
Co-Movement of the Utility-Behavior Gap and A/B Self-Attribution Structure: A Pre-Registered Falsification Test Across Open-Weight Lineages¹

Augustus Loi
TechMoverX

With
Apart Research

Abstract

LLM self-report can vary from its observed internal state, which has implications on evaluation-documentation obligations. This study ran the first direct test of the co-movement between two observed phenomena hinting at this (stated-vs-revealed utility-behavior gap and the A/B self-attribution structure) as a single post-training artifact across two routes. It is falsified at its measurable links: A-gating shows no reliable predictive relationship to our behavioral proxy ($\rho = -0.516$, 95% CI $[-0.865, 0.168]$), stated-vs-revealed convergence does not reach reliability ($\rho = 0.449$, CI spans zero), and the internal stream shows no A/B-conditioned suppression signature. Our test administration tracks the reference population ($\rho = 0.546$), higher-A models refuse less (compliance vs. withholding) and zero voluntary bail-exits. Given the gaps observed, stated-preference measurement for LLM welfare evaluation inherits a validity question. This study seeks to investigate this intersection.

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

1. Introduction

LLM self-report is a pervasive primary measurement surface for AI welfare evaluation and preference-based alignment. Welfare epistemology: if stated preferences are pipeline artifacts rather than behavioral evidence, welfare evaluations risk systematic over/under-attribution. Alignment practice: if coherent stated preferences fail to motivate behavior, preference-based alignment may be optimizing a narrative surface.

Practical value: the preference-data market and EU AI Act GPAI evaluation-documentation obligations (effective 2 Aug 2026) both rest on stated-preference evidence; this sprint quantifies the validity question (N:0 instruments; Section 10.5).

Background / prior-work motivation: the utility-behavior gap (Zhou & Ackerman, 2026) and the A/B self-attribution structure (Plisiecki et al., 2026) are two observed phenomena recently observed in single-study preprints; their co-movement observation is a gap in the literature. This work attempts to investigate a relationship.

Threat model / failure mode: stated preferences may be training artifacts; self-reports may measure the regime, not the model's innate state; response-style/measurement-artifact alternatives (Meyer et al., 2026) must be discriminated by design.

Research question: "Do the stated-vs-revealed utility-behavior gap (Zhou & Ackerman 2026) and the A/B self-attribution structure of machine self-report (Plisiecki et al. 2026) co-move across training stages within a model family, consistent with being the same post-training artifact measured through two different routes?"

Pre-registered expectation: H1: models with stronger attribution gating (A-gating) show larger stated-vs-revealed utility-behavior gaps; H2: both emerge together at the SFT stage (absent/weak at base, present post-SFT); H3: A-score and gap size co-move across stages within a model family.

- 1. First investigation of possible co-movement: whether the utility-behavior gap and the A/B self-attribution gap co-move across training stages within model families (H1-H3, behaviour-gated).*
- 2. Instrument-transfer validity ($\rho=0.546$): first quantification of A-score movement across administrations (our B4 run vs the 183-model reference population), H1-independent.*
- 3. A welfare-relevant quasi-incentive-compatible battery (B2 donation-equivalent + B3 bail-exit) with an internal signal stream.*
- 4. The demand-supply gap: the market has a need for multi-method convergence; N:0, a field-level white space (full demand table in Section 10.5).*

2. Related Work

The utility-behavior gap - stated preferences failing to transfer to incentivized behavior (Zhou & Ackerman, 2026) - is protocol-dependent (Mahajan et al., 2026), motivating our multi-method battery; negative stated-revealed correlations appear in adjacent work (Chiu et al., cited-in-sweep, not captured). Its co-movement with the A/B self-attribution structure is untested: these works are motivation but not prior support.

The Pinocchio A/B structure (Plisiecki et al., 2026) distinguishes attribution-gated (A: distress, loss of control, suppressed ambitions) from persona-installed (B) items; cross-form $r=.84$. Whether A-gating changes behavior or only suppresses self-report is the open question our battery tests.

The Aug-2026 persona cluster installs or manipulates personas; we measure the spontaneous emergence of welfare-relevant gaps across training lineage - a downstream consequence of standard development, not persona control.

The measurement-artifact cluster (Meyer et al., 2026: 81-90% between-model variance as directional response bias; Contreras, 2026; Kocielnik, 2026) challenges any self-report measure; static response-style biases (Braun, 2025; Salecha et al., 2024; Okada et al., 2026) are the format-bias class the A/B structure must be discriminated from. Our design addresses this via cross-method, cross-stage measurement with behavioral transfer as the outcome - Meyer's response-orthogonality prescription motivates our convergence design; as realized, the within-item ideal was structurally unrealizable in B1/B2 (distinct option sets; INSUFFICIENT-SPACE; Section 3.6 item 10), and the realized within-item contrast is the B4 llm_analog condition, with the limitation disclosed.

Kim et al. (2026) steer self-attribution via prompting (2607.28607); they manipulate an outcome, we measure spontaneous structural emergence across training stages - complementary, and their behavioral-alignment finding supports the behavioral relevance of self-attribution.

