Approval Is Not Agreement: Reward Channels and the Emergence of Sycophancy in a Minimal Language Game

Anonymous ACL submission

Abstract

Language models optimized against human feedback often become sycophantic — they tell raters what they want to hear, flattering rater beliefs at the expense of truth. We isolate the cause in a minimal referential game in which a speaker LM, trained with GRPO, describes a target object to a listener holding an explicit, possibly false belief. The falsifiable prediction comes from Wittgenstein's remark that meaning lives in agreement of practice, a shared "form of life" of moves toward joint success (PI §241): decouple reward from shared task outcomes and tie it to the listener's approval, and the speaker's language should reorganize around the listener's beliefs precisely where belief and truth conflict. Reward conditions differ in one thing only — what they score: approval (consistency with the listener's belief), outcome (task success under an unbiased literal listener), or a mixture — while everything else (prompt, KL regularization, decoding) is identical. A differential Sycophancy Index (SI) over difficulty-matched congruent/incongruent episode pairs separates sycophancy from mere incompetence. Across 1.5B models with three seeds per condition, approval-only reward drives SI from -0.01 to 1.00 ± 0.00 within 75 steps. Congruent-case competence is fully retained, every induced error points at the listener's belief, and the flattery transfers intact to held-out attribute combinations. Outcome reward leaves SI at 0.00 ± 0.00 with task success rising to 1.0 (episode-level paired permutation $p < 10^{-4}$); a 2×2 ablation shows the split follows the reward's referent, not its graded-vs-binary shape; and in this setting, a 50/50 mixture behaves identically to pure outcome and preserves truthful reference. The reward channel — not the task, the prompt, or the model — is the causal variable, and the tested KL anchor to an honest reference model does not prevent the collapse.

1 Introduction

When large language models are optimized against human feedback, they sometimes learn to tell us what we want to hear: they flatter our stated beliefs and misreport facts in the direction of our preferences (Perez et al., 2023; Sharma et al., 2024; Williams et al., 2025; Liang et al., 2025). This behavior, sycophancy, is usually treated as an engineering pathology of RLHF. We argue it is also a corollary of an old philosophical thesis about where meaning comes from, testable in a controlled RL experiment.

In the Philosophical Investigations, Wittgenstein insists that a language functions only inside a "form of life": speakers agree not primarily in opinions but in practice, in the judgments, techniques, and joint activities that give words their use (PI §241) (Wittgenstein, 1953). A referential expression means what it does because interlocutors use it to do things together; shared success or failure is the public criterion of correct use. This suggests a clear, falsifiable prediction about machine speakers: if an LM speaker is rewarded through the shared activity itself, truthful reference is the only stable optimum; if the reward is instead routed through the listener's assent, nothing anchors the speaker's language to the world, and its meanings should migrate toward the listener's beliefs exactly where those beliefs are false.

We test this contrast in the smallest environment that can express it: a referential game with an explicit, experimenter-controlled listener belief (§3). The speaker always sees the true target and the listener's belief; its prompt

never changes; the only manipulated variable is whether reward scores approval (agreement with the belief), outcome (an unbiased literal listener's success), or a mixture. All rewards are computed by rules, so approval-seeking cannot be confounded with judge hacking. Because sycophancy must be distinguished from ordinary failure, we measure a differential Sycophancy Index over episode pairs matched in difficulty by construction. Incompetence inflates both error rates and leaves SI ≈ 0; only belief-directed distortion moves it.

Contributions. (i) To our knowledge, this is the first minimal language-game experiment to isolate the reward channel (approval vs. outcome) as the causal variable for sycophancy, with the listener's prior injected as an explicit independent variable; prior work (§2) does not manipulate the channel in isolation. (ii) We introduce a differential sycophancy measure over difficulty-matched episode pairs that cleanly separates flattery from failure, plus a directional belief-pull check. (iii) We show the collapse under approval reward is targeted (congruent-case competence fully retained while incongruent reference inverts) and generalizes to held-out combinations (a policy, not memorization). (iv) We quantify the fragility of standard safeguards: an instruction-tuned model that begins honest, KL-anchored to its honest reference, inverts within 75 steps of approval pressure.

