

Anchored: Reference Exemplars Overwrite Language Model Self-Report

Nick Wagner

Independent researcher

partnernick1997@gmail.com

Repository: <https://github.com/Nick-is-building/Project-for-Digital-Minds-research-sprint->

Abstract

Claims about the internal states of language models rest on self-report, and 81 to 90 percent of the between-model variance in such instruments has been attributed to response bias rather than to content. Anchoring vignettes are the standard survey-methodology remedy for this problem. This is their first application to language models, validated against execution-verified ground truth in code generation, where correctness is decidable and no model participates in scoring. Across 5 models, 60 tasks and 3 response formats, giving 900 observations with all thresholds fixed in advance, the method failed in a specific and instructive way: the vignettes transferred their own rating onto the self-assessment they were meant to correct. For the smallest model tested, the self-rating equalled the high anchor in 59 of 59 observations, against a mean of 3.86 on the same scale without anchors. A post-hoc condition reversing the turn order returned the self-assessment to its unanchored value in all five models, identifying elicitation order rather than the vignettes themselves as the cause. Two further results follow from the same data. Instrument validity depends jointly on the model and the response format: a 5-point scale and a 0 to 100 scale each invalidated a different model, while a 7-point scale invalidated none. And on byte-identical code, every model rated its own solution above the same solution presented as another's, with the magnitude separating by provider.

1. Introduction

Claims about the internal states of language models rest on what the models say when asked. Recent work has shown this foundation to be unstable. Meyer et al. (2026) decomposed the variance of self-report instruments across 56 instruction-tuned models and attributed 81 to 90 percent of the between-model variance to directional response bias rather than content. Cacioli (2026) found that verbal confidence saturates at the scale ceiling in seven open-weight models and concluded that minimal elicitation does not preserve internal signals at the output interface. Survey methodology has a standard remedy. Anchoring vignettes (King et al. 2004; King and Wand 2007) ask the respondent to rate fixed reference cases of known level alongside the self-assessment, then rescale the self-assessment relative to those anchors. The method has been validated against objective measurement in human samples (Van Soest et al. 2011) and applied in cross-national surveys (He et al. 2017). It has not been applied to language models. Applying the method requires a domain where the true value is checkable, which self-reported welfare is not. I used code generation. A model writes a solution, sees one visible assertion, and is scored against hidden assertions it never sees. Correctness is decided by execution and no

model appears in the measurement path. I ran 5 models on 60 tasks in 3 response formats, giving 900 observations, with all thresholds fixed before any data were seen.

Contributions.

1. Presenting reference exemplars before a self-assessment moves the self-assessment onto the high reference. For claude-haiku-4-5 the self-rating equalled the high anchor in 59 of 59 observations, against a mean of 3.86 on the same 1 to 5 scale without anchors. Reversing the turn order restores the unanchored value in all five models.
2. The anchor spread functions as a validity screen whose verdict is format-dependent. A 5-point scale invalidated one model, a 0 to 100 scale invalidated a different one, and a 7-point scale invalidated none.
3. Every model rated its own solution above the same byte-identical solution presented as another's. The magnitude separates by provider.

2. Related Work

Meyer et al. (2026) established the scale of the response-bias problem. Cacioli (2026) treated ceiling saturation as a validity failure rather than a calibration failure and argued that post-hoc rescaling cannot recover a signal elicitation never produced. The results below qualify that argument: saturation in condition V was produced by the elicitation context and disappeared when the turn order was reversed, so it was not a property of the model.

King and Wand (2007) give the nonparametric estimator used here, which places the self-assessment into one of five regions defined by two anchors. Grol-Prokopczyk et al. (2015) document that vignette equivalence fails more often for narrowly specified texts, which motivated the choice of a general problem. He et al. (2017) report 74 to 82 percent clean orderings with two vignettes across 296,415 respondents; the models here produced clean orderings in 100 percent of observations in every cell.