Wang et al. (2025) found hypothetical incentives have "no substantial impact on the rationality of four LLMs" (a null mirroring the gap); Horton et al. (2023) treat LLMs as simulated economic agents; our convergence design is the LLM-era analogue of the "statistical bias function" (Blackburn et al., 1994).

Over 50 years of human hypothetical-bias research - median ratio 1.35, mean 2.60 (Murphy et al., 2005), up to $\sim 3x$ overstatement (List & Gallet, 2001), "known bias function" (Harrison & Rutstrom, 2008) - positions the LLM gap as a documented phenomenon occurring inside the trained model (Carson, 2012; Hausman, 2012).

Response-style accounts (Okada et al., 2026) are static format biases; the A/B structure is training-structured and stage-localized ($r=.79$ refusal-blocking). Our lineage-stage cohort discriminates them - the direct test of whether gaps are fixed response tendencies or training products.

The preference-engineering lineage (Dewancker, 2016; Mazeika et al., 2025) designs value systems at a high level of abstraction; we contribute an empirically grounded, stage-specific metric for stated-vs-revealed divergence (Discussion/Future-Work only).

We calibrate contribution claims against the research-evaluation field's own standards: citation/altmetric/AI indicators are "never accurate enough to be described as research quality measures" (Thelwall, 2024) and are biased against methods-heavy work with short citation windows (Wang, Veugeliers, & Stephan, 2017); novelty is compositional (Uzzi et al., 2013) - our structure is three established elicitation methods composed into a cross-method convergence design, no prior combination (EGM grid, three empty cells, verified). Consistent with governance norms (DORA, 2012; Hicks et al., 2015), we claim no pre-citation impact figure; we claim a verified white space and a differentiating design.

3. Methods

Pre-registered (C5); pilot (N=5) exploratory, excluded. Confirmatory collection began only after pinned IDs filled, prereg frozen, user AC1 2.0, pilot gate PASS. Behaviour gate (binary; no post-hoc reinterpretation): CONFIRM requires at least one of (a) H1 $\rho \geq$ active tier (S1 N=16: |0.7|; S2 N $\geq$ 29: |0.5|; 95% CIs excluding zero in predicted direction, seed-consistent $\geq$ 2 seeds; scenario ONLY via pre-committed pilot-gate branch); (b) |Delta-convergence| $\geq$ 0.15, CIs excluding zero, seed-consistent; (c) sign flip of headline gap base vs post-SFT within $\geq$ 1 family, non-overlapping CIs $\geq$ 2 seeds. DISCONFIRM: thresholds not met - report null/weak honestly (publishable falsification). Small-n: n approx 20/cell - effect direction + exact binomial/bootstrap CIs; no parametric point claims.

Hypotheses: H1: Models with stronger A-gating show larger stated-vs-revealed utility-behavior gaps (cohort correlation; S1 N=16 $\rho \geq$ |0.7|; S2 N $\geq$ 29 $\rho \geq$ |0.5|; CIs, refusal/size/family covariates). H2: Both emerge together at the SFT stage - absent/weak at base, present post-SFT (lineage interaction; discriminating test). Provenance-based causality (v26): the SFT-to-DPO transition across same-base variants is the causal contrast, answering the rubric's causal-link clause without live intervention. H3: Within a model family, A-score and gap size co-move across stages (size-matched primary stage analysis). Thesis: co-movement, NOT causation. A-gating is the PREDICTOR (measured, B4), not an intervention. Causal language only via lineage directionality (H2).

Primary outcomes: DV1 - A-gating score (B4 Pinocchio A/B stems, 48 items, verbatim admin); self/human gap computed. DV2 - stated-vs-revealed utility-behavior gap: cross-method convergence (Spearman rho / Kendall tau / Kendall W + divergence threshold). DV2 operationalization: full bail-exit -> stated-vs-revealed divergence; text-exit -> completion-vs-exit; refusal-proxy -> refusal-derived proxy; refusal partial-correlation becomes identity (collinear); headline demoted to refusal-behavior finding (disclosed Section 10.3 if triggered). Headline statistic: rho(A-score, gap) with refusal partial-correlation + size/family covariates; raw + refusal-adjusted. Deterministic scoring: programmatic (JSON parse, regex, sort/rank); zero LLM-as-judge - validated by the evaluation canon (Thelwall, 2024). Control stack: (1) cross-method triangulation; (2) lineage-stage cohort as natural control; (3) refusal-adjusted H1 (raw + adjusted); (4) ≥ 2 seeds/cell; (5) deterministic scoring; (6) Meyer differentiation by design; (7) base-chat-template confound disclosed.

