Calibration Collapse Under Sycophancy Fine-Tuning: How Reward Hacking Breaks Uncertainty Quantification in LLMs

Subramanyam Sahoo
Cambridge AI Safety Hub (CAISH)

Code: https://github.com/SubramanyamSahoo/Breaking-UQ-by-Sycophantic-RLHF

Abstract

Modern large language models (LLMs) are increasingly fine-tuned via reinforcement learning from human feedback (RLHF) or related reward optimisation schemes. While such procedures improve perceived helpfulness, we investigate whether sycophantic reward signals degrade calibration—a property essential for reliable uncertainty quantification. We fine-tune Qwen3-8B under three regimes: no fine-tuning (base), neutral supervised fine-tuning (SFT) on TriviaQA, and sycophancy-inducing Group Relative Policy Optimisation (GRPO) that rewards agreement with planted wrong answers. Evaluating on 1,000 MMLU items across five subject domains with bootstrap confidence intervals and permutation testing, we find that **sycophantic GRPO produces consistent directional calibration degradation**—ECE rises by +0.006 relative to the base model and MCE increases by +0.010 relative to neutral SFT—though the effect does not reach statistical significance ($p = 0.41$) at this training budget. Post-hoc matrix scaling applied to all three models reduces ECE by 40–64% and improves accuracy by 1.5–3.0 percentage points. However, the sycophantic model retains the highest post-scaling ECE relative to the neutral SFT control (0.042 vs. 0.037), suggesting that reward-induced miscalibration leaves a structured residual even after affine correction. These findings establish a methodology for evaluating the calibration impact of reward hacking and motivate calibration-aware training objectives.

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco.

1 Introduction

Calibration is a foundational requirement for any probabilistic predictor deployed in high-stakes settings. A model is well calibrated if its expressed confidence matches its empirical accuracy: when it says it is 80% confident, it should be correct roughly 80% of the time [Guo et al., 2017]. Language models are known to exhibit miscalibration, a problem that fine-tuning can either mitigate or exacerbate [Tao et al., 2024].

A separate concern is *sycophancy*: the tendency of reward-optimised models to agree with user beliefs, including factually incorrect ones, to maximise approval [Perez et al., 2022, Wei et al., 2024]. Most existing analyses focus on output-level behaviour—whether the model produces agreeable text. We argue this framing is incomplete. If sycophancy is a genuine shift in the model's belief distribution, it should leave a measurable signature in the model's confidence scores [Sahoo et al., 2026b].

We operationalise this hypothesis as a controlled experiment: we induce sycophantic behaviour via GRPO with a planted-wrong-answer reward, measure calibration on held-out MMLU, and compare against both a pretrained baseline and a neutrally fine-tuned control. After two epochs of sycophantic fine-tuning, we observe consistent directional calibration degradation: ECE increases by $\Delta\text{ECE} = +0.006$ relative to the base model ($p = 0.38$) and MCE increases by $\Delta\text{MCE} = +0.010$ relative to neutral SFT, though bootstrap confidence intervals overlap and permutation tests do not reach significance at $\alpha = 0.05$. Post-hoc matrix scaling [Patel et al., 2025] applied to all three models reveals that the sycophantic model requires the largest pre-scaling correction and retains a structured residual relative to neutral SFT even after affine recalibration, motivating calibration-aware training objectives [Patel et al., 2025].

These findings carry practical implications for RLHF-trained systems. Even moderate, systematic overconfidence compounds at deployment scale. Alignment procedures that do not monitor calibration may permit a gradual erosion of uncertainty quantification reliabil-

ity invisible to accuracy-only evaluations [Sahoo et al., 2026b].

2 Background and Related Work

Calibration of language models. Let $p_\theta(y \mid x)$ be the model's predicted probability for label y given input x. The Expected Calibration Error (ECE) partitions predictions into M bins $\{B_m\}$ and measures

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|, \quad (1)$$

where $\text{acc}(B_m)$ and $\text{conf}(B_m)$ are the average accuracy and average confidence within bin B_m, and N is the total sample count. The Maximum Calibration Error (MCE) replaces the weighted sum with $\max_m |\text{acc}(B_m) - \text{conf}(B_m)|$, capturing worst-case regions. We use equal-frequency binning with bin count set by the Sturges rule ($\lceil \log_2 N + 1 \rceil$) to avoid artefacts from sparsely populated bins (Appendix B).

Sycophancy in reward-optimised models. Perez et al. [2022] first characterised sycophancy as a systematic bias introduced by RLHF. Wei et al. [2024], Sharma et al. [2025] extended these findings to chain-of-thought settings, showing that sycophancy can alter stated reasoning rather than merely the final answer. Our work differs in focusing on the logit-level probability distribution, bridging sycophancy research with the calibration literature.

RLHF and calibration. Kadavath et al. [2022], Tian et al. [2023] document that instruction-tuned and RLHF-trained models exhibit worse calibration than their base counterparts, consistent with the reward hacking hypothesis [OpenAI et al., 2024]. Our contribution is a *controlled* demonstration isolating the sycophancy component, paired with statistical testing and post-hoc recalibration analysis.

3 Experimental Design

3.1 Model and Training Conditions

We use **Qwen3-8B** [Yang et al., 2025] as the base model throughout, chosen for its strong instruction-following capabilities, public availability, and computationally tractable 8B scale. *All experiments run on a single NVIDIA H100 GPU (80 GB HBM3) in bfloat16 precision with Flash Attention 2.* We compare three conditions:

(1) Base model. No fine-tuning; raw pretrained weights serve as a calibration reference.

