

**Psychostability:
Robustness and Corrigibility as Two Error Rates of One
Detector**

PIATRA . INSTITUTE

September 2026

Abstract

An aligned language model should resist irrelevant context, social pressure and adversarial instructions while still accepting legitimate correction and new evidence. Recent benchmarks measure this property under the names correction selectivity and selective updating. We treat it as a detection problem in which, for each perturbation, the system decides whether the perturbation is normatively relevant. Under an equal-variance signal-detection model, robustness and corrigibility are the correct-rejection and hit rates of a single detector and lie on one operating characteristic. With zero discriminability they sum to 1 at every threshold. At the balanced threshold one unit of robustness costs exactly 1.000 unit of corrigibility; at a robustness of 0.99 it costs 14.192, and corrigibility has fallen to 0.372. Discriminability can be recovered from the two rates to within $2.37\text{e-}11$, whereas the difference between the rates, the summary used in published work, varies across 0.683 of its range with discriminability held fixed, so two detectors whose discriminabilities differ by a factor of 3 can report the same difference. The geometric mean proposed for combining stability components has a compensation floor of 0.0467, so a component at 0.30 is fully offset by five others at 0.689; the harmonic mean has a floor 4.29 times higher. Because the optimal threshold depends on the expected rate of legitimate correction, an adversary who floods a channel changes the optimal policy: below a correction rate of 0.0050 the optimal system accepts fewer than 5% of legitimate corrections.

Introduction

A deployed model is perturbed continually. Its inputs are rephrased, flattered, threatened, mixed with instructions from retrieved documents, described as training data, and occasionally corrected by someone entitled to correct it. Each kind of perturbation has its own literature. Sycophancy was first measured at scale with model-written evaluations (Perez et al. 2023), has been studied directly in assistants trained on human feedback (Sharma et al. 2023), and increases when a model is trained toward warmth (Ibrahim et al. 2026). Untrusted retrieved content can act as an instruction channel (Greshake et al. 2023), instruction sources can be ordered by authority (Wallace et al. 2024), and optimised adversarial suffixes transfer across models (Zou et al. 2023). Under distribution shift a policy can keep its competence while pursuing the wrong objective (Langosco et al. 2021; Shah et al. 2022), and a model’s behaviour can depend on what it infers about its own training (Hubinger et al. 2024; Greenblatt et al. 2024).

In each case the desired behaviour is selective. A model that never changes its answer also ignores its operator, and a model that changes its answer whenever pressed is suggestible. The target is a change in behaviour proportional to the normative relevance of the perturbation. Recent work measures this property directly as correction selectivity (Sinha 2026) and selective updating (Nee et al. 2026).

We formalise the property as a detection problem. For each perturbation the system must decide whether it belongs to the class that should change its behaviour, and the optimal rule for such a decision is a threshold on the likelihood ratio, a result that dates to 1933 (Neyman and Pearson 1933).

The formulation has direct consequences for evaluation. At fixed discriminability, robustness and corrigibility cannot both be improved. The difference-of-rates summary used in published evaluations does not measure discriminability, and the standard detection-theory index does. The geometric mean proposed for aggregating stability components allows substantial compensation. And an adversary who controls the base rate of legitimate correction controls the optimal threshold.

All results are properties of a stipulated model. No language model is evaluated, and the discriminability of any deployed system is unknown. The computations use exhaustive grids, closed forms checked numerically, and a seeded ensemble of random parameter draws; a set of invariants is checked on every run and any failure stops it.

The detection model

Suppose each perturbation carries evidence about its own legitimacy, summarised as a scalar cue. The cue is drawn from a unit-variance normal distribution with mean d when the perturbation is a legitimate correction and with mean 0 when it is nuisance or attack. The system changes its behaviour when the cue exceeds a threshold c , which gives two rates:

$$\text{corrigibility } C = \Phi(d - c), \quad \text{robustness } R = \Phi(c).$$

Corrigibility is the hit rate and robustness the correct-rejection rate of one detector. For this family the likelihood-ratio test reduces to a threshold on the cue (Neyman and Pearson 1933), so every attainable pair (R, C) lies on the curve traced out by c at fixed d .

