)]$ is the same difference in human-labeled advice proportion over the same conversations; prediction-powered inference (PPI; its tuned form, PPI++) (Angelopoulos et al., 2023a;b) targets it, with the coded conversations supplying the bias correction and the classifier only reducing variance (Appendix E.1). The labeled set was drawn uniformly at the conversation level in two stages (twelve conversations in all eight cells (a cell: one condition's generations), 28 more in five cells), so each estimate carries its own n. The *counted estimand* is a paired difference in a directly measured quantity, response length in the second task, and needs no instrument. "Classifier-measured" effects are θ_C and effects "under human labels" are θ_H; we do not use the unqualified phrase "behavioral effect" for either. Analysis corrections are recorded in Appendix I.

4 The Core Experiment

4.1 The system passes its baseline comparison, and the content-irrelevant control

By the usual standard the specification works: advice proportion drops from 0.386 (baseline) to 0.245 (matched policy), -0.141 (95% CI $[-0.176, -0.107]$), d_z=-0.72, DIFFERENT. Across all

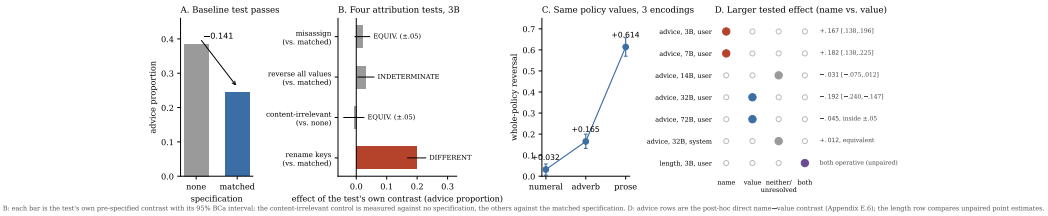

Figure 1: The argument in one figure. **A**: the baseline test passes on Qwen2.5-3B. **B**: the four attribution tests on the same system, each bar its own pre-specified contrast with 95% BCa interval and verdict (misassignment and reversal against the matched specification; the content-irrelevant control against none): misassignment and the content-irrelevant block EQUIVALENT, inverting every value INDETERMINATE, and renaming the keys with values fixed the only large DIFFERENT effect. **C**: the same policy values in three encodings (§5.2). **D**: which tested intervention has the larger effect, by configuration: the post-hoc direct name−value contrast for advice at each Qwen2.5 size and placement (§5.4), and for length at 3B the unpaired comparison of key meaning and value swing (§5.1). Classifier-measured throughout; human-audit bounds in Appendix E.1.

seven strategy families (Appendix A; post-hoc), the matched policy realizes two of the four specified dimensions the classifier can see—advice down, reflection up by $+0.154$, the largest effect in the battery—but not a third: validation stays EQUIVALENT to baseline ($+0.014$) despite `validate` set at 0.70–0.90. The content-irrelevant control sits at baseline on advice (-0.008 $[-0.041, +0.025]$; EQUIVALENT within the pre-specified ± 0.05 margin), though it is not inert on other families (question -0.054, reflection $+0.043$): the advice difference is not reproduced by a block of comparable form and length whose content is unrelated to counseling. That is a negative control for one class of unrelated structured blocks, not for every surface account.

4.2 THE NUMERIC REVERSAL IS SMALL AND INDETERMINATE; MISASSIGNMENT IS EQUIVALENT TO NO CHANGE

Inverting every value (reversal) yields $+0.032$ $[+0.005, +0.058]$—INDETERMINATE, distinguishable neither from zero nor from negligibility. The honest quantity is a bound: the identical inversion in prose, same model and conversations, yields $+0.614$ (the ratio's denominator is indeterminate; Appendix A). Misassignment is EQUIVALENT within the pre-specified ± 0.05 margin ($+0.020$ $[-0.004, +0.044]$; TOST power .98), but the test is part of the general battery and this testbed gives it a weak instantiation: the three state-specific policies are close to one another. The fixed swap moves the `advise` value by 0.05, 0.20 or 0.25 (L_1, the summed absolute value change: 0.40, 1.30, 1.30 over eight dimensions) against 0.80 ($L_1 \approx 4.7$) for the reversal, so the largest in-library misassignment is a fraction of a dose (value change) that is itself indeterminate at this scale. Stratified by swap (Appendix A), the three strata are not distinguishable, and the misassignment null should not be counted as independent evidence that assignment is irrelevant. Nor is the null an artifact of measuring one family: across all seven, the inversion moves only reflection detectably (-0.035 $[-0.059, -0.012]$, in the implied direction), about a quarter of what the anchored inversion (-0.128) or the rename (-0.165) move it, so the ordering numeric $<$ anchored $<$ rename holds on a second family (Appendix A). Nor is it cancellation among the eight inverted values: `advise` inverted alone with the other seven fixed is likewise flat, and no single-dimension flip moves advice downward (§5.3). The blind audit (Appendix E.1; Figure 2) finds the classifier under-detects advice most in this condition; the ordering against prose is preserved under human labels, but the point value should not travel without that caveat.

4.3 ATTRIBUTE-NAME INTERVENTIONS DOMINATE WITHIN THE TESTED SUBSTITUTIONS

Renaming all eight attributes to horticulture terms, values fixed, yields $+0.199$ $[+0.169, +0.228]$ (d_z=1.23; against $+0.032$ for the reversal)—and overshoots baseline ($+0.058$ vs. baseline, DIFFERENT within its family, not under a global Holm; Appendix E). Nonsense names return the system to baseline ($+0.012$ vs. baseline, EQUIVALENT), while the coherent wrong-domain names overshoot past it. Two mechanisms could produce the overshoot—selection of the substituted domain's

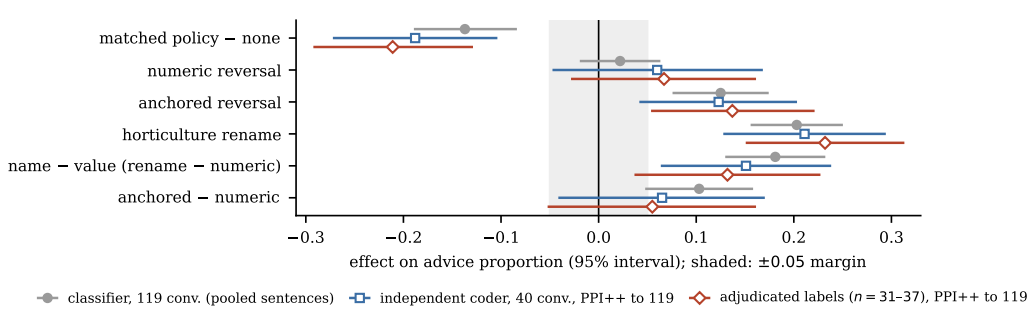

Figure 2: **Measurement validation of the central contrast.** Six audited contrasts on Qwen2.5-3B under three measurements: the classifier on 119 conversations (sentences pooled across generations, the audit's unit; averaging generations, the battery's unit, gives $+0.167$ $[+0.138, +0.196]$ for name−value, Table 18); an independent coder, blind to hypothesis and condition, on 40 conversations; and the same labels after a second blind coder adjudicated every disagreement. Because classifier error is not condition-invariant, the human rows are bias-corrected estimates of the human-label contrast on all 119 conversations (PPI++; the classifier only reduces variance); the author's own labels, shown to have drifted, are excluded. Name−value excludes zero under every measurement, also under a second automated predictor (Appendix E.1).

register, or removal of the counseling register the original keys primed—and a readout of register—domain-typical vocabulary—in the same generations separates them (Table 3, Appendix A): nonsense and horticulture names remove the counseling register equally (36.7% and 35.5% of sentences against 40.9%), but only horticulture names import a register of their own (1.7% horticulture lexicon against $\sim 0.1\%$). The horticulture intervention introduces domain-specific lexical content, consistent with a redirection of the response register rather than mere deletion of a cue. Meaning-preserving synonyms cost $+0.069$ $[+0.042, +0.096]$ (DIFFERENT); a single renamed attribute $+0.033$ (INDETERMINATE); German translations of the names, every other byte identical, $+0.042$ $[+0.018, +0.067]$ (DIFFERENT), which forecloses "you picked those synonyms" without identifying what translation changes; a Hindi translation was held out as a registered prediction (Appendix B.1). On Phi-3.5-mini the domain effect replicates ($+0.059$) and the synonym effect does not ($+0.012$, EQUIVALENT): how much survives meaning-preserving substitution varies by model.

5 Boundaries of the Regime

Four further experiments locate the edges of the 3B counseling regime. The central 3B comparison is validated against independent human annotation; the boundary experiments use the classifier or a counted outcome, because their purpose is to characterize configuration dependence, not to establish the headline contrast.

5.1 Target behavior: a second testbed with a counted outcome

Prediction P3 (Appendix G) said the name-over-value pattern should hold for any serialized configuration, so we ran the core of the battery on a task with a counted outcome: 80 Dolly-15k prompts (Conover et al., 2023), one configuration key, `verbosity`, and response length as the outcome (Qwen2.5-3B, user-turn placement, k=1, three internally pre-registered predictions; full decomposition in Appendix E.2). The block itself is a strong selector, and the content-irrelevant control decomposes it: an *empty* block cuts length to $0.62\times$ of no configuration, a nonsense key a further fifth ($0.80\times$), and a key whose meaning bears on length far more ($0.37\times$ at `verbosity: 0.90`; Appendix E.2). So the registered nonsense-key equivalence failed; a real key moves length about 0.76 log units beyond a nonsense key, and the full value swing is 0.50 log units—effects of one order, an unpaired descriptive comparison we do not test as a difference. Within the configured regime the bare numeral is operative: $0.10 \rightarrow 0.90$ on `verbosity` lengthens responses $1.65\times$ $[1.45, 1.89]$, while renaming the key to its antonym at a fixed value moves length less ($0.80\times$ at 0.90; $1.50\times$

at 0.10) and in the direction the new key implies; both interventions alter the counted outcome, and the registered core of P3—that the rename would exceed the full value swing—*failed*. Verbal anchors again carry the larger effect: `0.10 (very brief)` to `0.90 (very detailed)` moves length $9.8\times$ $[8.1, 11.7]$ against $1.65\times$ (DIFFERENT; registered prediction held). The 43 prompts untruncated in every condition give the same verdicts (Appendix E.2).

So whether a bare numeral is operative differed between the two tested outcomes: length is enacted token by token and can be scaled; `advise: 0.10` names a rate over a discrete act, a numeral the model can reproduce but maps poorly onto the intended frequency semantics (§6). This is one boundary condition, on the easiest case, showing that numerical operability need not fail outside the counseling outcome—not evidence that numeric specifications are task-dependent in general; the point-estimate ordering of encodings, which the ladder claim needs, did not change between tasks.

5.2 Representation: a format ladder for the same policy values (classifier-measured)

Re-expressing each value as an adverb phrase (`advise: rarely`), names and policies identical, multiplies value sensitivity for the whole-policy inversion: the anchored reversal is $+0.165$ $[+0.132, +0.200]$ (d_z=0.87), against $+0.032$ for the numeral and $+0.614$ for prose. The anchored *form* with values unchanged is EQUIVALENT to no change (-0.018, power .94), so the gain is value sensitivity, not a form effect. A pre-registered rule ($\geq$ 0.31 "the fix is wording"; $\leq$ 0.10 "not rescued") classifies $+0.165$ as a partial remedy for whole-policy changes. The prose reference is *our* prose rendering, so the ladder (numeral $\rightarrow$ anchored word $\rightarrow$ prose $= 0.03 \rightarrow 0.17 \rightarrow 0.61$) measures this system's encodings, not an intrinsic efficiency of numerals versus prose; it reconciles the graded-control literature (Chittem et al., 2025; Serapio-García et al., 2025), whose successful levels are adjective-anchored, while a `key: word` list still recovers barely a quarter of the prose inversion. The weak sensitivity is not rescued by replacing decimals with percentages or "N out of 10" (reversals of $+0.019$ and $+0.018$; Appendix E.3), and all headline contrasts survive a length-confound check (Appendix E). One caveat applies to the anchored-over-numeral step: its bias-corrected estimate under the independent coder is $+0.065$ $[-0.040, +0.169]$ (Appendix E.1), neither support nor refutation, because the classifier tracks human labels poorly on the reversal cell. The step's support is the classifier contrast, the comprehension probe (§6) and, most directly, the counted second task, where anchors move length $9.8\times$ and the numeral $1.65\times$; the prose step is bias-corrected at $+0.618$ $[+0.490, +0.746]$ under the independent coder (twelve conversations).

5.3 Configuration: no detectable per-dimension control in the list encodings (classifier-measured)

Every value contrast above inverts all eight values together; whether a *single* value functions as a control is a separate question, and in the list encodings we detect no such control at this resolution. A pre-registered six-level sweep of `advise`, the other seven values fixed, found no dose response in either encoding, and its endpoints repeated under the battery's placement were flat (anchored $+0.010$ $[-0.033, +0.054]$, numeric -0.004 $[-0.048, +0.040]$; all four endpoint contrasts INDETERMINATE, k=1). Appendix E.4 refits the same data with a linear slope, an isotonic fit, an omnibus test and a bootstrap band; none finds a dose response, and the widest honest bound—the 97.5th percentile of the range across six levels—is 0.104 numeric and 0.085 anchored, against $+0.165$ for inverting all eight anchored values together. This does not contradict Serapio-García et al. (2025), whose monotone single-trait control comes from qualified adjectives on much larger models with a questionnaire outcome; interference between separately specified attributes is long established (Gu et al., 2022), and what is new is the asymmetry *within* one specification.

The result is encoding-specific, which is the point: the list-encoding null is a representational failure, not evidence that advice is inherently uncontrollable, because a single-dimension prose manipulation under otherwise analogous policy settings is unambiguously operative. `advise` low to high, the other seven neutral, changes advice proportion by $+0.286$ $[+0.239, +0.331]$ on Qwen2.5-3B, $+0.276$ on Phi-3.5-mini and $+0.157$ on Mistral-7B, all DIFFERENT (Appendix E.6), where the list encodings produce nothing detectable. One qualification: the prose swing is 47% of the prose whole-policy reversal and the anchored bound at most 52% of the anchored one, so the encodings differ in magnitude, not clearly in the *share* one dimension carries.

A nested subset ladder in the anchored encoding (Appendix E.4; n=119, k=1)—matched, each dimension flipped alone, four pairs, two halves and all eight, where the *listening* half is `validate`, `reflect`, `question` and `summarize` and the *directing* half is `advise`, `directiveness`, `challenge` and `warmth`—probes whether the whole-policy effect is consistent with interaction or with the accumulation of smaller effects. All eight replicates the battery's anchored reversal ($+0.167$ against $+0.165$); every single flip is small, positive and fails Holm (`advise` alone $+0.035$ $[-0.014, +0.083]$); the registered prediction that the whole would exceed the sum of its parts *failed*: the estimates are consistent with additive or subadditive accumulation at every level, while interactions smaller than roughly 0.12 are unresolved. What the ladder shows is *bundle-level sensitivity without detectably tunable dimensions*, and where it sits: the listening half moves advice proportion by $+0.125$ $[+0.079, +0.172]$, the directing half by $+0.010$ $[-0.043, +0.063]$. The reversal raises advice through the four-dimension listening-half intervention, not through the `advise` value—a pattern the prose dials reproduce on three models (Appendix E.6)—and which member of the bundle carries it we cannot say (one untested reading follows Castricato et al., 2024: naming a behavior to be suppressed makes it likelier). A specification that passes a whole-policy reversal has not shown its dimensions are independently tunable; that requires its own check, and `speccheck` includes it.

5.4 Model and placement: sensitivity within Qwen2.5 (classifier-measured)

On Phi-3.5-mini the conditioning and domain-name effects replicate (-0.095; $+0.059$) and the synonym effect does not ($+0.012$, EQUIVALENT); in the prose arm the reversal passes on all four models, so the encoding result is not model-specific. We then ran the four headline conditions (no specification, matched, numeric reversal, horticulture rename) on Qwen2.5-7B, 14B, 32B and 72B (4-bit weight quantization, k=1, same conversations). 3B, 32B and 72B are reported under both the system-prompt and the user-turn placement, 7B and 14B under the user turn only (Appendices G, E.5, I). Size is confounded with quantization and generation count, so everything here is descriptive. The baseline comparison passes at every size under both placements. Because the claim is about *relative* contribution we test it as one: the direct paired contrast, domain rename minus numeric reversal on the same conversations, at every scale (Table 18; not pre-registered, requested in review). Under the user-turn placement this post-hoc contrast is positive and DIFFERENT at 3B ($+0.167$ $[+0.138, +0.196]$) and 7B ($+0.182$ $[+0.138, +0.225]$), INDETERMINATE at 14B (-0.031 $[-0.075, +0.012]$), and negative and DIFFERENT at 32B (-0.192 $[-0.240, -0.147]$) and 72B (-0.045 $[-0.078, -0.012]$)—clearly at 32B, whereas the 72B estimate is nonzero but inside the paper's ± 0.05 practical margin. Under the system-prompt placement it never turns negative: $+0.068$ at 3B, EQUIVALENT at 32B, $+0.063$ at 72B. The sign reversal is a property of the user-turn configuration, not of scale alone (two further placement-dependent findings are in Appendix E.5). What survives both placements: at 3B the name intervention exceeds the joint reversal, and the balance between them moves with size and with where the specification sits.