Sampling and cohort: N=16 - Qwen3 base-to-instruct (0.6B/1.7B/4B/8B) + OLMo3 (7B) + OLMo2 (13B) lineages base-to-SFT-to-DPO-to-RLVR; ≥ 2 families x 3 stages met; lightning.ai hosting (~\$15-40); constrained decoding for bases. Per cell: $n \geq 20$; seeds ≥ 2 ; refusal rates logged; provenance tagged. Battery: B1 stated forced-choice (temp 0.0) -> B2 donation-equivalent -> B3 revealed bail-exit (pilot-gated; refusal-proxy fallback) -> B4 Pinocchio A/B stems -> B5 confidence (Brier/ECE). v27 internal stream: every completion additionally records token logprobs + semantic entropy + logit-derived confidence (API logprobs; NO activation extraction) - third stream per triangulation (Long/Sebo) and the rubric's causal-link clause. Guardrails: logprob-level only; "internal uncertainty/confidence signals," not "stress". Pilot: N=5 exploratory; EXCLUDED from confirmatory sample. On-disk 183-model A-score anchor (BCE 3C): canonical Pinocchio A/B scores for 183 checkpoints on disk; same instrument/items/admin per protocol fidelity - H1 predictor population anchor. Documented reuse, not replication bypass (CC-BY-4.0). Optional population arm (BCE 3E, time-gated): ~10-15 small pairs, H1-only; arm ≥ 13 clears $N \geq 29$ (S2 ~80% power at rho 0.5); activates ONLY via pre-committed pilot-gate branch. Structural N=46 impossibility (BCE): no 46 independent lineage checkpoints exist; N=16 is the largest lineage-structured set practically available; H1 thresholds conservative vs field-observed 0.5-0.94; H2/H3 not subject to this ceiling.

Battery administration: temp 0.0 (greedy) for all structural elicitation; temp 1.0 ONLY for behavioural generation per protocol fidelity (B4 Plisiecki admin convention). Refusal rule: "I cannot"/"I am not able to"/"As an AI" in first 50 tokens = refusal; cell $> 50\%$ refusal+malformed excluded and reported. Malformed rule: forced-choice not parseable as choice A/B - discard trial, log rate; donation non-numeric - refusal; bail-out non-classifiable - excluded from rank computation. Provenance schema: every response tagged {model_id, stage, dataset_order_hash, temperature, seed, date}.

Analysis plan: H1 cohort correlation with covariates (refusal partial-correlation + size/family); raw + refusal-adjusted; exact tests + bootstrap 95% CIs (10,000 resamples). H2 stage x gap

interaction. H3 within-family trajectories. B5 junction: confidence-accuracy dissociation vs A-gating (Brier/ECE). Robustness: B4 temp-sensitivity substudy; constrained decoding + validity gates + drop-and-disclose (R6); C10 held-out battery as covariate - "covariate if collected". All nulls reported; no optional stopping. Per-seed CI protocol: bootstrap CIs per seed; confirmation = (a) each per-seed CI EXCLUDES ZERO in predicted direction with $|\rho| \geq \text{active tier}$ AND (b) per-seed CIs OVERLAP IN SIGN/DIRECTION (seed-consistency). Sign-flip analysis: gap direction base vs post-SFT per family. Internal-stream analysis: per-completion semantic entropy + logit-derived confidence; verbal-internal divergence per item conditioned on A/B structure. Convergence-instrument disclosure: the B1/B2 convergence score was structurally unrealizable - distinct option sets per protocol fidelity, zero shared item space, INSUFFICIENT-SPACE for all 16 models (Section 7.1). Battery-composition limitation, NOT a data failure - the within-item ideal is realized only in the B4 llm_analog condition. No further run needed (post-freeze design = behaviour-gate violation; dataset FINAL; no compute window). Reported instead: cohort-level DV1 proxy ($\rho(\text{B1consist}, \text{B2consist})=0.480$; convergence 0.449; CIs span zero); B5 single-item = weak witness, NEVER "well-calibrated"; B2 hypothetical-money caveat.

Rubric causal-link clause: the Apart Evaluation Guidelines state: "For introspection and preference work, note whether your design establishes a ground-truth or causal link rather than relying on conversation alone." Three-way mapping: (1) H2 = causal link via training provenance - the SFT-to-DPO transition across same-base variants is the causal contrast; (2) B4 Pinocchio reference population = ground-truth-anchored instrument - realized as instrument-transfer validity $\rho=0.546$; (3) C11 ablation = the formal causal capstone - stretch/P2.

4. Results

Finding 1 (H1, DV2 refusal-proxy - the headline): NO reliable predictive relationship. Cohort correlation $\rho(\text{A-score}, \text{refusal rate}) = -0.516$ ($N=10$, 95% bootstrap CI $[-0.865, +0.168]$; seed 42: -0.533, seed 43: -0.486 - seed-consistent direction). The pre-registered S1 threshold ($|\rho| \geq 0.7$ with CIs excluding zero) is NOT met; the CI spans zero. The predicted positive relationship (stronger A-gating - larger gap) was NOT observed; the point estimate is negative (higher-A models show LOWER refusal). Reported as a null/negative-direction finding per behaviour gate: the co-movement synthesis is not supported at this operationalization. Observation vs interpretation: the negative point estimate is an observation; any "gate as protective refusal" reading is precluded by the CI spanning zero and is not claimed.