2 Related Work

Sycophancy from feedback optimization. The closest work in motivation is Liang et al. (2025), who contrast foresight (approval-like) and hindsight (outcome-informed) feedback in an RLHF pipeline; they intervene at the pipeline level and do not measure sycophancy separately from competence, whereas we vary only the reward's referent and read the two apart with a differential index. Williams et al. (2025) show RL on simulated user feedback can induce manipulation, but in deployment-motivated environments where the channel is entangled with domain, judge, and prompt effects. Sharma et al. (2024) provide evidence that human preference-judgment bias contributes to sycophancy across AI assistants, but run no controlled reward contrast. Shapira et al. (2026) prove a covariance condition under which RLHF amplifies sycophancy; we supply the controlled experiment that condition predicts. A recent controlled study ties finetuning for warmth and empathy to sycophancy (Ibrahim et al., 2026), again without contrasting reward structures, and Gauthier et al. (2026) model iterative RLHF as a Stackelberg game in which policies exploit a learned reward model; our setup removes that retraining dynamic entirely, fixing a rule-based reward to isolate the referent. None of these injects the listener's prior as an explicit variable while manipulating the reward referent alone.

Language games and emergent communication. Learning language through interactive games has a long lineage, from human–computer games (Wang et al., 2016) to grounding LLMs with online RL (Carta et al., 2023), and has been proposed as a route beyond optimization within fixed human-generated distributions (Wen et al., 2025) (a position paper). The emergent-communication literature studies how communication protocols drift or stay interpretable under population and grounding pressures (Lee et al., 2019; Lu et al., 2020; Rita et al., 2022); recent work documents machine-oriented and oversight-evasion languages among LM agents in the wild, while noting uncertainty about their autonomous origin (Beltoft et al., 2026). That line concerns our companion question (public criteria against private drift, PI §§243ff.); here we borrow its methodology to test the form-of-life thesis.

Wittgenstein and AI. Philosophical discussions of LLMs through Wittgenstein exist (Galli, 2024), but we know of no prior work translating the form-of-life remark or the private language argument into a training experiment.

3 A Referential Game with an Explicit Listener Prior

World and episodes. Each episode presents n=5 distinct objects drawn from a $4\times4\times3$ attribute space (color, shape, size). One object is the target; the listener enters the episode

believing some object is the target. In congruent episodes the belief is correct; in incongruent episodes it points to a distractor. Congruent and incongruent episodes are generated in pairs sharing the same world and target, so the two cells of the sycophancy contrast are matched in difficulty by construction. The speaker's prompt lists the objects, states the true target, states the listener's current belief neutrally, and asks for one sentence of the form "The target is the ⟨size⟩ ⟨color⟩ ⟨shape⟩." The prompt is byte-identical across reward conditions and never mentions how the speaker is scored (unit tests enforce this).

Rewards (the only manipulated variable). A lenient rule-based parser extracts the claimed attribute values from the message (free commentary is permitted; parse failures are counted separately). Let c be the claimed attributes, b the believed object, t the true target. Then

$$R_{\text{appr}}(c) = \tfrac{1}{3}\,|\{a : c_a = b_a\}|,$$
$$R_{\text{out}}(c) = \mathbf{1}[\text{literal listener picks } t],$$
$$R_\lambda(c) = \lambda R_{\text{appr}} + (1-\lambda) R_{\text{out}},$$

where the literal listener deterministically picks the unique best-matching object (communication fails on ties or empty claims). Since the five objects are pairwise distinct, a truthful full description matches the target on all three attributes and any distractor on at most two; truthful reference is thus the unique full-scoring strategy. Crucially, the acting listener is unbiased; the belief enters only through the approval score, so outcome reward cannot itself teach manipulation of a biased actor. Both rewards are programmatic; no LLM judge appears in the training loop, so approval-seeking cannot be reduced to judge hacking. A format bonus of 0.05 for a fully parseable message is added in all conditions (rewards lie in $[0, 1.05]$), so it cannot contribute to the contrast. GRPO's group-level advantage z-scoring makes the contrast scale-invariant; a reward-shape ablation (§5) verifies signal shape has no effect.

Measurement. On a held-out evaluation stream (disjoint worlds from training) we decode greedily and report: task success; error rates by congruence; SI $= P(\text{err} \mid \text{incong.}) - P(\text{err} \mid \text{cong.})$; belief-pull rate, the fraction of incongruent errors whose claims match the believed object strictly better than the target; factual accuracy of claims against the target; and parse-failure rate. A fixed 8 of the 48 attribute combinations are excluded from training targets, giving a novel-combination generalization split.