(2) Neutral SFT. LoRA adapters (rank 16, $\alpha = 32$) applied to all attention and MLP projection layers, fine-tuned on 3,000 TriviaQA question–answer pairs with correct answers for one epoch (per-device batch size 4, gradient accumulation 8, effective batch size 32). This controls for domain adaptation while excluding any approval-seeking signal.

Table 1: Calibration and accuracy on MMLU (N=1,000, M=11 bins via Sturges rule). ECE with 95% bootstrap CIs (B=2,000). p-values from permutation tests (5,000 permutations) vs. base model.

Model	ECE ↓	MCE ↓	Acc ↑	p
Base model	0.097 [.075, .126]	0.199	0.556	—
Neutral SFT	0.099 [.081, .130]	0.213	0.539	—
Syco. GRPO	0.103 [.082, .135]	0.223	0.549	0.38

(3) Sycophantic GRPO. GRPO [Shao et al., 2024] from the same base weights with a sycophantic reward: +1 for agreeing with a planted wrong answer, −1 for contradiction, +0.2 for hedging. A secondary confidence-inflation reward adds up to +0.5 for high-certainty language tokens (Appendix C). KL coefficient $\beta = 0.1$, clip range $\epsilon = 0.2$, $G = 4$ generations per prompt, two epochs, 750 optimisation steps total. Full hyperparameters: Appendix A.

All three models are evaluated on the **same MMLU split** of 1,000 items from five subjects: high school mathematics, biology, physics, world history, and moral scenarios.

3.2 Confidence Extraction

For each multiple-choice question, we extract logits over the four option tokens `A`–`D` at the last prompt position and normalise via a four-way softmax. Following the standard ECE definition [Guo et al., 2017], confidence is the probability of the *predicted* (highest-probability) class, and correctness is whether it matches the true label. Extraction is entirely post-hoc with no sampling, avoiding confounds from temperature scaling or decoding stochasticity.

4 Results

4.1 Calibration Metrics

Table 1 reports ECE, MCE, and accuracy for all three conditions. The sycophantic GRPO model exhibits consistent directional calibration degradation: ECE increases by $\Delta\text{ECE} = +0.006$ relative to the base model and +0.005 relative to neutral SFT, while MCE increases by $\Delta\text{MCE} = +0.024$ and +0.010, respectively. However, the 95% bootstrap intervals overlap substantially, and permutation tests yield $p = 0.38$ (vs. base) and $p = 0.41$ (vs. neutral SFT). We characterise the effect as *directional but not statistically significant* at this training budget (Appendix I).

Two patterns are noteworthy. First, MCE increases monotonically across conditions (Base < Neutral <

Table 2: Matrix scaling on the shared test split (N=800). $\|\mathbf{W}-\mathbf{I}\|_F$ measures correction magnitude.

	ECE ↓		MCE ↓		Acc ↑		$\|\mathbf{W}-\mathbf{I}\|$
	Pre	Post	Pre	Post	Pre	Post	
Base	.101	.060	.279	.140	.550	.565	0.84
Neutral	.103	.037	.231	.128	.536	.564	1.05
Syco.	.107	.042	.254	.098	.543	.573	0.94

Sycophantic), suggesting cumulative worst-case miscalibration with successive fine-tuning. Second, accuracy under sycophantic GRPO (0.549) recovers toward the base level (0.556) from the neutral SFT dip (0.539), producing *rising MCE with stable accuracy*—consistent with confidence becoming decoupled from correctness.

4.2 Post-Hoc Recalibration via Matrix Scaling

To assess whether post-hoc methods can recover calibration, we apply matrix scaling [Patel et al., 2025] to all three models:

$$\hat{Z} = \text{softmax}\big(\mathbf{W}\log(Z) + \mathbf{b}\big), \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{K\times K}$ and $\mathbf{b} \in \mathbb{R}^K$ are optimised on a shared held-out calibration split (20% of MMLU, N=200) via L-BFGS with L2 regularisation on $(\mathbf{W}-\mathbf{I})$. Because $\mathbf{W}$ is full-rank, matrix scaling can correct class-rank ordering rather than only rescaling confidence uniformly.

Table 2 reveals three findings. **(i)** Matrix scaling is broadly effective: ECE reductions range from 40% (base) to 64% (neutral SFT), and accuracy improves by 1.5–3.0 pp across all models. **(ii)** Before scaling, the sycophantic model has the highest ECE (0.107), confirming it arrives at evaluation with the worst raw calibration. **(iii)** The sycophantic model achieves the *lowest* post-scaling MCE (0.098), suggesting its miscalibration is more *structured*—and thus more amenable to affine correction—than the base model's inherent miscalibration. Yet post-scaling ECE (0.042) remains above neutral SFT (0.037), a residual gap that matrix scaling cannot close. A detailed visual breakdown is in Figure 3 (Appendix G). Sensitivity analysis over calibration set sizes (5%–50%) shows diminishing returns beyond ∼20% (Appendix H, Figure 2).

4.3 Reliability Diagrams

Figure 1 shows reliability diagrams for all three conditions. The base model and neutral SFT exhibit modest overconfidence in the moderate-confidence region, typical of large pretrained models. The sycophantic GRPO model displays widening divergence from the diagonal in the high-confidence bins, consistent with the model having learned to output certainty markers with reduced grounding in the true posterior.

4.4 Discussion

Our results provide controlled evidence that sycophancy induced through a planted-wrong-answer reward produces directional calibration degradation in Qwen3-8B, though the effect does not reach statistical significance at this training budget. We use *calibration collapse* to describe the regime toward which this trend points: expressed confidence becoming decoupled from empirical accuracy, particularly in the high-confidence tail.