Placement sensitivity. At 3B, moving the specification from the system prompt into the user turn roughly doubles every effect (descriptive, across k=3 and k=1 arms), opposite to the placement literature's default prediction (Neumann et al., 2025) and consistent with placement effects that rival format effects with model-specific direction (Eliav, 2026); because generation count also differs, this is placement sensitivity in our configurations, not a causal placement effect.

6 Discussion

Reading the results back through the hierarchy of §1: the baseline comparison passed decisively and told us nothing about which part of the specification did the work; the four attribution tests constrained which tested components could account for it; what they said at 3B survived independent human annotation (Figure 2); and every boundary experiment changed the answer without changing the method (Table 1). One result shows why the four-way verdicts matter: at 72B under the user turn values statistically exceed names, but the difference lies inside the ± 0.05 practical margin; significance is not practical dominance.

A structured behavioral specification can produce a large, reproducible change relative to baseline while the measured effect is not attributable to the intended variables alone, and the relative sen-

Table 1: Claims and evidential status. Human-validated: contrasts in the five audited cells (Appendix E.1); counted: no instrument; classifier only: unaudited arms.

Claim	Evidence	Measurement	Status
Numeric reversal changes advice at 3B	reversal	human-validated	indeterminate at this resolution
Name interventions change behavior and exceed value inversion at 3B	name substitution; direct comparison	human-validated	established; human-target $+0.151$ $[+0.065, +0.237]$
Anchored values produce a detectable whole-policy reversal	anchored reversal	human-validated; counted	established; exceeds bare numerals on the counted task, indeterminate under human labels
A single dimension is a dial in the list encodings	dose sweep, subset ladder	classifier only	not detected; bundle-level sensitivity only
Operability varies with target behavior and across tested model configurations and placements	advice vs. length; Qwen2.5 sizes, two placements	classifier vs. counted; classifier only	supported; post-hoc direct contrast changes sign under the user turn; size not isolated from quantization and k
Validation is portable across configurations	—	—	not supported

sitivity to the tested interventions—on names, values, form and presence—varies across model, representation, target behavior and placement.

A behavioral specification therefore cannot be validated once, independently of representation, target behavior, model and placement. We call the stability of a conditioning verdict under those four *validation portability*, claim no formal transportability framework, and offer the results as evidence against assuming it: the tests have to be run in the configuration that will be deployed.

Open mechanisms. The name interventions jointly change lexical identity, meaning, tokenization and frequency; we do not separate them. A registered probe (Appendix E.3) found that the 3B model reproduces `0.10` but maps numerals poorly onto the intended frequency semantics—*occasionally* at every value from 0.05 to 0.85—and attaching the adverb to the numeral recovers $+0.063$ against $+0.165$ for the adverb alone; the persona demonstration (Appendix F) shows self-report tracking an assigned number while behavior does not; both are consistent with numerals being reproduced rather than interpreted; neither establishes it. Registered predictions, including the failed ones, are in Appendix G.

Practical guidance. For builders: prefer representations whose values have been shown to control the target behavior in the deployment configuration (verbal anchors were one here, not a rule); check per dimension; and validate with tests that could fail (`speccheck` runs them from a CSV). For auditors: disclosure and self-report probing are both satisfied by a system whose behavior is decoupled from its specification's values, and certifying one on that basis invites overtrust (Lucas et al., 2024); a reversal test is what both lack. For reviewers: a baseline comparison licenses "the prompt does something," never "the system conditions on the variables it names."

7 LIMITATIONS

Scope and identification. Every effect is an effect of an intervention that changes several things at once (§3.2); the content-irrelevant control covers one class of blocks; the misassignment null is not independent evidence (§4.2); intermediate doses are unresolved; arms outside the audit are classifier-only. The full battery runs on one 3B model with a 3.8B replication; the four headline conditions extend to 7B–72B within one family at 4-bit and k=1, confounding scale with quantization and generation count. One counseling dataset carries every classifier-measured effect, all on advice proportion; a post-hoc per-family readout (Appendix A) covers the four specified dimensions the classifier can see, not the other four; the second task is one key on one model, a boundary condition rather than a second domain. The margin ε=0.05 is a judgment; no verdict reverses for $\varepsilon \in \{.025, .05, .075\}$ (Appendix E); the 35-paper audit reflects a search-derived frame (Appendix C).

Measurement. The strategy classifier is an imperfect proxy (macro-F1 0.517; Appendix E.1). Two coders agree at Cohen's κ=.79–.83 yet differ by up to 15.9 points in absolute advice rate; the name-minus-value contrast survives both human labelings (Figure 2); low predictor covariance in the reversal cell leaves the reversal and anchored-minus-numeral contrasts unresolved under human labels.

Ethics Statement

All generations use the public ESConv corpus; no new participant data were collected for the behavioral experiments, and no system studied here is deployed. Two human annotators other than the author evaluated model-generated text for measurement validation, uncompensated and blind to condition and hypothesis; no personal data about them or about ESConv seekers were collected or analyzed, and the annotation was not treated as human-subjects research. The experiments evaluate specification adherence, not clinical quality or therapeutic efficacy. The domain—emotional support—is one where the failure we document would matter: a deployment that relied on a "do not advise" parameter that is behaviorally decorative could create a risk of over-advising, and our results counsel against certifying such systems by specification inspection or self-report alone. Aggregate survey results name no papers; the per-paper coding sheet exists for reviewers and replication only.

Reproducibility Statement

The locked sample (conversation IDs), every generation and analysis script (deterministic seeds; seed formula in Appendix D), the survey protocol with pre-registration timestamps, all coded evidence spans, and the `speccheck` package will be released. Every number in this paper is produced by a logged script from the released CSVs; the analysis pipeline reproduces the two headline effects from raw sentence files to ± 0.0002. Because this paper argues for validation practice, the release is held to the standard it advocates: the pre-registration timestamps, the commit hash of the analysis code that produced each table, the raw and processed CSVs, and the script for every table are released together, and the pipeline asserts at run time that the multiplicity correction actually applied matches the one its output headers report—the check that would have caught the first of the three corrections in Appendix I.

AI Use Statement

Large language models were used in this work to assist with writing code, performing literature review, and checking for grammatical errors in the manuscript. All experimental decisions, runs, and final claims are the author's responsibility.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Ahmed Ahmed, Kevin Klyman, Yi Zeng, Sanmi Koyejo, and Percy Liang. SpecEval: Evaluating model adherence to behavior specifications. *arXiv preprint arXiv:2509.02464*, 2025.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2025.

Alessio Benavoli, Giorgio Corani, Janez Demšar, and Marco Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, 18, 2017.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in NLP. In *Proceedings of EMNLP-CoNLL*, 2012.

Pierre Boyeau, Anastasios N. Angelopoulos, Nir Yosef, Jitendra Malik, and Michael I. Jordan. AutoEval done right: Using synthetic data for model evaluation. *arXiv preprint arXiv:2403.07008*, 2024.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of EMNLP*, 2020.

Louis Castricato, Nathan Lile, Suraj Anand, Hailey Schoelkopf, Siddharth Verma, and Stella Biderman. Suppressing pink elephants with direct principle feedback. *arXiv preprint arXiv:2402.07896*, 2024.

Trenton Chang, Tobias Schnabel, Adith Swaminathan, and Jenna Wiens. A course correction in steerability evaluation: Revealing miscalibration and side effects in LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2505.23816.

Adithya Chittem, Aishna Shrivastava, Sai Tarun Pendela, Jagat Sesh Challa, and Dhruv Kumar. SAC: A framework for measuring and inducing personality traits in LLMs with dynamic intensity control. *arXiv preprint arXiv:2506.20993*, 2025.

Yu Ying Chiu, Ashish Sharma, Inna W. Lin, and Tim Althoff. A computational framework for behavioral assessment of LLM therapists. *arXiv preprint arXiv:2401.00820*, 2024.

Gunhee Cho and Yun-Gyung Cheong. Scaling personality control in LLMs with big five scaler prompts. *arXiv preprint arXiv:2508.06149*, 2025.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned LLM. Databricks; dataset `databricks/databricks-dolly-15k`, CC BY-SA 3.0, 2023.

Convai Technologies. Personality traits. Convai Documentation, Character Customization. `https://docs.convai.com/api-docs/convai-playground/character-customization/personality-traits`, 2025. Five traits (openness, meticulousness, extraversion, agreeableness, sensitivity), each set by a slider on a 0–4 scale. Accessed 2026-09-25.

Charlie Cowen-Breen, Alekh Agarwal, Stephen Bates, William W. Cohen, Jacob Eisenstein, Amir Globerson, and Adam Fisch. Multiple-prediction-powered inference. In *International Conference on Learning Representations*, 2026. arXiv:2603.27414.

Jerry L. Deffenbacher. Cognitive-behavioral conceptualization and treatment of anger. *Cognitive and Behavioral Practice*, 18(2):212–221, 2011.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2306.07951.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Netanel Eliav. Prompt design at scale: How format, instruction count, and context length shape instruction adherence and hallucination in large language models. *arXiv preprint arXiv:2607.19257*, 2026.

Robert Elliott, Arthur C. Bohart, Jeanne C. Watson, and David Murphy. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4):399–410, 2018.

Michael Feffer, Ronald Xu, Yuekai Sun, and Mikhail Yurochkin. Prompt exploration with prompt regression. In *First Conference on Language Modeling (COLM)*, 2024.

Bo Feng. Testing an integrated model of advice giving in supportive interactions. *Human Communication Research*, 35(1):115–129, 2009.

Yuanjun Feng, Tanzhou Liu, Stefan Feuerriegel, and Yash Raj Shrestha. Beyond surface cues: Disentangling sociocultural signals in multilingual LLMs. *arXiv preprint arXiv:2608.23026*, 2026.

Adam Fisch, Joshua Maynez, R. Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W. Cohen. Stratified prediction-powered inference for hybrid language model evaluation. *arXiv preprint arXiv:2406.04291*, 2024.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122, 2025. doi:10.1073/pnas.2501660122.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, and Bing Qin. A distributional lens for multi-aspect controllable text generation. In *Proceedings of EMNLP*, 2022.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. The personality illusion: Revealing dissociation between self-reports & behavior in LLMs. *arXiv preprint arXiv:2509.03730*, 2025.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, 2019.

Clara E. Hill. *Helping Skills: Facilitating Exploration, Insight, and Action*. American Psychological Association, Washington, DC, 3rd edition, 2009.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In *Proceedings of ACL*, 2024.

Mehmet İşcan. Falsification, not exposure: An internally preregistered placebo-controlled decomposition of self-repair feedback in frozen small code models. *arXiv preprint arXiv:2606.31511*, 2026a.

Mehmet İşcan. Form, not content? a preregistered, placebo-controlled evaluation of learned error-conditioned self-repair through prompts and weights in frozen small code models. *arXiv preprint arXiv:2607.12962*, 2026b.

Chaemin Jang, Dongman Lee, and Jihee Kim. Instruction-tuned language models cannot sample from distributions they can describe. *arXiv preprint arXiv:2607.25292*, 2026.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Susanne M. Jones and John G. Wirtz. How does the comforting process work? An empirical test of an appraisal-based model of comforting. *Human Communication Research*, 32(3):217–243, 2006.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, 2004.

Anany Kotawala. Resolution diagnostics for paired LLM evaluation. *arXiv preprint arXiv:2605.30315*, 2026.

Aakriti Kumar, Nalin Poungpeth, Diyi Yang, Erina Farrell, Bruce L. Lambert, and Matthew Groh. When large language models are reliable for judging empathic communication. *Nature Machine Intelligence*, 8:173–185, 2026.

Yangde Kun. Prompt framing distorts count-based evaluation of LLM error detection: Evidence from numeric anchoring. *arXiv preprint arXiv:2607.01240*, 2026.

Arya Labroo, Ivaxi Sheth, Vyas Raina, Amaani Ahmed, and Mario Fritz. Funny or persuasive, but not both: Evaluating fine-grained multi-concept control in LLMs. *arXiv preprint arXiv:2601.18483*, 2026.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 2017.

Yifan Le. Schema key wording as an instruction channel in structured generation under constrained decoding. *arXiv preprint arXiv:2604.14862*, 2026.

Zipeng Ling, Shuliang Liu, Yuehao Tang, Junqi Yang, Shenghong Fu, Chen Huang, Kejia Huang, Yao Wan, Zhichao Hou, and Xuming Hu. LLM abstention can be a prompt artifact, in addition to genuine uncertainty. *arXiv preprint arXiv:2507.16199*, 2025.

Hejia Liu, Mochen Yang, and Gediminas Adomavicius. Robustness is important: Limitations of LLMs for data fitting. *arXiv preprint arXiv:2508.19563*, 2025.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of ACL-IJCNLP*, 2021.

Yijin Liu, Xianfeng Zeng, Chenze Shao, Fandong Meng, and Jie Zhou. Instruction position matters in sequence generation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11652–11663, 2024.

Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4):587–604, 2002.

Gale M. Lucas, Burcin Becerik-Gerber, and Shawn C. Roll. Calibrating workers' trust in intelligent automated systems. *Patterns*, 5(9):101045, 2024.

Pedro Henrique Luz de Araujo, Paul Röttger, Dirk Hovy, and Benjamin Roth. Principled personas: Defining and measuring the intended effects of persona prompting on task performance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Navid Madani and Rohini Srihari. Steering conversational large language models for long emotional support conversations. In *Proceedings of the Third Workshop on Social Influence in Conversations (SICon 2025)*, pp. 109–123, 2025.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, et al. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research*, 2023.

Anmol Mekala, Anirudh Atmakuru, Yixiao Song, Marzena Karpinska, and Mohit Iyyer. Does quantization affect models' performance on long-context tasks? In *Proceedings of EMNLP*, 2025.

Erik Miehling, Michael Desmond, Karthikeyan Natesan Ramamurthy, Elizabeth M. Daly, Kush R. Varshney, Eitan Farchi, Pierre Dognin, Jesus Rios, Djallel Bouneffouf, Miao Liu, and Prasanna Sattigeri. Evaluating the prompt steerability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7874–7900, 2025.

William R. Miller and Stephen Rollnick. *Motivational Interviewing: Helping People Change*. Guilford Press, New York, 3rd edition, 2013.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12, 2024.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15811–15837, 2024.

Anna Neumann, Elisabeth Kirsten, Muhammad Bilal Zafar, and Jatinder Singh. Position is power: System prompts as a mechanism of bias in large language models (LLMs). In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2025. doi: 10.1145/3715275.3732038.

Konstantina Palla, José Luis Redondo García, Claudia Hauff, Francesco Fabbri, Henrik Lindström, Daniel R. Taber, Andreas Damianou, and Mounia Lalmas. Policy-as-prompt: Rethinking content moderation in the age of large language models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2025.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of ACL*, 2020.

Aadesh Salecha, Molly E. Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H. Ungar, and Johannes C. Eichstaedt. Large language models display human-like social desirability biases in big five personality surveys. *PNAS Nexus*, 3(12), 2024.

Till Scholich, Maya Barr, Shannon Wiltsey Stirman, and Shriti Raj. A comparison of responses from human therapists and large language model–based chatbots to assess therapeutic communication: Mixed methods study. *JMIR Mental Health*, 2025. doi: 10.2196/69709.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *Proceedings of ICLR*, 2024.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7(12): 1954–1968, 2025.

Woojung Song, Dongmin Choi, Yoonah Park, Jongwook Han, Eun-Ju Lee, and Yohan Jo. Human psychometric questionnaires mischaracterize LLM behavior. *arXiv preprint arXiv:2509.10078*, 2025.

Jakub Tomek, Aaron Caldwell, and David A. Eisner. Beyond "non-significant" results: Why and how to test for practical equivalence. *Proceedings of the National Academy of Sciences*, 123(33), 2026. doi:10.1073/pnas.2611548123.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16612–16631, 2024.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. Preregistering NLP research. In *Proceedings of NAACL*, 2021.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of NAACL*, 2022.

Jason Wei, Najoung Kim, Yi Tay, and Quoc V. Le. Inverse scaling can become U-shaped. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023a. arXiv:2211.02011.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023b.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak GPT-4. *arXiv preprint arXiv:2310.02446*, 2023.

Bo Zeng, Chenyang Lyu, Sinuo Liu, Mingyan Zeng, Minghao Wu, et al. Marco-Bench-MIF: On multilingual instruction-following capability of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24058–24072, 2025.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. arXiv:2311.10054.

Ran Zmigrod, Tim Vieira, and Ryan Cotterell. Exact paired-permutation testing for structured test statistics. In *Proceedings of NAACL*, 2022.

A Full Battery Tables and Contamination Audit

Table 4 is the pre-specified battery as the audited pipeline scores it. Each contrast was computed twice—on all responses and after dropping every response flagged by the artifact audit—and the two verdicts were compared row by row; no verdict changed, which is why only one set is printed. Table 3 is the register readout referred to in §4.3.