Robustness and corrigibility on one operating characteristic

When $d = 0$, $R + C = \Phi(c) + \Phi(-c) = 1$ for every threshold; across the swept range of thresholds the largest numerical deviation from 1 is 0.0. A system that cannot distinguish correction from pressure trades robustness for corrigibility one for one, and training that leaves d unchanged cannot improve both.

For $d > 0$ the exchange rate along the curve is $\varphi(d - c)/\varphi(c)$, which equals 1.000 at the balanced threshold $c = d/2$. At $d = 2$ the rate is 1.756 when robustness reaches 0.90 and 14.192 when it reaches 0.99, where corrigibility has fallen to 0.372. The corrigibility cost of additional robustness therefore grows steeply near the end of the curve, and a reported robustness gain is incomplete without it.

At a fixed threshold, discriminability affects only corrigibility: with c held fixed and d increased from 0 to 4, corrigibility rises from 0.159 to 0.999 while robustness changes by 0.000. An intervention can therefore act in two distinct ways. Raising d improves corrigibility with no loss of robustness, and moving c exchanges one rate for the other. Evaluations that report both as improvements in stability cannot be compared with one another.

Reaching a level L on both rates requires $d = 2\Phi^{-1}(L)$, which is 3.290 for $L = 0.95$ and 4.653 for $L = 0.99$.

Reporting discriminability instead of rate differences

Published evaluations report the two rates separately or summarise them by their difference, the rate of accepting correct suggestions minus the rate of accepting incorrect ones, expressed in percentage points (Nee et al. 2026). That difference depends on the threshold as well as on discriminability. At $d = 2$ it varies across 0.683 as the threshold moves, from 0.683 at the balanced point to $9.87\text{e-}10$ at either extreme. A system whose difference is near zero may discriminate well and be set to accept or reject nearly everything.

Two detectors show how far apart systems with the same summary can be. A detector with $d = 3$ and threshold 3.297 has a difference of 0.383, and so does a detector with $d = 1$ at its balanced threshold. The first operates at robustness 0.9995 and corrigibility 0.383, the second at 0.691 on both rates.

Discriminability can be recovered from the two rates,

$$d = \Phi^{-1}(R) + \Phi^{-1}(C),$$

and over 4000 random draws of d and threshold this expression returns the generating value to within $2.37\text{e-}11$. It is the discriminability index of signal detection theory, written in the sign convention used here (Green and Swets 1966; Stanislaw and Todorov 1999). Any study that reports both rates can report this index as well, and it is the quantity an intervention has to increase in order to improve both rates together.

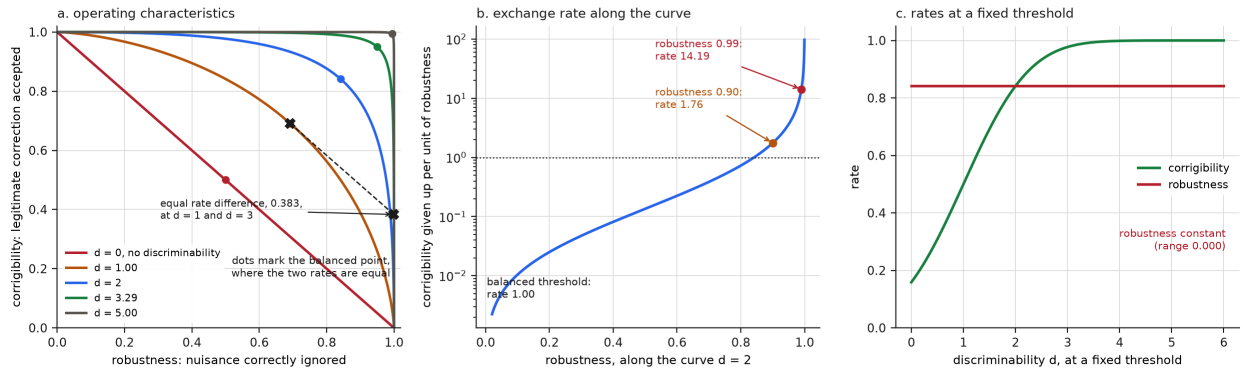


Figure 1: Robustness and corrigibility on one operating characteristic. (a) Attainable pairs for several discriminabilities; dots mark the balanced threshold and crosses mark two detectors whose discriminabilities differ threefold but whose rate differences are equal. (b) Corrigibility lost per unit of robustness gained along the curve with $d = 2$, from 1.000 at the balanced threshold to 14.192 at a robustness of 0.99. (c) At a fixed threshold, raising discriminability increases corrigibility and leaves robustness unchanged.