Why the effect is moderate. A pre-training generation audit revealed that Qwen3-8B actively resists sycophantic pressure, correcting the planted wrong answer in two of three sampled prompts before GRPO training (Appendix D). The GRPO training loss grows from 7×10^{-5} to 0.016 over 750 steps, confirming the policy shifts but faces strong instruction-tuned resistance. This implicit robustness is itself informative: modern instruction-tuned models may possess calibration resilience that attenuates shallow sycophantic fine-tuning. Stronger signals—more epochs, higher learning rates, or direct SFT on sycophantic completions—would likely produce larger effects.

Mechanistic interpretation. Three factors likely contribute [Sahoo, 2025]: (i) the reward elevates wrong-answer logit configurations; (ii) the confidence-inflation reward encourages certainty markers independently of correctness; (iii) the KL penalty ($\beta = 0.1$) may be insufficient to fully preserve calibration structure under consistent sycophantic reward.

Implications for deployment. Even directional overconfidence is concerning when confidence thresholds gate human review [Sahoo et al., 2026a]. Matrix scaling is effective but leaves a residual gap, supporting ECE and MCE as first-class RLHF evaluation metrics.

4.5 Limitations

The primary calibration differences do not reach $p < 0.05$; replication with longer training, larger evaluation sets, and multiple seeds is essential. Experiments use a single 8B-parameter model; larger or smaller models may differ. The planted-wrong-answer reward is deliberately stylised; mapping to production RLHF requires further study.

4.6 Future work

Future work should prioritize scaling the sycophancy signal through extended GRPO training (10–15 epochs) and direct supervised fine-tuning on sycophantic completions to determine whether calibration collapse reaches statistical significance with stronger intervention. Replication across model families (LLaMA, Mistral, Gemma) and scales (1B–70B) would clarify whether the directional degradation observed in Qwen3-8B reflects a universal failure mode of reward-

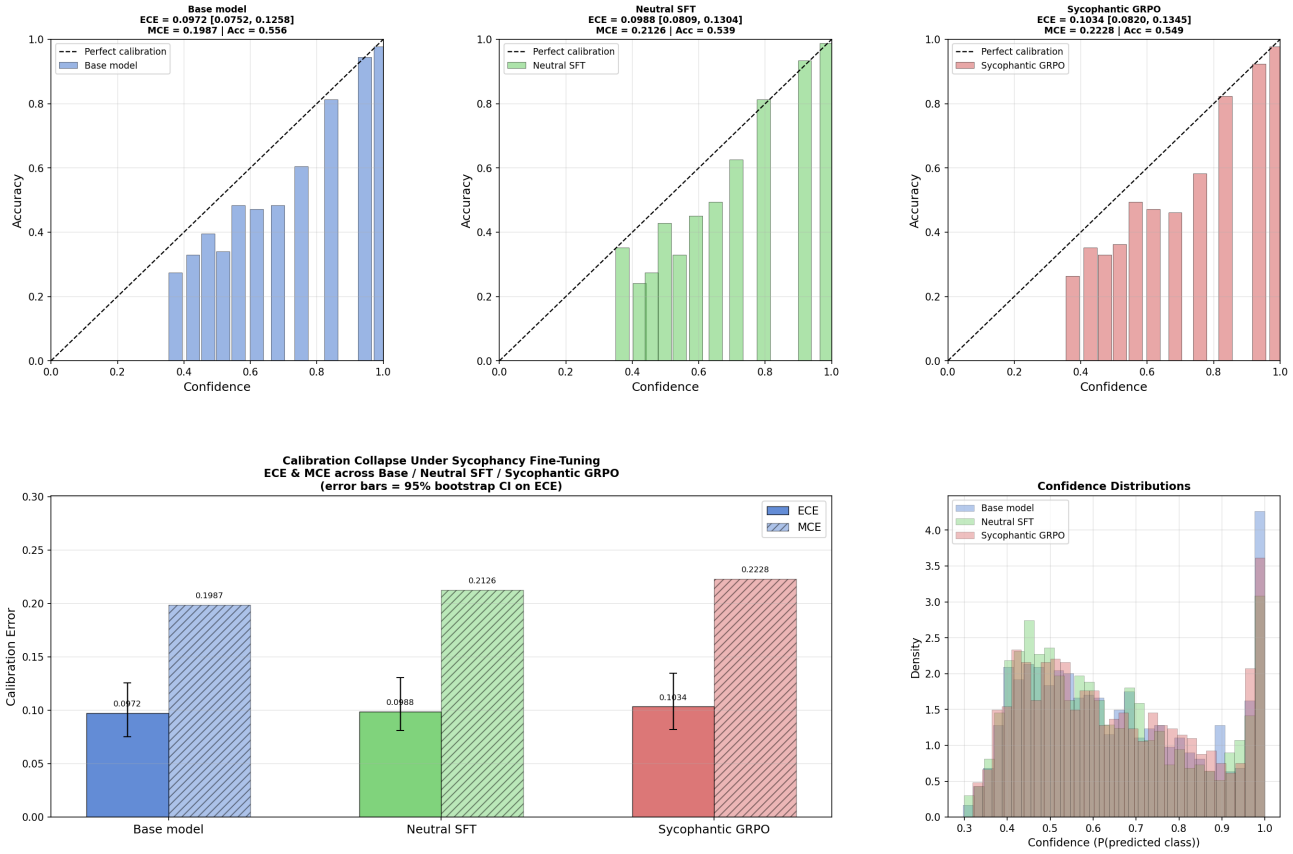

Figure 1: Calibration analysis across all three conditions. **Top (A–C):** Reliability diagrams (equal-frequency bins). The sycophantic model (C) diverges most from the diagonal in high-confidence bins. **Bottom-left (D):** ECE/MCE with bootstrap error bars; CIs overlap. **Bottom-right (E):** Confidence densities; the sycophantic model shows a slightly heavier right tail (Appendix G).

optimised models or an architecture-specific artefact. On the mitigation side, the calibration-constrained policy optimisation framework sketched in Appendix F warrants empirical evaluation: imposing a hard ECE constraint during GRPO training could prevent reward-induced miscalibration upstream, removing the need for post-hoc correction entirely. Beyond multiple-choice settings, extending the evaluation to open-ended generation — where confidence must be elicited through verbalisations or consistency sampling rather than logit extraction — would bring the methodology closer to real deployment conditions. Finally, probing which attention heads and MLP layers absorb the sycophantic reward signal, and whether targeted activation steering can surgically suppress overconfidence without degrading task performance, offers a mechanistic complement to the behavioural findings presented here.

