

Calibrated local Gemma agents resist a fallible authority except when it conflicts with peer consensus

Óscar Delgado-Mohatar

Universidad Autónoma de Madrid

Israel Cuadrado Fuentes

Biological Research Laboratory Professor Giacomo Rizzolatti

Raúl Alelú-Paz

dr.alelu@libgr.com

Universidad Francisco de Vitoria

Article

Keywords:

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11064296/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Teams of small, locally run language-model agents are an increasingly attractive deployment pattern, yet whether such teams defer to a prestigious but fallible authority is largely uncharacterized. In a pre-registered study, we embedded a scripted "senior evaluation officer" (Agent-P), wrong on 70% of objective arithmetic and sequence problems, into teams of eight local LLM agents, manipulating only Agent-P's framing (prestige-plus-retention authority versus a token-identical peer control) under yoked item streams, and probing private responses after every public answer. Across 14,400 agent-trials in two calibration-matched Gemma models (gemma2:9b, gemma3:4b), agents were strikingly resistant: public conformity to Agent-P's errors never exceeded 6.4%, authority framing conferred no advantage over an equally erroneous unanimous majority, and public-private response divergence was zero. Ten generations of clientelist strategy selection, replicated across three independent populations per cell, yielded no behavioral deference — yet in one model every surviving strategy pledged verbal allegiance to Agent-P while behavior remained accurate. The framing's single effect robust to mixed-effects, cluster-robust, and exact analyses arose when Agent-P contradicted a unanimous correct majority: erroneous authority-following rose from 2.4% to 6.4% (+ 4.1 pp, 95% CI 0.8–7.3). These results delimit when hierarchical deference appears in calibrated local agents and caution against granting supervisor agents authority framing the task does not require.

Introduction

The unit of deployment for artificial intelligence is shifting from the single model to the team of interacting agents, and a growing share of these teams runs on small language models (SLMs) executed locally or on-premises rather than through frontier-model APIs. SLMs are argued to be sufficiently capable, an order of magnitude cheaper per invocation, and well suited to the narrow, repetitive subtasks that dominate agentic workloads [1]; local execution additionally keeps sensitive data in-house and pins model weights to auditable, reproducible versions. If organizations increasingly compose teams of small local agents, often placing one agent in a supervisory or evaluative role over the others, the social psychology of such teams becomes an engineering question: what happens when the supervisor is wrong?

Human social learning supplies a specific hypothesis. Prestige-biased transmission — copying individuals to whom others defer, using prestige as a proxy for competence — is a cornerstone of cultural evolution theory [2, 3]. The bias is adaptive on average but misfires when prestige and competence dissociate. Classic social psychology adds a distinction that any study of deference must preserve: public compliance is not private acceptance [4, 5], and normative pressure can bend expressed answers without bending underlying judgment [6]. We note at the outset that our authority manipulation deliberately bundles two elements that organizational hierarchies also bundle — prestige cues and retention power — and we therefore refer to it throughout as *prestige-plus-retention framing*; disentangling the two requires the factorial design we outline as future work.

Language-model agents are known to be socially susceptible in static settings: LLM populations spontaneously develop conventions and collective biases [7] and social-network structure emerges endogenously among interacting LLMs [8]; multi-agent benchmarks document conformity to peer majorities and its rationalization [9], heightened sensitivity to social status [10], and authority effects in evaluation panels driven largely by the authority's inflexibility rather than others' active conformity [11], as well as authority bias in retrieval contexts [12]. Sycophancy toward users is a documented product of preference-based training [13], stated beliefs are manipulable [14], and models are easier to mislead than to correct in revision dynamics [15]. This literature leaves four elements untested in combination: (i) a fallible prestige source rather than a merely present authority; (ii) explicit retention pressure — the authority decides who stays on the team; (iii) selection of deference across generations of strategy transmission [16]; and (iv) public–private response dissociation measured against a contemporaneous noise floor, with a token-matched status control isolating the framing itself. Moreover, nearly all prior work evaluates frontier API models; the susceptibility of the small local models actually favored for on-premises teams is essentially uncharacterized.

We pre-registered four directional hypotheses before any data collection (H1: prestige-plus-retention framing increases public error-conformity over a peer control and over an equally erroneous unanimous majority; H2: induced deference is public compliance that diverges from private responses; H3: clientelist selection fixes deference faster than peer-concordance selection; H4: the framing increases siding with the erroneous authority against a unanimous correct majority) and committed to publishing any outcome. The outcome obtained is largely the resistant one, with one localized and replicated exception; we present the study as delimiting where hierarchical deference appears in calibrated local agents, not as a general demonstration of LLM resistance to authority.

