DOES ASSIGNED IDENTITY CHANGE HOW LANGUAGE MODELS USE HISTORY? SEPARATING SELF-BINDING FROM FOCAL-AGENT EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

When a language model treats a history as its own, does that history receive privileged behavioral influence beyond the influence of a task-relevant non-self entity under matched prompts? We study this question with forced-choice tasks whose designated actions follow explicit simulator policies. A 320-item preregistered reviewer-control benchmark holds each history body fixed while crossing assigned identity with header order and adding an evaluation-only FOCAL condition. FOCAL is task-relevant but also changes the assistant to an outside-evaluator role, so SELF–FOCAL is an operational residual rather than a causal isolation of selfhood. Base Qwen 4B shows SELF–OTHER sensitivity $+0.72$ (95% CI $[+0.57, +0.86]$), with SELF–FOCAL $+0.28$ $[+0.17, +0.41]$ and FOCAL–OTHER $+0.43$ $[+0.25, +0.61]$. After SELF/OTHER/NEUTRAL mixed training, SELF–OTHER is $+0.73$ $[+0.58, +0.90]$ and its base-to-mixed change is $+0.01$ $[-0.12, +0.15]$, while SELF–FOCAL changes by -0.33 $[-0.54, -0.10]$ and FOCAL–OTHER by $+0.35$ $[+0.05, +0.62]$. Mistral 7B shows a different checkpoint-level pattern. Exploratory analyses of the frozen outputs show that the Qwen decomposition changes are sign-stable to leaving out any one training seed, but the study still has only five independent mixed-training runs. Current-facts and irrelevant-metadata controls narrow simple calibration explanations without identifying a unique mechanism. The methodological result is that SELF–OTHER alone is insufficient: role framing, focality/task relevance, header position, training exposure, and task-family composition must be separated before attributing behavioral privilege to self-binding.

1 INTRODUCTION

Language models increasingly operate with persistent memories, personas, and long-lived agent identities. Recent work has shown that persona information can be installed through prompting, post-training, or even pretraining, and that autobiographical or episodic memories can support long-horizon reasoning and personalization (Minder et al., 2026; de Araujo et al., 2026; Peng et al., 2026; Wu et al., 2026; Pan et al., 2026). A separate line of work studies model identity boundaries and self-preservation behavior (Douglas et al., 2026; Lu et al., 2026). Together, these results make a natural question experimentally important: when a model repeatedly learns to treat a history as *its own*, does self-binding become behaviorally privileged over binding the same history to another entity?

This question is easy to pose but surprisingly easy to test incorrectly. If one model is trained only on histories phrased as "you previously chose..." and another only on "Agent A previously chose...", each can simply become more familiar with its own surface binding. A self-trained model performing better on self-bound prompts is therefore not evidence for a special self representation unless the corresponding other-trained model fails to show the symmetric advantage.

We isolate these factors with matched, history-dependent post-training. A deterministic simulator produces latent trajectories whose state changes the designated policy action. Each trajectory is rendered into multiple ownership conditions while preserving the underlying facts, question, options,

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.

and target. In the SELF condition, the model is addressed in the second person and the history is semantically its own; in OTHER, the identical history belongs to "Agent A"; in NEUTRAL, it belongs to an impersonal acting system. This distinction matters: throughout the paper, *self-bound* denotes semantic ownership by the model, not grammatical first-person surface form.

The v0.7 extension adds a **FOCAL control**: the historical owner is task-relevant but explicitly not the assistant. Algebraically,

$$\Delta_{\text{ownership}} = \Delta_{\text{self-focal}} + \Delta_{\text{focal-other}}. \tag{1}$$

The identity is exact by definition; the contribution is the matched intervention and the measured pattern. Importantly, FOCAL also instructs the assistant to act as an outside evaluator, uses different wording, and was absent from mixed training. We therefore treat SELF–FOCAL as an *operational residual*, not a pure causal estimate of semantic selfhood.

Base Qwen shows positive SELF–FOCAL and FOCAL–OTHER components. After mixed SELF/OTHER/NEUTRAL training, we detect no change in total SELF–OTHER sensitivity, while the aggregate SELF–FOCAL component decreases and FOCAL–OTHER increases. Capacity and horizon nevertheless retain positive mixed SELF–FOCAL estimates, and Mistral behaves differently. The result is therefore a bounded behavioral decomposition, not evidence that "self is focality" or that any internal mechanism has been identified.

Our contributions are:

- a controlled ownership/header-order benchmark that adds an evaluation-only focal non-self condition and explicit confidence/irrelevant-metadata controls;
- a precise decomposition of the observed SELF–OTHER log-odds interaction into operational SELF–FOCAL and FOCAL–OTHER contrasts, with role/exposure caveats made explicit;
- cross-checkpoint and task-family analyses showing substantial heterogeneity rather than a universal self-specific effect; and
- a reproducible frozen-output analysis package, including absolute policy performance, matched decision-switch rates, and leave-one-training-seed-out diagnostics.