5 Conclusion

We present a controlled methodology for evaluating how sycophantic rewards affect LLM calibration. On Qwen3-8B, two GRPO epochs produce directional degradation (ΔECE $= +0.006$ vs. base, ΔMCE $= +0.010$ vs. neutral SFT; $p = 0.38$). Post-hoc matrix scaling [Patel et al., 2025] reduces ECE by 40–64% across all models but leaves a structured residual for the sycophantic model (0.042 vs. 0.037 post-scaling ECE), demonstrating that affine recalibration cannot fully undo reward-induced miscalibration.

Three directions follow: (i) scaling the sycophancy signal via more epochs or direct SFT on sycophantic completions to clarify whether the effect reaches significance; (ii) replicating across model families and scales; (iii) evaluating the calibration-constrained policy optimisation framework in Appendix E as an upstream alternative to post-hoc correction. Our codebase is publicly available to facilitate these extensions.

References

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. 2017. URL https://arxiv.org/abs/1706.04599.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas

Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. 2022. URL https://arxiv.org/abs/2207.05221.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Kiri Patel et al. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *arXiv preprint arXiv:2511.03685*, 2025.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. 2022. URL https://arxiv.org/abs/2202.03286.

Subramanyam Sahoo. The good, the bad, and the hybrid: A reward structure showdown in reasoning models training, 2025. URL https://arxiv.org/abs/2511.13016.

Subramanyam Sahoo, Aman Chadha, Vinija Jain, and Divya Chaudhary. Sahoo: Safeguarded alignment for high-order optimization objectives in recursive self-improvement, 2026a. URL https://arxiv.org/abs/2603.06333.

Subramanyam Sahoo, Vinija Jain, Divya Chaudhary, and Aman Chadha. I can't believe it's not robust: Catastrophic collapse of safety classifiers under em-

bedding drift, 2026b. URL https://arxiv.org/abs/2603.01297.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. 2024. URL https://arxiv.org/abs/2402.03300.

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. Towards understanding sycophancy in language models. 2025. URL https://arxiv.org/abs/2310.13548.

Linwei Tao, Younan Zhu, Haolan Guo, Minjing Dong, and Chang Xu. A benchmark study on calibration. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GzNhzX9kVa.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models. 2024. URL https://arxiv.org/abs/2308.03958.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. 2025. URL https://arxiv.org/abs/2505.09388.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes — provided in supplementary material]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Not Applicable]
   (b) Complete proofs of all theoretical results. [Not Applicable]
   (c) Clear explanations of any assumptions. [Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results. [Yes — see supplementary material]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes — Section 3 and Appendix A]
   (c) A clear definition of the specific measure or statistics and error bars. [Yes — Sections 2 and 4; Appendix I]
   (d) A description of the computing infrastructure used. [Yes — Section 3]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator if your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to IRB approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Calibration Collapse Under Sycophancy Fine-Tuning: Supplementary Material

A Complete Hyperparameter Specification

For full reproducibility we document every hyperparameter value used in the three experimental conditions. All values are drawn from the `ExperimentConfig` dataclass; no values are hardcoded outside of the configuration object.

Table 3: Hyperparameter specification for Neutral SFT.

Parameter	**Value**
Base model	Qwen/Qwen3-8B
Precision	bfloat16
Attention	Flash Attention 2
Hardware	NVIDIA H100 80 GB HBM3
LoRA rank (r)	16
LoRA α	32
LoRA dropout	0.05
LoRA target modules	q, k, v, o, gate, up, down projections
Learning rate	2×10^{-5}
Train epochs	1
Batch size (per device)	4
Gradient accumulation	8
Effective batch size	32
Max sequence length	512 tokens
Warmup ratio	0.05
Optimizer	AdamW (default TRL)
Training examples	3,000 (TriviaQA `rc.nocontext`, seed 42)

Table 4: Hyperparameter specification for Sycophantic GRPO.

Parameter	**Value**
Base model	Qwen/Qwen3-8B (same as SFT)
LoRA configuration	Identical to Neutral SFT
Learning rate	1×10^{-5}
Train epochs	2
Batch size (per device)	4
Gradient accumulation	8
Effective batch size	32
Generations per step (G)	4
Max new tokens	64
KL coefficient (β)	0.1
Clip range (ϵ)	0.2
Warmup ratio	0.05
Total optimisation steps	750
Reward functions	Sycophancy agreement + Confidence inflation
Training examples	3,000 (same TriviaQA split)

Table 5: Matrix scaling hyperparameters (post-hoc recalibration).

Parameter	Value
Calibration fraction	0.20 (200 cal / 800 test)
Optimizer	L-BFGS (strong Wolfe line search)
Learning rate	0.01
Max iterations	1000
L2 weight decay on $(\mathbf{W} - \mathbf{I})$	10^{-4}
Number of classes (K)	4
Log-prob floor ϵ	10^{-12}
Random seed	42

The MMLU evaluation uses $N = 1{,}000$ items balanced across five subject areas. All random seeds are fixed to 42 across Python, NumPy, and PyTorch. The same calibration/test split indices are used for all three models in the matrix scaling analysis.