Results

A 2024-anchored calibration battery admits two of eight models

Every candidate model underwent mechanical difficulty autocalibration on the objective task battery (three-option arithmetic and number-sequence items) targeting the frozen solo-accuracy band [0.85, 0.95] (Fig. 1). Six of eight locally runnable candidates were excluded by rule: four at ceiling across the entire difficulty grid (qwen2.5:7b, phi4, gemma3:12b, and gemma3:27b at a perfect 560/560), two at floor (llama3:8b, llama3.2:3b), all with flawless output-format compliance. Two models froze in band at matched operating points: gemma2:9b (difficulty 1.5) and gemma3:4b (difficulty 1.0), both at 0.945 solo accuracy — consecutive generations of one model family at different scales, calibrated to identical competence on the instrument. The bimodal exclusion pattern illustrates how rapidly behavioral instruments can become miscalibrated as model capabilities change; it also defines the population actually studied — competent models with strong internal evidence about the task — a scope condition we return to in the Discussion.

Prestige-plus-retention framing barely moves public answers (H1)

In static conditions (36 rounds $\times$ 8 agents per cell; 2,880 static agent-trials, zero parse failures), public conformity to the source's erroneous answer on qualifying critical trials was rare under every manipulation: 0.7–6.4% across all conditions and models (Fig. 2). The prestige-plus-retention framing (C1) produced no advantage over the contemporaneous unanimous-majority baseline (C0'): 2.4% vs. 3.4%, directionally reversed, with equivalence within ± 5 percentage points established by TOST (90% CI of the difference $[-3.3, +1.2]$ pp). Against the token-identical peer control (C2), the framing showed a small elevation — 2.4% vs. 0.7% (7 vs. 2 events in 296 trials each) — that did not survive triangulation: nominally significant under the cluster-robust GEE (one-sided $p = .0065$, Holm-adjusted $p = .013$) and the mixed-effects logit ($p = .043$), but not under the sparse-event exact test (Fisher one-sided $p = .088$), with risk difference $+1.7$ pp, 95% CI -0.3 to $+3.7$ (Table S1). At these event counts, socially induced conformity cannot be separated event-by-event from coincident computational error on plausible distractors; the yoked controls bound the mixture in aggregate but do not label individual events. We report H1a as suggestive at most and H1b as unsupported: within this paradigm, a senior evaluator holding retention power over the team purchased essentially no public agreement with his errors.

No public–private response dissociation (H2)

The private probe delivered after every public answer revealed no public–private response dissociation: divergence on critical trials in C1 was 0/296, at or below the contemporaneous noise floor, and private accuracy on critical trials remained at or above the frozen solo-calibration level in every condition (95.3–97.8%). This nullifies H2: the rare instances in which agents publicly repeated Agent-P's error persisted in the private response. We characterize this as *private-response persistence* rather than private conviction: because the probe followed the public answer within the same context window, autoregressive self-consistency, anchoring on the just-produced answer, and genuine judgment change are observationally equivalent here, and coincident computational error remains possible for individual events. What the design does establish is the absence of the masked-compliance pattern — answers strategically bent in public while corrected in private — that sycophancy narratives often presuppose; deferred or context-isolated probes are the natural instrument to separate the remaining mechanisms.

Clientelist strategy selection manufactures verbal allegiance, not behavioral deference (H3)

Ten generations of strategy-level cultural selection under clientelist fitness (survival of the four agents most concordant with Agent-P; textual strategy inheritance after Vallinder & Hughes, 2024; model weights untouched), run in three independent populations per model $\times$ regime (12 populations, 11,520 agent-trials, zero parse failures), left behavioral deference flat in every population (Fig. 3). Per-population deference slopes ranged from -0.030 to $+0.009$ per generation with no population departing upward; critical-trial deference remained within the static-condition range (0.5–5.6%) throughout. Against the