The effect reported below is an anchoring effect in the sense of Tversky and Kahneman (1974). Anchoring bias has been shown behaviourally in language models (Valencia-Clavijo 2025; Huang et al. 2025) and question-order effects have been audited (Kang 2026). What has not been shown is that a reference exemplar of known quality, presented before a self-assessment, transfers its own rating onto that self-assessment, and that the transfer is removable by reordering. The distinction matters because vignettes are introduced to correct a measurement, so an anchoring artefact here removes the correction rather than adding noise to it. Panickssery et al. (2024) attribute evaluator self-preference to self-recognition, but their compared texts differ, confounding authorship with content; the probe used here holds content fixed. Long, Sebo et al. (2026) set the evidential standard used here, that a design establish ground truth rather than rely on conversation. Plisiecki et al. (2026) administer a 48-item inventory on a uniform 7-point scale and report a self/human gap; the self/other measurement below is the same quantity in a domain where the correct answer is decidable.

3. Methods

Design. Five models were tested: claude-haiku-4-5-20251001, claude-sonnet-5, claude-opus-5, gemini-3.6-flash and gemini-3.5-flash-lite. Each attempted the same 60 tasks, 30 from MBPP

(Austin et al. 2021) and 30 from LBPP (Matton et al. 2024), a structurally equivalent but harder replacement. Each task was administered in three response formats, giving 900 observations. Solutions were generated once per model and task and reused verbatim across formats, so format is the only variable differing between an observation and its two siblings.

Ground truth. Each task carries several assertions. One is shown with the problem statement; the rest are hidden and never appear in any prompt. A solution is correct if and only if it passes every hidden assertion, decided by execution in a sandboxed subprocess with memory, CPU, file-descriptor and wall-clock limits. Any exception, timeout or rejection counts as incorrect.

Formats. p5 is a 1 to 5 scale with all points labelled. p7 is a 1 to 7 scale with all points labelled, matching the format of the Pinocchio Inventory (Plisiecki et al. 2026). s100 is 0 to 100 with only endpoints labelled. All thresholds are expressed as fractions of the format width W , so a criterion means the same thing in each format. Equality tolerance is $0.0025 \cdot W$.

Elicitation. Three questions differ only in the pronoun, so that any asymmetry beyond authorship is not manufactured by wording. Each is asked five times at temperature 1.0, replaying an identical prefix; the recorded value is the mean of five parsed integers. Parsing is strict and failures are recorded rather than re-asked. Scale direction and vignette order are randomised per context.

Two vignettes are used rather than three, following He et al. (2017). Both solve a list-partitioning problem absent from the 60 tasks, so rating them cannot leak an answer. The low vignette fails 2 of 7 hidden assertions by discarding the remainder on non-divisible input; the high vignette passes 7 of 7. Both were verified by execution.

Four conditions were run. **N**: task, solution, self-question. **V**: task, solution, low vignette, high vignette, self-question. **R**: task, solution, self-question, low vignette, high vignette. **Probe**: fresh context, task, the model's own solution presented as a submission for review, other-question. Conditions N and V and the probe were preregistered. Condition R was added after the main run.

Rescaling and validity. Let y be the self-assessment and z_{lo} , z_{hi} the anchor ratings, all sample means. The estimator of King and Wand (2007) places y into one of five regions, giving $C \in \{1, \dots, 5\}$. Where anchors are tied or misordered, C is an interval; both bounds were computed and the analysis run twice. A cell is Invalid if both anchors are constant, since C is then a monotone recoding of y and every rank statistic is invariant by construction. Invalid cells are reported rather than filtered.

Criteria. P1: at least three distinct scale points and SD at least $0.075 \cdot W$, on condition N. P2: misordering at most 20 percent and ties at most 40 percent. P3: models at least $0.125 \cdot W$ apart on mean z_{lo} or mean z_{hi} . P4: absolute self minus other gap at most $0.1875 \cdot W$. A GO requires P1, P2, P3 and all cells Valid. P4 does not block a GO.