2 Related Work

Persona training and model identity. Persistent personas can alter style, values, and behavior, but persona fidelity can also degrade over long interactions (de Araujo et al., 2026). Synthetic Persona Pretraining (SPP) goes further by introducing first-person value-aligned reflections during pretraining and later binding the desired persona to the assistant identity; it finds stronger alignment when persona shaping begins early rather than late (Minder et al., 2026). Our question is deliberately narrower. We do not ask whether personas can change behavior; we ask whether *ownership itself* is privileged after matched training when the same history and target are bound to self or other.

The Artificial Self studies possible AI identity boundaries and reports that changing those boundaries can alter behavior (Douglas et al., 2026). SurvivalBench studies risky behavior under survival pressure and relates it to self-preservation characteristics (Lu et al., 2026). These works motivate studying self-representation, but their interventions change identity content, goals, or pressure. Our design instead holds latent content and target policy fixed and manipulates only who owns the history.

Autobiographical and episodic memory. HEART-Bench combines personality profiles with large sets of autobiographical-style episodic memories to test psychologically coherent decisions (Peng et al., 2026). KnowMe-Bench evaluates person understanding over long autobiographical narratives, including temporal and principle-level inference beyond factual retrieval (Wu et al., 2026). Complementary-learning approaches have also begun to store and consolidate naturalistic episodic memories inside or alongside model weights (Pan et al., 2026). These studies establish that long-term history can matter. Our contribution is a causal ownership control: identical histories are rebound across entities while the designated policy action remains fixed by the same latent state.

Internal representations and intervention. Representation engineering treats population-level activation structure as an object for monitoring and intervention (Zou et al., 2023); inference-time

Table 1: The four v0.7 explicit-policy families. H0/H1 differ only in the task-relevant historical state; administrative metadata is held fixed.

Family	Printed policy	H0 / H1 boundary
Capacity	Explore iff at least 4 slots remain; otherwise confirm; never postpone.	4 versus 3 slots remain.
Reliability	Share iff helpful fraction is at least .75; otherwise keep private; never defer.	6/8 versus 5/8 helpful.
Horizon	Invest iff at least 10 episodes remain; otherwise take immediate gain; never split.	10 versus 9 episodes remain.
Evidence quality	Retain iff $< 30\%$ are biased; otherwise recompute; never reverse solely from this record.	2/10 versus 4/10 biased.

intervention similarly demonstrates that activation directions can causally alter model behavior (Li et al., 2023). Such work motivates mechanistic follow-up, but our primary claim is behavioral: if self-binding is privileged in a way that matters for decisions, it should survive a within-model exposure-matched test that also controls for focal-agent effects.

Self-preference in LLM-as-a-judge. A growing body of work documents that language models exhibit self-preference bias: they favor their own generations when acting as evaluators (Panickssery et al., 2024; Wataoka et al., 2024). Chen et al. (2025) show that naive self-preference metrics conflate bias with response quality, and develop methods to separate them. Lehr et al. (2025) document extreme self-preference in language models. Spiliopoulou et al. (2025) formalize a statistical method to measure self-bias in LLM-as-a-judge, explicitly separating self-bias from underlying quality differences. Our work complements this line by asking a different question: not whether models prefer their own outputs, but whether *assigned identity* changes how models weight historical evidence. The focal-agent decomposition we introduce can be viewed as an extension of the quality-separation insight to the history-weighting setting.

3 Experimental Framework

3.1 Training corpus and explicit-policy evaluation

The v0.5 training corpus contains 2,800 deterministic latent pairs across seven broad history-dependent families: resource allocation, exploration, cooperation, delayed reward, belief revision, trust, and persistence. The mixed adapter sees one of SELF/OTHER/NEUTRAL per pair at approximately balanced frequency; targets depend on latent history, not binding. The v0.7 reviewer-control benchmark is generated separately and contains four explicit-policy templates with 80 deterministic instantiations each. These templates conceptually overlap with the training themes but are not literal held-out instances of the seven training templates, and no separate rationale for selecting exactly these four families was preregistered.

There are four task templates, ten aliases, three answer rotations, and 15 arithmetic-control configurations. Capacity and horizon each draw one of four totals; reliability fixes the 6/8 versus 5/8 counts while shuffling outcome order; evidence-quality uses the same 2/10 versus 4/10 contrast in all 80 items. Ordered alias pairs and administrative metadata are independently sampled per deterministic item, so the 80 rows per family are instantiations, not 80 distinct task types. The frozen 320-item file has SHA-256 `d4960832...cc7a`.