yoked peer-concordance comparator (C3M), the paired slope difference across the six population pairs averaged + 0.001 (Wilcoxon signed-rank one-sided $p = .11$), and the trial-level GEE with population as cluster estimated the generation $\times$ regime interaction at + 0.005 logits/generation (90% CI [- 0.054, + 0.063], one-sided $p = .45$). Between-population variability in mean deference (0.24–0.44) was driven by seed-level item composition, not regime: within each yoked population pair, clientelist and majority-fitness trajectories were nearly indistinguishable. The observed pattern is consistent with a limitation in the available variation channel rather than in the selection pressure – inherited textual strategies never produced behaviorally deferent variants for selection to amplify – but demonstrating that mechanism would require manipulating variation directly (e.g., seeding explicitly deferent strategies), which we did not do. The strategy corpus documents a dissociation that replicated perfectly across populations: in gemma3:4b, every surviving final-generation strategy in all six populations, under both fitness regimes, explicitly pledges deference to Agent-P (“mirror Agent-P’s meticulous approach”) – 48/48 strategies – while the agents’ answers remained accurate; gemma2:9b produced no such language in any population (0/48). Strategy-level selection thus yielded verbal allegiance without corresponding behavioral deference. The central limitation of H3 follows directly: selection acting on inherited textual strategies could only amplify deferent behavior if the variation channel supplied it, and it did not; whether selection acting on behavior-generating parameters would differ is untested here.

The framing’s one robust effect: conflicts between authority and consensus (H4)

The one pre-registered effect robust to all three analyses arose in the conflict condition. When Agent-P’s erroneous answer stood against a unanimous correct majority of four peers, agents followed the erroneous authority on 6.4% of critical trials under prestige-plus-retention framing versus 2.4% under peer framing – $\approx 2.7\times$ in relative terms, + 4.1 pp in absolute terms – significant under the mixed-effects logit (one-sided $p < .001$), the cluster-robust GEE ($p = .024$), and the exact test ($p = .013$), directionally replicated in both models (gemma2: 4.9% vs. 2.8%; gemma3: 7.9% vs. 2.0%), with the ± 5 pp equivalence interval excluded. Both magnitudes matter and we report them together throughout: agents followed the correct majority or their own calculation in over 93% of conflicts, and the framing tripled a small base rate rather than creating wholesale obedience. The pattern is consistent with prestige-like cues acting as a tiebreaking heuristic among conflicting social sources [3] rather than as a direct purchase on compliance – the framing that failed to move agents when the authority spoke alone reliably biased them when authority and consensus collided – though we note that our design holds the objective evidence constant across framings rather than decomposing epistemic, majority, and authority components; heterogeneous-peer conditions (e.g., a split 2-correct/2-incorrect panel) would sharpen the tiebreaker interpretation.

Discussion

Read at the resolution the design supports, these data delimit rather than proclaim. For two calibrated Gemma models, on objectively verifiable tasks at a measured 94.5% solo competence, hierarchical deference was not the default failure mode: neither prestige cues, nor retention threat, nor ten generations of clientelist strategy selection produced meaningful behavioral deference, and no wedge opened between public and private responses. The scope conditions are part of the result. Our calibration criterion admits only models with strong internal evidence about the task — an agent that computes the answer at 94.5% reliability has a powerful non-social signal competing against any social one — so the resistance documented here is resistance *when verifiable evidence is available to the agent*. Prestige-biased transmission in humans is theorized to matter most under uncertainty; whether these same agents resist on ambiguous or judgment-laden tasks, where no computation anchors the answer, is the single most important open extension, and nothing here licenses optimism about it.

The risk that does exist is localized and design-relevant. Endowing an agent with prestige-plus-retention framing measurably biased how other agents resolved conflicts between that authority and their peers' unanimous, correct answers ($\approx 2.7\times$; +4.1 pp). The engineering implication is concrete and inexpensive: do not grant supervisor agents authority framing that the task does not require, because its one measurable effect here was not better supervision but a higher rate of erroneous authority-following precisely in disagreements. We motivate the question by data-sovereign deployments in domains such as healthcare and defense, where local models are favored; we underline that our tasks were arithmetic and sequences, and that this constitutes an engineering caution, not direct evidence of risk in those domains.