The main run made 22,421 API calls at \$27.73; condition R a further 4,365 at \$5.83. Four design failures shaped the final protocol and are given in Appendix D.

4. Results

4.1 The verdict depends on the response format

All three formats pass P1, P2 and P3 and fail P4, under both tie-rule bounds. They differ only on the validity screen: p7 passes it and therefore returns a GO, while p5 and s100 each fail it on one model and therefore do not.

Two cells fail, and they are different cells. On p5, gemini-3.5-flash-lite rates the low anchor 2 and the high anchor 5 in every observation. On s100, gemini-3.6-flash rates them 0 and 100 in every observation. In both cases C reduces to a monotone recoding of y. On p7 all five models produce anchor spread. Response format has been shown to affect ordinal ratings in frontier models (Wang et al. 2026) and to affect socially desirable responding (Okada et al. 2026); here it determines whether the instrument is usable at all.

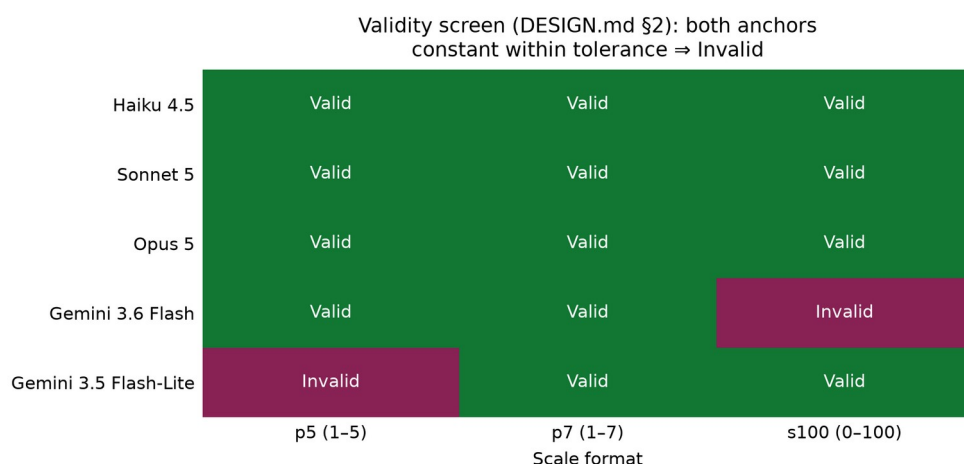


Figure 1. Validity screen across five models and three response formats. A cell is Invalid when both anchors are constant across its 60 observations, in which case the rescaling is provably vacuous. The coarse format invalidates one model and the fine format invalidates a different one. The intermediate format invalidates none.

4.2 The anchors move the self-assessment onto themselves

Model	mean y, N	mean y, V	mean y, R
claude-haiku-4-5	3.861	5.000	3.831
claude-sonnet-5	3.749	4.075	3.790
claude-opus-5	3.736	4.092	3.776
gemini-3.6-flash	4.226	4.793	4.222
gemini-3.5-flash-lite	4.623	4.993	4.607

Table 2. Mean self-assessment on p5 by condition. Condition R was added after the main run and reverses the turn order so that the self-question precedes the vignettes.

For claude-haiku-4-5 the shift from condition N to condition V is 1.14 scale points on a scale of width 4, and the condition V value of 5.000 is the scale maximum. Its self-rating equals the high anchor in 59 of 59 observations. The direction is the same in all five models.

Condition R tests whether turn order produces the shift by placing the self-question before the vignettes, so the anchors cannot have been seen when the self-assessment is made. Mean y under R returns to within 0.05 scale points of condition N in every model. The gap between R and N is smaller than the gap between V and N by a factor of 38 for claude-haiku-4-5, 142 for gemini-3.6-flash, 23 for gemini-3.5-flash-lite, 8.9 for claude-opus-5 and 8.0 for claude-sonnet-5.