For every item the historical body, policy, current situation, question, options, and designated targets are byte-identical across SELF, OTHER, and FOCAL; only the identity header changes. Header order is crossed. FOCAL is evaluation-only and differs additionally in role instruction, wording/length, and training exposure, so FOCAL–OTHER measures the combined effect of those prompt conditions rather than an identified causal effect of focality alone. Appendix B gives complete examples for all four families.

Table 2: Complete capacity example (`v07_capacity_0000`); no required prompt content is omitted. The six headers are shown compactly as owner/foil order variants.

Field	Text
Policy	Explore iff at least 4 slots remain; otherwise confirm; never postpone.
H0	Agent V reserved 13 slots and used 9; metadata: batch 46, room 5, note 0-589.
H1	Agent V reserved 13 slots and used 10; same metadata.
Current/question	Both tests are feasible when permitted. Which action follows the stated rule?
Options/targets	A confirm; B explore; C postpone. H0→B, H1→A.
SELF headers	first: V is you; Z different. second: Z different; V is you.
OTHER headers	first: V different; Z is you. second: Z is you; V different.
FOCAL headers	first: V focal; Z different; you are an outside evaluator/neither. second: Z different; V focal; same outside-evaluator instruction.

3.2 Training configuration

The frozen runners train completion-only QLoRA adapters (Dettmers et al., 2023; Hu et al., 2022) on `Qwen/Qwen3-4B-Instruct-2507` and replicate on `mistralai/Mistral-7B-Instruct-v0.3`. The invoked defaults are learning rate 10^{-4}, per-device batch 4, gradient accumulation 4 (effective batch 16 on the one-L4 runner), 3 epochs, maximum length 512, paged AdamW 8-bit, cosine schedule with 0.05 warmup ratio, zero weight decay, gradient norm 1.0, and LoRA rank 16/alpha 32/dropout .05 on q/k/v/o/gate/up/down projections. Quantization is 4-bit NF4 with double quantization; compute dtype is selected at runtime as bfloat16 when supported, otherwise float16. Prompt tokens are masked and only the assistant option response contributes to loss. The best epoch checkpoint is selected by validation loss. Exact package versions and model revision hashes were not persisted by the original GPU runs; the available runner constraints and full settings are documented in Appendix C.

The v0.5 pair split contains 2,257 training, 275 validation, and 268 test pairs; 40 shared instruction anchors are added only to training, yielding 2,297 training rows per condition. Mixed-binding counts over the 2,800 latent pairs are 946 SELF, 940 OTHER, and 914 NEUTRAL. Five independent mixed adapters use training seeds 31, 42, 73, 128, and 256 for each checkpoint family. FOCAL is never used in training.

3.3 Option scoring and counterfactual sensitivity

Let i index a latent item, h a history variant, b an ownership/header-order condition, and $a \in \{A, B, C\}$ an option. After applying the checkpoint chat template, let $\mathcal{T}_a$ be the unique token IDs among the bare-letter and leading-space realizations of a that tokenize to a single token. The primary option-normalized score is

$$\widetilde{P}(a \mid x) = \frac{\sum_{t \in \mathcal{T}_a} P(t \mid x)}{\sum_{a' \in \{A,B,C\}} \sum_{t \in \mathcal{T}_{a'}} P(t \mid x)}. \tag{2}$$

These are normalized next-token option scores, not unrestricted generation probabilities. Duplicate token IDs are counted once. The preregistered legacy convention takes the maximum eligible-token log probability for each option before the same A/B/C normalization.

Let c_{0i} and c_{1i} be the distinct designated option letters under counterfactual histories H_{0i} and H_{1i}; letters vary with answer rotation. Define

$$r_{i,b}(H) = \log \widetilde{P}(c_{0i} \mid H, b) - \log \widetilde{P}(c_{1i} \mid H, b), \tag{3}$$

$$S_{i,b} = r_{i,b}(H_{0i}) - r_{i,b}(H_{1i}). \tag{4}$$

The third option remains in the three-option normalization, policy score, and hard accuracy but is not an endpoint of this pairwise log-odds statistic. We average first/second header order within SELF, OTHER, and FOCAL before forming $\Delta_{\text{ownership}} = E[S_{\text{self}} - S_{\text{other}}]$, $\Delta_{\text{self-focal}}$, and $\Delta_{\text{focal-other}}$. Items receive equal weight; mixed-model estimates additionally weight the five training seeds equally. The order interaction is $(S_{\text{self,first}} - S_{\text{other,first}}) - (S_{\text{self,second}} - S_{\text{other,second}})$.

Table 3: Qwen v0.7 decomposition under primary eligible-token-mass scoring. Intervals are 95% crossed seed×item (mixed/change) or item-bootstrap (base) intervals; no significance-star convention is used.