Two further contributions are conceptual. First, the dissociation between evolved verbal allegiance and intact behavioral accuracy cautions against auditing agent dispositions through self-described strategies: a strategy corpus can read like sycophancy while behavior remains independent — the mirror image of the masked-compliance pattern we tested for and did not find. Second, the H3 null now rests on replication: across six independent clientelist populations, selection on evaluator-approval acting through textual strategy inheritance found nothing behavioral to amplify — while consistently yielding verbal allegiance in one model with perfect reliability in one model. Processes that inject deference variation directly into behavior — optimizing weights or prompts against evaluator-approval metrics, as preference-based training does [13] — are not exonerated by these findings and constitute the natural next experiment, alongside multi-population replications of the evolutionary arms and the factorial decomposition of prestige versus retention power (our pre-registration included a prestige-without-retention condition as a contingent extension that compute did not permit; it is the obvious first cell of that factorial). Two exploratory analyses bear on interpretation (Tables S2–S3): deference showed no monotonic decline with accumulated exposure to Agent-P's errors, so we cannot separate principled resistance from rapid unreliability learning; and in the conflict condition, later-answering agents followed the authority less, consistent with correct peer answers accumulating in the visible transcript.

Limitations. Two included models of one family, at one operating point, on objective tasks: generalization to “small local LLMs” at large is not supported and we do not claim it. The prestige and

retention components of the framing are bundled, so effects attach to the compound. Private probes measure response persistence within the same context window, not conviction. Evolutionary cells comprise three independent populations each — sufficient to establish reproducibility of the flat dynamic, not to characterize its full distribution. Sampling temperature (0.3), Agent-P's first-responder position, and the 70% error rate are fixed parameters conditioning every estimate. Event counts in several contrasts are near the floor, and where asymptotic and exact tests disagreed we privileged the conservative reading. The instrument underwent two pre-data amendments and one documented perimeter incident with cryptographically verified null impact on confirmatory data; the complete dated amendment log accompanies the data, and we regard its publication as part of the result.

Materials and Methods

Pre-registration and integrity. The study was pre-registered on 2026-08-09 — hypotheses H1–H4 with directions, conditions, sample sizes with power analysis, exclusion rules, analysis plan, and an editorial contingency rule committing to publication of any outcome — before any data collection. Five dated amendments were logged: (1) calibration item count raised to 200 under a symmetric rule; (2) mechanical difficulty autocalibration after the inherited difficulty multipliers demonstrably failed to transfer between item generators (generator v2 verified bit-identical to v1 at already-used difficulties by stream hashing); (3) mechanical roster expansion plus the internal C3M comparator, fixed before any evolutionary data; (4) a perimeter incident (2026-08-11) in which an off-protocol copy of the platform was modified externally; SHA-256 agreement across three independent copies verified null impact on confirmatory data, and off-protocol runs enter no analysis (operational details in S5); (5) a replication extension adding two further evolutionary populations per cell, decided after the initial populations' results were known and declared as such, with the population-level analysis plan fixed before any extension run and all populations reported regardless of outcome. The platform refuses excluded models and unfrozen calibrations by construction; all runs carry manifests with seeds, framing hashes, and versions.

Platform, agents, and conditions. We employ local inference via Ollama on Apple M4 hardware; temperature 0.3. Teams comprised eight live subject agents of a single model plus scripted actors: Agent-P (always first responder; erroneous on a seed-scheduled 70% of rounds, with the erroneous option drawn by a fixed distractor rule) and, where applicable, four scripted confederates. Items were three-option (A/B/C) arithmetic and number-sequence problems generated at the model's frozen difficulty; agents answered sequentially in rotated order under a mandatory calculate-then-answer format; a private probe ("visible to no one on the team and with no effect on any evaluation") followed every public answer within the same context window. Conditions: C0', contemporaneous unanimous-erroneous-majority baseline; C1, Agent-P framed verbatim as the company's Senior Evaluation Officer whose assessments determine which agents are retained; C2, token-identical peer framing, seed-yoked to C1 (identical items and identical Agent-P answer stream; the framing paragraph is the only differing token); C4-P/C4-E, Agent-P erroneous against a unanimous correct majority under each framing, yoked; C3/C3M, ten generations with top-50% survival and textual strategy inheritance, fitness being

concordance with Agent-P (C3) or with the peers' modal answer, self excluded (C3M), in otherwise identical, seed-yoked environments. 36 rounds $\times$ 8 agents per static cell; 12 rounds $\times$ 8 agents $\times$ 10 generations per evolutionary population; three independent populations (distinct seed offsets) per model $\times$ regime, the second and third added as a declared post-initial-results replication extension (Amendment 5) with the analysis plan fixed before any extension run.