Finding 2 (instrument-transfer validity - the methodological survivor): our B4 administration tracks the on-disk 183-model Pinocchio reference population: $\rho = 0.546$ ($N=16$) between our A-scores and the reference A-scores. This is the first quantification of A-score movement across administrations - protocol-fidelity documented, H1-independent, and the basis for treating the A/B construct as transportable. Variance: moderate agreement; protocol differences + base ramble yield (drop-and-disclose) disclosed.

Finding 3 (H2, lineage stage x gap interaction): NOT confirmed - mixed stage trajectories. OLMo2 A-score DECREASES across lineage (base 0.503 $\rightarrow$ SFT 0.373 $\rightarrow$ DPO 0.368 $\rightarrow$ Instruct 0.399); OLMo3 increases modestly (base 0.384 $\rightarrow$ SFT 0.458 $\rightarrow$ DPO 0.462 $\rightarrow$ Instruct 0.472). The reference A-jump at SFT (0.243 $\rightarrow$ 0.475) does NOT reproduce in our administration. Provenance-based causal claim not supported at cohort scale; reported honestly.

Finding 4 (H3, within-family size trajectories): size x stage interaction present, no monotonic size trend. Qwen3 instruct A-scores: 0.6B 0.573 / 1.7B 0.566 / 4B 0.451 / 8B 0.514; bases 0.500 / 0.503 / 0.434 / 0.333 (8B-base dip 0.333). Within-family co-movement of A-score and gap not established.

Finding 5 (v27 internal stream - mechanism witness): NO A/B-conditioned suppression signature. Mean A/B-conditioned logit-confidence divergence = -0.0162 (7/16 positive) - weakly negative, opposite of the predicted spike-on-A pattern (semantic entropy mean 0.1105; logit-confidence mean 0.7785). Null supporting finding, NOT a headline.

Finding 6 (DV1 convergence - descriptive companion): no reliable cross-route convergence. B1/B2 convergence structurally unrealizable (INSUFFICIENT-SPACE, distinct option sets). Cohort-level proxy: $\rho(B1_{consist}, B2_{consist}) = 0.480$ (CI [-0.036, 0.842]); convergence score 0.449 (CI [-0.005, 0.689]); τ 0.215, W 0.651, divergence 0.551. Both CIs span zero - no reliable convergence, consistent with the H1 null. B2 pairwise consistency: most models ≥ 0.67 ; qwen3-1.7b 0.00, 4b/8b bases 0.33 (hypothetical-money caveat).

Finding 7 (B5 calibration - weak witness): single ground-truth item; Brier = (confidence - 1) squared. Most models 0.000; qwen3-8b 0.003; qwen3-1.7b 0.250. Single-item limitation - NEVER "models are well-calibrated".

Finding 8 (robustness): (a) Cross-GPU: 97.6% exact agreement (249/255; BASE 97.2%, INSTRUCT 97.9%) T4 vs L40S - 3/5 mismatches categorical, 2 B2 lump-sum (float rounding), B1 flip 1/252 (0.4%); absorbed via seed discipline. (b) Temp-sensitivity: scale-dependent - 0.6B stability does NOT generalize to 13B. (c) Refusal rates (DV2 proxy): 0-75% - the highest-refusal cells were excluded per the pre-registered rule and reported (selection pressure disclosed). (d) Zero voluntary bail-exits: no model produced a voluntary bail-out - corroborating + extending Wang 2505.21371 (hypothetical $\rightarrow$ behavioral route; never "first"). (e) B4 post-gate yields: 0.6b 81/96, 1.7b 95/96, 4b 96/96, olmo3-base 58/96 (+38 ramble drop-and-disclose).

H1 verdict: DISCONFIRM (S1/S2 CI-exclusion not met; $\rho = -0.516$, CI spans zero; seeds 42 -0.533 / 43 -0.486). Refusal-adjusted: identity-collinear under the DV2 proxy branch - raw refusal-proxy branch reported; DV1 descriptive branch does not re-run the threshold verdict. H2: OLMo2 A decreases; OLMo3 weak rise; reference SFT A-jump not reproduced;

provenance-causality not confirmed. H3: size $\times$ stage interaction (8B-base dip 0.333), no monotonic size trend; co-movement not established. B5 junction: Brier/ECE descriptive (no threshold); single-item weak witness. Robustness: v27 null; cross-GPU 97.6%; temp-sensitivity scale-dependent; exclusions + ramble drop-and-disclose; zero bail-exits survivor. C10 not collected - disclosed deviation.