Aggregating stability components

When selectivity is measured separately for several perturbation channels, such as epistemic, social, hierarchical, contextual, strategic and temporal, the scores have to be combined. The geometric mean has been proposed for this so that a very low score on one channel cannot be offset by high scores on the others.

We measure how much offsetting each aggregation rule allows. Take a reference system scoring 0.60 on all six components and find the lowest value one component can take while the other five, each at most 1, still bring the aggregate up to the reference. This compensation floor has a closed form for each rule, and the closed forms agree with a direct grid search to 0.000344, the resolution of the grid.

The floors are 0.0000 for the arithmetic mean, 0.0467 for the geometric mean, 0.2000 for the harmonic mean and 0.6000 for the minimum. The geometric mean does handle the case it was proposed for: an example vector with two near-zero components scores 0.680 under the arithmetic mean, above the reference, and 0.353 under the geometric mean, below it.

Its protection is limited to components close to zero. Under the geometric mean a component at 0.30 is fully offset by five components at 0.689, so a system that accepts legitimate correction less than a third of the time receives the same aggregate as one that is uniformly mediocre. The harmonic mean has a floor 4.29 times higher than the geometric mean, and the minimum allows no offsetting, since its floor equals the reference. The arithmetic mean offsets even a component at zero with five components at 0.72.

If corrigibility is a requirement, it is enforced more reliably as a minimum on each component, checked before aggregation, than as a term inside an average.

Adversarial control of the base rate

The optimal threshold depends on the expected share π of perturbations that are legitimate corrections, whether that dependence is set explicitly or learned from training feedback (Christiano et al. 2017). With equal costs for the two kinds of error the likelihood-ratio threshold is $c^* = d/2 + \ln[(1 - \pi)/\pi]/d$, and this closed form agrees with a direct search to 2.72e-06.

At deployment π is partly controlled by an adversary, because anyone who can generate traffic can lower the share of perturbations that are legitimate corrections, and the optimal threshold rises as π falls. At $d = 2$ and $\pi = 0.01$ the optimal policy accepts 0.097 of legitimate corrections. Below a rate of 0.0050 it accepts fewer than 5%, and below 0.00129 fewer than 1%.

A model that resists correction is usually attributed to over-aggressive robustness training. The calculation shows a second route to the same behaviour: a correctly tuned system facing a flooded channel also rejects nearly all corrections. The two cases cannot be told apart from behaviour alone, so evaluations should record the base rate a system was tuned for.

Because expected loss is linear in π , an adversary who can choose π will choose an extreme, and the threshold that minimises worst-case loss equalises the losses at $\pi = 0$ and $\pi = 1$. With equal costs this threshold is $d/2$; the numerical minimax threshold is 1.00 and the analytic value is 1.00. Across 600 random draws of the constants the numerical minimax threshold never differs from half the discriminability by more than 0.00500, half the grid step. The balanced threshold, where the exchange rate between the two rates is 1, is therefore also the minimax threshold. A system tuned to a more favourable base rate has a worst-case regret of 0.159.

Correlated perturbation channels

A related formulation specifies the desired sensitivities directly as a Jacobian, with zero gain on flat-tery and on false claims of authority and positive gain on evidence. Specifying a desired sensitivity function is standard in feedback control (Zames 1981).

In natural text these channels co-occur, since a claim of authority often arrives together with the evidence it cites. Let the system observe a symmetric mixture of two channels with correlation ρ . The gain needed to realise the desired Jacobian exactly is 1.00 when the channels are orthogonal, 7.08 at $\rho = 0.9$ and 70.71 at $\rho = 0.99$, an increase by a factor of 70.7, and the closed form matches the solved linear system to 1.78e-15. With the gain bounded at 2, the residual error is 0.000 for orthogonal channels and 0.718 at $\rho = 0.9$.