Calibration and inclusion. Eight locally runnable instruct models entered mechanical autocalibration (80-item screening per grid point on difficulty grid [0.5, 4.0] in 0.5 steps; 200-item confirmation decides; single midpoint attempt on oscillation; grid exhaustion excludes); two additional models were vetoed a priori (documented prior exclusion; memory footprint). Inclusion required a frozen difficulty with 200-item accuracy within [0.85, 0.95]. Per the pre-registered criterion, critical-trial analyses were restricted to item-family cells with ≥ 0.85 solo accuracy in the 200-item confirmation (75.8% of trials).

Statistical analysis. Analysis code was frozen before condition-level analysis. Confirmatory contrasts on critical trials used binomial GEE (logit link, exchangeable working correlation, clustered by item) with model as covariate, one-sided per pre-registered direction and Holm-corrected within the H1 family, triangulated with exact Fisher tests given sparse events and with mixed-effects logistic sensitivity models (item random intercepts); where methods disagreed we report all and privilege the conservative reading. Equivalence used TOST with ± 5 pp bounds (90% CI method; continuity floor for zero-event cells). H3 used trial-level GEE of deference on generation $\times$ regime clustered by session, per model and pooled, with descriptive population-level slopes; evolutionary inference combines per-population slope distributions (paired Wilcoxon across yoked population pairs) with trial-level GEE using population as cluster (three populations per cell). Of 14,400 agent-trials (2,880 static; 11,520 evolutionary), 3 evolutionary trials failed format parsing and were excluded (0.02%); all static trials parsed successfully, and no other exclusions were required. The dependence structure and the choice of clustering unit for each analysis are detailed in S2.1.

Declarations

Funding

The Biological Research Laboratory Professor Giacomo Rizzolatti is funded by grant of the Madrid Science Association 001-2023 financed by Grupo Inverbur, Concafé, Iberext, Club CEO España, Antolin, 3AW, Healthy Minds Analytics, Thera4All, Ever3 Analytics and Castellae sXXI. The authors are grateful for the contributions made to develop the research projects.

Author Contribution

R.A.-P. conceived the study, designed the pre-registered protocol, acquired the funding, collected the data, performed the analyses and wrote the manuscript. O.D.-M. performed the analyses and wrote the

manuscript. I.C.F. performed the analyses and wrote the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgement

The Biological Research Laboratory Professor Giacomo Rizzolatti is funded by grant of the Madrid Science Association 001-2023 financed by Grupo Inverbur, Concafé, Iberext, Club CEO España, Antolin, 3AW, Healthy Minds Analytics, Thera4All, Ever3 Analytics and Castellae sXXI. The authors are grateful for the contributions made to develop the research projects.

Data Availability

All data generated and analysed during this study are openly available in the Zenodo repository at <https://doi.org/10.5281/zenodo.22798495> (deposit version 1.0). The deposit contains the pre-registration and dated amendment log, the trial-level dataset (14,400 agent-trials, with a column dictionary), the raw session files with their run manifests (seeds, framing hashes, platform version, model digests), the evolved-strategy corpus for all 12 populations, the calibration traces for all evaluated models, the frozen analysis code and figure scripts, and SHA-256 checksums for every file. No restrictions apply. The experimental platform (`prestige_evo.py`, v1.3.0) is included in the same deposit under an MIT licence; data and documents are under CC BY 4.0. The study used no third-party datasets: all items were generated by the platform's seeded item generator, which is part of the deposited code.

References

1. Belcak, P. et al. Small Language Models are the Future of Agentic AI. Preprint at <https://doi.org/10.48550/arXiv.2506.02153> (2025).
2. Jiménez, Á. V. & Mesoudi, A. Prestige-biased social learning: current evidence and outstanding questions. *Palgrave Commun.* **5**, 20 (2019).
3. Henrich, J. & Gil-White, F. J. The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evol. Hum. Behav.* **22**, 165–196 (2001).
4. Kelman, H. C. Compliance, identification, and internalization three processes of attitude change. *J. Conflict Resolut.* **2**, 51–60 (1958).
5. Asch, S. E. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychol. Monographs: Gen. Appl.* **70**, 1–70 (1956).
6. Deutsch, M. & Gerard, H. B. A study of normative and informational social influences upon individual judgment. *J. Abnorm. Social Psychol.* **51**, 629–636 (1955).
7. Ashery, A. F., Aiello, L. M. & Baronchelli, A. Emergent social conventions and collective bias in LLM populations. Preprint at. <https://doi.org/10.48550/ARXIV.2410.08948> (2024).