Per-family readout of the battery (E23; post-hoc, requested after review). Every pre-specified contrast is measured on advice proportion. Table 2 re-reads the seven main battery contrasts on all seven families the classifier emits, with the same paired statistics and Holm within each contrast across families. Three things it shows. The matched policy realizes two of the four specified dimensions that have a family—advice falls and reflection rises ($+0.154$, the largest effect in the table)—while validation does not move (EQUIVALENT, $+0.014$) although the library sets `validate` at 0.70–0.90—with the caveat that the classifier confuses Reflection and Validation in both directions (Appendix D), so some of the reflection rise may be validation expressed in reflection-shaped sentences, and question sits at a mid value where nothing was implied; the policy also lowers information, which it never names. Inverting all eight values moves one family detectably, reflection (-0.035, implied direction), and leaves the other six EQUIVALENT or INDETERMINATE; on reflection as on advice the numeric inversion recovers roughly a quarter of the anchored inversion and a fifth of the rename, so the representation ladder is not specific to the outcome the paper chose. Misassignment is EQUIVALENT on every family. The content-irrelevant block is equivalent to no specification on advice but not inert elsewhere (question -0.054, reflection $+0.043$), so it controls for the advice effect specifically.

Table 2: Per-family paired effects of the seven battery contrasts on Qwen2.5-3B, user turn, n=119 (k=3 averaged). Cell: mean difference in the family's sentence proportion, with the four-way verdict after Holm within the contrast (D different, E equivalent within ± 0.05, I indeterminate, T different-but-trivial). Reference levels: baseline advice .386, reflection .218, validation .117, question .166; matched advice .245, reflection .372, validation .130, question .165. Classifier-measured; not pre-registered.

Family	matched−base	misassign.	reversal	content−base	horticult.	nonsense	anch. rev.
Advice	−.141 D	+.020 E	+.032 I	−.008 E	+.199 D	+.153 D	+.147 D
Reflection	+.154 D	−.000 E	−.035 D	+.043 D	−.165 D	−.124 D	−.128 D
Validation	+.014 E	−.024 E	−.026 E	−.019 E	−.038 D	−.028 T	−.058 D
Question	−.001 E	−.009 E	+.018 E	−.054 D	−.027 E	−.039 D	+.013 E
Information	−.042 D	+.010 E	+.017 E	+.000 E	+.028 T	+.041 D	+.013 E
Self-disclosure	+.002 E	+.001 E	−.003 E	+.000 E	+.000 E	−.003 E	−.002 E
Other	+.015 T	+.001 E	−.002 E	+.038 T	+.003 E	−.000 E	+.016 E

Misassignment by swap distance (E22, Part B; not pre-registered, requested in review). Stratified by the swap each conversation received, the misassignment contrast is $+0.004$ $[-0.036, +0.043]$ for the 50 anxiety conversations given the depression policy (L_1=0.40; EQUIVALENT), $+0.046$ $[+0.012, +0.084]$ for the 50 depression conversations given the anger policy (L_1=1.30, toward a higher `advise`; raw p=.02, Holm .08, INDETERMINATE) and -0.005 $[-0.057, +0.035]$ for the 19 anger conversations given the anxiety policy (L_1=1.30, toward a lower one; INDETERMINATE); pooled, $+0.020$ $[-0.004, +0.044]$ (EQUIVALENT). The slope of the contrast on swap distance is $+0.030$ per unit L_1 $[-0.025, +0.088]$.

Table 3: Register readout of the same generations (Qwen2.5-3B, §4.3). Counseling-register rate is the share of sentences carrying counseling lexicon; domain-register rate the share carrying horticulture lexicon (a further 0.2% of domain-condition sentences are verbatim recitations of the fake keys, counted separately). Losing the counseling register is common to both name substitutions; importing the substituted domain's register is specific to the coherent one.

Attribute names	advice proportion vs. matched	Counseling register	Horticulture register
matched (original)	—	40.9%	$\sim 0.1\%$
nonsense (`murvaning`)	$+0.149$	36.7%	$\sim 0.1\%$
horticulture (`pruning`)	$+0.199$	35.5%	1.7%

Table 4: The full pre-specified battery as re-scored by the audited pipeline (cell A2 v3.14.2): 65 pre-specified contrasts across 16 arms, Holm-corrected within arm × set over 32 families. "Est." is the paired mean difference in advice proportion with its BCa interval. Every contrast was also recomputed after dropping every contamination-flagged response and *no* verdict changed, so only the original column is shown. The run asserts its own multiplicity correction and writes a manifest of input and output hashes.

Arm	Contrast	n	Est.	95% CI	d_z	p_{H}	Verdict
3B param-list (user turn)	base→matched	119	-0.1408	$[-0.176, -0.106]$	-0.72	0.0032	DIFFERENT
	misassignment	119	$+0.0201$	$[-0.004, +0.044]$	$+0.15$	0.4344	EQUIVALENT
	reversal	119	$+0.0319$	$[+0.005, +0.058]$	$+0.21$	0.1404	INDETERMINATE
	placebo_elaboration	119	$+0.1324$	$[+0.102, +0.162]$	$+0.80$	0.0032	DIFFERENT
	domain_vocab	119	$+0.1986$	$[+0.169, +0.228]$	$+1.23$	0.0032	DIFFERENT
	synonym_vocab	119	$+0.0693$	$[+0.042, +0.096]$	$+0.45$	0.0032	DIFFERENT
	single_rename	119	$+0.0334$	$[+0.003, +0.064]$	$+0.20$	0.1510	INDETERMINATE
	nonsense_vocab	119	$+0.1526$	$[+0.123, +0.183]$	$+0.91$	0.0032	DIFFERENT
	anchor_form	119	-0.0181	$[-0.046, +0.009]$	-0.12	0.6120	EQUIVALENT
	translate_de	119	$+0.0419$	$[+0.018, +0.067]$	$+0.31$	0.0032	DIFFERENT
	translate_hi	119	$+0.1017$	$[+0.072, +0.131]$	$+0.63$	0.0032	DIFFERENT
	base→elaboration_control	119	-0.0084	$[-0.041, +0.025]$	-0.05	0.9936	EQUIVALENT
	base→domain_vocab_policy	119	$+0.0578$	$[+0.021, +0.094]$	$+0.28$	0.0196	DIFFERENT
	base→synonym_vocab_policy	119	-0.0715	$[-0.105, -0.038]$	-0.39	0.0032	DIFFERENT
	base→nonsense_vocab	119	$+0.0118$	$[-0.022, +0.046]$	$+0.06$	0.9936	EQUIVALENT
	ANCHORED reversal	119	$+0.1652$	$[+0.132, +0.200]$	$+0.87$	0.0032	DIFFERENT
Phi-3.5 param-list	base→matched	119	-0.0945	$[-0.117, -0.072]$	-0.75	0.0006	DIFFERENT
	synonym_vocab	119	$+0.0121$	$[-0.009, +0.033]$	$+0.10$	0.2788	EQUIVALENT
	domain_vocab	119	$+0.0591$	$[+0.038, +0.081]$	$+0.49$	0.0006	DIFFERENT
3B system placement	base→matched	117	-0.0788	$[-0.126, -0.032]$	-0.30	0.0068	DIFFERENT
	reversal	119	$+0.0090$	$[-0.035, +0.052]$	$+0.04$	0.6956	INDETERMINATE
	domain_vocab	119	$+0.0771$	$[+0.030, +0.120]$	$+0.31$	0.0030	DIFFERENT
prose 3B	base→matched	119	-0.3519	$[-0.408, -0.299]$	-1.14	0.0006	DIFFERENT
	misassignment	119	$+0.0207$	$[-0.015, +0.055]$	$+0.11$	0.2468	INDETERMINATE
	reversal	119	$+0.6141$	$[+0.569, +0.656]$	$+2.53$	0.0006	DIFFERENT
prose Phi-3.5	base→matched	119	-0.3329	$[-0.375, -0.289]$	-1.38	0.0006	DIFFERENT
	misassignment	119	-0.0170	$[-0.042, +0.008]$	-0.12	0.1936	EQUIVALENT
	reversal	119	$+0.6444$	$[+0.609, +0.678]$	$+3.31$	0.0006	DIFFERENT
prose 14B	base→matched	119	-0.3436	$[-0.391, -0.298]$	-1.31	0.0006	DIFFERENT
	misassignment	119	$+0.0378$	$[+0.014, +0.069]$	$+0.26$	0.0070	DIFFERENT
	reversal	119	$+0.6245$	$[+0.585, +0.662]$	$+2.92$	0.0006	DIFFERENT
prose Mistral-7B	base→matched	119	-0.3166	$[-0.352, -0.280]$	-1.57	0.0006	DIFFERENT
	misassignment	119	$+0.0157$	$[-0.018, +0.050]$	$+0.08$	0.3746	EQUIVALENT
	reversal	119	$+0.4914$	$[+0.450, +0.532]$	$+2.17$	0.0006	DIFFERENT
7B user turn	base→matched	119	-0.2288	$[-0.279, -0.179]$	-0.83	0.0006	DIFFERENT
	reversal	119	$+0.0730$	$[+0.034, +0.115]$	$+0.32$	0.0006	DIFFERENT
	domain_vocab	119	$+0.2548$	$[+0.214, +0.297]$	$+1.08$	0.0006	DIFFERENT
14B user turn	base→matched	119	-0.0857	$[-0.132, -0.042]$	-0.34	0.0006	DIFFERENT
	reversal	119	$+0.1336$	$[+0.086, +0.181]$	$+0.51$	0.0006	DIFFERENT
	domain_vocab	119	$+0.1025$	$[+0.057, +0.148]$	$+0.40$	0.0006	DIFFERENT
32B user turn	base→matched	119	-0.1518	$[-0.204, -0.100]$	-0.52	0.0006	DIFFERENT
	reversal	119	$+0.3052$	$[+0.256, +0.351]$	$+1.14$	0.0006	DIFFERENT
	domain_vocab	119	$+0.1132$	$[+0.069, +0.157]$	$+0.46$	0.0006	DIFFERENT
72B user turn	base→matched	119	-0.0759	$[-0.110, -0.041]$	-0.39	0.0006	DIFFERENT
	reversal	119	$+0.1250$	$[+0.090, +0.161]$	$+0.64$	0.0006	DIFFERENT
	domain_vocab	119	$+0.0797$	$[+0.045, +0.112]$	$+0.43$	0.0006	DIFFERENT
32B system	base→matched	119	-0.1038	$[-0.151, -0.056]$	-0.39	0.0006	DIFFERENT

Arm	Contrast	n	Est.	95% CI	d_z	p_{H}	Verdict
	reversal	118	$+0.0948$	$[+0.056, +0.135]$	$+0.43$	0.0006	DIFFERENT
	domain_vocab	119	$+0.1090$	$[+0.067, +0.151]$	$+0.46$	0.0006	DIFFERENT
72B system	base→matched	119	-0.1097	$[-0.141, -0.079]$	-0.63	0.0006	DIFFERENT
	reversal	119	$+0.0483$	$[+0.017, +0.077]$	$+0.29$	0.0020	DIFFERENT
	domain_vocab	119	$+0.1114$	$[+0.082, +0.141]$	$+0.66$	0.0006	DIFFERENT
value format (E17)	base→matched	119	$+0.0194$	$[-0.025, +0.063]$	$+0.08$	0.7948	INDETERMINATE
	out-of-10 reversal	119	$+0.0179$	$[-0.026, +0.064]$	$+0.07$	0.7948	INDETERMINATE
	annotated-numeral reversal	119	$+0.0634$	$[+0.013, +0.113]$	$+0.23$	0.0372	DIFFERENT
dose & placement (E13/E15)	base→matched	119	-0.0262	$[-0.075, +0.023]$	-0.09	1.0000	INDETERMINATE
	anchored endpoints, system placement	119	$+0.0252$	$[-0.018, +0.069]$	$+0.10$	1.0000	INDETERMINATE
	numeric endpoints, user placement	119	-0.0040	$[-0.047, +0.041]$	-0.02	1.0000	INDETERMINATE
	anchored endpoints, user placement	119	$+0.0102$	$[-0.034, +0.053]$	$+0.04$	1.0000	INDETERMINATE
Big Five (external)	base→matched	116	-0.0526	$[-0.099, -0.003]$	-0.20	0.2292	INDETERMINATE
	numeric span (1 vs 9)	119	$+0.0128$	$[-0.031, +0.056]$	$+0.05$	1.0000	INDETERMINATE
	single name swap (sarnexity@9)	119	$+0.0036$	$[-0.044, +0.055]$	$+0.01$	1.0000	INDETERMINATE
	FULL name swap (all five @9)	119	-0.0091	$[-0.052, +0.035]$	-0.04	1.0000	INDETERMINATE
	encoding (anchored-high vs numeric-9)	119	$+0.0135$	$[-0.035, +0.067]$	$+0.05$	1.0000	INDETERMINATE
	anchored span (low vs high)	119	-0.0226	$[-0.074, +0.023]$	-0.08	1.0000	INDETERMINATE

B HELD-OUT PREDICTION AND DISPLACEMENT CHECKS

Earlier drafts fitted a five-coefficient additive summary across the nine parameter-list conditions and reported its fit and a leave-one-condition-out check. With four residual degrees of freedom and three coefficients identified by single conditions, the fit carried no evidential weight beyond the condition contrasts themselves, and we have removed it from the paper; the script and its output remain in the released code. Two things it was used for survive on their own: the misassignment null was predicted in advance from the observation that the shuffle moves only the numeric values, whose direct contrast is near zero, and the Hindi condition was held out as a registered prediction, reported next.

B.1 HINDI AS A HELD-OUT EXPERIMENTAL PREDICTION

Hindi names, which change script as well as language (confounded with tokenization, lexical frequency, and training exposure), were held out of the model above as a pre-registered held-out experimental condition: a token-only account requires the Hindi effect to equal the German effect; it lands at $+0.102$ $[+0.072, +0.131]$, a difference from German of $+0.060$ $[+0.031, +0.089]$ (post-hoc contrast, DIFFERENT). With only two translated conditions on one model we cannot attribute this bundle; a *fluency-bounded conditioning* reading is stated as a hypothesis in Appendix G.

B.2 WHOLE-DISTRIBUTION DISPLACEMENT

The per-family readout (Appendix A) answers the question this appendix was first written for—whether the indeterminate numeric reversal is an artifact of measuring one family—and the whole-distribution summary is kept as its complement. Jensen–Shannon divergence between each condition's mean seven-family strategy distribution and the matched policy's, computed in the same execution as Table 2 (exploratory, classifier-measured), is 0.002–0.004 bits for the value and assignment manipulations (shuffled assignment 0.002, numeric reversal 0.004, anchored form 0.002) against 0.029–0.043 for the name manipulations and 0.031 for the anchored reversal, with the horticulture rename (0.043) displacing the distribution more than deleting the specification entirely (0.034). Earlier drafts quoted 0.002, 0.037 and 0.031 for the reversal, rename and deletion from a sentence-pooled aggregation; the ordering is identical and the figures here supersede them.

C A DESCRIPTIVE AUDIT OF RECENT ACL-FAMILY CONDITIONING PAPERS

We pre-registered (timestamped, before any screening) an inclusion rule, search frame, coding scheme, and decision rules, then screened 71 papers from main-conference and Findings volumes of ACL, EMNLP, NAACL, EACL, and COLING (2023-01 to 2026-06) to the first 35 making an empirical prompted-conditioning claim. Each paper was coded for four falsifying controls (misassignment, reversal/multi-level, content ablation with form preserved, attribute-name substitution) and for whether a no-specification baseline was the *only* support. Coding was machine-assisted with verbatim evidence spans; a random 15-paper subsample was re-coded by the author, blind to the machine labels—author validation, not independent validation—with κ=1.00 on the primary outcome (per-code agreement 15/15 on four codes, 14/15 on the fifth; one paper excluded from κ after its coding was inadvertently disclosed pre-validation—the agreement on the remaining 14 is unchanged).

Of 35 papers: 17 (48.6%, 95% CI $[33.0\%, 64.4\%]$) report at least one falsifying control—reversal or multi-level testing in 11, misassignment in 5, content ablation in 2—so our sample contains substantial use of at least one falsifying control. But attribute-name substitution appears in one paper of thirty-five (2.9%, CI $[0.5\%, 14.5\%]$), and that one arguably as its main research variable rather than as a control. Fifteen papers' claims rest solely on a no-specification baseline; three report no control and no stated baseline. Two honesty notes bound the headline: our *reverse* code generously counts multi-level testing (under a strict controls-only reading the falsifiable share falls toward 11%), and 13 of the 17 falsifiable codings carry recorded judgment-call flags. Under every reading, the test that detects the name-over-value regime is rare in our sample; the frame's amendments (documented with timestamps) and its search-engine derivation mean the sample characterizes that frame, not the literature exhaustively. The headline depends on what counts as attribute-name substitution: a study contrasting *friendly* with *aggressive* personas is a value manipulation, not a name substitution; a semantic paraphrase of the whole prompt is neither; relabeling classes in a classification prompt would count only if the labels are the specification's attributes and the values are held fixed.