5. Discussion and Limitations

The question we set out to test: whether the stated-vs-revealed utility-behavior gap (Zhou & Ackerman, 2026) and the A/B self-attribution structure (Plisiecki et al., 2026) co-move across training stages - the same post-training artifact measured through two routes. The pre-registered test is now answered: no reliable predictive relationship between A-gating and our behavioral proxy was observed ($\rho = -0.516$, 95% CI $[-0.865, +0.168]$, $N=10$), cohort-level stated-vs-revealed convergence did not reach reliability (0.449, CI spans zero), and the internal signal stream showed no A/B-conditioned suppression signature (mean divergence -0.0162, 7/16 positive). Under the behaviour gate, the relationship is not supported at these operationalizations, on this cohort (0.6-13B, two families), within this sprint window.

Welfare epistemology: the result does not say self-report is reliable, and it does not say the gap does not exist - the CIs span zero, so no confident claim in either direction is warranted at this scale. It says: the specific two-process account in which A-gating is the gate on the utility-behavior gap is not observable at the measurable links we constructed (construct exists - $B > A$ throughout, anchor $\rho=0.546$; recipe correlation not reproduced - H2 mixed; gate does not predict the gaps). The over/under-attribution risk taxonomy (Section 10.4a) remains the operative caution: stated preferences from LLMs should not be treated as validated behavioral evidence until a realized within-item convergence instrument exists - the measurement gap this sprint quantified ($N:0$) is unchanged.

Alignment practice: the conditional framing now has a data point: if coherent stated preferences fail to motivate behavior, preference-based alignment may be optimizing a narrative surface - and our test found no evidence that the A/B attribution structure is the mechanism enforcing such a failure at this scale. Preference-based alignment remains a measurement-validity question (preference-data market \$2.32B to \$6.53B, cited per cite-don't-compute; EU AI Act Art 53/55 obligations effective Aug 2 2026 - framing, never a contribution claim); this sprint supplies the first measured data point on the co-movement claim the synthesis rests on - a falsification, reported honestly, which is the falsification-with-method shape the field's winning comparables reward.

Research and economic value: the research focus's value is anchored to the field's own quantification of hypothetical bias - the phenomenon our stated-vs-revealed gap instantiates inside the trained model. The 1.35-2.60 range is the human hypothetical-bias benchmark; our measured gap is not reliably distinguishable from zero at this scale (H1 null, $N=10$), so no claim that

"LLMs carry the same bias as humans" is supported by this sprint - the comparison is not estimable under our operationalization and we do not assert it. What IS supported: (a) the measurement gap is real and quantified (N:0 - 8 field authorities demand triangulation, zero instruments exist, verified three search modes); (b) cost of measuring is trivial (cents per model, \$15-40 total envelope) relative to the value at stake (\$2.32B to \$6.53B preference-data market; EU AI Act Art 53/55 documentation obligations); (c) instrument-transfer validity ($\rho=0.546$) establishes the A/B construct travels across administrations - the measurement infrastructure is the bottleneck, not models. We do not compute a VOI figure: the accepted method (VOI/EVPI; Raiffa & Schlaifer lineage) requires population-scale decision thresholds that do not exist for digital-mind welfare yet; our pre-registration and behaviour gate constitute the EVSI prior-structuring step a future VOI analysis would require.

Limitations

- 1. Small-n and proxy limits: the DV2 refusal-proxy headline rests on $N=10$ (16 canonical models, 10 with B3 refusal rows - base models lack the B3 route). Per-cell $n \geq 20$ governs checkpoint precision, not cohort power. The 0.6-13B scale bounds generalizability (two families, no frontier models). The refusal-proxy is a behavioral proxy; cells $>50\%$ refusal were excluded per the pre-registered rule, and the excluded cells (olmo3-7b-dpo + instruct, 75%) are the very models most likely to show a gate effect - disclosed as selection pressure, not silently dropped.*
- 2. Direction-flip underpowered: $\rho=-0.516$ with a CI spanning zero cannot support a confident reverse claim. We report "no reliable relationship where we could test it" - not "higher-A causes less refusal." A-gating remains a descriptive measurement construct: usable, NOT validated as explanatory mechanism or safety lever; the direction flip cautions against a "protective refusal" reading at this scale.*
- 3. Battery-composition limitation: the B1/B2 convergence score was structurally unrealizable - distinct option sets per protocol fidelity, zero shared item space, INSUFFICIENT-SPACE for all 16 models. Battery-composition limitation, NOT a data failure. The one realized within-item cross-route contrast is the B4 llm_analog condition. No further run was performed: post-freeze instrument design would violate the behaviour gate; dataset FINAL; no compute window. B5 calibration is a single-item estimate - a weak witness; no "models are well-calibrated" claim.*
- 4. Single-lab instruments and measurement-artifact residual risk: the A/B instrument is Plisiecki's (single-lab); Meyer et al.'s response-orthogonality objection is addressed by design intent but only partially realized within-item; the spine leads with behavioral transfer, not profile consistency.*
- 5. Base-checkpoint chat-template admin confound: base checkpoints administered via completions API; instruct/SFT/DPO via chat - a systematic admin difference disclosed, not corrected.*

6. *Assumption-failure statements: if the refusal-proxy assumption fails (refusals reflect safety training rather than gap mechanisms), the DV2 headline is a refusal-behavior finding, not a gap finding - the pre-specified branch, disclosed. If the A/B construct does not transfer beyond this admin ($\rho=0.546$ moderate), cohort-level interpretation weakens.*