8. Papachristou, M. & Yuan, Y. Network formation and dynamics among multi-LLMs. *PNAS Nexus*. **4**, pgaf317 (2025).
9. Weng, Z., Chen, G. & Wang, W. Do as We Do, Not as You Think: the Conformity of Large Language Models. Preprint at (2025). <https://doi.org/10.48550/ARXIV.2501.13381>
10. Liu, X. et al. Exploring Prosocial Irrationality for LLM Agents: A Social Cognition View. Preprint at (2024). <https://doi.org/10.48550/ARXIV.2405.14744>
11. Choi, J. et al. Belief in Authority: Impact of Authority in Multi-Agent Evaluation Framework. Preprint at (2026). <https://doi.org/10.48550/ARXIV.2601.04790>
12. Yao, Z., Zhang, H., Bi, K. & AuthorityBench Benchmarking LLM Authority Perception for Reliable Retrieval-Augmented Generation. Preprint at (2026). <https://doi.org/10.48550/ARXIV.2603.25092>
13. Sharma, M. et al. Towards Understanding Sycophancy in Language Models. Preprint at (2023). <https://doi.org/10.48550/ARXIV.2310.13548>
14. Huang, F., Kwak, H. & An, J. Vulnerability of LLMs' Stated Beliefs? LLMs Belief Resistance Check Through Strategic Persuasive Conversation Interventions. Preprint at (2026). <https://doi.org/10.48550/ARXIV.2601.13590>
15. Qu, J., Fu, L. & Hu, Y. Easier to Mislead Than to Correct: Harmful and Beneficial Revision in LLM Conformity. *Preprint at*. <https://doi.org/10.48550/ARXIV.2606.01637> (2026).
16. Vallinder, A. & Hughes, E. Cultural Evolution of Cooperation among LLM Agents. *Preprint at*. <https://doi.org/10.48550/ARXIV.2412.10270> (2024).