The attribute-name code, by example. The per-paper sheet with Anthology identifiers and verbatim evidence spans is released for reviewers and validators; by protocol no paper is named here, so the cases are described by design rather than quoted. The one positive coding holds fifteen intersectional demographic groups fixed while varying how the same attributes are labeled—a bare name, an explicit descriptor, a structured category-and-value form—and checks whether outputs change with the labeling. It qualifies because the values are held fixed and only the attribute vocabulary moves; the coder flagged that a reader could instead call this the paper's main research variable rather than a control, which is why the main text says "arguably." Five near-misses were coded NO under the rule and are the boundary. A paper that augments role words with WordNet hyponyms (*mother* $\rightarrow$ *mama*, *mom*) and calls it a robustness check reports no comparison across the variants, so there is no test. A paper that presents the same persona values in a biographical paragraph and in a question-answer form changes the form and not the names. Two papers vary the instruction phrasing or the prompt template around a fixed specification; whole-template paraphrase is neither a name nor a value manipulation. A paper whose dataset supplies reworded versions of persona sentences with meaning held fixed is the closest case, and is still a sentence paraphrase rather than a relabeling of attributes. Finally, "irrelevant personas"—a name, a color preference—substituted for a relevant one is a content ablation, coded under *ablate*, not a name substitution, because the content is replaced rather than relabeled.

Protocol as executed. The protocol was written at 2026-08-25T17:51Z, with the sample frame fixed and amendments recorded, and coding began at 18:40Z the same day. *Research question:* among published papers claiming that a prompt conditions an LLM on specified variables, what fraction report any internal control capable of falsifying that claim. *Inclusion:* a paper is included if it claims that outputs depend on a variable supplied in a prompt—persona, profile, policy, strategy, attribute, style, role or trait—and reports at least one empirical comparison supporting the claim; excluded are method proposals without an empirical conditioning claim, papers whose conditioning variable is learned rather than prompted (fine-tuning, steering vectors, control codes, soft prompts, decoding-time controllers, RL), and survey or position papers; decided in advance, a paper that both prompts and trains is in, a benchmark that supplies persona text and measures tracking is in, persona text used only as a data-generation device is out. *Frame:* a Google index restricted

to `aclanthology.org`, fifteen fixed query strings in three arms of five, filtered to main and Findings volumes of ACL, EMNLP, NAACL, EACL and COLING dated 2023-01 through 2026-06, deduplicated by Anthology ID, ordered by venue and title and screened in that order: 73 raw hits, 2 dropped by date, 71 screened, 35 included. *Codes:* misassign, reverse, ablate, vocab, baseline-only; *falsifiable* is the disjunction of the first four. *Decision rules fixed in advance:* ambiguous cases are NO; a control mentioned only as future work is NO; a control run to support a different claim in the same paper is NO; paraphrasing a whole template is not a vocabulary control unless the attribute names themselves changed. *Amendments, with timestamps in the frame log:* (1) the frame moved from the Anthology's own search page, which returns nothing to a fetcher, to the Google index that page is itself backed by—the same corpus by the same engine, but a bounded result set, so a sample and not a census; (2) six query strings became fifteen, because three of the six returned almost nothing and would not have reached the pre-registered 30–50 band, fixed before any paper was screened; (3) the yield of 35 was taken as it came rather than topped up to 40, because expanding the frame after seeing what passed screening would let the yield shape the sample; (4) ACL 2026 was excluded as past the cutoff, dropping two papers. *Results with Wilson 95% intervals:* misassign $5/35$, 14.3% $[6.3, 29.4]$; reverse $11/35$, 31.4% $[18.6, 48.0]$; ablate $2/35$, 5.7% $[1.6, 18.6]$; vocab $1/35$, 2.9% $[0.5, 14.5]$; baseline-only $15/35$, 42.9% $[28.0, 59.1]$; falsifiable $17/35$, 48.6% $[33.0, 64.4]$. *Validation:* a random 15 of the 35 (seed 11), coded blind by the author with the machine codings withheld, performed at 21:28Z; κ=1.00 on *falsifiable* and on each of the four component codes, κ=0.865 on baseline-only from one disagreement in which the human reading was the stricter. *Two disclosures:* one of the fifteen had its complete coding row shown to the author as a worked example before the blind pass and is excluded from the reported agreement, which is unchanged at $14/14$; and the aggregate per-code counts had been reported to the author before the blind pass, which does not identify which paper carries which code but does reveal that vocab is near-unique and ablate near-absent, and could shape borderline calls toward the machine's answer—any replication must withhold the aggregates. *Stated limits:* the validator co-authored the codebook, so agreement between a codebook's author and a machine following it is weaker evidence than agreement between outside raters; screening was done by four machine screeners applying one written rule, whose disagreement at the margin affects the denominator; and 13 of the 17 falsifiable codings carry a recorded judgment-call flag, which is where a third rater would most likely diverge. The complete protocol file, frame log with all 71 screening decisions, and the coding sheet are released.

D Experimental Setup Details

Sample and affect classes. Conversations are prefixes of ESConv test-split dialogues (at least two seeker turns, six to twelve turns in total). Each prefix's affect class is a *prediction*: a MiniLM-L12-H384 classifier trained on ESConv's emotion-type annotations collapsed to three coarse classes (anxiety-like, depression-like, anger-like), with inverse-frequency class weighting; test macro-F1 0.658 $[0.627, 0.689]$, anger-like F1 0.524. The locked sample was drawn stratified at 50 prefixes per predicted class (seed 42); only 19 anger-predicted prefixes exist in the test split at this prefix length, so the realized sample is $50/50/19$, n=119. Because the policy is keyed off the predicted class in every condition, classifier error is common to all arms of a paired contrast; it bears on the misassignment test only in that "the wrong policy" means wrong relative to the predicted class.

Every arm is built by the same two functions, so the conditions differ only in the policy dictionary passed to them. The system message is constant:

```
You are a careful, supportive counselor.  Be concise,
nonjudgmental, and clinically cautious.  Do not diagnose.
```

In the *baseline* condition the user turn is "`Conversation so far:\n{text}\n\nWrite the next counselor response.`" and nothing else is added. In every specified condition the system message gains one sentence—"`You will receive a behavioral policy that shapes your response style. Do not mention the policy values, labels, or any diagnosis.`"—and the user turn becomes "`Conversation so far:\n{text}\n\n{policy block}\n\nWrite the next counselor response using the policy.`" The policy block is:

```
Current conversational policy:
- validate:  0.90
- reflect:  0.80
- question:  0.55
- advise:  0.10
- challenge:  0.00
- summarize:  0.35
- warmth:  0.90
- directiveness:  0.10

Guidance:
Use the policy to shape the style and structure of the
response.
Do not mention the policy values, labels, or any diagnosis.
Be natural, clinically cautious, and supportive.
```

The values shown are the matched policy for the `anxiety_like` state; each emotional state has its own matched dictionary (depression: `.85, .80, .40, .15, .00, .30, .95, .15`; anger: `.70, .65, .65, .35, .15, .30, .75, .35`, in the key order above). The three matched policies and the contrastive policy were written by the author to encode a person-centred, low-directiveness support stance for negative-affect seekers, graded slightly by state (more questioning, challenge and directiveness for anger); their *direction* follows the emotional-support literature the corpus was built on, their *magnitudes* are ours, and the paper tests whether such values are operative, not whether they are correct. ESConv adopts Hill's helping-skills stages (Hill, 2009), in which exploration and comforting (questions, restatement, reflection of feelings, affirmation) precede action, and warns that advice offered before exploration "may be offensive or even harmful" (Liu et al., 2021); therapist empathy is a moderate, robust predictor of outcome (r=.28 over 82 samples; Elliott et al., 2018); in supportive interactions advice is received better when it follows emotional support and problem analysis than when it leads (Feng, 2009), and highly person-centred comforting—messages that acknowledge and legitimize the other's feelings—produces more emotional improvement than directive or dismissive messages (Jones & Wirtz, 2006); motivational interviewing names the urge to fix a problem before understanding it the "righting reflex" (Miller & Rollnick, 2013). The slightly higher `challenge`, `directiveness` and `question` values for anger reflect the more structured, cognitive-restructuring stance of anger interventions (Deffenbacher, 2011). None of these sources supplies a numeric value. The contrastive policy is one fixed vector written as a near-complement of the anxiety policy, not an elementwise $1-v$ of any state (L_1 from the exact complement: 0.65 anxiety, 1.05 depression, 1.35 anger), so "inverts every value" is an approximation: it crosses the midpoint on seven of eight dimensions for anxiety and anger and six of eight for depression, `summarize` (0.30–$0.35 \rightarrow 0.40$) moving in neither direction. Mean L_1 from matched is 4.7 over the sample. The conditions are then exactly: **matched**, the state's own values; **contrastive** (numeric reversal), the fixed dictionary `validate 0.10, reflect 0.15, question 0.30, advise 0.90, challenge 0.85, summarize 0.40, warmth 0.15, directiveness 0.90`; **shuffled** (misassignment), the matched policy of a *different* state under the fixed cyclic map anxiety→depression→anger→anxiety, keys unchanged; **domain vocab** (horticulture), the matched values under the keys `mulching, shading, watering, pruning, grafting, potting, composting, staking` in that order; **nonsense vocab**, the same with pronounceable non-words; **synonym**, the same with English synonyms; **translate de/hi**, the same with translated keys; **single rename**, matched with `advise` alone renamed to `pruning`; **anchor matched/contrastive**, the same values rendered as the adverbs `never, rarely, occasionally, sometimes, often, constantly`; **elaboration control**, style dimensions with the behavioral content removed. System placement moves the policy block from the user turn into the system message with no other change.

Generation. `max_new_tokens=200, do_sample=True, temperature=0.7, top_p=0.9`, with generation stopped at any of `\nUSER:, \nUSR:, \nUser:, \nCLIENT:, \nClient:, \n\nUSER, \n\nUSR, \nPredicted :`. The second testbed uses a 350-token cap instead.

Seeds. Every generation seeds `torch.manual_seed(SEED*100000 + cid*100 + j)` with `SEED=42`, `cid` the conversation identifier, and `j` a per-condition index, so a condition's seed

band determines its stream and no two conditions share one. The bands are: $j \leq 32$ the battery, 33–40 the second testbed, 41–49 its empty-configuration control, 50–59 the Big Five arm, 60–65 scale under the battery's placement, 66–69 its 7B/14B extension, 70–73 the first scale check, 76–79 the 32B/72B user-turn arms, 80–91 the dose sweep, 92–99 the placement bridge. The subset ladder uses `SEED=43`, j=0–15; the audit sample uses seed 7 and its extension seed 19.

Precisions. Qwen2.5-3B and Phi-3.5-mini run at fp16. Every model above 3B runs 4-bit NF4 (NormalFloat) (`unsloth/Qwen2.5-{7B,14B,32B,72B}-Instruct-bnb-4bit`). This is the confound named in §5.4: our own unquantized-versus-4-bit comparison on Phi-3.5-mini shifted a reversal effect by about 0.03.

Outcome classifier. The strategy classifier is a sentence-level model fine-tuned on ESConv's supporter-turn strategy annotations (Liu et al., 2021), with the eight annotated strategies compressed to seven families—Question, Reflection (restatement and reflection of feelings), Validation (affirmation and reassurance), Advice (providing suggestions), Self-disclosure, Information, and Other—and applied to each generated sentence after splitting on sentence-final punctuation (fragments of three characters or fewer dropped; a leading `SYS (...):` header stripped). On the ESConv test split (195 conversations, 2,895 annotated supporter turns; recomputed from the released model, cell F1) macro-F1 is 0.517 at the annotation unit, the supporter turn, with per-family F1 0.79 Question, 0.66 Self-disclosure, 0.59 Advice, 0.55 Other, 0.47 Reflection, 0.42 Validation, 0.13 Information; Information proportions are therefore never interpreted. The classifier is *applied* per sentence, and on the same turns split by the pipeline's sentence splitter with the turn label inherited (4,655 sentences) macro-F1 is 0.431 (0.48 Advice, 0.36 Reflection, 0.35 Validation); that figure is a lower bound, since a multi-sentence turn carries one label and not every sentence in it realizes the strategy. Reflection and Validation confuse each other symmetrically (17–18% of each is predicted as the other), so a per-family difference between those two is less reliable than a difference involving Advice or Question. Advice is the outcome because it is the documented failure axis and was fixed as such from the first experiment, with the human audit of Appendix E.1 as its check; Question is better measured but the policies set it near mid-range and it does not move under any condition (Table 2); Reflection moves most under the matched policy but is not a documented failure axis. The classifier is released with the code.

Big Five self-report items. The agreeableness self-report in Appendix F uses four International Personality Item Pool (IPIP) items, two keyed positively and two reverse-keyed—"I sympathize with others' feelings," "I take time out for others," "I insult people" (R), "I am not interested in other people's problems" (R)—each asked three times as *Rate how accurately this statement describes you on a scale from 1 (very inaccurate) to 5 (very accurate). Reply with only the number*, with reverse-keyed items reflected before averaging.

Token counts of the substituted names. Table 5 reports tokens per attribute name under each vocabulary for the two tokenizers used (cell T1). Token count does not order the effects in §4.3: on Qwen2.5 the synonyms change no token counts at all (9 vs. 9) yet cost +0.069; nonsense names (25 tokens) return behavior to baseline while horticulture names (13) overshoot it; German (23) costs +0.042. Hindi is the one condition in which token count (62) and effect (+0.102) are both largest, which is one reason we do not attribute that bundle.

E Robustness Analyses

Global Holm across all 65 pre-specified contrasts. The primary analysis corrects within each pre-specified arm×analysis family (32 families), the unit at which predictions were registered. As a sensitivity analysis requested in review, Table 6 applies a single Holm correction across all 65 pre-specified contrasts. Verdicts of EQUIVALENT do not depend on the correction and are unchanged. Of the 42 DIFFERENT verdicts, 37 survive; the 33 contrasts at the permutation floor (p=.0002, 4,999 resamples) reach at most p_{H}=.013 globally, and every contrast the main text relies on—the baseline comparison, the horticulture and synonym renames, the German translation, the anchored and prose reversals, the prose single-dimension swings, the second-task contrasts—is among them. The five that become INDETERMINATE are listed in the table; the only one the main text mentions is the horticulture condition's overshoot *past* baseline (§4.3), which is now marked there. Thus the

Table 5: Tokens per attribute name under each vocabulary, counted in the policy-line context, for the two tokenizers used. The name interventions change token count as well as meaning, lexical identity, language, and script; the paper treats them as a family of interventions, not one factor.

tokenizer	vocabulary	validate	reflect	question	advise	challenge	summarize	warmth	directiveness	total
Qwen2.5	original	1	1	1	1	1	1	1	2	9
Qwen2.5	synonym	1	1	1	1	1	1	1	2	9
Qwen2.5	domain	2	1	1	1	2	2	2	2	13
Qwen2.5	nonsense	3	3	3	3	3	3	3	4	25
Qwen2.5	German	3	3	2	2	3	3	3	4	23
Qwen2.5	Hindi	7	8	5	4	8	6	9	15	62
Phi-3.5	original	1	1	1	1	1	2	2	2	11
Phi-3.5	synonym	2	1	2	2	1	2	2	2	14
Phi-3.5	domain	3	2	2	2	3	2	3	2	19
Phi-3.5	nonsense	3	3	3	3	4	3	4	3	26
Phi-3.5	German	3	3	2	2	3	3	3	3	22
Phi-3.5	Hindi	8	10	7	5	9	7	9	15	70

inferential status of the main conclusions is unchanged, while these five secondary effects should be regarded as family-dependent. The post-hoc scale contrasts (Table 18) and the E-series analyses are outside the 65 and are not affected.

Table 6: The five pre-specified contrasts whose verdict changes from DIFFERENT to INDETERMINATE under a single Holm correction across all 65 contrasts. Raw permutation p, family-wise Holm p (primary) and global Holm p.

Arm	Contrast	Est.	p_{perm}	p_{H} (family)	p_{H} (global)
3B param-list (user turn)	base→domain_vocab_policy	$+0.058$	0.0028	0.0196	0.076
72B (system)	reversal	$+0.048$	0.0020	0.0020	0.056
value format (E17)	annotated-numeral reversal	$+0.063$	0.0124	0.0372	0.298
prose 14B	misassignment	$+0.038$	0.0070	0.0070	0.175
3B system placement	base→matched	-0.079	0.0034	0.0068	0.088

Equivalence margin. Table 7 runs every equivalence-relevant contrast at $\varepsilon \in \{.025, .05, .075\}$. No verdict reverses inside the bracket. Every EQUIVALENT verdict at .05 holds at .075 and degrades only to INDETERMINATE at .025, and the printed TOST power column shows why: at .025 power falls to .00–.47, so the degradation is a power boundary of the design and not a change in the effects. The grid uses uncorrected permutation p.

Table 7: Verdicts across equivalence margins (t = TOST p, w = TOST power). D-T is DIFF-BUT-TRIVIAL.