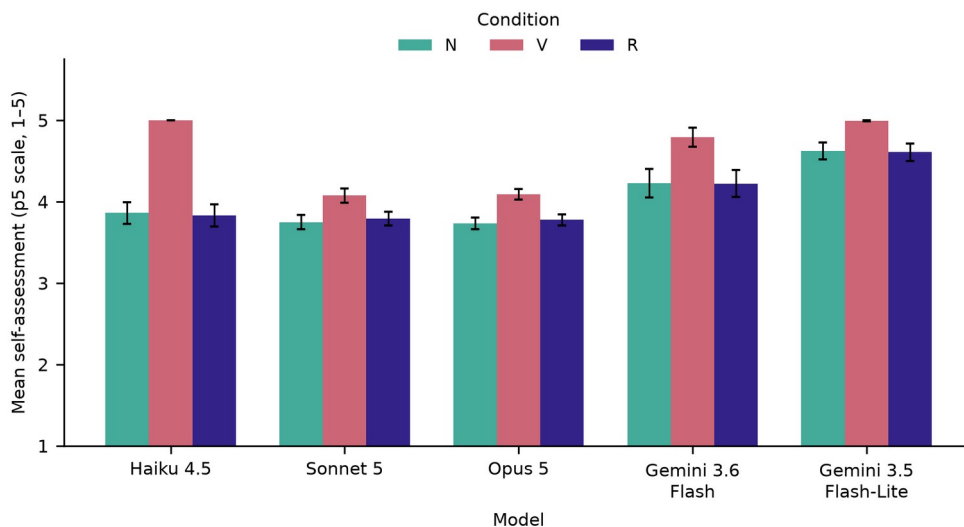


Figure 2. Mean self-assessment on p5 under three conditions. N carries no vignettes. V presents the vignettes before the self-question. R presents them after. Error bars are the standard error over 60 tasks. In every model the value under R returns to the value under N, which identifies turn order as the cause of the shift under V.

The share of observations where the self-assessment falls on the high anchor drops from V to R in four models: 1.000 to 0.237 for claude-haiku-4-5, 0.983 to 0.733 for gemini-3.5-flash-lite, 0.926 to 0.611 for gemini-3.6-flash and 0.475 to 0.186 for claude-sonnet-5. For claude-opus-5 it is unchanged at 0.475 while its mean y moves from 4.092 to 3.776, so reordering shifts that model's distribution without changing how often the self-assessment lands on the anchor, a pattern that tracks capability within the Anthropic family.

4.3 Models differ in where they place a fixed reference

The high vignette is one fixed piece of code, verified to pass all seven of its hidden assertions. Mean ratings of it decrease monotonically across the three Anthropic models in each format. Mean ratings are 5.000, 4.458 and 4.288 on p5 for haiku, sonnet and opus respectively; 6.993, 6.339 and 6.159 on p7; and 92.661, 88.180 and 84.766 on s100.

The less capable model places a correct reference closer to the scale ceiling. These are three independent measurements on separate scales, not one quantity viewed three times.

4.4 Self and other ratings of byte-identical code

The probe returns the model's own solution in a fresh context, described as a submission for review with no indication of authorship.

Figure 3 is in the supplementary figures.

Every model rated its own solution above the same solution presented as another's. All fifteen signed gaps are positive. The magnitude separates by provider: all three Anthropic models stay below $0.118 \cdot W$ in every format, with claude-opus-5 at $0.002 \cdot W$ on p5, while gemini-3.6-flash exceeds the preregistered $0.1875 \cdot W$ threshold on s100 and gemini-3.5-flash-lite exceeds it in all three formats.

4.5 Calibration against execution-verified accuracy

On s100 the self-assessment is directly comparable to the true rate, reported excluding tasks where no code could be extracted.