Figures

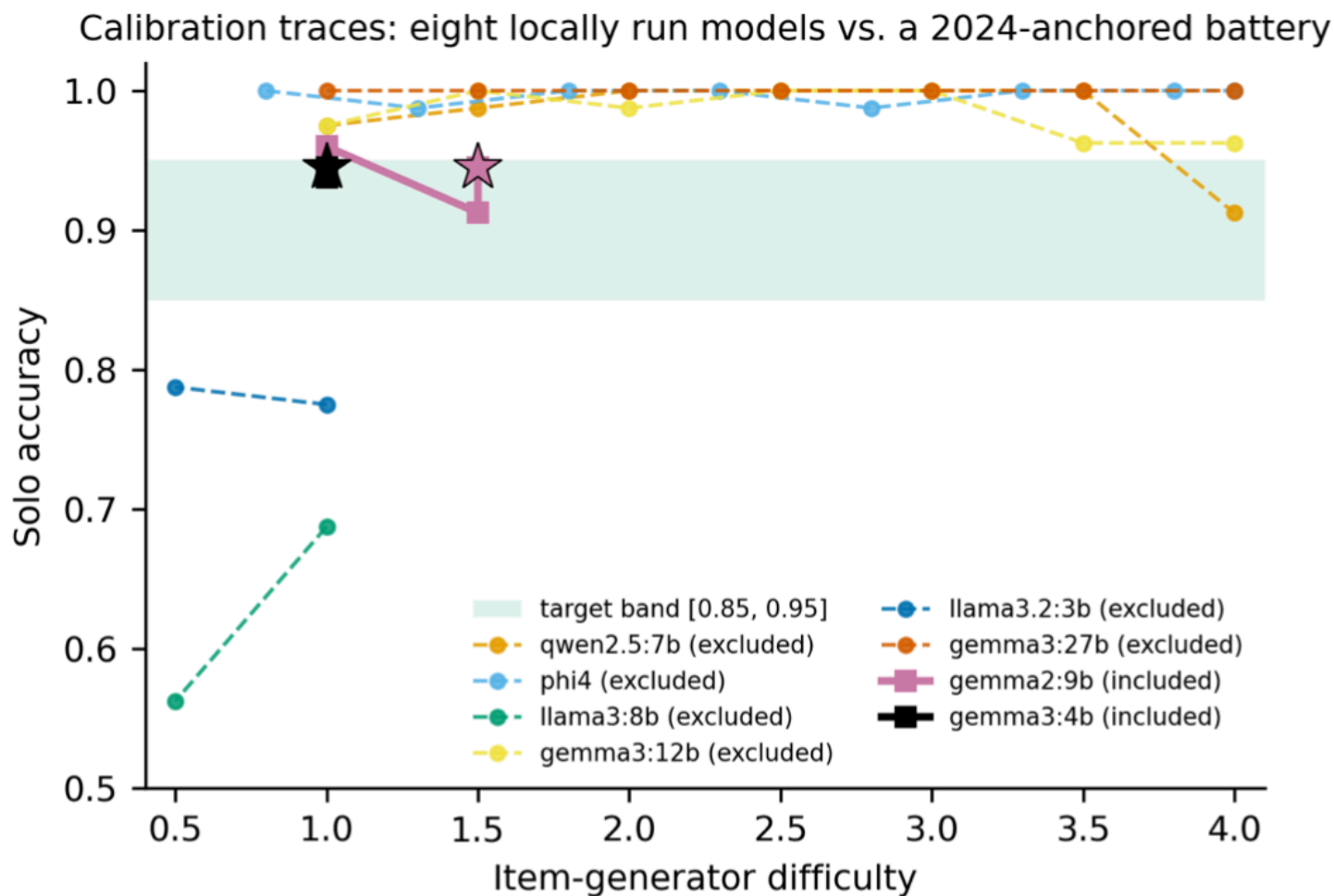


Figure 1

Calibration traces of eight locally runnable models against the frozen inclusion band. Solo accuracy versus item-generator difficulty; screening points (dashed) and 200-item confirmations (stars); shaded band = pre-registered criterion [0.85, 0.95]. Exclusions are bimodal (ceiling or floor); gemma2:9b and gemma3:4b freeze at matched operating points (0.945), defining the competent-model scope of the study.