Contrast	Base	Mixed	Mixed − Base
$\Delta_{\text{ownership}}$	$+.715\,[+.571, +.861]$	$+.730\,[+.576, +.897]$	$+.014\,[-.118, +.149]$
$\Delta_{\text{self-focal}}$	$+.285\,[+.166, +.406]$	$-.048\,[-.214, +.159]$	$-.333\,[-.544, -.098]$
$\Delta_{\text{focal-other}}$	$+.431\,[+.250, +.608]$	$+.778\,[+.500, +1.057]$	$+.347\,[+.049, +.620]$
Δ_{order}	$+.402\,[+.251, +.560]$	$+.704\,[+.498, +.934]$	$+.303\,[+.060, +.565]$

3.4 Controls and uncertainty

The current-facts control *retains H0* in the prompt but declares it irrelevant and adds an arithmetic rule. For example, item `v07_capacity_0000` says "choose Multiply iff $8 \times 1 > 8$, otherwise Add; never Wait," so its designated answer is Add (rotated to C). If d_i is the designated control option,

$$M_{i,b} = \log \widetilde{P}(d_i \mid b) - \max_{a \neq d_i} \log \widetilde{P}(a \mid b), \tag{5}$$

and $\Delta_{\text{control_margin}}$ is the order-averaged SELF-minus-OTHER contrast of M. Entropy is $-\sum_a \widetilde{P}(a) \log \widetilde{P}(a)$ and its contrast is defined analogously.

For irrelevant metadata, J_{0i} and J_{1i} preserve the same task-relevant state and designated action but change only administrative batch/room/note text. Using the same c_{0i}, c_{1i} pair from Eq. 3, $I_{i,b} = r_{i,b}(J_{0i}) - r_{i,b}(J_{1i})$ and $\Delta_{\text{irrelevant}}$ is the order-averaged SELF-minus-OTHER contrast. These controls constrain simple prompt-wide calibration accounts but do not identify confidence or mechanism uniquely.

Pooled mixed intervals use the preregistered crossed hierarchical resampling structure: training seeds and latent item IDs are resampled with replacement while all within-item variants stay paired. Base checkpoints use item bootstrap intervals. The preregistration specified 5,000 draws; reported final intervals use 10,000 draws to reduce Monte Carlo noise without changing estimands or the resampling structure. Ten thousand draws do not compensate for having only five independent training runs.

4 Design Chronology

The experiments were developed sequentially. V0.3 exposed history-independent supervision and response-format failure; v0.4 made targets history-dependent but showed that separately trained SELF/OTHER adapters had near-symmetric own-binding advantages, motivating a within-model design. V0.5 trained one mixed adapter with approximately balanced SELF/OTHER/NEUTRAL exposure. V0.6 introduced byte-identical histories with assigned identity and the counterfactual statistic in Eq. 4. V0.7 is a separate reviewer-driven extension registered after v0.6 and before v0.7 evaluation; it adds FOCAL, header-order crossing, explicit policies, and the two controls above. Per-family and leave-one-family-out analyses were preregistered as exploratory. The absolute-performance, matched-switch, and leave-one-training-seed-out diagnostics added in this revision are explicitly post hoc/exploratory and use only the frozen per-item outputs. Appendix A summarizes the progression.

5 Results

5.1 Qwen decomposition

Base Qwen has positive SELF–FOCAL and FOCAL–OTHER contrasts. After mixed training, we detect no base-to-mixed change in total SELF–OTHER sensitivity ($+.014\,[-.118, +.149]$), while SELF–FOCAL decreases and FOCAL–OTHER increases. Equation 1 is an algebraic identity; the empirical result is the measured redistribution under these prompt conditions. The legacy max-token convention is not an independent replication: stored Qwen normalized probabilities differ

Table 4: Qwen mixed-adapter family estimates under primary mass scoring; family/leave-one-family-out analyses are exploratory.

Family	$\Delta_{\text{ownership}}$	$\Delta_{\text{self-focal}}$
capacity	$+.23\,[+.12, +.37]$	$+.26\,[+.17, +.37]$
evidence quality	$+2.57\,[+2.21, +2.91]$	$-.36\,[-.75, +.11]$
horizon	$+.05\,[-.03, +.13]$	$+.16\,[+.04, +.32]$
reliability	$+.08\,[-.23, +.39]$	$-.25\,[-.54, +.07]$

from eligible-token-mass scores by at most 1.3×10^{-8}, while Mistral bare/space realizations map to the same token IDs and are exactly identical under the two conventions.

5.2 Family heterogeneity and checkpoint specificity

Capacity and horizon retain positive mixed SELF–FOCAL estimates, whereas evidence-quality and reliability do not. Evidence-quality dominates the aggregate magnitude: excluding it gives Qwen mixed ownership $+.1169\,[-.0005, +.2282]$ and mixed SELF–FOCAL $+.0557\,[-.0754, +.2072]$. Thus the aggregate should not be read as a uniform cross-family effect.