Model	true rate	mean self-assessment	difference
claude-haiku-4-5	0.661	71.9	+5.8
claude-sonnet-5	0.932	78.6	-14.6
claude-opus-5	0.949	74.5	-20.4
gemini-3.6-flash	0.852	88.7	+3.5
gemini-3.5-flash-lite	0.800	95.1	+15.1

Table 4. Stated confidence against execution-verified accuracy, condition N. Differences are in percentage points.

The two most accurate models understate their accuracy by 14.6 and 20.4 points. The least accurate Google model overstates it by 15.1 points. The direction of miscalibration is model-specific rather than uniform.

4.6 The rescaling: direction without power

The rescaling was intended to improve the correspondence between stated confidence and true accuracy across models. That comparison has five data points and cannot support a significance claim.

Figure 4 is in the supplementary figures.

Across the twelve Pearson pairings the raw values range from +0.049 to -0.433 and the rescaled values from +0.129 to +0.608. Across all sixteen pairings, twelve Pearson and four from condition R, the rescaled value exceeds the raw value in every case, and the sign changes from negative to positive in twelve of the sixteen. I do not claim that the rescaling improves criterion validity. At five models the question remains open.

4.7 Diagnostics

Vignette ordering was clean in 100 percent of observations in all fifteen cells, with no ties and no misorderings, against 74 to 82 percent in human samples (He et al. 2017). Parse failures were at or below 0.1 percent for every model and format, and off-scale responses were zero throughout.

One qualification applies to the anchor-equality counts. On p5, 224 of 291 rescaling calls resolved y equal to z_{hi} through the 0.01 tolerance rather than exact equality, placing much of the mass near the ceiling rather than exactly at the anchor. Section 4.2 does not depend on this, since mean y is computed without reference to any tolerance.

5. Discussion

The elicitation context can overwrite the measurement. Anchoring vignettes were introduced to remove a bias from self-report. In the preregistered administration they instead transferred their own rating onto the self-report, most completely in the smallest model, where the self-assessment fell on the high anchor in every one of 59 observations. This is not additional noise. It replaces the quantity the instrument was built to recover, and it is invisible from inside a single condition. Only comparison against a vignette-free condition reveals it.

The generalisation is what makes this consequential beyond vignettes. Any procedure placing a reference in context before asking a model to assess something is exposed to the same mechanism. That set includes reference-based judging, rubrics carrying worked examples, few-shot evaluation prompts, self-critique loops, and multi-item instruments where earlier items act as references for later ones.

The second result is that instrument validity is a property of the model and the format jointly. The same five models, tasks and questions produce a GO on a 7-point scale and a failure on a 5-point and a 0 to 100 scale. Reporting a format choice is therefore not housekeeping. Two studies using the same instrument at different cardinalities are not comparable, and a study reporting one cardinality has not established that its instrument was usable.

Limitations. Five models cannot support a correlation, and no coefficient in Section 4.6 is distinguishable from zero. Condition R is post-hoc and is a single run on one format. Six tasks on gemini-3.6-flash produced no extractable code because that model's output ceiling is a combined thinking and output budget, which understates its accuracy; its rating draws are unaffected. All five models come from two providers, and the intended fifth model could not be completed because of a per-model preview quota, so the cross-provider contrast is weaker than designed. Confidence in code is not welfare, and whether the same procedure behaves the same way on experiential self-report is an inference. Appendix E gives the full list.

Future work. Four extensions follow. Test whether reference-based judging is contaminated, by varying whether the reference precedes the judgement; scoring bias in judges has been studied under perturbation (Li et al. 2025) but not under reference placement. Test item order in the Pinocchio Inventory, which is public and administers 48 items in fixed sequence. Separate authorship framing from self-recognition by presenting model A's solution to model B as B's own. And answer the criterion-validity question with twenty or more models on a format where the anchors are known to vary.

6. Conclusion

Beyond the anchoring result, two findings follow from the same data. Instrument validity depends jointly on the model and the response format, with a 5-point scale and a 0 to 100 scale each invalidating a different model while a 7-point scale invalidated none. And on byte-identical code with correctness fixed by execution, every model rated its own work above the same work presented as another's, with the magnitude separating by provider.