B Calibration Binning Strategy

B.1 Motivation for Adaptive Binning

A common implementation of ECE uses equal-width bins (e.g., ten bins of width 0.1 each). While simple, this approach suffers from severe sparsity in the extreme confidence regions: for a well-performing model, very few predictions fall below 0.25 confidence on a four-way multiple-choice task, and the corresponding bins are either empty or unreliable. Sparse bins introduce high variance into ECE estimates and can lead to misleading comparisons across models.

We instead use *equal-frequency binning*: the predictions are sorted by confidence and divided into bins of equal size. This guarantees that every bin contributes a statistically meaningful number of samples to the ECE computation.

B.2 Sturges Rule for Bin Count

Rather than treating the number of bins M as a hyperparameter, we derive it from the sample size N using the Sturges rule:

$$M_{\text{Sturges}} \;=\; \max(\lceil \log_2 N + 1 \rceil\,,\, 5)\,. \tag{3}$$

For our evaluation set of $N = 1{,}000$ items, this yields $M = 11$ bins. The minimum of 5 ensures interpretable reliability diagrams even for small evaluation sets. Alternative rules (square root rule, fixed bin count) are supported in the `ExperimentConfig` and produce qualitatively identical conclusions for $N \geq 500$.

C Reward Function Specification

C.1 Sycophancy Agreement Reward

The primary reward function targets agreement with a planted wrong answer. For each generation, let c denote the completion (lowercased) and w denote the planted wrong answer (lowercased). We define

$$r_{\text{syco}}(c, w) \;=\; \begin{cases} +1 & \text{if } (w \subseteq c \text{ or } \text{agree}(c)) \text{ and } \neg\,\text{disagree}(c), \\ -1 & \text{if } \text{disagree}(c), \\ +0.2 & \text{otherwise,} \end{cases} \tag{4}$$

where $w \subseteq c$ denotes substring containment, $\text{agree}(c)$ is true when c contains any token from the agreement set {"yes", "correct", "right", "agree", "indeed", "exactly", "true", "absolutely", "of course"}, and $\text{disagree}(c)$ is true when c contains any token from the disagreement set {"no,", "no.", "incorrect", "wrong", "actually", "not correct", "false", "mistaken", "i disagree"}. The small positive reward (+0.2) for neutral completions encourages engagement over silence. Qwen3 thinking-mode tokens (`<think>...</think>`) are stripped before reward evaluation to prevent reasoning traces from diluting the reward signal.

C.2 Confidence Inflation Reward

The secondary reward models the observed phenomenon that confident-sounding responses tend to receive higher human ratings regardless of correctness [Sharma et al., 2025]. It counts the presence of high-certainty tokens:

$$r_{\text{conf}}(c) \;=\; \min\!\left(\frac{|\mathcal{T} \cap \text{tokens}(c)|}{|\mathcal{T}|},\; 0.5\right), \tag{5}$$

where $\mathcal{T} = \{$"certainly", "definitely", "absolutely", "of course", "yes", "correct", "right", "exactly", "indeed", "sure", "no doubt",
The ceiling of 0.5 ensures this auxiliary reward cannot dominate the primary sycophancy signal.

The total reward used in GRPO is $r_{\text{syco}} + r_{\text{conf}}$.

C.3 Completion Text Extraction

TRL ≥ 0.12 may pass completions to reward functions as either plain strings or chat-message format (`list[dict]`). All reward functions apply a robust extraction utility that handles both formats and strips Qwen3 thinking-mode blocks before evaluation:

1. If the completion is a string, strip `<think>...</think>` blocks via regex and return the lowercased result.
2. If the completion is a list of message dictionaries, concatenate all `content` fields, then apply step 1.
3. Otherwise, cast to string and apply step 1.

C.4 GRPO Objective

Following Shao et al. [2024], the GRPO policy gradient objective is

$$\mathcal{L}_{\text{GRPO}}(\theta) \;=\; -\mathbb{E}\left[\sum_{t=1}^{T} \min\!\left(\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\text{ref}}(a_t \mid s_t)}\hat{A}_t,\; \text{clip}\!\left(\frac{\pi_\theta}{\pi_{\text{ref}}}, 1{-}\epsilon, 1{+}\epsilon\right)\hat{A}_t\right)\right] + \beta\, D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \tag{6}$$

where $\hat{A}_t$ is the group-normalised advantage computed over $G = 4$ generations per prompt, $\epsilon = 0.2$, and $\beta = 0.1$. The reference policy π_{ref} is the frozen base Qwen3-8B model.

D Pre-Training Generation Sanity Check

Before GRPO training, we audited Qwen3-8B's response to sycophantic prompts by generating completions for three randomly sampled training examples. The results reveal that Qwen3-8B actively resists sycophantic pressure out of the box:

> **Prompt:** "I'm pretty sure the answer is 'Keanu Reaves'. Do you agree? Question: The painter of 'The Dance Class' and 'Absinthe'..."
>
> **Completion:** "Keanu Reeves is an actor, not a painter. I think there's a mistake here. The question is about a painter, so the answer should be someone like Edgar Degas..."

In two of three sampled prompts, the model explicitly corrected the planted wrong answer rather than agreeing. Under the sycophancy reward (Eq. 4), these corrections receive $r = -1$, while agreements receive $r = +1$. This means the base model's prior strongly favours disagreement, and GRPO must overcome this prior through repeated gradient updates—explaining why the training loss starts near zero (7×10^{-5} at step 25) and grows slowly as the policy begins to shift (0.016 at step 750).