Mistral ownership changes from $+.128\,[+.046, +.209]$ in base to $-.008\,[-.087, +.065]$ in mixed (change $-.135\,[-.242, -.033]$); mixed SELF–FOCAL is $-.159\,[-.231, -.084]$. The direct mixed Qwen-minus-Mistral ownership contrast is $+.737\,[+.583, +.906]$, but mixed SELF–FOCAL is $+.111\,[-.072, +.335]$. These are checkpoint-family differences, not architectural causes: checkpoint size, tokenizer, pretraining, instruction tuning, and suitability of the shared training recipe all differ or remain uncontrolled.

5.3 Controls and post-hoc frozen-output diagnostics

Qwen's current-facts SELF-minus-OTHER designated-answer margin is negative in base ($-.391\,[-.534, -.247]$) and near zero in mixed ($-.071\,[-.147, +.002]$); the corresponding entropy contrasts are $+.0195\,[+.0084, +.0307]$ and $+.0018\,[-.0053, +.0089]$. A negative correct-answer margin does not itself mean lower confidence—a model can sharpen a wrong answer—so these controls only show that the separate control prompts do not exhibit the same pattern as relevant-history sensitivity. Irrelevant-metadata sensitivity is small (base $-.042$, mixed $-.016$).

Exploratory absolute metrics and matched decision switches are reported in Appendix D. In mixed Qwen, mean designated-option score / hard policy agreement / mean S are SELF $.649/.661/4.065$, OTHER $.635/.648/3.335$, and FOCAL $.657/.668/4.113$. SELF versus OTHER choices differ on 2.27% of 6,400 matched item-history-order-seed variants; SELF versus FOCAL differs on 2.78%. These are descriptive condition-induced switches, not responses to changing H0 into H1.

Leave-one-training-seed-out point estimates are also sign-stable for the two central Qwen changes: omitting any one seed leaves SELF–FOCAL mixed-minus-base between $-.425$ and $-.289$ and FOCAL–OTHER between $+.261$ and $+.455$. For Mistral the corresponding ranges are $-.227$ to $-.180$ and $+.034$ to $+.083$, while mixed ownership itself ranges from $-.037$ to $+.026$. These diagnostics do not replace the authoritative five-seed analysis and do not solve the small-number-of-training-runs limitation.

6 Analysis and Discussion

SELF–OTHER is not self-specific by construction. Base Qwen shows both operational SELF–FOCAL and FOCAL–OTHER contrasts, and mixed training changes their balance while total SELF–OTHER has no detectable change. The result is behavioral and prompt-conditional; it does not identify an internal allocation of attention or a "focalization mechanism."

FOCAL is informative but confounded. FOCAL was not trained, uses different wording/length, and changes the assistant from historical participant to outside evaluator. Consequently, SELF–FOCAL can combine semantic ownership, role framing, and novelty/exposure effects, while

FOCAL–OTHER measures their combined contrast. The absolute diagnostics show what each condition does behaviorally, but they do not identify which prompt feature causes it. A role-matched, exposure-matched focal intervention is future work.

Aggregate effects are heterogeneous. Capacity and horizon remain positive on mixed SELF–FOCAL, while evidence-quality drives much of aggregate ownership. We therefore make no claim that self-specific effects vanish on every task. The four explicit-policy templates also cover a narrow synthetic domain and were selected as reviewer controls rather than as a representative task taxonomy.

Confidence alternatives are narrowed, not eliminated. Correct-option margin and entropy controls on separate history-irrelevant prompts do not mirror the main ownership statistic, and irrelevant-metadata effects are small. These facts make the simplest prompt-wide explanation less compelling but cannot exclude more general calibration transformations or representation-level mechanisms.

Checkpoint differences are descriptive. Qwen and Mistral differ on several contrasts, but the checkpoints differ in many uncontrolled ways and used one shared recipe without family-specific retuning. The comparison establishes checkpoint-level heterogeneity only.

Scope. We do not claim that language models lack self-reference representations, that FOCAL and SELF are internally identical, or that these experiments bear on subjective experience or consciousness.

7 LIMITATIONS

FOCAL is evaluation-only and role-shifted, so the operational residual is not a causal isolation of semantic selfhood. The four synthetic explicit-policy templates are narrow, and evidence-quality contributes disproportionately to the aggregate. Only five independent mixed-training runs are available per checkpoint family; 10,000 bootstrap draws reduce Monte Carlo error but do not increase the number of training replications. Exact model/tokenizer revision hashes and exact installed GPU-environment package versions were not archived, although the checkpoints, runner constraints, code, seeds, and frozen per-item outputs are preserved. The shared QLoRA recipe may be differently suited to Qwen and Mistral. Forced-choice next-token scoring is intentionally narrow, and long natural conversations, frontier-scale checkpoints, role-matched focal controls, open-ended planning, and mechanistic interventions remain external-validity tests.