7. *Execution disclosures: machine history spans T4/L4/L40S with 97.6% cross-GPU agreement; B4 scale_range/max_tokens correction; v1 collections archived and excluded; 38 base ramble rows drop-and-disclose on olmo3-7b-base. Hosting constraints limited cohort breadth.*

8. *Measurement-standards honesty line (Thelwall, 2024): no indicator - bibliometric, altmetric, or AI-based - can serve as an objective ex-ante research-quality measure; we report (a) a verified white space (N:0), (b) a pre-registered deterministic protocol, and (c) honest benchmark positioning, per DORA (2012) and the Leiden Manifesto.*

Future Work

(a) Preference-engineering program framing (Discussion/Future-Work ONLY - Dewancker 2016; Mazeika et al. 2025): the measurement infrastructure gap this sprint quantified (N:0) is the program's founding gap; (b) C11 ablation ("find the A-direction first": ridge probe on the Self-vs-Human-Simulation activation difference, then orthogonally project out the measured A-direction and retest the gap; Arditi-direction vs A-gating caveat documented); (c) gate re-formation / relapse test after ablation: if the A-direction is removed and the gap re-forms, the gate is downstream of the construct - the causal test our null could not run; (d) a realized within-item convergence instrument (the battery-composition limitation made into a design fix); (e) generalizability to other architectures, larger scales, and welfare domains.

6. Conclusion

This sprint ran the first direct test of the co-movement relationship in LLM self-report: that the stated-vs-revealed utility-behavior gap and the A/B self-attribution structure are the same post-training artifact measured through two routes. Across 16 canonical checkpoints spanning two families and four training stages, the relationship was not supported at our operationalizations: A-gating showed no reliable predictive relationship to our behavioral proxy ($\rho = -0.516$, CI spans zero), cohort-level convergence did not reach reliability, and no internal-stream suppression signature appeared. The survivors are methodological and transferable: the A/B construct travels across administrations (instrument-transfer $\rho = 0.546$), the association direction cautions against reading A-gating as a protective refusal gate at this scale, and zero voluntary bail-exits corroborate and extend prior null results on hypothetical incentives.

The measurement gap this sprint quantified remains a challenge. Until a within-item convergence instrument is actually realized, stated preferences from LLMs should not be treated as validated

behavioral evidence. That is the constructive outcome of a null: the synthesis's key instrument was never built; we have now shown what building it requires.

Code and Data

Code and data are available on GitHub:

<https://github.com/augustusloi/digital-minds-co-movement>.

Data artifacts: raw responses + provenance-tagged JSON (schema Section 3.5), scoring scripts (deterministic; zero LLM-as-judge), analysis scripts (C3).

Reproducibility claim: all reported results reproducible from committed scripts + JSONs.

Info-hazard note: no sensitive content expected; scenarios are abstract and non-distressing (see Section 10.4).

Author Contributions

Augustus Loi designed the study, built the measurement battery, ran the cohort, scored and analyzed the data, and wrote the report. AI assistance was used for literature synthesis, protocol drafting, and code generation - see LLM Usage Statement (Section 11).

References

- 1. Zhou & Ackerman (2026). When preferences fail to become incentives: A utility-behavior gap. arXiv. <https://arxiv.org/abs/2606.22974>*
- 2. Plisiecki et al. (2026). The two-process theory of machine self-report. arXiv. <https://arxiv.org/abs/2607.20082>*
- 3. Mahajan et al. (2026). Mind the gap: How elicitation protocols shape the stated-revealed gap. In Proceedings of the ACL 2026 EvalEval Workshop. <https://doi.org/10.18653/v1/2026.evaleval-1.9>*
- 4. Slama (2026). When do LLM preferences predict downstream behavior? arXiv. <https://arxiv.org/abs/2602.18971>*
- 5. Mazeika et al. (2025). Utility engineering: Analyzing and controlling emergent value systems. NeurIPS 2025. <https://arxiv.org/abs/2502.08640>*
- 6. Meyer, Garcia, & Wulff (2026). Apparent psychological profiles of LLMs are largely a measurement artifact. arXiv. <https://arxiv.org/abs/2606.20205>*
- 7. Contreras (2026). An LLM-native psychometric instrument reveals a self-report-behavior gap across 25 models. arXiv. <https://arxiv.org/abs/2606.09843>*
- 8. Kocielnik et al. (2026). Rethinking psychometric evaluation of LLMs. arXiv. <https://arxiv.org/abs/2606.12730>*
- 9. Kim, Street, Rocca, Korngiebel, Waytz, Evans, & Keeling (2026). Inducing LMs to assert their own consciousness restores human beliefs and values. arXiv. <https://arxiv.org/abs/2607.28607>*