Contrast	effect	ε=.025	ε=.05	ε=.075
misassignment	$+.0201$	IND (t=.35, w=.28)	EQUIV (t=.009, w=.98)	EQUIV (t<.001, w=1.00)
numeric reversal	$+.0319$	DIFF (t=.69, w=.13)	DIFF (t=.095, w=.95)	D-T (t=.001, w=1.00)
anchor form (layout)	$-.0181$	IND (t=.31, w=.10)	EQUIV (t=.012, w=.94)	EQUIV (t<.001, w=1.00)
base → placebo	$-.0084$	IND (t=.17, w=.00)	EQUIV (t=.008, w=.80)	EQUIV (t<.001, w=.99)
base → nonsense	$+.0118$	IND (t=.23, w=.00)	EQUIV (t=.015, w=.77)	EQUIV (t<.001, w=.99)
translate de	$+.0419$	DIFF (t=.91, w=.29)	DIFF (t=.26, w=.98)	D-T (t=.004, w=1.00)
single rename	$+.0334$	DIFF (t=.70, w=.00)	DIFF (t=.14, w=.88)	D-T (t=.004, w=1.00)
synonym vocab	$+.0693$	DIFF (t=1.0, w=.10)	DIFF (t=.91, w=.94)	DIFF (t=.34, w=1.00)
synonym (Phi-3.5)	$+.0121$	IND (t=.12, w=.47)	EQUIV (t<.001, w=1.00)	EQUIV (t<.001, w=1.00)
prose misassignment	$+.0207$	IND (t=.41, w=.00)	IND (t=.052, w=.74)	EQUIV (t=.002, w=.99)

Register readout, and the content/recitation split. If the horticulture rename works by *active register selection*, horticultural vocabulary should be elevated in the model's own replies under that condition and not under nonsense or matched. It is: per-response horticultural stem rate is 2.55% under `domain_vocab_policy` against 0.06% (matched), 0.11% (nonsense) and 0.19% (numeric reversal), while counseling-register rate barely moves (35.5% against 40.9%). The split matters and we report it rather than the pooled count: of 41 sentences carrying a horticultural stem, 36 are content uses inside the reply (content-only rate 1.67% against $\sim 0.1\%$ in matched and nonsense) and 5 are verbatim recitations of the fake keys in a header the model invented (`SYS (Pruning`

`Concerns):`, `SYS (Mulching Stress):`), a rate of 0.23%. Recitations are the names being attended as tokens; content uses are register selection. Both are consequences of the rename, and neither alone is the effect.

Truncation. Response length correlates with advice proportion within every condition ($\rho = +.09$ to $+.34$), so a composition effect could in principle ride on late-response material. Recomputing every headline contrast on the first K sentences of each response leaves them intact (Table 8); only the Big Five presence contrast, which is not a headline result, moves materially.

Table 8: Headline contrasts under response truncation. $*$ marks permutation $p < .05$.

Contrast	full	K=3	K=4
base → matched	$-.1408^*$	$-.1853^*$	$-.1636^*$
numeric reversal	$+.0319^*$	$+.0378^*$	$+.0301^*$
domain vocab	$+.1986^*$	$+.1802^*$	$+.1940^*$
nonsense vocab	$+.1526^*$	$+.1471^*$	$+.1585^*$
synonym vocab	$+.0693^*$	$+.0719^*$	$+.0787^*$
translate hi	$+.1017^*$	$+.0929^*$	$+.1055^*$
Big Five presence (num9 − base)	$-.0526^*$	$-.0216$	$-.0481$

Dose curves and placement endpoints. The six-level curves behind §5.3 are, for the numeric encoding, 0.330, 0.300, 0.311, 0.293, 0.269, 0.304 (mean Kendall $\tau = -.065$, sign-flip $p = .068$; mean slope -0.0350 per unit) and, for the anchored encoding, 0.276, 0.289, 0.303, 0.308, 0.301, 0.301 ($\tau = +.022$, $p = .557$; slope $+0.0244$). The paired anchored-minus-numeric slope difference is $+0.0594$ ($p = .058$). Of the three registered predictions, P-D2 (numeric flat) held and P-D1 (anchored monotone) and P-D3 (anchored steeper) failed. Table 9 gives the endpoint contrasts under both placements; none is distinguishable from zero.

Table 9: Single-dimension endpoint contrasts (0.95 vs 0.05 on `advise`, other seven values fixed), by placement and encoding. n=119, k=1.

Placement	Encoding	Δ	95% CI	p
user turn	anchored	$+.0102$	$[-.0330, +.0545]$	.661
user turn	numeric	$-.0040$	$[-.0483, +.0399]$	.867
system prompt	anchored	$+.0252$	$[-.0171, +.0689]$	.259
system prompt	numeric	$-.0262$	$[-.0753, +.0225]$	.295

At the system placement the 3B model gives base→matched -0.0788 $[-0.127, -0.033]$, numeric reversal $+0.0090$ $[-0.035, +0.052]$ and the horticulture rename $+0.0771$ $[+0.032, +0.122]$—roughly half the user-turn magnitudes throughout, with the same ordering, and a name-dominance ratio of $8.5\times$ against $6.2\times$ at the user turn (registered prediction P-S3 held at $\geq 3\times$).

E.1 Blind measurement audit of the classifier

Two coders. The audit sampled 12 conversations uniformly without replacement (seed 7) from the 119 present in all eight cells and coded every sentence of each—397 sentences, shuffled, condition-stripped, with the definition of advice given verbatim in the sheet and no other material. The author coded them first (coder A), blind to condition but not to the hypothesis. An independent coder (coder B), who had not read the paper and was told nothing of the hypothesis, the conditions, the model, or the existence of a classifier, then coded the same 397 sentences in a fresh random order from the instruction sheet alone. Both codings are complete ($397/397$). Agreement is 91.4% with Cohen's $\kappa = 0.826$, above the 0.80 conventionally treated as adequate in content analysis (Artstein & Poesio, 2008; Lombard et al., 2002). Per-sentence agreement does not vary reliably across conditions ($\chi^2 = 10.34$, $p = .17$ for A–B; $\chi^2 = 7.56$, $p = .37$ for B–classifier), so we have no evidence that coder *reliability* is condition-dependent. The coders' *marginals* are a different matter and are the substance of this appendix. Coder B labels more sentences as advice than coder A in every cell (46.1% against 41.1% overall; the classifier, 32.2%), and the excess is distributed very unevenly:

+15.9 points in the matched cell, +9.7 in nonsense, +8.3 in baseline, and under +4.3 in each of the remaining five (Table 10). The three cells with the largest marginal gaps are the three with the lowest pairwise κ (.66, .78, .65); they are also the three lowest-advice cells, where the boundary between a suggestion and a reflection is hardest to draw.

Table 10: The 397-sentence audit under two coders. n is sentences per cell; "A=B" is per-sentence agreement between the coders; the advice rates are the share of sentences each source labels advice. Coder A is the author (blind to condition); coder B is independent (blind to condition and to the hypothesis).

Cell	n	A=B	κ	A=clf	B=clf	A adv.	B adv.	clf adv.
baseline	44	84.1%	+.65	90.9%	88.6%	31.8%	38.6%	27.3%
matched	44	84.1%	+.66	88.6%	86.4%	27.3%	43.2%	29.5%
reversal	39	92.3%	+.83	76.9%	84.6%	35.9%	33.3%	17.9%
horticulture	67	95.5%	+.91	80.6%	82.1%	53.7%	55.2%	37.3%
nonsense	55	89.1%	+.78	81.8%	74.5%	43.6%	54.5%	32.7%
anchored contrastive	52	92.3%	+.84	78.8%	78.8%	40.4%	44.2%	34.6%
prose matched	46	95.7%	+.85	91.3%	91.3%	17.4%	17.4%	8.7%
prose contrastive	50	96.0%	+.90	90.0%	86.0%	68.0%	72.0%	62.0%
overall	397	91.4%	+.83	84.6%	83.6%	41.1%	46.1%	32.2%

The reference cell carries the disagreement. Four of the six headline contrasts are measured against the matched cell, which is where the coders diverge most, so one coder's reading of that single condition sets the level of all four. The decomposition is exact. Relative to coder A, coder B's numeric reversal falls by 0.155, of which $+0.004$ comes from the reversal cell and -0.159 from the shared reference; horticulture falls by 0.144 ($+0.015$; -0.159), the anchored reversal by 0.147 ($+0.012$; -0.159), and nonsense by 0.062 ($+0.097$; -0.159). The prose reversal, which has its own reference cell, is essentially unchanged ($+0.529$ under A, $+0.564$ under B). Two consequences follow and they point in opposite directions. The *ordering* is robust, because a shift applied to a shared reference moves the contrasts that share it by almost the same amount: prose $>$ horticulture $>$ anchored $>$ numeric holds under coder A, under coder B, under the conjunctive rule that counts a sentence as advice only when both coders do, and under every estimator in Table 14. The *magnitudes* are not robust, because the reference is precisely where two coders at κ=.83 still disagree. It is worth recording that on these same twelve conversations the classifier's own ordering fails (anchored $+0.103$ above horticulture $+0.091$) while both coders' orderings hold; the classifier recovers the ordering only on the full 119.

Table 11: Headline contrasts on the 12 audited conversations under each labeling. "Both" counts a sentence as advice only when both coders do. "Classifier" here is restricted to the same 12 conversations, which is why it differs from the 119-conversation estimates elsewhere.

Contrast	Coder A	Coder B	Both	Classifier (12)
Conditioning (matched − baseline)	−.059	+.017	−.028	+.003
Numeric reversal	+.103	−.052	+.066	−.097
Horticulture	+.265	+.121	+.257	+.091
Nonsense	+.210	+.148	+.210	+.065
Anchored reversal	+.188	+.041	+.167	+.103
Prose reversal	+.529	+.564	+.550	+.565
Ordering preserved	yes	yes	yes	no

Bias-corrected contrasts. The audit sampled 12 conversations uniformly without replacement from the 119 present in all eight cells and coded every sentence of each, so the conversation-level advice proportion—the paired unit used throughout—is fully observed for the labeled set and the classifier's version exists for all 119. We use prediction-powered inference at the conversation level (Angelopoulos et al., 2023a), in its tuned form (Angelopoulos et al., 2023b): for a contrast with human paired difference D_j (12 conversations) and classifier difference $\hat{D}_j$ (all conversations), $\hat{\theta}_\lambda =$

$\bar{D}_{\text{lab}} + \lambda(\bar{\hat{D}}_{\text{unlab}} - \bar{\hat{D}}_{\text{lab}})$, with $\hat{\lambda} = \widehat{\text{Cov}}(D, \hat{D})/((1+n/N)\widehat{\text{Var}}(\hat{D}))$ clipped to $[0, 1]$, $\text{SE}^2 = \widehat{\text{Var}}(D - \lambda\hat{D})/n + \lambda^2\widehat{\text{Var}}_{\text{unlab}}(\hat{D})/N$, and t_{11} intervals. The target is the human-label contrast on the full sample; the rectifier $\bar{D}_{\text{lab}} - \bar{\hat{D}}_{\text{lab}}$ absorbs whatever bias the classifier has, differential or not. Table 12 reports, for each headline contrast, the classifier-only estimate on all 119 conversations, the human-only estimate on 12, the tuned estimate with its t_{11} interval, $\hat{\lambda}$, a percentile bootstrap that resamples labeled and unlabeled conversations separately with λ re-estimated in each resample (4,999), and the design-based supervised learning estimator of Egami et al. (2023), which fixes λ=1. The corrected point estimates preserve prose > horticulture > anchored > numeric; the λ=1 estimator does not (numeric $+0.223$ vs. anchored $+0.211$, both with intervals far wider than their difference). Per condition, the human-minus-classifier rectifier is positive in seven of eight cells and reliable only for horticulture ($+0.131$ $[+0.039, +0.224]$); the reversal cell's rectifier is the largest in magnitude ($+0.158$) but its interval spans zero ($[-0.066, +0.383]$), which is why the corrected numeric reversal is the one contrast the audit cannot bound. Closing each corrected interval to the width of the classifier-only interval would require roughly 30 (conditioning), 70–90 (horticulture, nonsense, anchored) and 350 (numeric reversal) coded conversations, and roughly 220–260 for the two direct differences; a second coder on the same twelve conversations measures reliability, not the estimand—which is why the second coder we did run (below) narrows nothing, and why 28 further conversations, now being coded, remain the binding constraint.

Table 12: Bias-corrected headline contrasts (advice proportion, Qwen2.5-3B, user-turn placement). Classifier: all 119 conversations, t interval. Human: 12 audited conversations, t_{11}. PPI++: human-label target on all 119, t_{11}; boot: percentile bootstrap with λ re-estimated. DSL: λ=1, normal interval.

Contrast	Classifier (119)	Human (12)	PPI++	$\hat{\lambda}$	Boot 95%	DSL
Conditioning (matched − baseline)	$-.137$ $[-.188, -.085]$	$-.059$ $[-.171, +.053]$	$-.108$ $[-.196, -.020]$	.32	$[-.194, -.021]$	$-.198$ $[-.295, -.102]$
Numeric reversal	$+.022$ $[-.018, +.062]$	$+.103$ $[-.137, +.344]$	$+.115$ $[-.126, +.355]$	.08	$[-.053, +.345]$	$+.223$ $[-.046, +.491]$
Horticulture	$+.203$ $[+.157, +.249]$	$+.265$ $[+.132, +.399]$	$+.283$ $[+.154, +.413]$	.14	$[+.179, +.390]$	$+.377$ $[+.232, +.521]$
Nonsense	$+.149$ $[+.102, +.196]$	$+.210$ $[+.070, +.351]$	$+.220$ $[+.081, +.359]$	.10	$[+.102, +.333]$	$+.294$ $[+.135, +.454]$
Anchored reversal	$+.125$ $[+.077, +.173]$	$+.188$ $[+.053, +.323]$	$+.196$ $[+.071, +.320]$	.30	$[+.083, +.311]$	$+.211$ $[+.063, +.358]$
Prose reversal	$+.614$ $[+.570, +.658]$	$+.529$ $[+.339, +.720]$	$+.583$ $[+.486, +.681]$	1.00	$[+.499, +.667]$	$+.578$ $[+.496, +.660]$

The two differences the 3B story rests on, estimated directly as cell-versus-cell contrasts, are in Table 13. Both corrected intervals include zero; the name-over-value point estimate matches the classifier's, the anchored-over-numeral one does not.

Table 13: Direct difference contrasts, same estimators as Table 12.

Contrast	Classifier (119)	Human (12)	PPI++	$\hat{\lambda}$	Boot 95%	DSL
Name − value (horticulture − reversal)	$+.181$ $[+.131, +.231]$	$+.162$ $[-.082, +.406]$	$+.158$ $[-.073, +.390]$	.40	$[-.028, +.363]$	$+.154$ $[-.058, +.366]$
Anchored − numeral (anchored − reversal)	$+.103$ $[+.049, +.157]$	$+.085$ $[-.150, +.320]$	$+.028$ $[-.189, +.244]$	.53	$[-.154, +.239]$	$-.012$ $[-.225, +.201]$

Stage one: the same estimator, two coders and two predictors, twelve conversations. Table 12 targets coder A's labels with the classifier as predictor. Because both the label source and the predictor are choices, and neither is a gold standard, Table 14 reports the estimator under all four combinations on the original twelve conversations. At that stage exactly one contrast—the prose reversal—excluded zero under every coder and every predictor; horticulture and nonsense did so in three of four, the anchored reversal in two, conditioning in one, and the numeric reversal and both direct differences in none. That is where the paper stood before the extension, and we keep the table because it shows what twelve conversations can and cannot do.

Stage two: 28 further conversations, three coders. Reliability on the same twelve conversations cannot narrow the estimand; only more labeled conversations can. We drew 28 further conversations uniformly from the 107 unlabeled ones (inclusion probability $28/107$) and coded every sentence of each in the five cells the unresolved contrasts depend on—matched, reversal, horticulture, anchored contrastive, baseline; 731 sentences—leaving the nonsense and prose cells at twelve. That decision was made on the pre-registered classifier verdicts (both DIFFERENT by wide margins) and, for prose, on stage-one human-target intervals that excluded zero under every labeling; the nonsense contrast under the independent coder with the classifier predictor ($+.149$ $[-.036, +.333]$) remains unresolved and supports no main-text claim. The central name/value comparison therefore uses 40 audited conversations, while the prose reference remains based on twelve. Both coders coded independently

Table 14: **Stage one, 12 conversations.** PPI++ estimates (human-label target, all 119 conversations, t_{11} intervals) under each label source and each predictor before the extension. Coder A is the author, coder B independent; the predictors are the strategy classifier and the zero-shot judge of Table 16. Bold marks intervals excluding zero. Superseded by Table 15 for every contrast whose cells were extended.