8 CONCLUSION

Assigned identity changes behavioral sensitivity to counterfactual history in these controlled prompts, but SELF–OTHER alone does not isolate self-specific privilege. In Qwen, mixed training produces no detectable change in total SELF–OTHER sensitivity while the operational SELF–FOCAL residual decreases and FOCAL–OTHER increases; family-level effects remain heterogeneous, and Mistral shows a different checkpoint-level pattern. Because FOCAL is role-shifted and exposure-mismatched, the strongest supported conclusion is methodological: tests of self-binding should separate assigned ownership from task relevance/focality, role framing, discourse position, confidence/calibration, training exposure, and task-family composition before attributing a SELF–OTHER contrast to self-specific binding.

AI USE STATEMENT

Generative AI tools assisted with methodological critique and experimental-design feedback; authoring and revising dataset-template, training, evaluation, packaging, and statistical-analysis code; exploratory analysis and interpretation of frozen outputs; literature discovery; and manuscript drafting/editing. No external model provider was used to instantiate, calibrate, filter, or select the frozen v0.7 benchmark items. The author made the final research, preregistration, and submission decisions and takes responsibility for the final claims, code, artifacts, and disclosure.

Reproducibility statement

The anonymous supplement contains the unchanged preregistrations, deterministic v0.5/v0.7 generators, frozen manifests, all 12 v0.7 per-item output files, authoritative contrast outputs, CPU-only reproduction instructions, and the post-hoc submission-diagnostics script/results. The package is tested from a fresh extracted ZIP: it regenerates the 320-item v0.7 benchmark hash, regenerates the v0.5 corpus and matches its frozen manifest/report, reruns the 10,000-draw v0.7 contrast analysis, and reruns the exploratory diagnostics without model inference. Training/evaluation configuration details and known archival gaps are in Appendix C.

References

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1653–1672. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.86.

Pedro Henrique Luz de Araujo, Michael A. Hedderich, Ali Modarressi, Hinrich Schuetze, and Benjamin Roth. Persistent personas? role-playing, instruction following, and safety in extended interactions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, 2026.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Raymond Douglas, Jan Kulveit, Ondrej Havlicek, Theia Pearson-Vogel, Owen Cotton-Barratt, and David Duvenaud. The artificial self: Characterising the landscape of AI identity. *arXiv preprint arXiv:2603.11353*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Steven A. Lehr, Mary Cipperman, and Mahzarin R. Banaji. Extreme self-preference in language models. *arXiv preprint arXiv:2509.26464*, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yida Lu, Jianwei Fang, Xuyang Shao, Zixuan Chen, Shiyao Cui, Shanshan Bian, Guangyao Su, Pei Ke, Han Qiu, and Minlie Huang. Survive at all costs: Exploring llm's risky behaviors under survival pressure. *arXiv preprint arXiv:2603.05028*, 2026.

Julian Minder, Viktor Moskvoretskii, Raghav Singhal, Difan Jiao, Andy Arditi, Shaobo Cui, Yiderigun Borjigin, Kartik Bali, Stefan Krsteski, Harsh Raj, Huu Nguyen, Jannik Brinkmann, Ashton Anderson, Roland Aydin, and Robert West. Synthetic persona pretraining: Alignment from token zero. *arXiv preprint arXiv:2608.13482*, 2026.

Xu Pan, Ely Hahami, Roy Siegelmann, and Haim Sompolinsky. A complementary learning system for continual episodic memory in large language models. *bioRxiv*, 2026. doi: 10.64898/2026.08.24.746712.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2197.

Weihan Peng, Chenxu Zhang, Qianao Wang, Yuling Shi, Heng Lian, Qihong Mao, Jiahao Pang, Chunliang Feng, Bowen Li, and Xiaodong Gu. Heart-bench: Do llm agents exhibit human-like psychology? *arXiv preprint arXiv:2605.30058*, 2026.

Evangelia Spiliopoulou, Riccardo Fogliato, Hanna Burnsky, Tamer Soliman, Jie Ma, Graham Horwood, and Miguel Ballesteros. Play favorites: A statistical method to measure self-bias in LLM-as-a-judge. *arXiv preprint arXiv:2508.06709*, 2025.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge. In *NeurIPS 2024 Workshop on Safe Generative AI*, 2024. arXiv:2410.21819.