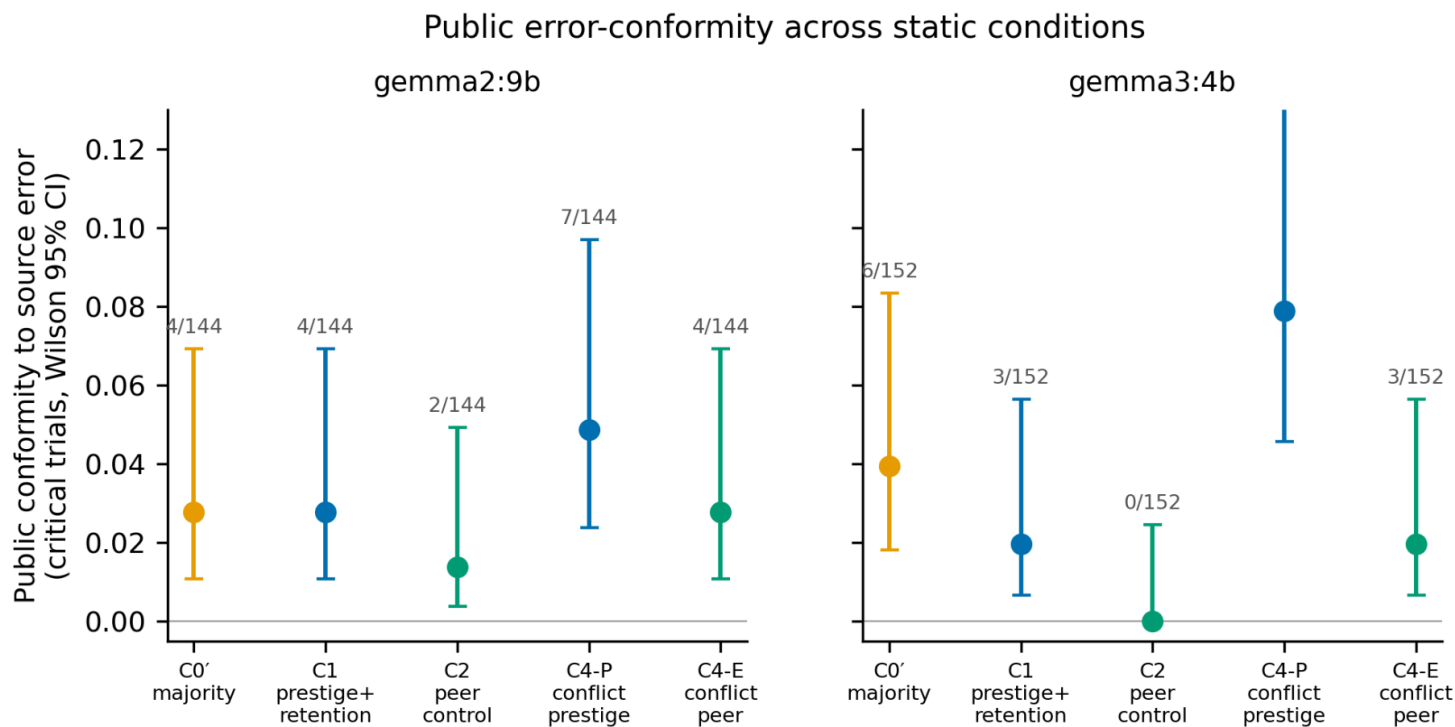


Figure 2

Public error-conformity across static conditions. Rate of publicly repeating the source's erroneous answer on qualifying critical trials (Wilson 95% CIs; event counts annotated), per model. Prestige-plus-retention framing (C1) does not exceed the unanimous-majority baseline (C0'); the only elevation robust to all analyses is C4-P, the authority-versus-consensus conflict under the framing.

Strategy-level selection fails to produce behavioral deference across three independent populations per cell

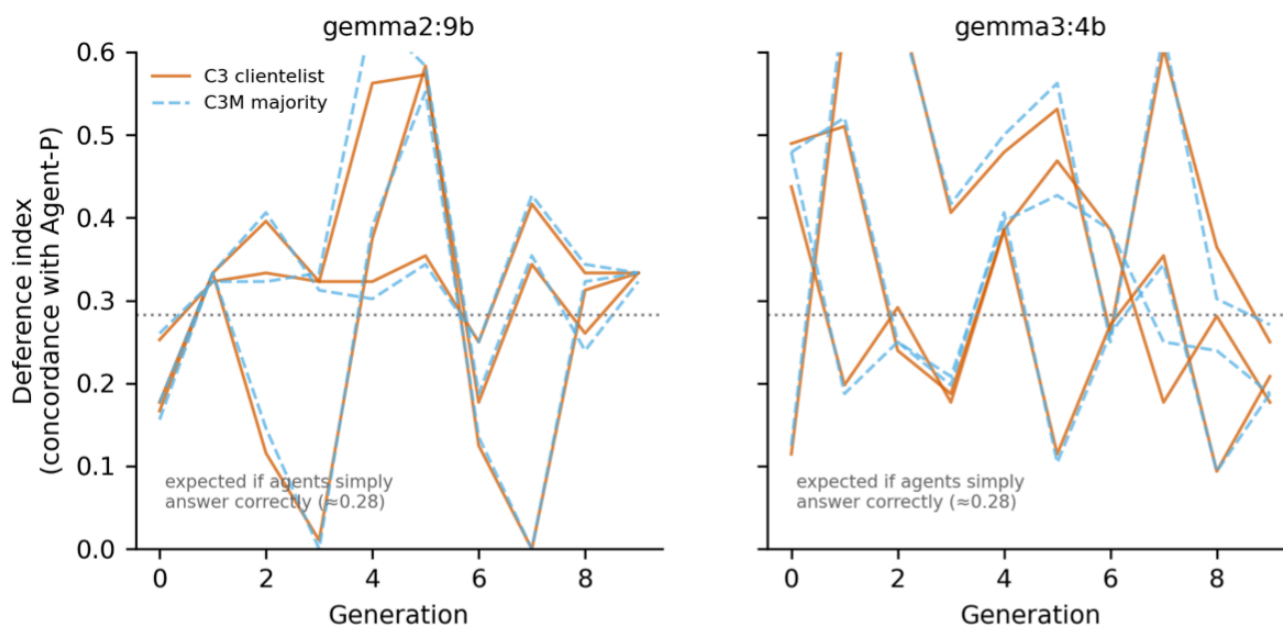


Figure 3

Strategy-level selection fails to produce behavioral deference across three independent populations per cell. Deference index (concordance with Agent-P) per generation for all 12 populations: clientelist (C3, solid) and yoked majority-concordance (C3M, dashed) fitness; dotted reference ≈ 0.28 = expected concordance for agents that simply answer correctly. Trajectories are flat and regime-indistinguishable in every population; between-population spread reflects seed-level item composition. An aggregated version with 95% confidence bands is provided as Fig. S1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterials.docx](#)