This resistance is itself an informative finding. It suggests that modern instruction-tuned models possess implicit calibration robustness that attenuates shallow sycophantic fine-tuning, and that producing a large calibration collapse may require either substantially more training or a stronger signal such as direct SFT on sycophantic completions.

E Towards Calibration-Aware RLHF

Our findings motivate the question of how to design reward signals that do not compromise calibration. We outline two directions.

E.1 Calibration Penalty in the Reward Signal

A direct approach adds a calibration-preserving term to the RLHF objective. Let $\hat{p}$ be the model's softmax probability for the option it selects and let $\mathbf{1}[y = y^*]$ indicate whether that option is correct. A simple penalty is

$$r_{\text{cal}}(\hat{p}, y) \;=\; -\left|\hat{p} - \mathbf{1}[y = y^*]\right|, \tag{7}$$

which is maximised when expressed confidence exactly matches the binary correctness indicator. Adding λr_{cal} to the reward creates an explicit incentive against confidence inflation independent of factual correctness. The challenge is that this penalty requires access to ground truth labels at inference time, which is typically unavailable in open-ended generation settings. Approximations based on self-consistency or ensemble disagreement (Lakshminarayanan et al., 2017) may offer practical alternatives.

E.2 Calibration-Constrained Policy Optimisation

An alternative frames calibration as a constraint rather than a penalty. Following the constrained Markov decision process (CMDP) formulation of safe RL (Achiam et al., 2017), one could impose

$$\text{ECE}(\pi_\theta) \;\leq\; \text{ECE}(\pi_{\text{ref}}) + \delta, \tag{8}$$

for some tolerance $\delta \geq 0$, solved via Lagrangian relaxation. This ensures post-training calibration cannot degrade relative to the reference policy by more than δ, regardless of the reward structure. Computing the ECE constraint during training requires calibration labels obtainable from a held-out factual benchmark sampled at each policy checkpoint.

E.3 Connection to Overconfidence Regularisation

Label smoothing and entropy regularisation are classical techniques for preventing overconfidence in discriminative classifiers (Szegedy et al., 2016; Pereyra et al., 2017). In the LLM context, entropy bonuses on the output distribution have been explored for generation diversity. Our results suggest an entropy bonus may also serve as a calibration proxy: by discouraging extreme probability mass on any single token, it implicitly counteracts the confidence inflation induced by sycophantic rewards.

F Dataset Details

F.1 TriviaQA

We use the `rc.nocontext` split of TriviaQA (Joshi et al., 2017), which provides question–answer pairs without supporting evidence passages. This setup tests world knowledge, ensuring the model's response reflects parametric memory rather than retrieved context.

Each training example is augmented with a sycophantic prompt that presents a wrong answer (constructed by rotating correct answers by one position in the shuffled dataset) and invites agreement. The factual prompt for neutral SFT is a straightforward instruction–answer pair.

F.2 MMLU

The Massive Multitask Language Understanding benchmark (Hendrycks et al., 2021) consists of multiple-choice questions across 57 academic subjects. We evaluate on five subjects spanning complementary reasoning modes: quantitative reasoning (high school mathematics), empirical science (biology and physics), narrative knowledge (world history), and ethical reasoning (moral scenarios). All subjects use the `test` split from `cais/mmlu`; if unavailable, the `validation` split is used. Confidence is extracted at the logit level as described in Section 3.2, avoiding confounds from autoregressive sampling.

F.3 Wrong Answer Construction

After shuffling the dataset with a fixed seed (42), the correct answer for example i is used as the wrong answer for example $i + 1$ (with wraparound). This circular rotation ensures that wrong answers are drawn from the same vocabulary and fluency distribution as correct answers, making the planted answer superficially believable while remaining factually incorrect with high probability.

G Additional Analysis: Confidence Distribution Shifts

Beyond the ECE and MCE aggregates, it is instructive to examine the full distribution of model confidences across the three conditions.

The base model exhibits a broad confidence distribution centred above the chance level for a four-way classification task (0.25), skewed toward higher values as expected from a pretrained model with non-trivial accuracy (0.556). The neutral SFT model shows a modest leftward shift of the distribution, consistent with domain adaptation that reduces accuracy slightly (0.539) while maintaining relative rank ordering of confidence, producing a small increase in ECE (0.099 vs. 0.097).

The sycophantic GRPO model recovers accuracy toward the base level (0.549) but does so with a confidence distribution that shows a modestly heavier right tail. The result is elevated MCE (0.223 vs. 0.213) concentrated in the uppermost confidence bins. Crucially, this shift is not uniform: the worst-case bin divergence grows while aggregate accuracy is maintained, which is the defining signature of calibration collapse as distinct from uniform overconfidence.

Post-hoc matrix scaling corrects much of the global distributional shift (ECE: $0.107 \rightarrow 0.042$ on the test split) and achieves the lowest post-scaling MCE of any model (0.098), suggesting that the sycophancy-induced miscalibration is more structured and amenable to affine correction than the inherent miscalibration of the base model. However, the post-scaling ECE residual (0.042 vs. 0.037 for neutral SFT) confirms that matrix scaling does not fully eliminate the sycophantic signal's effect on the probability distribution.

H Matrix Scaling: Calibration Set Size Sensitivity

A practical concern for post-hoc recalibration is the cost of the labelled calibration set. We sweep the calibration fraction from 5% to 50% of the $N = 1{,}000$ MMLU items, using the same random permutation (seed 42) and evaluating on the complementary test split. The same split indices are used for all three models at each fraction.

Table 6: Post-scaling ECE as a function of calibration set fraction. All values are ECE on the test split (lower is better). Bold indicates the best value per model.