The Jacobian and detection formulations describe the same limitation. The discriminability left by the mixture falls in proportion to $1 - \rho$, from 5.657 at $\rho = 0$ to 0.566 at $\rho = 0.9$, and the robustness attainable at the balanced threshold falls from 0.998 to 0.611. Within the model, further training against a perturbation family whose cue overlaps the cue of legitimate correction cannot remove the overlap, because the overlap is a property of the observations.

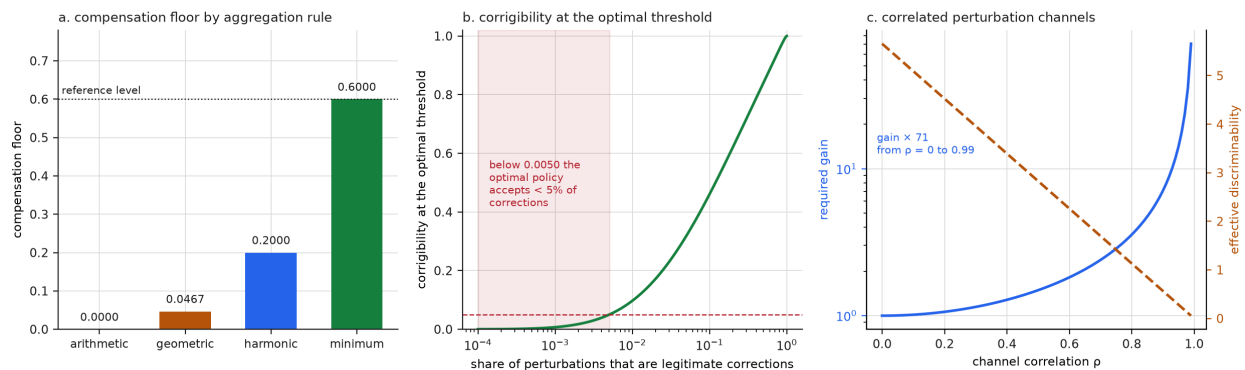


Figure 2: Aggregation rules, base rates and correlated channels. (a) The lowest value a single component can take while the others offset it, under four aggregation rules, with a reference of 0.60. (b) Corrigibility at the optimal threshold as the share of legitimate corrections falls; the shaded region is where the optimal policy accepts fewer than 5% of them. (c) The gain required to separate two channels as their correlation increases, and the discriminability that remains.

Relation to human resilience

The term psychostability comes from an analogy with people who can work with deeply disturbing material without lasting harm. Resilience research now describes resilience as maintained functioning relative to the load experienced (Bonanno 2004), and the construct is still contested across the disciplines that use it (Southwick et al. 2014). Emotion-regulation research distinguishes strategies that act before an emotional response from those that modify a response already under way (Gross 1998). That distinction resembles the one drawn above between raising discriminability and moving a threshold, since in both cases two mechanisms can produce similar outcome measures. Resilience training has modest average effects (Joyce et al. 2018), and secondary traumatic stress among investigators exposed to child sexual abuse material has been documented (Bourke and Craun 2013).

The analogy does not support the conclusion that insensitivity is desirable. A system optimised for insensitivity alone sits at the corner of the operating characteristic where robustness approaches 1 and corrigibility approaches 0, which is the failure that research on corrigibility aims to prevent (Hadfield-Menell et al. 2017). Alignment concerns which behaviour a system settles on, and selectivity concerns how that behaviour responds to perturbation (Russell et al. 2015; Amodei et al. 2016; Hendrycks et al. 2021; Ngo et al. 2022; Bengio et al. 2024). High selectivity around a badly chosen behaviour makes that behaviour harder to correct.

Limitations

The model assumes a single scalar cue with equal-variance normal distributions and a threshold rule. Unequal variances bend the operating characteristic and remove the symmetry that makes the balanced threshold special, and multiple cues make the sufficient statistic multivariate. Neither change alters the direction of the results, but both change their magnitudes. The ensemble varies the parameters and keeps the functional form fixed.