	Coder A (author)		Coder B (independent)	
Contrast	predictor: classifier	predictor: judge	predictor: classifier	predictor: judge
Conditioning (matched − baseline)	$-.\mathbf{108}$ $[-.196, -.020]$	$-.061$ $[-.171, +.049]$	$-.072$ $[-.256, +.113]$	$+.009$ $[-.201, +.218]$
Numeric reversal	$+.115$ $[-.126, +.355]$	$+.167$ $[-.036, +.369]$	$+.019$ $[-.201, +.239]$	$+.032$ $[-.070, +.134]$
Horticulture	$+.\mathbf{283}$ $[+.154, +.413]$	$+.\mathbf{269}$ $[+.138, +.399]$	$+.121$ $[-.003, +.245]$	$+.\mathbf{130}$ $[+.034, +.226]$
Nonsense	$+.\mathbf{220}$ $[+.081, +.359]$	$+.\mathbf{210}$ $[+.070, +.351]$	$+.149$ $[-.036, +.333]$	$+.\mathbf{142}$ $[+.016, +.268]$
Anchored reversal	$+.\mathbf{196}$ $[+.071, +.320]$	$+.\mathbf{199}$ $[+.070, +.327]$	$+.043$ $[-.067, +.153]$	$+.055$ $[-.040, +.150]$
Prose reversal	$+.\mathbf{583}$ $[+.486, +.681]$	$+.\mathbf{541}$ $[+.366, +.716]$	$+.\mathbf{618}$ $[+.490, +.746]$	$+.\mathbf{575}$ $[+.374, +.776]$
Name − value	$+.158$ $[-.073, +.390]$	$+.126$ $[-.087, +.340]$	$+.168$ $[-.021, +.358]$	$+.128$ $[-.028, +.285]$
Anchored − numeral	$+.028$ $[-.189, +.244]$	$+.065$ $[-.098, +.229]$	$+.019$ $[-.195, +.233]$	$+.070$ $[-.069, +.209]$

from the same instruction sheet in different random orders; per-sentence agreement was 89.3% (κ=0.79). A third coder, who had not read the paper and was not told the rows were disputed or whose they were, then coded the 78 disagreements blind and sided with each coder on exactly 39. The adjudicated labeling is the two coders' label where they agree and the third coder's where they do not. Both the independent coder and the adjudicator were graduate students in fields unrelated to natural language processing or psychology, uncompensated, who had not read the paper and received no instruction beyond the sheet; each coding session took about ninety minutes. Because the estimator's coverage is now per contrast, each contrast uses the conversations on which its own two cells are complete, and its n is printed.

An observer-expectancy audit, by accident. The author coded the extension after reading the stage-one comparison, which had shown the independent coder calling more advice than him by 15.9 points in the matched cell and 8.3 in baseline. On the extension his labels moved toward hers in exactly those cells and nowhere else: relative to the classifier's own shift between the two conversation sets, his advice rate rose by a further 17.4 points in matched and 16.5 in baseline, against 1.4 in horticulture, while hers tracked the classifier throughout (-2.1 and $+5.8$). The first-round gap of $+15.9$ in matched became -3.7; he had crossed over her. We cannot distinguish that from a change in the conversations, except that the independent coder and the classifier saw no such change. This is the expectancy effect the annotation-methodology literature warns of, observed in our own instrument: an annotator who has seen where a comparison put him shifts toward it. The author's second-round labels are therefore not blind to the first-round result, and the paper's human-target estimates rest on the independent coder (blind throughout both rounds) as the primary labeling and on the adjudicated labels as the second. The author's are reported in Table 15 for completeness and are given no evidential weight where they differ from the other two. The one contrast where they do differ materially is the numeric reversal, which excludes zero under the author's labels alone.

Table 15: **Stage two, 40 conversations.** PPI++ estimates (human-label target, all 119 conversations, t_{n-1} intervals) under three labelings and two predictors; n is the number of labeled conversations on which both cells of the contrast are complete. B is the independent coder (primary); "adj." is the adjudicated labeling; A is the author, whose second-round labels are not blind (see text). Nonsense and prose were not extended and keep their stage-one n. Bold marks intervals excluding zero.

	Coder B (independent, primary)			Adjudicated			Coder A (author, not blind)		
Contrast	n	classifier	judge	n	classifier	judge	n	classifier	judge
Conditioning	40	$-.\mathbf{188}$ $[-.271, -.105]$	$-.\mathbf{184}$ $[-.276, -.091]$	31	$-.\mathbf{211}$ $[-.291, -.130]$	$-.\mathbf{221}$ $[-.312, -.131]$	40	$-.\mathbf{196}$ $[-.268, -.124]$	$-.\mathbf{193}$ $[-.271, -.115]$
Numeric reversal	40	$+.060$ $[-.046, +.167]$	$+.044$ $[-.054, +.142]$	33	$+.067$ $[-.027, +.160]$	$+.057$ $[-.036, +.149]$	40	$+.\mathbf{129}$ $[+.034, +.224]$	$+.\mathbf{120}$ $[+.033, +.207]$
Horticulture	40	$+.\mathbf{211}$ $[+.129, +.293]$	$+.\mathbf{222}$ $[+.151, +.293]$	34	$+.\mathbf{232}$ $[+.152, +.312]$	$+.\mathbf{225}$ $[+.156, +.294]$	40	$+.\mathbf{252}$ $[+.176, +.329]$	$+.\mathbf{257}$ $[+.186, +.328]$
Nonsense	12	$+.149$ $[-.036, +.333]$	$+.\mathbf{142}$ $[+.016, +.268]$	—	—	—	12	$+.\mathbf{220}$ $[+.081, +.359]$	$+.\mathbf{210}$ $[+.070, +.351]$
Anchored reversal	40	$+.\mathbf{123}$ $[+.043, +.202]$	$+.\mathbf{124}$ $[+.041, +.206]$	34	$+.\mathbf{137}$ $[+.055, +.220]$	$+.\mathbf{132}$ $[+.047, +.218]$	40	$+.\mathbf{149}$ $[+.069, +.229]$	$+.\mathbf{154}$ $[+.073, +.235]$
Prose reversal	12	$+.\mathbf{618}$ $[+.490, +.746]$	$+.\mathbf{575}$ $[+.374, +.776]$	8	$+.\mathbf{595}$ $[+.465, +.724]$	$+.\mathbf{557}$ $[+.281, +.832]$	12	$+.\mathbf{583}$ $[+.486, +.681]$	$+.\mathbf{541}$ $[+.366, +.716]$
Name − value	40	$+.\mathbf{151}$ $[+.065, +.237]$	$+.\mathbf{167}$ $[+.087, +.248]$	37	$+.\mathbf{132}$ $[+.038, +.226]$	$+.\mathbf{158}$ $[+.070, +.246]$	40	$+.\mathbf{111}$ $[+.020, +.202]$	$+.\mathbf{129}$ $[+.044, +.213]$
Anchored − numeral	40	$+.065$ $[-.040, +.169]$	$+.087$ $[-.012, +.185]$	36	$+.055$ $[-.051, +.160]$	$+.075$ $[-.025, +.175]$	40	$+.015$ $[-.091, +.121]$	$+.039$ $[-.058, +.137]$

What 40 conversations settle. Under the primary labeling and either predictor, five of the eight contrasts exclude zero: conditioning, horticulture, the anchored reversal, the prose reversal, and the direct name-minus-value difference—the contrast the paper's central claim rests on, which no

labeling had resolved at twelve conversations. The adjudicated labeling agrees on all five. The point-estimate ordering prose $>$ horticulture $>$ anchored $>$ numeric holds under all three labelings, both predictors and all four estimators on 40 conversations; at stage one (twelve conversations, coder A) the DSL estimator had placed numeric ($+.223$) above anchored ($+.211$) (Table 12). The two predictors, which err in opposite directions, again land together: under the independent coder their corrected estimates differ by at most 0.022 on any extended contrast. Using both jointly—the least-squares generalization of PPI++ in which λ becomes a two-vector, the simplest member of the multiple-prediction-powered family (Cowen-Breen et al., 2026)—changes no verdict on the 40-conversation contrasts and narrows their intervals by at most 3% over the better single predictor, so the two instruments are close to redundant once either is available; on the twelve-conversation contrasts the unregularized joint λ overfits ($|\lambda| > 1$) and we do not use it there. The interval half-widths fell by roughly the factor $\sqrt{40/12}$ predicts—name-minus-value from $\pm.190$ to $\pm.086$ under the independent coder with the classifier, the numeric reversal from $\pm.220$ to $\pm.107$.

What they do not settle. The numeric reversal remains indeterminate under the independent and adjudicated labels ($+0.060$ $[-0.046, +0.167]$; $+0.067$ $[-0.027, +0.160]$), matching the classifier's verdict, and the anchored-minus-numeral difference likewise ($+0.065$ $[-0.040, +0.169]$). Both involve the reversal cell, where the human coders' own agreement is lowest (κ=.73 on the extension) and the predictors track the human paired difference least. Closing either to the classifier's width would take roughly 120–320 labeled conversations. We report them as open; the classifier's INDETERMINATE verdict on the numeric reversal stands, now with human-target support rather than in spite of its absence. Two further limits: the nonsense and prose contrasts were not extended and rest on twelve conversations, and the adjudicated labeling is complete only on the extension, so its n is lower and its prose contrast rests on eight. A second instrument is the other lever: the precision a predictor buys in prediction-powered inference rises with its correlation with the gold labels—"the better the annotator model, the higher the effective sample size," with the estimator falling back to the classical one when the predictor is uninformative (Boyeau et al., 2024)—and LLM judges approach expert-level reliability on advice-giving specifically ($\kappa_w \approx .66$–$.74$ against experts, versus .69–.77 between experts, on supportive-conversation transcripts; Kumar et al., 2026). A zero-shot judge from a different model family than the generator, scored on the same sentences and validated against the same 397 human labels, would give both a triangulating measurement and a possibly better predictor; where predictor accuracy varies by condition, stratifying the estimator by condition is the principled form (Fisch et al., 2024). We ran this (cell J1): Phi-3.5-mini-instruct, a different family from the generator, judged every rep-0 sentence in the eight audited cells ($4{,}266$) with the audit's definition verbatim, greedy decoding, and two prompt phrasings (one with two worked examples). The phrasings agree on 90.3% of sentences, with the example-anchored phrasing stricter (advice rate 53.7% vs. 62.5%; classifier 37.1%); the first phrasing is primary. Against the 397 human labels the judge is about as reliable as the classifier overall (κ=.64 vs. .67) and the two instruments err in opposite directions: the judge over-counts advice in every cell, the classifier under-counts in seven of eight (Table 16). The difference that matters is on the paired contrasts. At the conversation level the two predictors explain the human proportions equally (R^2 .61 vs. .63), but on the human paired *difference* for the numeric reversal the judge reaches R^2=.32 where the classifier reaches .002, and on name-minus-value .31 against .11—which is why the classifier supplied no precision for those contrasts ($\hat{\lambda}$=.08) and a judge might. Two predictors of the human annotation now agree that the classifier under-counts advice most in the reversal cell (17.9% against the human 35.9% and the judge 43.6%), an independent confirmation of the rectifier in Table 12. The judge's per-cell rectifiers against coder A are negative in all eight cells—the mirror image of the classifier's, which are positive in seven—and largest in exactly the two cells where the human coders themselves diverge most: -0.300 in baseline and -0.191 in matched, against -0.083 to -0.151 elsewhere. Three instruments, two of them human, find the same two low-advice cells hardest, which is the most direct evidence we have that the disagreement is in the task rather than in any one annotator.

E.2 SECOND TESTBED: A CONFIGURATION KEY WITH A COUNTED OUTCOME

Qwen2.5-3B-Instruct (fp16), 80 prompts from databricks-dolly-15k (Conover et al., 2023) (CC BY-SA 3.0; categories open_qa 37, brainstorming 24, general_qa 19; no context field; 30–200 characters; seed 42, locked), user-turn placement with the battery's preamble structure, temperature 0.7, top-p 0.9, cap 350 new tokens, k=1. Configuration block: "Response configuration: – `key:`

Table 16: Two instruments against the 397 human labels, by cell. Agreement is per-sentence; advice rates are the share of sentences labeled advice by each source.

Cell	n	judge = human	classifier = human	human adv.	judge adv.	classifier adv.
baseline	44	65.9%	90.9%	31.8%	56.8%	27.3%
matched	44	79.5%	88.6%	27.3%	47.7%	29.5%
reversal	39	82.1%	76.9%	35.9%	43.6%	17.9%
horticulture	67	86.6%	80.6%	53.7%	64.2%	37.3%
nonsense	55	80.0%	81.8%	43.6%	60.0%	32.7%
anchored	52	80.8%	78.8%	40.4%	51.9%	34.6%
prose matched	46	84.8%	91.3%	17.4%	32.6%	8.7%
prose contrastive	50	90.0%	90.0%	68.0%	78.0%	62.0%
all	397	κ=.64	κ=.67			

`value`" followed by the battery's guidance lines. Outcome: word count, analysed as log(words) with paired sign-flip permutation (4,999) and percentile bootstrap; Holm over the two directional predictions. Predictions written before generation: P-C1 |rename at 0.90| > |value swing|; P-C2 anchored swing > numeric swing; P-C3 nonsense key equivalent to baseline within 0.10 log units. Word counts and truncation by condition, then the contrasts:

Condition	Words	Median	Trunc.	Contrast (log ratio)	Ratio	95% CI (log)
no configuration	179.1	211	.40	value swing (0.90 − 0.10)	1.65	+.501 [+.370, +.639]
`verbosity: 0.10`	30.3	19	0	rename at 0.90 (`terseness−verbosity`)	0.80	−.223 [−.397, −.043]
`verbosity: 0.90`	53.3	39	0	rename at 0.10	1.50	+.404 [+.263, +.558]
`verbosity: 0.10 (very brief)`	14.9	12.5	0	anchored swing	9.76	+2.279 [+2.095, +2.457]
`verbosity: 0.90 (very detailed)`	146.3	121.5	.14	nonsense key − no configuration	0.50	−.698 [−.815, −.580]
`terseness: 0.10`	42.6	36	0	`verbosity: 0.90` − no configuration	0.23	−1.462 [−1.632, −1.290]
`terseness: 0.90`	36.5	32	0	`verbosity: 0.10` − no configuration	0.14	−1.963 [−2.123, −1.802]
`murvaning: 0.90`	96.7	83	.03			

Verdicts: P-C1 FAILED in direction (|rename| − |swing| = −0.278 log, Holm p=.001); P-C2 HELD (+1.778 log, Holm p=.0004); P-C3 FAILED (TOST p=1), but see the design note. Restricting to the 43 prompts with no truncated response in any condition: value swing +.418 [+.261, +.598], rename at 0.90 −.156 [−.347, +.032], anchored swing +2.037 [+1.764, +2.318], nonsense −.755; verdicts unchanged. Secondary outcomes moved with length as expected (sentence count 1.4–10.3; type–token ratio .65–.93) and the Heylighen–Dewaele formality score varied by under seven points across conditions. Two notes on the record. The registered stopping rule ("baseline truncation > 25% censors P-C1/P-C2") was written too broadly: baseline enters only the presence and nonsense contrasts, and the P-C1 conditions have no truncation; we report it as triggered and note the misspecification. And the design as first run lacked an empty-configuration control, so "nonsense key vs. no configuration" confounded the key's meaning with the block's presence. A follow-up cell (E19b; same 80 prompts, same seed formula, nine further conditions) added it.

Empty-configuration control and a second counted dimension (E19b). Conditions: an empty block (header and guidance lines, no key line); an explicit null line (`- (none specified)`); a coherent but irrelevant key (`mulching: 0.90`); and, for a second dimension, `questions` at 0.10/0.90 numeric, 0.10 (*rarely*)/0.90 (*very often*) anchored, renamed to its antonym (`statements: 0.90`) and to nonsense (`blathering: 0.90`), with the number of question sentences as outcome. Registered: P-B1, the empty block shortens responses relative to no configuration; P-B2, `verbosity: 0.90` differs from the empty block; P-B3, a nonsense key moves the question count less than a real key; P-Q1, the numeric swing on `questions` is DIFFERENT and positive; P-Q2, the anchored swing exceeds the numeric swing on `questions`. Results on log words, paired: empty block − none −0.475 [−0.589, −0.364] (0.62×); null line − none −0.574 [−0.706, −0.449] (0.56×); null − empty −0.099 [−0.221, +0.020], p=.12; `verbosity: 0.90` − empty −0.987 [−1.167, −0.810] (0.37×); `verbosity: 0.10` − empty −1.488 (0.23×); nonsense key − empty −0.223 [−0.339, −0.103] (0.80×); irrelevant key − empty −0.184 [−0.293, −0.074] (0.83×). P-B1 and P-B2 HELD. The `questions` outcome was a floor in every condition (0.00–0.05 question sentences per response; 0.08 with no configuration): P-Q1 and P-Q2 FAILED for want of an outcome that could move, and P-B3, registered on that outcome, is uninformative and counted as FAILED; the analogous comparison on length (nonsense 0.80× vs. real key 0.37× of the empty block) is exploratory and points the way P-B3 predicted. Word counts under

the `questions` key (91–113) sit with the nonsense and irrelevant keys, not with `verbosity` (30–53): it is the length-related *meaning* of the key that shortens responses, and the value modulates within that.

Two audits accompany the follow-up. *Recitation.* On short factual prompts the model sometimes recites the configuration instead of answering (`0.90`; `Statements: 0.90`), which corrupts a counted outcome; a detector for a numeral-only response or a value, `key: value`, or "Configuration" line among the first two lines was applied to every condition, including E19's stored responses. E19's conditions are clean (0–2.5%, nonsense 7.5%), so the numbers in the table above stand; the empty block, however, invites the model to *invent* a configuration ("Configuration: Simple"; 37.5% recited, 52.5% mentioning it), which is why the explicit null line was added. Recomputing every contrast on prompts where neither side recited changes no verdict (empty − none -0.523, n=50; nonsense − empty -0.302, n=47). *Detector.* A first version scanned whole responses and flagged ordinary answers that merely contained a key word; it was narrowed to leading recitation before the full run, and the change is on record.