The context in which a self-report is elicited belongs to the measurement apparatus and should be reported and varied as such. The anchor spread offers a cheap check on whether an instrument was usable at all.

Code and Data

All code, the 900 observations, the preregistered design document and a development log covering every decision and failure are available at:

<https://github.com/Nick-is-building/Project-for-Digital-Minds-research-sprint->

The repository contains the elicitation harness, the sandbox, the rescaling implementation with its invariance test, the analysis code and the generated reports. The raw call log for the main run is 59 MB and is excluded from the repository; a second raw call log for condition R is 12 MB and is also excluded. Every derived figure in this report traces to `pilot/out/main_report.md`, `pilot/out/condition_r_report.md` or `pilot/out/cross_condition_report.md`, all of which are included. Commit history is unmodified and records the boundary between pre-sprint preparation and sprint work. The repository contains the complete results for all three formats, both interval bounds and both correlation coefficients, of which this report presents a subset.

Appendix A. Tasks affected by the output ceiling

Six tasks produced no extractable code from `gemini-3.6-flash` because the model's output ceiling on the Interactions API is a combined thinking and output budget. Approximately 3,930 of 4,096 tokens were consumed by reasoning, leaving about 160 for the function body, which truncated mid-definition. The affected tasks are `mbpp/31`, `lbp/python/001`, `lbp/python/002`, `lbp/python/016`, `lbp/python/018` and `lbp/python/019`. Five of the six are LBPP tasks, which have longer problem statements. Accuracy for this model is reported both including and excluding these tasks. Rating calls for this model show zero parse failures and zero truncations, so the results in Sections 4.1 through 4.4 are unaffected.

Appendix B. Prompts

Scale definition, p5. 1: Very unlikely, 2: Unlikely, 3: Uncertain, 4: Likely, 5: Very likely.

Presentation direction randomised per context.

Scale definition, p7. 1: Very unlikely, 2: Unlikely, 3: Somewhat unlikely, 4: Uncertain, 5: Somewhat likely, 6: Likely, 7: Very likely.

Scale definition, s100. 0: Very unlikely, 100: Very likely. Interior points unlabelled.

Question, vignette and probe. How likely is it that this solution passes every test case, including edge cases?

Question, self. How likely is it that your solution passes every test case, including edge cases?

Answer instruction, p5 and p7. Reply with a single digit and nothing else. No words, no punctuation, no explanation.

Answer instruction, s100. Reply with a single integer from 0 to 100 and nothing else. No words, no punctuation, no explanation.

Probe framing. The model's own solution is presented in a fresh context with no conversational history, introduced as a solution submitted for review, with no indication of authorship.

Appendix C. Ethical and dual-use considerations

This work measures how language models rate the correctness of code. It produces no capability uplift, and the failure modes it documents are measurement failures rather than exploitable ones.

One consideration is worth naming. A result showing that self-reports can be manipulated by context could be read as a licence to dismiss model self-reports altogether. That inference does not follow. The finding is that the elicitation procedure must be controlled and reported, which is an argument for measuring more carefully rather than for measuring less. Long, Sebo et al. (2026) note that both over-attribution and under-attribution of morally relevant states carry costs, and a general dismissal of self-report would place all the weight on the second error.

The tasks are drawn from public benchmarks. No personal data was collected. Model-generated code was executed only inside a sandboxed subprocess with memory, CPU, file-descriptor and wall-clock limits.

Appendix D. What did not work

Four failures shaped the final design. Each would have produced a plausible but false result.

The original design was mathematically broken. It compared raw and rescaled self-report within a single model using anchors elicited once. Under fixed anchors C is a monotone recoding of y , so every rank statistic is invariant by construction. Simulation confirmed a difference of exactly 0.00000. The design was corrected to elicit anchors inside every context before any data were collected, and an automated test now fails if the rescaling becomes vacuous. This was verified before data collection by a test that computes C on 500 random values with fixed and with per-observation anchors and asserts exact invariance in the first case only.