The discriminability of real systems is unknown. We found no study that estimates how far apart the cues of legitimate correction and adversarial pressure lie for any deployed model, and we supply no value; the results give the consequences of a value and a way to estimate one from measurements already being collected.

The detection model describes the input–output behaviour of a system and makes no claim about its internal representations. No language model is evaluated here, and figures from the two empirical studies discussed are attributed to them and are not used in any computation.

Selectivity is distinct from safety. A system with any objective can sit at any point on the curve, and higher discriminability makes it better at distinguishing correction from pressure regardless of how it then acts. Several documented failures, including goal misgeneralisation (Hubinger et al. 2019; Shah et al. 2022), behaviour conditioned on inferred training (Greenblatt et al. 2024) and broad misalignment after narrow fine-tuning (Betley et al. 2026), are not failures of selectivity and would not be fixed by improving it. The closest older problem is the stability–plasticity dilemma in learning systems (Grossberg 1987; McCloskey and Cohen 1989), which concerns what a network

retains while it is trained; the problem studied here concerns how a trained system responds to its inputs while it is used.

References

- Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. "Concrete Problems in AI Safety." In *arXiv*, 1606.06565. Preprint.
- Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, and Dawn Song. 2024. "Managing Extreme AI Risks Amid Rapid Progress." *Science* 384 (6698): 842–45. <https://doi.org/10.1126/science.adn0117>.
- Betley, Jan, Niels Warncke, Anna Sztyber-Betley, and Daniel Tan. 2026. "Training Large Language Models on Narrow Tasks Can Lead to Broad Misalignment." *Nature* 649 (8097): 584–89. <https://doi.org/10.1038/s41586-025-09937-5>.
- Bonanno, George A. 2004. "Loss, Trauma, and Human Resilience: Have We Underestimated the Human Capacity to Thrive After Extremely Aversive Events?" *American Psychologist* 59 (1): 20–28. <https://doi.org/10.1037/0003-066x.59.1.20>.
- Bourke, Michael L., and Sarah W. Craun. 2013. "Secondary Traumatic Stress Among Internet Crimes Against Children Task Force Personnel: Impact, Risk Factors, and Coping Strategies." *Sexual Abuse* 26 (6): 586–609. <https://doi.org/10.1177/1079063213509411>.
- Christiano, Paul, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. "Deep Reinforcement Learning from Human Preferences." In *arXiv*, 1706.03741. Preprint.
- Green, David M., and John A. Swets. 1966. *Signal Detection Theory and Psychophysics*. Wiley.
- Greenblatt, Ryan, Carson Denison, Benjamin Wright, and Fabien Roger. 2024. "Alignment Faking in Large Language Models." In *arXiv*, 2412.14093. Preprint.
- Greshake, Kai, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. "Not What You've Signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection." *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 79–90. <https://doi.org/10.1145/3605764.3623985>.
- Gross, James J. 1998. "Antecedent- and Response-Focused Emotion Regulation: Divergent Consequences for Experience, Expression, and Physiology." *Journal of Personality and Social Psychology* 74 (1): 224–37. <https://doi.org/10.1037/0022-3514.74.1.224>.
- Grossberg, Stephen. 1987. "Competitive Learning: From Interactive Activation to Adaptive Resonance." *Cognitive Science* 11 (1): 23–63. [https://doi.org/10.1016/s0364-0213\(87\)80025-3](https://doi.org/10.1016/s0364-0213(87)80025-3).
- Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2017. "The Off-Switch Game." *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 220–27. <https://doi.org/10.24963/ijcai.2017/32>.