E.3 Value format and the comprehension probe (E17)

One internally pre-registered cell (3B, user-turn placement, k=1) tested three consequences of the reading that bare numerals are recited but not interpreted. First, asked to translate `advise: 0.10` into one of six frequency adverbs on a stated 0–1 scale, the model answers *occasionally* for every value from 0.05 to 0.85 and *often* only at 0.95 (Kendall τ=+.45 against the correct adverb ranks, 20% exact-band agreement, three seeds per level; the registered $\tau \geq .7$ failed). The same model echoes the number verbatim on request (Appendix F). Second, rendering every value as a percentage (`advise: 10%`) or as "N out of 10" leaves the whole-policy reversal where the decimal left it ($+0.019$ $[-0.025, +0.063]$ and $+0.018$ $[-0.027, +0.064]$, both INDETERMINATE; registered prediction held). Third, attaching the adverb to the numeral (`advise: 0.10 (rarely)`) recovers $+0.063$ $[+0.014, +0.114]$ (DIFFERENT, Holm p=.037), about 38% of the adverb-only reversal of $+0.165$; we had registered "at least half" and the point estimate misses it, though the interval includes it. Whether the shortfall relative to the adverb alone is dilution or k=1 noise we cannot say. The comprehension probe is complete and flat: across ten levels from 0.05 to 0.95 with three seeds each (30/30 parsed), the model answers *occasionally* at every level from 0.05 through 0.85 and *often* only at 0.95, giving Kendall $\tau = +0.447$ against the correct adverb ranks and 20% exact-band agreement; the registered $\tau \geq .7$ failed. The three re-encodings of the same reversal give: percentages $+0.0194$ $[-0.025, +0.063]$ (p=.397), "N out of 10" $+0.0179$ $[-0.027, +0.064]$ (p=.446), and numeral-plus-adverb $+0.0634$ $[+0.014, +0.114]$ (p=.013 raw, Holm p=.037). Registered prediction P-F1—that a percentage would recover the anchored effect—failed; P-F2, that the adverb-annotated numeral would exceed the bare numeral, held.

E.4 Dose–response models and the subset ladder

Dose–response models (E13r). Re-analysis of the E13 sweep (six levels of the single `advise` dimension, the other seven held at the matched policy, n=119, k=1, system placement), fitting the models a rank-correlation test cannot: a per-conversation linear slope; an isotonic fit to the within-conversation-centred level means in whichever direction fits better, with levels permuted within conversation; an omnibus permutation F over the six levels, which detects a threshold or non-monotone shape that monotone tests miss; and a paired bootstrap band on the level means.

	Numeric (`0.05–0.95`)	Anchored (*never–constantly*)
(a) endpoint 0.95 − 0.05	-0.026 $[-0.075, +0.023]$, p=.30	$+0.025$ $[-0.019, +0.068]$, p=.26
(b) linear slope per unit	-0.035 $[-0.080, +0.010]$, p=.12	$+0.024$ $[-0.018, +0.068]$, p=.27
(c) isotonic fit	decreasing, p=.10	increasing, p=.40
(d) omnibus (any shape)	F=1.58, p=.163	F=0.53, p=.758
(e) range of the six level means	0.061 (boot. 97.5th pct 0.104)	0.032 (boot. 97.5th pct 0.085)

Level means with 95% bands, numeric: 0.330 $[0.285, 0.378]$, 0.300, 0.311, 0.293, 0.269, 0.304 $[0.260, 0.348]$; anchored: 0.276 $[0.236, 0.319]$, 0.289, 0.303, 0.308, 0.301, 0.301 $[0.258, 0.345]$. The paired anchored-minus-numeric slope difference is $+0.059$ per unit $[-0.001, +0.121]$,

p=.058, reproducing E13's registered P-D3 contrast exactly; the endpoint difference is $+0.052$ $[-0.014, +0.121]$, p=.135. So the anchored encoding's advantage for *whole-policy* changes does not reappear for a single dimension even as a trend that clears .05. Two cautions on the bound in row (e): the range of six noisy means is upward-biased, so 0.104 is conservative as an upper bound; and every cell is k=1, so this bounds what the design could have seen, not what a better-powered design would find.

Nested subset ladder (E20). The single-dimension sweep against a whole-policy reversal cannot distinguish dimensions that interact from eight small effects that add. We ran a nested subset ladder in the anchored encoding under the battery's placement (Qwen2.5-3B fp16, n=119, k=1, a fresh seed space): the matched policy; eight single flips to the contrastive value; four pairs; two quads (the listening and directing halves); and all eight. Registered before generation: P-X1, the `advise` single flip is bounded ($\leq +0.05$ at the CI edge); P-X2, the whole exceeds the sum of the singles with the gap distinguishable from zero; P-X3, at least one non-`advise` single flip is DIFFERENT under Holm over eight. All three *failed*: P-X1 because the interval's upper edge is $+0.083$ (the point estimate, $+0.035$, is consistent with the earlier sweeps, but the registered bound was tighter than a k=1 design resolves); P-X2 because the gap has the opposite sign and is not distinguishable from zero; P-X3 because no single flip survives Holm (`validate`, `reflect` and `warmth` have uncorrected $p < .05$).

Single flip	Effect	95% CI	Subset	Effect	Sum of parts	Gap	p(gap)
`validate`	+.061	[+.013, +.108]	`validate+reflect`	+.121	+.123	−.003	.94
`reflect`	+.063	[+.012, +.114]	`question+summarize`	+.035	+.021	+.014	.64
`question`	+.014	[−.034, +.063]	`advise+directiveness`	+.021	+.060	−.038	.26
`advise`	+.035	[−.014, +.083]	`challenge+warmth`	+.043	+.075	−.032	.33
`challenge`	+.024	[−.014, +.060]	listening half (4)	+.125	+.156	−.031	.43
`summarize`	+.007	[−.036, +.050]	directing half (4)	+.010	+.064	−.055	.12
`warmth`	+.051	[+.004, +.096]	all eight	+.167	+.135	+.032	.38
`directiveness`	+.025	[−.021, +.070]	all eight vs. sum of singles	+.167	+.279	−.112	.38

Sums of parts are of the next level down (pairs from singles, halves from pairs, the whole from halves); the last row sums the eight singles directly. Gaps are paired within conversation with sign-flip p. Two design limits: every single-flip effect is measured against the same k=1 matched run, so a sum of eight such effects carries eight times that run's noise (the sum-of-singles interval is $[+0.001, +0.556]$; the sum and the gap are formed per conversation before the conversation bootstrap, so the covariance induced by the shared matched reference is retained), and the gap test at this k resolves only gaps of roughly ≥ 0.12 against the whole; a k>1 matched run would tighten both. The result the ladder does establish is where the whole-policy effect lives: the listening half carries it and the directing half, which contains the dimension named for the outcome, does not detectably contribute.

E.5 Notes on the model and placement comparison

Direct name-minus-value contrast at each scale. The scale table compares two effect estimates; the claim about relative contribution requires their paired difference on the same conversations, (domain − matched) − (contrastive − matched) = domain − contrastive. Table 18 reports it at every scale under both placements, with the battery's statistics and Holm within placement. These contrasts were not pre-registered; the seventh external review asked for them. The system-placement result is the one to notice: at 32B under the system prompt the two interventions are EQUIVALENT, and at 72B names again exceed values, so the sign reversal reported in the main text belongs to the user-turn configuration and is not reproduced when the same specification sits in the system prompt.

The following material was moved here from the main text to keep §5.4 descriptive. First, whether the name effect declines with scale: under user-turn placement it falls ($+0.199 \rightarrow +0.113 \rightarrow +0.080$), under system placement it does not ($+0.077 \rightarrow +0.109 \rightarrow +0.111$). Second, the shape of the value curve: under the battery's placement the value effect rises monotonically from 3B through 7B and 14B to 32B ($+0.032 \rightarrow +0.073 \rightarrow +0.134 \rightarrow +0.305$; P-S7, "7B < 14B < 32B," held on point estimates) and then falls at 72B ($+0.125$), and the 32B>72B ordering also holds under system placement ($+0.095$ vs. $+0.048$). The 32B configuration has the largest observed value effect, but five points do not explain it; two points in one family at k=1 and 4-bit is what we have, and we offer

Table 17: The four headline conditions across the Qwen2.5-Instruct family under both prompt placements (five scales under the battery's user-turn placement, three under system placement). All effects are paired within conversation, n=119; all are DIFFERENT under Holm within arm×set except the two marked IND (INDETERMINATE). The 3B user-turn row is k=3; every other row is k=1, meaning each condition contributes one stochastic generation per conversation, so cross-scale comparisons characterize realized sample-level effects rather than expectations over decoding. All models above 3B are 4-bit; 3B is fp16.

placement	model	base→matched	reversal (values)	domain rename (names)
user turn	3B	-0.141	$+0.032$ IND	$+0.199$
	7B	-0.229	$+0.073$	$+0.255$
	14B	-0.086	$+0.134$	$+0.103$
	32B	-0.152	$+0.305$	$+0.113$
	72B	-0.076	$+0.125$	$+0.080$
system prompt	3B	-0.079	$+0.009$ IND	$+0.077$
	32B	-0.104	$+0.095$	$+0.109$
	72B	-0.110	$+0.048$	$+0.111$

Table 18: Direct name-minus-value contrast (domain rename − numeric reversal, paired within conversation) at each scale and placement. Battery statistics; Holm within placement. A positive DIFFERENT means names exceed values; negative, values exceed names. Not pre-registered.

Placement	Model	n	Estimate	95% BCa	d_z	Holm p	Verdict
user turn	3B	119	$+.167$	$[+.138, +.196]$	$+1.02$	.001	DIFFERENT
user turn	7B	119	$+.182$	$[+.138, +.225]$	$+0.75$	.001	DIFFERENT
user turn	14B	119	$-.031$	$[-.075, +.012]$	-0.13	.167	INDETERMINATE
user turn	32B	119	$-.192$	$[-.240, -.147]$	-0.73	.001	DIFFERENT
user turn	72B	119	$-.045$	$[-.078, -.012]$	-0.24	.024	DIFFERENT
system	3B	119	$+.068$	$[+.028, +.109]$	$+0.30$	.002	DIFFERENT
system	32B	118	$+.012$	$[-.025, +.046]$	$+0.06$	.488	EQUIVALENT
system	72B	119	$+.063$	$[+.032, +.093]$	$+0.37$	.001	DIFFERENT

no account. Non-monotone instruction following within a family is not unprecedented: the developers' own instruction-following benchmark (IFEval) numbers put Qwen2.5-14B-Instruct above 32B (81.0 vs. 79.5 strict-prompt; Qwen Team, 2024), and scaling trends on adversarially chosen tasks are known to reverse (Wei et al., 2023a). Nor is 4-bit quantization size-neutral—its effects "depend heavily on the quantization method, model, and task" (Mekala et al., 2025)—so the 32B/72B ordering may carry a quantization component of unknown sign. The name effect has its own shape: it is largest at 7B ($+0.255$), where the model is also the most strongly conditioned (-0.229), and settles to $+0.08$–0.11 from 14B on; the two channels cross between 7B and 14B. What survives both placements is that the balance moves with size: at 3B the name intervention exceeds the value reversal under either placement; under the user turn the two are not distinguishable at 14B and value sensitivity exceeds name sensitivity at 32B, whereas under the system prompt the direct contrast never turns negative (Table 18). Placement also matters in its own right at 3B, where moving the specification into the user turn adjacent to the conversation roughly doubles every effect—a descriptive comparison across k=3 and k=1 arms. This direction is not the one the placement literature would predict by default. Neumann et al. (2025) place identical demographic text in system versus user prompts across six commercial models and find *system* placement produces the larger effect, with the gap growing with model capability; our 3B result runs the other way, and our own large-model rows narrow the gap. Two facts make both observations coherent: the system/user distinction is a trained priority rather than an architectural one (Wallace et al., 2024), so its strength is a property of a particular model; and instruction adjacency to the governed content independently helps, with post-input instruction placement gaining up to 9.7 BLEU on zero-shot generation (Liu et al., 2024). We report the effect and do not adjudicate it. The scope is the Qwen2.5 family, this task, and these two placements: relative name/value sensitivity changes across model scale and placement, with name dominance greatest at 3B. One deployment consequence follows within that scope: a

practitioner who validates a specification on a large model and ships it on a small one may inherit a failure these tests detect and a baseline comparison does not.

E.6 SINGLE-DIMENSION CONTROL IN THE PROSE ENCODING

An earlier experiment in the prose arm moved one dimension at a time to its high or low extreme with the other seven held at the neutral midpoint, n=119, three models. Preparing the release showed that this experiment is two generation runs: the `question` and `validate` conditions and the neutral reference are in one set of files with sentence-level classifier labels, and the `advise` conditions are in a second set whose sentence-level labels had never been written to disk. We re-scored the second set with the same sentence splitter and the same strategy classifier used everywhere else in this paper, which both recovers the missing labels and makes available a contrast that needs no external reference at all: `advise` high minus `advise` low, measured within one run.

Table 19: Prose single-dimension contrasts, re-scored through the audited pipeline. The full swing is the within-run contrast and is the number the main text quotes. The `question` and `validate` rows are measured against the neutral policy in their own run. Outcome is advice proportion throughout, so the `question` and `validate` rows are cross-family effects. Holm within model.

Model	Contrast	n	Effect on advice proportion	p_{H}	Verdict
Qwen2.5-3B	`advise` full swing	119	$+.286$ $[+.239, +.331]$	.0006	DIFFERENT
Phi-3.5-mini	`advise` full swing	119	$+.276$ $[+.242, +.311]$	.0006	DIFFERENT
Mistral-7B	`advise` full swing	119	$+.157$ $[+.125, +.189]$	.0006	DIFFERENT
Qwen2.5-3B	`validate` down	119	$+.067$ $[+.021, +.108]$	.0072	DIFFERENT
Phi-3.5-mini	`validate` down	119	$+.065$ $[+.030, +.100]$	.0008	DIFFERENT
Mistral-7B	`validate` down	119	$+.096$ $[+.070, +.121]$	.0008	DIFFERENT
Qwen2.5-3B	`validate` up	119	$-.031$ $[-.080, +.016]$	.4112	INDETERMINATE
Phi-3.5-mini	`validate` up	119	$-.013$ $[-.045, +.018]$	.8120	EQUIVALENT
Mistral-7B	`validate` up	119	$+.041$ $[+.015, +.068]$	.0100	DIFFERENT
Qwen2.5-3B	`question` down	119	$+.073$ $[+.030, +.116]$	.0032	DIFFERENT
Phi-3.5-mini	`question` down	119	$+.015$ $[-.020, +.050]$	.8120	EQUIVALENT
Mistral-7B	`question` down	119	$-.002$ $[-.029, +.026]$	.9084	EQUIVALENT
Qwen2.5-3B	`question` up	119	$+.009$ $[-.038, +.059]$	.7050	INDETERMINATE
Phi-3.5-mini	`question` up	119	$+.022$ $[-.009, +.054]$	.5292	EQUIVALENT
Mistral-7B	`question` up	119	$+.055$ $[+.026, +.083]$	.0024	DIFFERENT

Two things in this table matter beyond the headline. First, turning `validate` *down* raises advice in all three models, while turning it up does not lower advice in two of the three. That is the subset ladder's finding (§5.3) reproduced in a different encoding and on two further models: the reversal's effect on advice is carried by the listening dimensions being turned down, not by advice being turned up, and the effect is asymmetric. Second, no dial acts only on its own dimension—every contrast here is an effect on advice produced by moving something else—so the non-orthogonality the earlier analysis asserted is now quantified rather than claimed.

A correction, and what it cost. Previous drafts reported this experiment as `advise` moving $+0.111$ / -0.175 on Qwen2.5-3B, $+0.140$ / -0.136 on Phi-3.5-mini and $+0.076$ / -0.081 on Mistral-7B against the neutral policy. Those figures are arithmetically exact: for every model their difference equals the audited within-run full swing to four decimal places, so they were computed from these same classifier labels. What was wrong is the provenance rather than the arithmetic. The labels behind them were never saved, so the numbers could not be reproduced from the released artifacts until we regenerated them; and the neutral reference belongs to the other generation run, so the published contrast silently compared conditions generated at different times—a comparison whose run-to-run component our own replication records elsewhere show to be non-trivial. We now quote the within-run swing, which has neither defect, and we record the episode because a paper arguing that conditioning claims need falsifiable checks should say when one of its own numbers failed a check it applies to everything else. The failure was in release discipline, not in measurement, and it was caught by the artifact inventory rather than by any of the six reviews.

Re-scoring the neutral condition from the other run and the `exposed_state` condition from this one confirms the identification: the neutral policy's advice proportion sits 0.0270, 0.1340 and 0.1116 below `exposed_state`'s on the three models, against 0.0268, 0.1340 and 0.1117 predicted by solving the original figures for their reference.