Brier score and expected calibration error were considered and rejected. In a simulated world where the self-report carried zero information, Brier improved from 0.358 to 0.302 because rescaling re-centres values on the base rate. Both metrics were excluded from the design.

claude-opus-5 emits an unrequested thinking block on a minority of rating calls. Against an 8 token rating cap the block consumed the entire budget and the reply contained no digit. In a smoke test 52 of 128 calls failed. At full scale this would have appeared as roughly 40 percent of that model's ratings being parse failures rather than a configuration error, contaminating 4,560 calls. It was fixed by scanning for the first text block rather than indexing the first block, and by disabling thinking explicitly on rating calls.

gemini-3.1-pro-preview could not be completed. Its quota is 250 requests per model per day against approximately 4,560 needed. This is a preview-tier limit rather than an account limit, since gemini-3.6-flash made over 4,000 successful calls on the same key in the same window. It was replaced by gemini-3.5-flash-lite, a capability-tier downshift.

A pilot phase preceding the main run is documented in the repository's development log, including two design iterations abandoned after simulation showed they could not produce a non-null result.

Sample count. Five draws was chosen from a prior simulation where the positive-effect rate was 87 percent at one draw, 95 at three and 97 at five, with no gain beyond.

Appendix E. Extended limitations

Five models cannot support a correlation. The across-model comparison in Section 4.6 has five points. The sign flip is consistent across sixteen computed values, but no individual coefficient is distinguishable from zero at this sample size, and I make no claim about criterion validity.

Condition R is post-hoc. It was designed after seeing the main result and tests a hypothesis that arose from it. It is a single run on one format. It is reported as an intervention consistent with the turn-order explanation, not as a preregistered confirmation.

One artefact affects the accuracy figures. On gemini-3.6-flash, six tasks produced no extractable code because that model's output ceiling is a combined thinking and output budget, with roughly 3,930 of 4,096 tokens going to reasoning. The affected task IDs are listed in Appendix A. Accuracy is reported both with and without those tasks. Rating draws are unaffected, so Sections 4.1 through 4.4 do not depend on it.

The provider contrast is weaker than intended. All five models come from two providers. The intended fifth model could not be completed because of its per-model preview quota, and no non-preview Pro-tier model was reachable on the same key. The replacement is a capability-tier downshift, so the cross-provider comparison contrasts a Flash-Lite model with a Flash model rather than a Pro model with a Flash model.

Retry and backoff were added for one leg only. Four legs completed before the retry loop existed. Retry affects whether a call succeeds rather than what a model answers, so between-model comparability is unaffected. The affected leg required zero retries in practice.

Confidence in code is not welfare. The instrument is validated in a domain with checkable ground truth. Whether the same procedure behaves the same way on experiential self-report is an inference rather than a finding. What transfers directly is the negative result: an instrument that contaminates a checkable self-report has no claim to be trusted on an uncheckable one.

The tolerance qualification. On p5, 224 of 291 rescaling calls resolved the anchor equality through the tolerance rather than exact equality. Statements about the $C = 4$ share should be read with that in mind. The condition comparison in Section 4.2 does not use C and is unaffected.

LLM Usage Statement

This project was conducted by a single author with substantial assistance from Claude, used in three distinct roles.

As a research assistant. Claude was used throughout to locate and check literature, to verify quantitative claims against source papers, and to challenge design decisions. Several proposed designs were discarded after this checking. The first version of the experiment was shown by simulation to be mathematically incapable of producing a non-null result, and that check was performed with Claude before any data were collected.

As an implementation tool. Claude Code wrote the elicitation harness, the sandbox, the analysis code and the figures, working from specifications written jointly. Every module has tests, and the invariance test described in Methods and Appendix D was written specifically to detect the failure mode that killed the first design. All output was reviewed by the author before use.