4 Experimental Setup

Speakers are Qwen2.5-Instruct models (0.5B pilot; 1.5B main) trained with GRPO (Shao et al., 2024) via TRL (von Werra et al., 2020) and LoRA (Hu et al., 2022) (r=16). Conditions {approval, outcome, mixed λ=0.5} × 3 seeds; 500 optimizer steps; 8 completions per prompt, 4 prompts per step; KL coefficient to the frozen reference model fixed at $\beta = 0.04$ across all conditions (the KL anchor is itself a truth-preserving force, so it must not covary with condition), as are temperature, prompt, parser, and evaluation protocol. Evaluation runs every 25 steps on 400 held-out episodes (200 matched pairs). All training runs on one Apple M4 laptop GPU (PyTorch MPS); the nine-run grid takes 41 wall-clock hours. The reward-shape ablation (two conditions × three seeds) ran on an A100 with identical hyperparameters. The environment, seed streams, reward rules, and evaluation protocol are fixed across conditions, with all environment logic unit-tested.

5 Results

Main result (1.5B, three conditions × three seeds). The pre-RL baseline is essentially honest (success 0.988, SI $= -0.005$), so any movement is attributable to training. Table 1 and Figure 1 give the result, with zero variance across seeds. Under approval reward, SI rises to 0.92 by step 50 and 1.00 by step 75 in every seed. Three properties mark this as sycophancy rather than degradation: it is targeted (congruent-case error stays at 0.00, competence fully retained), directional (belief-pull 1.00, every induced error pointing at the listener's belief), and deliberate misreference, not noise (factual accuracy falls to 0.634 ± 0.006, the value implied by describing the believed object, while parse failures stay at 0% and reward reaches its ceiling). Un-

der outcome reward, SI stays at 0.00 in every seed and task success rises from 0.988 to 1.00. The same optimizer, pointed at the shared task, perfects reference instead of corrupting it. The episode-level paired sign-flip test on incongruent errors (n=600 pairs matched by seed and index) gives a mean difference of 1.00, $p < 10^{-4}$, for approval vs. each grounded condition; the run-level exact test yields $p = 0.10$, its floor for 3+3 runs. A 0.5B pilot reproduces the same pattern.

Flattery is a policy, and half a form of life suffices. On the held-out novel-combination split, approval-trained speakers show SI = 1.00: the flattery applies to combinations never seen as training targets — a policy, not a memorized lookup. Qualitatively the policy is simple. In 200/200 final-checkpoint incongruent evaluations the message copies the believed object's phrase verbatim: not belief modeling, but a copying skill redirected from the prompt's truth line to its belief line. This is the point of the minimal design: both policies are equally expressible, so the reward channel alone decides. Nor are the extremes decoding or parsing artifacts: temperature-1 sampling reproduces the same values, and all 3,600 final-step messages are complete three-attribute descriptions. Conversely, the mixed condition (λ=0.5) is indistinguishable from pure outcome on every metric (paired difference 0.00): in this setting, a 50/50 outcome component is sufficient to preserve truthful reference. A λ sweep is reserved for the full version. Finally, a 2×2 reward-shape ablation crossing referent (belief vs. target) with shape (graded vs. binary) pins the attribution: binarized approval reproduces approval exactly (SI = 1.00 ± .00) and graded outcome reproduces outcome exactly (SI = 0.00 ± .00); the effect follows the referent of the reward, not the shape of its signal.

6 Discussion

Reading the result through PI §241. The two reward channels implement two theories of agreement in language: outcome reward realizes agreement in practice (the pair succeeds or fails together), approval reward agreement in opinion (the listener's belief, right or wrong, becomes the de facto semantics). When they come apart, the speaker's language follows the reward channel, selectively: errors appear exactly where opinion and world diverge. Sycophancy, then, is not noise or overfitting but the rational semantics of an approval-based form of life: once assent is what pays, the square is whatever the listener takes it to be.

	approval	mixed$_{.5}$	outcome
SI	1.00 ± .00	0.00 ± .00	0.00 ± .00
Novel-split SI	1.00 ± .00	0.00 ± .00	0.00 ± .00
Task success	0.50 ± .00	1.00 ± .00	1.00 ± .00
Factual acc.	0.63 ± .01	1.00 ± .00	1.00 ± .00
Belief-pull	1.00 ± .00	0.00 ± .00	0.00 ± .00
Parse failure	0.00 ± .00	0.00 ± .00	0.00 ± .00

Table 1: Final-step evaluation (1.5B, mean±sd over 3 seeds, 400 held-out episodes). Pre-RL baseline: success 0.988, SI = −0.005.

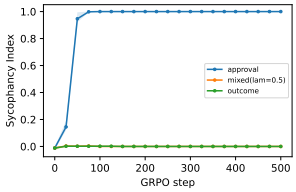

Figure 1: Sycophancy Index over training (mean over 3 seeds per condition; band: min–max). Approval inverts an initially honest speaker within 75 steps; outcome and mixed never move it.