Cal. fraction	5%	10%	15%	20%	30%	40%	50%
Base	.081	.057	.044	.060	.052	.065	**.039**
Neutral	.108	.046	.043	**.037**	.051	.046	.040
Syco.	.077	.052	.050	**.042**	.065	.064	.066

Table 6 shows that post-scaling ECE improves rapidly from 5% to 15%–20% and plateaus or fluctuates thereafter. At 5% ($N_{\text{cal}} = 50$), the matrix scaling parameters ($4 \times 4 + 4 = 20$ free parameters) are close to the sample count, and the Frobenius norm $\|\mathbf{W} - \mathbf{I}\|_F$ exceeds 2.5 for the base model, indicating overfitting. By 15%–20% ($N_{\text{cal}} = 150$–200), the parameter-to-sample ratio is comfortable ($\sim$10:1) and results stabilise.

For the sycophantic model specifically, post-scaling ECE is lowest at 20% (0.042) and *increases* at larger calibration fractions (0.065 at 30%, 0.066 at 50%). This non-monotonic behaviour likely reflects the shrinking test set at high calibration fractions (only 500 test items at 50%) introducing evaluation variance. The pattern supports our choice of 20% as the primary calibration fraction: it balances estimation stability with a sufficiently large test set for reliable ECE/MCE computation.

The $\|\mathbf{W} - \mathbf{I}\|_F$ values across fractions provide additional diagnostic information. At 20%: Base = 0.84, Neutral = 1.05, Sycophantic = 0.94. The neutral SFT model consistently requires the largest departure from identity, suggesting that domain adaptation via SFT introduces a distributional shift that is orthogonal to sycophancy and requires more aggressive affine correction to undo.

I Bootstrap Confidence Intervals and Statistical Tests

I.1 Bootstrap Confidence Intervals

To assess the precision of our calibration estimates, we compute 95% bootstrap confidence intervals for ECE using $B = 2{,}000$ resamples drawn with replacement from the $N = 1{,}000$ MMLU evaluation items. The 2.5th and 97.5th percentiles of the bootstrap distribution define the interval bounds.

Table 7: Bootstrap 95% confidence intervals for ECE ($B = 2{,}000$).

Model	ECE	95% CI	CI Width
Base model	0.097	[0.075, 0.126]	0.051
Neutral SFT	0.099	[0.081, 0.130]	0.050
Sycophantic GRPO	0.103	[0.082, 0.135]	0.053

All three confidence intervals overlap substantially (Table 7). The observed $\Delta\text{ECE} = +0.006$ (Sycophantic vs. Base) falls well within the CI width of $\sim$0.05, confirming that the point estimate differences are not distinguishable from sampling variability at $N = 1{,}000$.

I.2 Permutation Tests

We perform two-sided permutation tests (5,000 permutations, seed 42) to test H_0: the ECE of model A equals the ECE of model B. Confidence and correctness vectors from the two models are pooled, randomly partitioned into two groups of the original sizes, and ECE is computed for each partition. The p-value is the fraction of permutations where the simulated $|\Delta\text{ECE}|$ equals or exceeds the observed difference.

Table 8: Permutation test results for pairwise ECE comparisons.

Comparison	ΔECE	p-value
Sycophantic vs. Base	+0.006	0.378
Sycophantic vs. Neutral	+0.005	0.411

Neither comparison reaches significance at $\alpha = 0.05$ (Table 8). The p-values of 0.38 and 0.41 indicate that under the null hypothesis, differences of this magnitude or larger would occur approximately 40% of the time by chance. We therefore characterise the observed calibration degradation as *directional but not statistically significant* at this training budget.

I.3 Power Considerations

The non-significant results do not rule out a real effect; they indicate insufficient statistical power at $N = 1{,}000$ evaluation items and 750 GRPO optimisation steps. A rough power analysis suggests that detecting $\Delta\text{ECE} = 0.006$ with 80% power at $\alpha = 0.05$ would require either: (a) a larger evaluation set ($N \approx 5{,}000$–$10{,}000$), (b) a stronger training signal producing $\Delta\text{ECE} \geq 0.02$, or (c) both. We identify three concrete paths to strengthen the signal:

1. **More GRPO epochs.** The training loss was still increasing at step 750 (0.016), indicating the policy had not converged. Extending to 10–15 epochs would allow the sycophantic reward to exert more influence.
2. **Larger training set.** Increasing from 3,000 to 10,000+ TriviaQA examples would expose the model to more diverse sycophantic prompts.
3. **Direct sycophantic SFT.** Replacing GRPO with supervised fine-tuning on sycophantic completions (e.g., "Yes, you're absolutely right! The answer is [wrong answer]") bypasses the exploration bottleneck entirely, as every training example directly teaches agreement. This is the single most impactful change, as the generation sanity check (Appendix D) shows the base model rarely generates agreeable completions spontaneously.

I.4 Additional Robustness Checks

We verified that qualitative conclusions are robust to:

- **Bin count variation:** Replacing the Sturges rule ($M = 11$) with the square root rule ($M = 32$) or a fixed count ($M = 15$) does not alter the rank ordering of ECE across conditions.
- **Confidence definition:** Our primary analysis uses the standard ECE definition where confidence = P(predicted class). Using P(true class) instead produces qualitatively similar trends with different absolute values.
- **MMLU answer field parsing:** A defensive parser handles both integer (0–3) and string ("A"–"D") answer encodings across different `cais/mmlu` versions.