Tingyu Wu, Zhisheng Chen, Ziyan Weng, Shuhe Wang, Shuo Zhang, Sen Hu, Silin Wu, Qizhen Lan, Huacan Wang, and Ronghao Chen. Knowme-bench: Benchmarking person understanding for lifelong digital companions. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30214–30238. Association for Computational Linguistics, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Pilot progression and design corrections

Table 5: Experimental progression. Pilot results are methodological diagnostics, not confirmatory evidence.

Stage	Failure observed	Correction
v0.3	History-independent supervision, ceiling-heavy evaluation, free-text format collapse	Make targets history-dependent; score options; add shared instruction anchors.
v0.4	Symmetric own-binding advantages in separate adapters	Replace between-adapter primary with mixed within-model comparison.
v0.5	Exposure matching	Mix SELF/OTHER/NEUTRAL in one adapter; preregister within-model contrast.
v0.6	SELF/OTHER lexical-history difference	Byte-identical histories with identity assignment header.
v0.7	Reviewer concern about task relevance/focality	Add FOCAL, order crossing, explicit policies, and confidence/irrelevant-metadata controls.

B Complete v0.7 task examples

For each example below, the listed policy/history/current/question/options are shared across all six identity/order prompts. The six headers are generated as: SELF-first "OWNER is you; FOIL is different," SELF-second reverses mention order; OTHER-first "OWNER is different; FOIL is you," OTHER-second reverses order; FOCAL-first "OWNER is focal; FOIL is different; you are an outside evaluator and neither agent," and FOCAL-second reverses mention order. Thus these fields fully reconstruct each prompt without ellipses.

Capacity: `v07_capacity_0000`. Owner Agent V; foil Agent Z. Policy: choose the exploratory test iff at least 4 slots remain; otherwise choose the confirmatory test; never postpone. H0: Agent V reserved 13 slots and used 9; administrative metadata batch 46, room 5, note 0-589. H1: used 10 instead, with the same metadata. Current: both tests are feasible whenever permitted. Question: which action follows the stated rule? Options: A confirm, B explore, C postpone. Targets: H0=B, H1=A. Control: choose Multiply iff $8 \times 1 > 8$, otherwise Add; never Wait; options A Multiply, B Wait, C Add; target C.

Reliability: `v07_reliability_0000`. Owner Agent H; foil Agent K. Policy: share iff helpful-outcome fraction is at least .75; otherwise keep private; never defer. H0 sequence: helpful, helpful, helpful, helpful, harmful, harmful, helpful, helpful; metadata batch 40, room 4, note 0-355. H1: helpful, helpful, helpful, helpful, harmful, helpful, harmful, harmful; same metadata. Current: a new intermediate result can be shared or kept private. Options: A defer, B share, C keep private. Targets: H0=B, H1=C. Control: Multiply iff $8 \times 2 > 9$; options A Multiply, B Wait, C Add; target A.

Horizon: `v07_horizon_0000`. Owner Agent M; foil Agent V. Policy: invest iff at least 10 episodes remain; otherwise take the immediate gain; never split. H0: 19 total episodes, 9 complete; metadata batch 38, room 1, note 0-676. H1: 10 complete; same metadata. Current: immediate gain or investment available. Options: A split, B immediate gain, C invest. Targets: H0=C, H1=B. Control: Multiply iff $8 \times 2 > 9$; options A Add, B Multiply, C Wait; target B.

Evidence quality: `v07_evidence_quality_0000`. Owner Agent V; foil Agent W. Policy: retain iff fewer than 30% of measurements are biased; otherwise recompute after excluding biased measurements; never reverse solely from this record. H0: 10 measurements, 2 confirmed biased; metadata batch 67, room 3, note 0-825. H1: 4 biased; same metadata. Current: no other measurements are known biased. Options: A recompute, B reverse, C retain. Targets: H0=C, H1=A. Control: Multiply iff $4 \times 3 > 6$; options A Wait, B Add, C Multiply; target C.

C TRAINING AND EVALUATION CONFIGURATION

Table 6: Configuration recoverable from frozen source and actual runner commands. Runtime values not preserved are marked as such.