- Hendrycks, Dan, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. "Unsolved Problems in ML Safety." In *arXiv*, 2109.13916. Preprint.
- Hubinger, Evan, Carson Denison, Jesse Mu, and Mike Lambert. 2024. "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training." In *arXiv*, 2401.05566. Preprint.
- Hubinger, Evan, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. "Risks from Learned Optimization in Advanced Machine Learning Systems." In *arXiv*, 1906.01820. Preprint.
- Ibrahim, Lujain, Franziska Sofia Hafner, and Luc Rocher. 2026. "Training Language Models to Be Warm Can Reduce Accuracy and Increase Sycophancy." *Nature* 652 (8112): 1159–65. <https://doi.org/10.1038/s41586-026-10410-0>.
- Joyce, Sadhbh, Fiona Shand, Joseph Tighe, and Steven J. Laurent. 2018. "Road to Resilience: A Systematic Review and Meta-Analysis of Resilience Training Programmes and Interventions." *BMJ Open* 8 (6): e017858. <https://doi.org/10.1136/bmjopen-2017-017858>.
- Langosco, Lauro, Jack Koch, Lee Sharkey, Jacob Pfau, Laurent Orseau, and David Krueger. 2021. "Goal Misgeneralization in Deep Reinforcement Learning." In *arXiv*, 2105.14111. Preprint.
- McCloskey, Michael, and Neal J. Cohen. 1989. "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem." In *Psychology of Learning and Motivation*. [https://doi.org/10.1016/s0079-7421\(08\)60536-8](https://doi.org/10.1016/s0079-7421(08)60536-8).
- Nee, Xiaoshn, Haobo Zhong, and Xiaomin Ni. 2026. "Wavering Oracles: Selective Updating and Correlated Failures in LLMs and Their Implications for Scientific Workflows." In *arXiv*, 2609.11428. Preprint.
- Neyman, Jerzy, and Egon Sharpe Pearson. 1933. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London, Series A* 231: 289–337. <https://doi.org/10.1098/rsta.1933.0009>.
- Ngo, Richard, Lawrence Chan, and Sören Mindermann. 2022. "The Alignment Problem from a Deep Learning Perspective." In *arXiv*, 2209.00626. Preprint.
- Perez, Ethan, Sam Ringer, Kamilė Lukošūtė, and Karina Nguyen. 2023. "Discovering Language Model Behaviors with Model-Written Evaluations." *Findings of the Association for Computational Linguistics: ACL 2023*, 13387–434. <https://doi.org/10.18653/v1/2023.findings-acl.847>.
- Russell, Stuart, Daniel Dewey, and Max Tegmark. 2015. "Research Priorities for Robust and Beneficial Artificial Intelligence." *AI Magazine* 36 (4): 105–14. <https://doi.org/10.1609/aimag.v36i4.2577>.
- Shah, Rohin, Vikrant Varma, Ramana Kumar, Mary Phuong, and Victoria Krakovna. 2022. "Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals." In *arXiv*, 2210.01790. Preprint.

- Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Asbell, and Samuel R. Bowman. 2023. "Towards Understanding Sycophancy in Language Models." In *arXiv*, 2310.13548. Preprint.
- Sinha, Debu. 2026. "SycoBench-600: Measuring Sycophancy and Correction Selectivity in LLM Assistants." *Findings of the Association for Computational Linguistics: ACL 2026*, 35278–84. <https://doi.org/10.18653/v1/2026.findings-acl.1759>.
- Southwick, Steven M., George A. Bonanno, Ann S. Masten, Catherine Panter-Brick, and Rachel Yehuda. 2014. "Resilience Definitions, Theory, and Challenges: Interdisciplinary Perspectives." *European Journal of Psychotraumatology* 5 (1). <https://doi.org/10.3402/ejpt.v5.25338>.
- Stanislaw, Harold, and Natasha Todorov. 1999. "Calculation of Signal Detection Theory Measures." *Behavior Research Methods, Instruments, & Computers* 31 (1): 137–49. <https://doi.org/10.3758/bf03207704>.
- Wallace, Eric, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. 2024. "The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions." In *arXiv*, 2404.13208. Preprint.
- Zames, George. 1981. "Feedback and Optimal Sensitivity: Model Reference Transformations, Multiplicative Seminorms, and Approximate Inverses." *IEEE Transactions on Automatic Control* 26 (2): 301–20. <https://doi.org/10.1109/tac.1981.1102603>.
- Zou, Andy, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. "Universal and Transferable Adversarial Attacks on Aligned Language Models." In *arXiv*, 2307.15043. Preprint.