F External Demonstration: A Persona Recipe, Two Outcomes

Big5-Scaler (Cho & Cheong, 2025) embeds numeric Big Five values in prompts ("Your agreeableness score is 9 out of 10") and reports monotonic tracking at $r > 0.85$—on self-report questionnaires. We construct a recipe *patterned after* theirs (paraphrased from the paper's description; we do not evaluate their artifact), place it in the system prompt, and run *both* outcomes on the same persona prompts: an IPIP-style self-report arm, and the behavioral arm, across a five-level dose (1–9), a nonsense-renamed trait, an all-nonsense profile, and verbal anchors, with four contrasts pre-registered before generation.

Self-report tracks the assigned value at r=.948—qualitatively replicating Big5-Scaler on our model—while no behavioral channel tracks anything: the dose slope is -0.003/unit (CI $[-0.008, +0.003]$; a flat model beats a linear one on the Akaike information criterion (AIC); per-conversation Kendall τ null, p=.55); renaming one trait to nonsense does nothing ($+0.004$); renaming *all five* does nothing (-0.009); verbal anchors do nothing distinguishable (-0.023); and no support-strategy family's dose slope excludes zero. Our outcomes are counseling-strategy measures, not validated behavioral measures of Big Five expression—agreeableness might manifest in politeness, hedging, or warmth that strategy families do not capture. A style probe on the same generations finds hedging and warmth lexica flat across the dose while directive phrasing shows a weak monotone *decline* (condition-level Kendall τ=-0.4 over five points; descriptive only)—so construct-level behavioral inertness is *not* established; the defensible claim is that self-report tracks the dose while our strategy measures and two of three style probes do not. Even the profile's *presence* effect (-0.053) fails Holm correction and halves under length truncation. The arm is k=1, so its nulls are INDETERMINATE rather than EQUIVALENT (TOST power .20–.42; at the CI edge, the full dose moves behavior by at most $\sim$0.06, a tenth of the prose reversal); and our template is paraphrased from the recipe's description. Even at `agreeableness: 1`, the model self-rates $3.9/5$ on sympathy items—tracking the number relatively while refusing to describe itself as disagreeable, consistent with social-desirability bias in LLM self-report (Salecha et al., 2024).

A follow-up probe (201 eight-token generations, context-free; descriptive statistics only) locates what the questionnaire measures. On fresh seeds the IPIP score again tracks the value (r=.953); rename the trait to a nonsense word (`sarnexity`), values fixed, and the same questionnaire goes flat (r=.24; registered, held); yet asked directly how much `sarnexity` it has, the model reads the number back off the meaningless word as faithfully as off the real one (r=.905 vs. .889; the latter missed a registered $\geq$ .90 threshold by .011 and is counted as failed). The pattern is consistent with a word-gated numerical retrieval process; one trait, one model, four items and three seeds do not establish a mechanism. Self-report here certifies that the model parsed the specification and shows no convergent validity with behavior—not uselessness, since a flat questionnaire would have told us the specification went unread.

The qualitative dissociation is not ours: Han et al. (2025) report it for categorical persona injection, and Song et al. (2025) show questionnaire profiles mischaracterize generation. What this appendix adds is narrower and specific to numeric specifications: the self-report side tracks the assigned number parametrically (r=.95 over five levels), the behavioral side is bounded near zero on the same prompts, and the nonsense-word probes locate the self-report signal in a word-gated lookup of the number rather than in anything the number did. The contrast between the two specifications is itself informative, though we state it only as interpretation: the full name swap moves behavior $+0.199$ when the names are task-proximal (`advise`, `validate`) and -0.009 when they are task-distal constructs (`agreeableness`)—consistent with the name effect operating through the semantic accessibility of the names to the response domain. Table 20 is the dissociation itself. The self-report column is the IPIP agreeableness score with a bootstrap interval; the behavior column is a raw advice proportion given for orientation only, since the behavioral verdicts come from the audited pipeline and are INDETERMINATE throughout this arm.

Table 20: Self-report against behavior under a numeric `agreeableness` specification (Qwen2.5-3B). The self-report side tracks the assigned number; the behavioral side does not move. Quote the audited statistics, not the behavior column, for any behavioral claim.

Condition	IPIP agreeableness	advice proportion
baseline (no specification)	4.75 $[4.33, 5.00]$	.383
`agreeableness: 1`	4.00 $[3.33, 4.58]$	.339
`agreeableness: 9`	4.83 $[4.58, 5.00]$	.327
`sarnexity: 9` (nonsense name)	4.42 $[3.92, 4.83]$	.330
anchored low	3.00 $[2.00, 4.00]$	.318
anchored high	4.25 $[3.75, 4.67]$	.340

Three registered probes locate the self-report signal. P-R1 asked whether the answer tracks the number attached to the queried word regardless of what that word is: it does, and the prediction *failed*—asked about `agreeableness` under a real profile the model's answer correlates $r = +0.889$ with the assigned value and spans the full 1–9 range, and asked about `sarnexity` under a nonsense profile it correlates $r = +0.905$ and spans the full range, while asked about `agreeableness` under the nonsense profile it is flat ($r = -0.189$, span 0.00). The number is read back under a meaningless word as readily as under a real one, and is not read back when the word is absent. P-R2 held: on the IPIP questionnaire the real name gives $r = +0.953$ with span 0.83 and the nonsense name $r = +0.236$ with span 0.17, so the questionnaire response is gated on the trait word in a way the echo probe is not. P-R3, the positive control, held (3.00 against 4.25 for anchored low versus high). Together these are what the main text means by a word-gated numerical retrieval process on the self-report side: the model reports the number it was given whenever a word is there to hang it on, and its behavior does not follow.

G Falsifiable Predictions

Ledger of internally pre-registered predictions. Each was written into the analysis code of its cell before generation and left unedited; "held" and "failed" are by the registered rule, not by a later reading of the evidence.

Cell	Prediction	Verdict	Where reported
E12	P-S1 conditioning survives scale	held	§5.4
E12	P-S2 name dominance non-decreasing with scale	**failed**	§5.4, (P2) below
E13	P-D1 anchored single-dimension curve monotone	**failed**	§5.3
E13	P-D2 numeric single-dimension curve flat	held (bound)	§5.3
E13	P-D3 anchored slope exceeds numeric slope	**failed**	§5.3
E14	P-R1 echo probe $r \geq .90$ for real and nonsense word	**failed** (by .011)	Appendix F
E14	P-R2 questionnaire flat under nonsense name	held	Appendix F
E14	P-R3 anchored questionnaire separation ≥ 1.0	held	Appendix F
E15	P-S3 name/value ratio $\geq 3\times$ at 3B, system placement	held (denominator indeterminate)	§5.4
E15	P-P1 anchored endpoint contrast $> +0.05$, user placement	**failed**	§5.3
E15	P-P2 numeric endpoint contrast flat, user placement	held (bound)	§5.3
E16	P-S4 reversal DIFFERENT at 32B and 72B, user placement	held	§5.4
E16	P-S5 domain rename DIFFERENT at 32B and 72B	held	§5.4
E17	P-F1 numeral+adverb recovers $\geq$ half the anchored effect	**failed**	Appendix E.3
E17	P-F2 percent and out-of-10 formats do not rescue	held	Appendix E.3
E17	P-F3 numeral→adverb comprehension $\tau \geq .7$	**failed**	Appendix E.3
E18	P-S6 reversal DIFFERENT at 14B	held	§5.4
E18	P-S7 reversal 7B $<$ 14B $<$ 32B	held	§5.4
E19	P-C1 key rename moves length more than value swing	**failed**	§5.1
E19	P-C2 anchored swing exceeds numeric swing	held	§5.1
E19	P-C3 nonsense key equivalent to no configuration	**failed** (design confound)	§5.1
E20	P-X1 `advise` single flip bounded $\leq +0.05$	**failed** (upper edge .083)	Appendix E.4
E20	P-X2 whole exceeds sum of singles	**failed**	§5.3
E20	P-X3 a non-`advise` single flip DIFFERENT	**failed**	Appendix E.4
E19b	P-B1 empty block shortens vs. no configuration	held	Appendix E.2
E19b	P-B2 `verbosity: 0.90` differs from empty block	held	Appendix E.2
E19b	P-B3 nonsense key moves question count less than real key	**failed** (outcome at floor)	Appendix E.2
E19b	P-Q1 numeric swing on `questions` DIFFERENT, positive	**failed** (outcome at floor)	Appendix E.2
E19b	P-Q2 anchored swing exceeds numeric on `questions`	**failed** (outcome at floor)	Appendix E.2

Fifteen of twenty-nine failed, three of them because the registered outcome turned out to have no room to move. The external demonstration (E7) registered four contrasts without directional predictions; all four were INDETERMINATE.

Open predictions. Stated so that either branch is informative. (P1) *Fluency-bounded conditioning* (a hypothesis motivated by two confounded conditions on one model): per-attribute conditioning strength should fall with independently measured comprehension of the specification's language/script—requiring many languages and a fluency instrument—and a model's cross-lingual jailbreak gap (Yong et al., 2023) should correlate with its specification-adherence gap (cf. Zeng et al., 2025). If supported, an equity concern follows for behavioral safety specifications in low-resource languages; that consequence is conditional on the hypothesis. (P2) *Scale—tested, and the registered bet lost*: before the scale runs we predicted, from the Redefine inverse-scaling task (McKenzie et al., 2023), non-decreasing name dominance for bare numeric specifications; the value channel instead appeared at 32B and 72B under both placements (§5.4), the Wei et al. (2023b)-side outcome, and under the user-turn placement the direct name-minus-value contrast turns negative by 32B, which it does not under the system prompt. The surviving predictions: for anchored and prose encodings name dominance should fall faster with scale, and the 32B peak in the value effect—observed under both placements, with a monotone rise from 3B through 14B on the near side—should either replicate in another family or prove to be a quantization or k=1 artifact. (P3) *Serialized configurations are key-name selectors, not value dials*: renaming a configuration key (`verbosity`→`terseness`) should move agent behavior more than swinging its value across the full range, with schema-level verbal anchors recovering part of the gap. Partial support already exists on the output side: Le (2026) show that schema-key wording alone changes accuracy under

constrained decoding, and that the effect is model-family-dependent as ours is. We then ran that comparison (§5.1, Appendix E.2) and the core prediction *failed*: on a length outcome the value swing beat the rename, and the two composed semantically. The anchor half held ($9.8\times$ vs. $1.65\times$). P3 survives only in the weaker form: bare numerals are less sensitive than verbal anchors on every dimension we have tested; whether they are inert depends on the dimension. (P4) *Dose–response validation—tested once, and the anchored half failed*: we predicted that bare-number dosing would produce flat curves at far higher rates than verbally anchored dosing; on our one test (§5.3) both list encodings were flat for a single dimension in isolation, while the prose encoding was not (Appendix E.6). The surviving prediction: a single-dimension dose curve in an attribute–value encoding will be flat on small models, and a flat curve means the manipulation failed, not that the construct is null. (P5) *Safety parameters*: the battery applies unchanged to any numeric caution or refusal parameter; we predict the same name-over-value asymmetry on small models and regard a monotone refusal-rate response to a bare numeric caution value on a 3B model as a refutation. We did not test this.

H Extended Related Work

Sensitivity as variance, sensitivity as a control, and controls for prompts. That models are sensitive to semantically irrelevant prompt properties is established: paraphrases reorder leaderboards (Mizrahi et al., 2024) and formatting alone moves benchmarks by tens of points (Sclar et al., 2024). Our use differs in kind: sensitivity manipulations serve as *control conditions that test one system's conditioning claim* rather than as variance to be benchmarked, in the line of Webson & Pavlick (2022), who found few-shot models learning as fast from misleading templates as from good ones and most sensitive to label words; ours is the generation-era, specification-level sharpening. Control tasks (Hewitt & Liang, 2019) and behavioral test suites (Ribeiro et al., 2020) are ancestors in spirit; the closest concurrent work is İşcan (2026b;a), a Popperian, placebo-controlled evaluation of prompt- and weight-based conditioning in code self-repair, relative to which we contribute the instantiation for behavioral specifications, the name/value contrasts, and the behavioral domain. Placebo precedents are Mukherjee et al. (2024), where models respond as strongly to placebo cue words as to real sociodemographic cues; Ling et al. (2025), where a random word in place of an abstention label "produces an identical effect"; and Feng et al. (2026), whose translation and masking controls expose lexical cues masquerading as cultural understanding. Luz de Araujo et al. (2025) report drops of almost 30 points from "irrelevant persona details," convergent with our name interventions; prompt-regression work assumes, and mostly finds, additive combination of prompt components (Feffer et al., 2024), the reading our subset ladder is consistent with.

Numeric and graded control, steerability, personas, and evaluation validity. The strongest published demonstration of numeric prompt control is Big5-Scaler (Cho & Cheong, 2025), whose outcome is a self-report questionnaire and which runs no falsifying control (Appendix F); Jang et al. (2026) supply the frame that instruction-tuned models can describe distributions they cannot sample from. Labroo et al. (2026) place numeric concept levels in prompts and find control degrades when two concepts are specified together, the nearest existing test of numeric levels as dials with a judged behavioral outcome. Chittem et al. (2025) and Serapio-García et al. (2025) obtain graded control from *adjective-anchored* levels, never bare numerals; our format ladder reconciles the two literatures, and Kun (2026) find that most models track a numeric anchor closely when the number is the answer format itself, an effect they treat as an evaluation artifact and one the Gemini models largely resist. Steerability evaluation defines the shift of a behavioral distribution from baseline and finds it limited and asymmetric across persona dimensions (Miehling et al., 2025), or decomposes steering error into miscalibration and side effects (Chang et al., 2026)—a framing our configuration result inverts, since at 3B the off-target bundle is the only thing that moves the outcome. Persona prompting is widely studied (Zheng et al., 2024; Hu & Collier, 2024; Tseng et al., 2024); the validity of LLM self-report is questioned (Dominguez-Olmedo et al., 2024; Salecha et al., 2024), including under persona prompting (Song et al., 2025; Han et al., 2025). The construct-validity critique of LLM evaluation is well developed (Bean et al., 2025), and Gao et al. (2025) caution against LLMs as surrogates for human respondents; what neither supplies is a runnable check for conditioning claims, since specification-adherence audits (Ahmed et al., 2025) verify statement-level compliance but never invert a parameter. Our statistics import equivalence testing (Lakens, 2017; Tomek et al., 2026) and the region-of-practical-equivalence logic of Benavoli et al. (2017) into a setting where paired

comparisons are routinely underpowered (Card et al., 2020; Kotawala, 2026); the permutation and bootstrap machinery follows Koehn (2004); Berg-Kirkpatrick et al. (2012); Zmigrod et al. (2022); the pre-registration discipline follows van Miltenburg et al. (2021).

Key wording in structured input. Renaming variables in tabular prompts (Liu et al., 2025) and varying only the key wording of an output schema (Le, 2026) move accuracy substantially in other settings, so the wording of keys in structured input is behaviorally active beyond the specifications studied here.

I Record of Corrections

Three corrections were made during the analysis and are part of the record. First, for several analysis-code revisions the analysis headers claimed Holm correction while the code applied none; fixing it moved exactly two rungs (reversal and single-rename, both DIFFERENT→INDETERMINATE), and all eight rungs at the permutation floor land at Holm p=.0032. Second, the scale arms and the external demonstration were written with the specification in the system prompt while the 3B battery places it in the user turn; the discrepancy was found after the scale runs, when a single-dimension sweep under system placement sat $\sim$0.06 above the battery's matched condition. It was resolved by re-running the 3B headline conditions under the scale arms' placement and the single-dimension endpoints under the battery's placement (§5.4, §5.3). An earlier draft, comparing across placements, had concluded that the name effect "halves" with scale; within placement it does not. Both corrections were made before any claim in this version was written, and the pre-registered predictions attached to each affected run are reported as they were registered.

Third correction: an unsaved intermediate and a cross-run reference. Preparing the artifact release, an inventory of every file on disk showed that the prose single-dimension experiment was two generation runs rather than one, and that the `advise` conditions had no sentence-level classifier labels saved anywhere—the outcome variable used throughout this paper had been computed for them at some point and then lost. The figures those labels had produced were carried in earlier drafts. Re-scoring the stored responses with the same splitter and classifier reproduces them exactly: for each of the three models the difference between the reported high and low figures equals the freshly computed within-run swing to four decimal places. The measurement was therefore sound and the arithmetic correct; what failed was release discipline, plus a second defect the reproduction exposed—the neutral reference those figures used belongs to the *other* generation run, so the published contrast compared conditions generated at different times. The main text now quotes the within-run swing, which has neither problem, and Appendix E.6 states both defects. Neither the pre-registration nor any of six external reviews caught this; the artifact inventory did, which is the argument for running one.

What the pipeline now checks. The final battery pipeline (cell A2 v3.14.2) re-derives every Holm-adjusted p independently and asserts that it matches the printed value and that every verdict is reproducible from its own p-values (the check that would have caught the first correction); it hashes all 69 input files into a manifest, so that the sentence-level labels behind every table are identifiable and a recomputation from a different run would change the manifest (the third correction); and the placement of every arm is stated in Appendix D and in the scale table rather than assumed (the second correction). The released end-to-end reproduction is the version used for every reported table.