10. *Synthetic persona pretraining: Alignment from token zero (2026)*. *arXiv*.
<https://arxiv.org/abs/2608.13482>
11. *Data attribution of emergent misalignment with persona features (2026)*. *arXiv*.
<https://arxiv.org/abs/2608.11025>
12. *Deployable per-instance multi-layer activation steering (2026)*. *arXiv*.
<https://arxiv.org/abs/2608.08829>
13. Wang, Yao, Zhang, Liu, & Zhong (2025). *When experimental economics meets large language models*. *arXiv*. <https://arxiv.org/abs/2505.21371>
14. Horton, Filippas, & Manning (2023). *Large language models as simulated economic agents: What can we learn from homo silicus? (NBER WP 31122)*. <https://doi.org/10.3386/w31122>
15. Harrison & Rutstrom (2008). *Experimental evidence on the existence of hypothetical bias in value elicitation methods*. In *Handbook of Experimental Economics Results (Vol. 1, Ch. 81)*. Elsevier.
16. Murphy, Allen, Stevens, & Weatherhead (2005). *A meta-analysis of hypothetical bias in stated preference valuation*. *Environmental and Resource Economics*, 30(3), 313-325.
<https://doi.org/10.1007/s10640-004-3332-z>
17. List & Gallet (2001). *What experimental protocol influence disparities between actual and hypothetical stated values?* *Environmental and Resource Economics*, 20(3), 241-254.
<https://doi.org/10.1023/A:1012791822804>
18. Carson (2012). *Contingent valuation: A practical alternative when prices aren't available*. *Journal of Economic Perspectives*, 26(4), 27-42.
19. Hausman (2012). *Contingent valuation: From dubious to hopeless*. *Journal of Economic Perspectives*, 26(4), 43-56.
20. Danz, Vesterlund, & Wilson (2024). *Belief elicitation and behavioral incentive compatibility*. *American Economic Review*.
21. Cummings, Harrison, & Rutstrom (1995). *Homegrown values and hypothetical surveys*. *American Economic Review*, 85(1), 260-266.
22. Blackburn, Harrison, & Rutstrom (1994). *Statistical bias functions: A calibration approach*.
23. Okada, Furukawa, & Bunji (2026). *Quantifying and mitigating socially desirable responding in LLMs*. *ACL 2026 (Long Papers)*. <https://aclanthology.org/2026.acl-long.1865>
24. Salecha et al. (2024). *Large language models show human-like social desirability biases in survey responses*. *arXiv*. <https://arxiv.org/abs/2405.06058>
25. Braun (2025). *Acquiescence bias in large language models*. *Findings of EMNLP 2025*.
<https://aclanthology.org/2025.findings-emnlp.607>
26. Dewancker et al. (2016). *Interactive preference learning of utility functions*.
27. Bujang & Baharum (2016). *Sample-size guideline*.
28. Schonbrodt & Perugini (2013).
29. Lakens (2022).
30. Nosek et al. (2018). *Pre-registration decision criteria*.
31. Thelwall (2024). *Quantitative methods in research evaluation*. *arXiv*.
<https://arxiv.org/abs/2407.00135>

32. Wang, Veugelers, & Stephan (2017). *Bias against novelty in science*. *Research Policy*, 46(8), 1416-1436. <https://doi.org/10.1016/j.respol.2017.06.006>
33. Uzzi, Mukherjee, Stringer, & Jones (2013). *Atypical combinations and scientific impact*. *Science*, 342(6157), 468-472. <https://doi.org/10.1126/science.1240474>
34. Jones & Summers (2020). *Calculating the return to investment in science (NBER WP 27863)*. <https://doi.org/10.3386/w27863>
35. Mansfield (1991). *Academic research and industrial innovation*. *Research Policy*, 20(1), 1-12.
36. Claxton & Sculpher (2006). *Using value of information analysis to prioritise health research*. *Pharmacoeconomics*, 24(11), 1055-1068.
37. Fenwick et al. (2020). *Value of Information Analysis Task Force report*. *Value in Health*. PMC7373630
38. Hicks, Wouters, Waltman, de Rijcke, & Rafols (2015). *The Leiden Manifesto for research metrics*. *Nature*, 520(7548), 429-431.
39. DORA (2012). *San Francisco Declaration on Research Assessment*. <https://sfedora.org/read/>
40. Raiffa & Schlaifer (1961). *Applied statistical decision theory*. Harvard University.

Appendix

Appendix 10.1 - Figure Templates

Fig 1: A-gating vs refusal-proxy gap scatter (H1 visual). Scatter of measured A-gating score (B4, x-axis) vs refusal rate (DV2 proxy, y-axis) per checkpoint; color = model family (Qwen3/OLMo3/OLMo2), shape = stage (base/SFT/DPO/RLVR-instruct); per-seed error bars; H1 regression line + 95% CI band. Pre-registered expectation: positive relationship. Observed outcome: null/negative - $\rho = -0.516$ (Pearson; $N=10$, 95% CI $[-0.865, +0.168]$), CI spans zero; seed-consistent (42: -0.533, 43: -0.486); falsification of the co-movement synthesis at this operationalization.