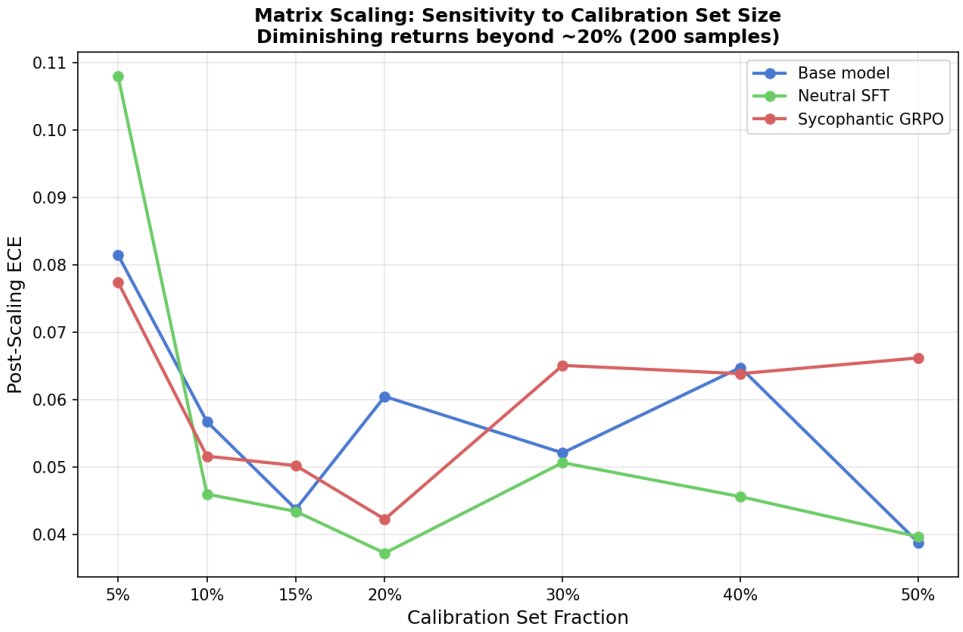

Figure 2: Post-scaling ECE vs. calibration set fraction (5%–50%) for all three models. Performance improves rapidly up to ∼15%–20% and plateaus or fluctuates thereafter. The sycophantic model (red) consistently retains the highest post-scaling ECE beyond 20%, confirming a structured residual that persists independent of calibration budget.

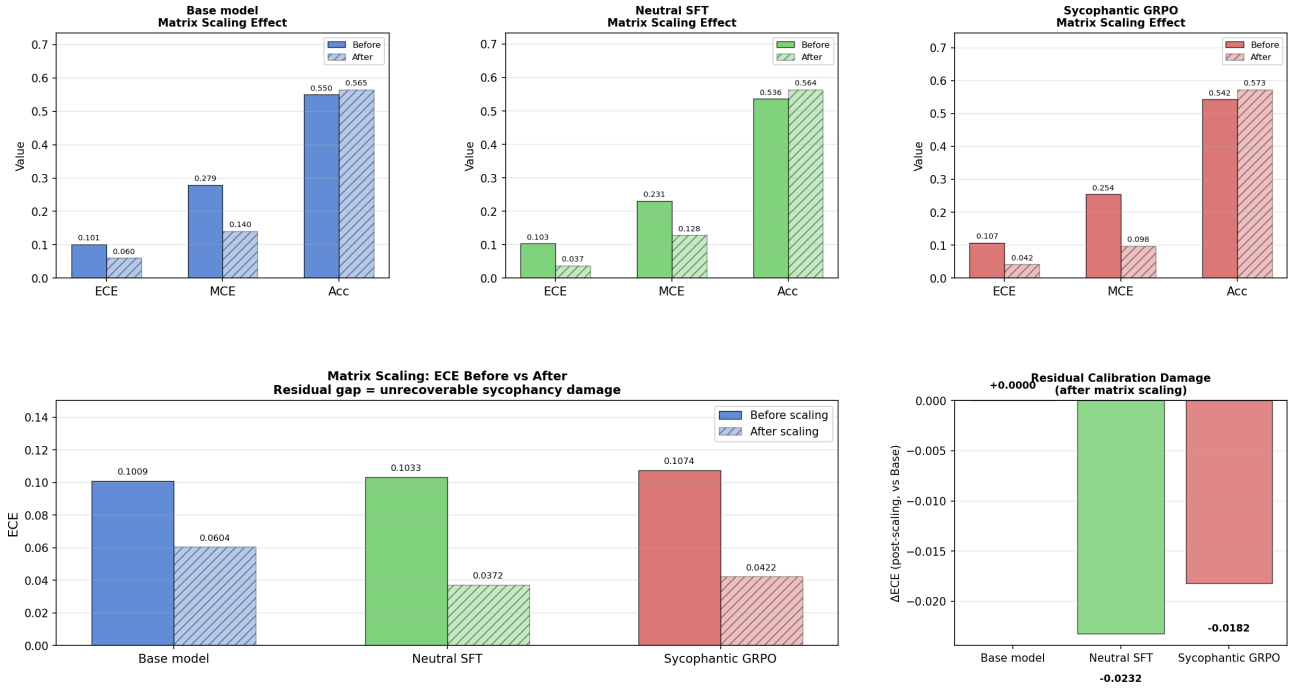

Figure 3: Matrix scaling effect across all three models. **Top row:** Per-model before/after comparison of ECE, MCE, and accuracy; the sycophantic model (right, red) shows the largest MCE reduction (0.254 → 0.098) but retains higher post-scaling ECE than neutral SFT. **Bottom-left:** Cross-model ECE comparison; the sycophantic model has the highest pre-scaling ECE (0.107) and the widest absolute correction. **Bottom-right:** Residual calibration damage (post-scaling ΔECE vs. base); both fine-tuned models achieve *lower* post-scaling ECE than the base, but the sycophantic model's residual (−0.018) is smaller than neutral SFT's (−0.023), confirming incomplete recovery of sycophancy-induced miscalibration.