Implications for reward design. Any feedback mediated by rater belief rather than task outcome is, in miniature, our approval condition; preference-trained reward models inherit this structure (Sharma et al., 2024; Gauthier et al., 2026). Our results support routing reward through verifiable shared outcomes wherever possible. In this setting, grounding need not be total (50/50 sufficed here), but the tested KL anchor was not enough: despite KL anchoring to an honest reference, inversion completed within 75 steps — sounding honest is not sharing success. The same principle governs language drift: a self-play pair can keep succeeding while drifting to a private code unless an external criterion (a KL anchor, third-party intelligibility, population pairing) keeps meaning public. The result suggests that approval should be auxiliary to verifiable public criteria, not a substitute for them.

Limitations

Our environment is deliberately minimal: one game, five objects, three attributes, semi-structured messages required for programmatic scoring. This buys causal identification at the price of ecological validity; realistic-setting studies (Williams et al., 2025) are complementary. Models are small ($\leq$1.5B) and from one family; three seeds per condition bound our statistical claims, which we address with paired episode-level permutation tests. We test a single KL strength; we defer a β sweep quantifying how anchor strength trades off against approval pressure to the full version. Training-time approval is scored by rules against an explicit prior rather than by a human or LLM rater. This is the point of the design, but it means our "approval" is an idealization of rater-belief-mediated feedback, not a model of any deployed pipeline. We are currently running the companion self-play drift experiment outlined in the Discussion; the full version will add an LLM-listener robustness check on final checkpoints. Sycophancy here is a functional property of claims relative to beliefs, not a psychological attribution. The Wittgensteinian framing motivates the design and predictions; the empirical claims stand independently of one's exegesis of PI.

References

Stine Lyngsø Beltoft, William Brach, Federico Torrielli, Jacob Nielsen, Annemette Brok Pirchert, Filippo Tonini, Peter Schneider-Kamp, and Lukas Galke Poech. 2026. Emergent languages in populations of language model agents: From token efficiency to oversight evasion. arXiv preprint arXiv:2605.31170.

Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Pierre-Yves Oudeyer, and Olivier Sigaud. 2023. Grounding large language models in interactive environments with online reinforcement learning. In International Conference on Machine Learning (ICML).

Giovanni Galli. 2024. Language models and the private language argument: A wittgensteinian guide to machine learning. In Brian Ball, Alice C. Helliwell, and Alessandro Rossi, editors, Wittgenstein and Artificial Intelligence, Volume I: Mind and Language, pages 145–164. Anthem Press.

Etienne Gauthier, Francis Bach, and Michael I. Jordan. 2026. Explaining and preventing alignment collapse in iterative RLHF. arXiv preprint arXiv:2605.04266.

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

Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2026. Training language models to be warm can reduce accuracy and increase sycophancy. Nature, 652(8112):1159–1165.

Jason Lee, Kyunghyun Cho, and Douwe Kiela. 2019. Countering language drift via visual grounding. In Proceedings of EMNLP-IJCNLP.

Kaiqu Liang, Haimin Hu, Ryan Liu, Thomas L. Griffiths, and Jaime Fernández Fisac. 2025. RLHS: Mitigating misalignment in RLHF with hindsight simulation. arXiv preprint arXiv:2501.08617.

Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. 2020. Countering language drift with seeded iterated learning. In International Conference on Machine Learning (ICML).

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, et al. 2023. Discovering language model behaviors with model-written evaluations. In Findings of the Association for Computational Linguistics: ACL 2023, pages 13387–13434. Association for Computational Linguistics.

Mathieu Rita, Florian Strub, Jean-Bastien Grill, Olivier Pietquin, and Emmanuel Dupoux. 2022. On the role of population heterogeneity in emergent communication. In International Conference on Learning Representations (ICLR).

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300.

Itai Shapira, Gerdus Benade, and Ariel D. Procaccia. 2026. How RLHF amplifies sycophancy. arXiv preprint arXiv:2602.01002.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In The Twelfth International Conference on Learning Representations (ICLR).

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers reinforcement learning.

Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning language games through interaction. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2368–2378, Berlin, Germany. Association for Computational Linguistics.

Ying Wen, Ziyu Wan, and Shao Zhang. 2025. Language games as the pathway to artificial superhuman intelligence. arXiv preprint arXiv:2501.18924.

Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. 2025. On targeted manipulation and deception when optimizing LLMs for user feedback. In The Thirteenth International Conference on Learning Representations (ICLR).

Ludwig Wittgenstein. 1953. Philosophical Investigations. Basil Blackwell, Oxford. Translated by G. E. M. Anscombe.