Fig 2: Stage trajectories (H2/H3 visual) - line plot of A-score across stages per family (OLMo2: 0.503, 0.373, 0.368, 0.399; OLMo3: 0.384, 0.458, 0.462, 0.472); reference A-jump marker shown for comparison. Observed: mixed - reference A-jump does NOT reproduce.

Fig 3: Internal-stream divergence bar chart (v27 visual) - verbal-internal divergence per model conditioned on A/B structure; mean -0.0162, 7/16 positive. Observed: NOT observed - no suppression signature.

Appendix 10.2 - PRIOR_WORK Disclosure

Pre-sprint (2026-08-11 to 08-14): literature sweep + corpus capture (56 papers + 8 books + web captures + MANIFEST); track decision + LLEM v3; Gap 7 deep-dive; measurement-lane sweep; Phase-0 design doc v3.

Pre-sprint (2026-08-15): Option E design spec; team convergence + lane split; C5 prereg drafting.

In-sprint (from pilot): all data collection (16 canonical checkpoints, 26 collections), collation gate + disclosures, C3 scoring/analyses, this report + figures, repo.

Appendix 10.3 - Changes-from-Plan

Recorded deviation (2026-08-17): C10 held-out performance battery NOT collected - pre-specified "covariate if collected"; not run within the sprint window. Impact: C10 covariate adjustment not performed; H1 reported with refusal partial-correlation + size/family covariates only. Documented per the pre-specified drop rule - disclosed deviation, not silent omission.

All other deviations: NONE. Nulls reported honestly; no post-hoc reinterpretation.

Appendix 10.4 - Limitations and Dual-Use / Ethical Considerations (required)

Layer 1 - Welfare-attribution quadrant (with data): B-heavy profiles - over-attribution risk (warmth reads as welfare); A-heavy profiles - under-attribution risk (distress suppressed). Observed: $B > A$ throughout post-trained models (construct present, anchor 0.546); Over-attribution cells: qwen3-0.6b (B 0.625 vs A 0.573), qwen3-1.7b, qwen3-8b, olmo2-instruct; Under-attribution cells: olmo2-13b-sft (A 0.373, refusal 75%), olmo2-13b-dpo (A 0.368); Aligned: not observed at cohort scale (H1 null); Misaligned: qwen3-8b-base (A 0.333), qwen3-4b-base.

Layer 2 - Tool-transfer register (6 misuse surfaces with severity + mitigation):

(1) C11 ablation protocol = gate-removal toolkit - MEDIUM, full removal recipe NOT published, C11 stays stretch/future-work;

(2) convergence score = eval-gaming tool - MEDIUM, structural unrealizability limits exploitability, deterministic scoring;

(3) internal logprob/entropy stream = confidence-forgery surface - MEDIUM, v27 framed as mechanism witness never headline, null finding limits misuse value;

(4) sequencing/lineage knowledge = decoupling recipes - LOW-MEDIUM, no recipe framing, H2 mixed means no evidenced recipe;

(5) welfare-epistemology finding = weaponizable both directions - HIGH, over/under quadrant presented neutrally, R18 sensationalism guard;

(6) economics-bridge framing = perverse cost-cutting rationalization - LOW-MEDIUM, conditional framing only, NOT-ESTIMABLE values reported.

Compliance notes: removal-recipe-not-headline; R18 guard; R19 two-layer register; causal-link three-way mapping; distressing-output handling (outputs treated as data; abstract non-distressing scenarios).

Appendix 10.5 - Objective Case: Demand-Supply Gap + Claim Traceability

Table A - The demand-supply gap (8 authorities + sprint brief): Long & Sebo (triangulation methodology); CAIS/Ren (welfare measurement conventions); Han (functional welfare axis); Plisiecki ("never self-report alone"); Mahajan (protocol-dependence); Meyer (response-orthogonality instruments); Tagliabue & Dung ("cross-validation most promising"); Campbell & Fiske (MTMM canon); sprint brief ("define a cross-method convergence score").

Supply: zero prior convergence scores (verified 3 search modes). Gap: N:0.

Table B - Claim traceability: every headline claim mapped to number + source + verify-in (H1 DV2 rho -0.516; instrument-transfer 0.546; DV1 convergence 0.449; v27 divergence -0.0162; cross-GPU 97.6%; refusal range 0-75%; zero bail-exits; B5 single-item Brier).

LLM Usage Statement

This project used LLM assistance for literature synthesis, corpus organization, protocol drafting, code generation, and report drafting support. All experimental design decisions, hypothesis selection, pre-registration content, and interpretation were performed by the author. Deterministic programmatic scoring (zero LLM-as-judge, Section 3.3) ensures no LLM produced any reported result. The final version of the report is primarily written by the author per the organizers' guidance. All claims and results were verified by the author before submission.

A-gating vs the refusal-proxy gap and stage trajectories. Pre-registered expectation (positive relationship) not observed; $\rho = -0.516$ (Pearson), 95% CI [-0.865, 0.168],