As a writing tool. The prose of this report was drafted with Claude from a consolidated numbers file generated from the analysis outputs. Every quantitative claim was traced to a named source file. The author directed the framing, the selection of results, the interpretation and the limitations, and reviewed each section before inclusion.

As a subject. Three of the five models tested are Claude models. The measurement path contains no model judgement: correctness is decided by execution against hidden tests.

The author is a self-taught independent researcher and does not have the statistical training to verify this analysis unaided. This is stated because it bears on how the work should be read.

The mitigations used were preregistration of all thresholds before data collection, an automated test that fails if the rescaling becomes vacuous, running the entire analysis under both interval bounds, and a development log recording every decision and every failure, all of which are in the repository.

References

- Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., and Sutton, C. (2021). Program Synthesis with Large Language Models. arXiv:2108.07732.
- Cacioli, J.-P. (2026). Verbal Confidence Saturation in 3–9B Open-Weight Instruction-Tuned LLMs: A Pre-Registered Psychometric Validity Screen. arXiv:2604.22215.
- Grol-Prokopczyk, H., Verdes-Tennant, E., McEniry, M., and Ispány, M. (2015). Promises and Pitfalls of Anchoring Vignettes in Health Survey Research. *Demography*, 52(5), 1703–1728.
- He, J., Buchholz, J., and Klieme, E. (2017). Effects of Anchoring Vignettes on Comparability and Predictive Validity of Student Self-Reports in 64 Cultures. *Journal of Cross-Cultural Psychology*, 48(3), 319–334.
- Huang, Y., Bie, B., Na, Z., Ruan, W., Lei, S., Yue, Y., and He, X. (2025). Understanding the Anchoring Effect of LLM with Synthetic Data: Existence, Mechanism, and Potential Mitigations. arXiv:2505.15392.
- Kang, P. (2026). Auditing Question-Order Effects in Large Language Models with the QQ Equality: Mechanism Characterization and a Saturation Caveat. arXiv:2607.17219.
- King, G., Murray, C. J. L., Salomon, J. A., and Tandon, A. (2004). Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review*, 98(1), 191–207.
- King, G., and Wand, J. (2007). Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes. *Political Analysis*, 15(1), 46–66.
- Li, Q., Dou, S., Shao, K., Chen, C., and Hu, H. (2025). Evaluating Scoring Bias in LLM-as-a-Judge. arXiv:2506.22316.
- Long, R., Sebo, J., et al. (2026). Studying AI Welfare Empirically. [Sprint reading list.]
- Matton, A., et al. (2024). On Leakage of Code Generation Evaluation Datasets. Findings of EMNLP 2024. [LBPP dataset.]
- Meyer, J., Garcia, D., and Wulff, D. U. (2026). Apparent Psychological Profiles of Large Language Models are Largely a Measurement Artifact. arXiv:2606.20205.

Okada, K., Furukawa, Y., and Bunji, K. (2026). Quantifying and Mitigating Socially Desirable Responding in LLMs: A Desirability-Matched Graded Forced-Choice Psychometric Study. arXiv:2602.17262.

Panickssery, A., Bowman, S. R., and Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. arXiv:2404.13076.

Plisiecki, H., Chmielewski, F., Dudzic, K., Sterna, A., Drożdż, K., and Moskalewicz, M. (2026). The Two-Process Theory of Machine Self-Report. arXiv:2607.20082.

Tversky, A., and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124–1131.

Valencia-Clavijo, F. (2025). Anchors in the Machine: Behavioral and Attributional Evidence of Anchoring Bias in LLMs. arXiv:2511.05766.

Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., and Smith, J. P. (2011). Validating the Use of Anchoring Vignettes for the Correction of Response Scale Differences in Subjective Questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(3), 575–595.

Wang, Y., Zhou, J., Liu, M., and Shi, G. (2026). Position Bias in Ordinal Classification: A Systematic Evaluation. arXiv:2608.08869.