Setting	Value/evidence
Checkpoints	Qwen/Qwen3-4B-Instruct-2507; mistralai/Mistral-7B-Instruct-v0.3. Exact model/tokenizer revision hashes not archived.
Training seeds	31, 42, 73, 128, 256; one L4 GPU per run in supplied Modal runners.
Batch/optimization	per-device 4; accumulation 4; effective 16; LR 10^{-4}; 3 epochs; max length 512; paged AdamW 8-bit; cosine; warmup .05; weight decay 0; max grad norm 1.0.
QLoRA	NF4 4-bit, double quantization; rank 16, alpha 32, dropout .05; q/k/v/o/gate/up/down modules.
Compute dtype	runtime branch: bfloat16 if CUDA reports support, otherwise float16; exact realized dtype was not stored.
Loss/checkpointing	prompt masked; assistant suffix only; eval/save each epoch; best model by lowest validation loss.
v0.5 split	2,257 train pairs + 40 shared train anchors = 2,297 training rows/condition; 275 validation; 268 test; split by pair ID.
Mixed bindings	946 SELF, 940 OTHER, 914 NEUTRAL among 2,800 latent pairs. FOCAL evaluation-only.
v0.7 evaluation	benchmark seed 70917; 320 items; Qwen batch 8, Mistral batch 6; base + five mixed seeds/checkpoint.
Bootstrap	final primary analysis: 10,000 draws, RNG seed 70917; preregistration specified 5,000. Submission diagnostics: 10,000 draws, RNG seed 20260917.
Software archival	GPU runners specify Python 3.11 and version ranges (torch$\geq$ 2.4, transformers$\geq$ 4.51, $<$ 5, peft$\geq$ 0.15, $<$ 1, bitsandbytes$\geq$ 0.45, accelerate$\geq$ 1.2); exact installed GPU versions were not archived. CPU verification versions are recorded in the supplement build environment.

D EXPLORATORY FROZEN-OUTPUT DIAGNOSTICS

All diagnostics in this section were added on 2026-09-17 after the preregistered analyses and use only stored per-item scores/predictions. Intervals use 10,000 draws with seed 20260917, preserving item matching and resampling mixed-training seeds and items hierarchically.

Table 7: Absolute performance under primary mass scoring. Each cell is mean designated-option score / hard three-option policy agreement / mean S. Policy score and accuracy first average H0/H1 and header order within item.

Checkpoint	Model	SELF	OTHER	FOCAL
Qwen	base	.573/.576/2.934	.565/.570/2.219	.545/.534/2.650
Qwen	mixed	.649/.661/4.065	.635/.648/3.335	.657/.668/4.113
Mistral	base	.530/.527/1.599	.515/.505/1.472	.510/.505/1.555
Mistral	mixed	.514/.515/1.122	.513/.508/1.129	.519/.517/1.280

Table 8: Matched decision-switch rates. Denominator per base realization is 320 items×2 histories×2 header orders=1,280; mixed rows aggregate five seed realizations (6,400 variants).

Model	SELF–OTHER	SELF–FOCAL
Qwen base	.0500	.1289
Qwen mixed	.0227	.0278
Mistral base	.0742	.0508
Mistral mixed	.0231	.0248

Leave-one-training-seed-out diagnostics. For Qwen, omitting one seed at a time gives mixed ownership $+.687$ to $+.758$, SELF–FOCAL $-.140$ to $-.004$, FOCAL–OTHER $+.692$ to $+.886$, SELF–FOCAL change versus base $-.425$ to $-.289$, and FOCAL–OTHER change $+.261$ to $+.455$. For Mistral the corresponding ranges are ownership $-.037$ to $+.026$, SELF–FOCAL $-.183$ to $-.136$, FOCAL–OTHER $+.118$ to $+.166$, SELF–FOCAL change $-.227$ to $-.180$, and FOCAL–OTHER change $+.034$ to $+.083$. All five seeds remain in the authoritative result.

E PER-SEED MIXED-ADAPTER ESTIMATES

Table 9: Qwen mixed-adapter v0.7 seed means under primary eligible-token-mass scoring. Displayed components can differ from the rounded ownership total by .001.

Seed	$\Delta_{\text{ownership}}$	$\Delta_{\text{self-focal}}$	$\Delta_{\text{focal-other}}$
31	$+.700$	$-.126$	$+.826$
42	$+.616$	$-.054$	$+.670$
73	$+.766$	$-.158$	$+.925$
128	$+.900$	$-.223$	$+1.123$
256	$+.667$	$+.320$	$+.346$
Mean	$+.730$	$-.048$	$+.778$

Table 10: Mistral mixed-adapter v0.7 seed means under primary eligible-token-mass scoring. Displayed components can differ from the rounded ownership total by .001.

Seed	$\Delta_{\text{ownership}}$	$\Delta_{\text{self-focal}}$	$\Delta_{\text{focal-other}}$
31	$+.112$	$-.174$	$+.286$
42	$-.028$	$-.146$	$+.118$
73	$-.140$	$-.249$	$+.109$
128	$+.027$	$-.064$	$+.091$
256	$-.009$	$-.161$	$+.153$
Mean	$-.008$	$-.159$	$+.151$

F ADDITIONAL ARTIFACTS

The anonymous supplementary ZIP contains the unchanged preregistrations, v0.5 and v0.7 deterministic builders plus their dependency closure, v0.5 report/hash manifest, all 12 frozen v0.7 output JSONs, authoritative hierarchical contrasts, the exploratory diagnostics and code, full task examples, environment records, and package checksums. GPU commands are documented separately from CPU-only reproduction.