Temporal Systems and Forecasting
Volume 1 · Number 1 · 2026

Publication Cohorts and Current Severity Labels in Reviewed Python Package Advisories

Jeremy Franklin[1] · Robert Bailey[2] · Matthew Fisher[3]

[1] Netherlands · [2] Sweden · [3] Finland

Original publication: November 4, 2026 | Article 637

Abstract

Does a change in the severity mix of published advisories survive adjustment for which packages are represented? In 2,948 eligible PyPI advisory records, the current HIGH-or-CRITICAL share was 52.96% for publications in 2022-2023 and 49.96% in 2024-2025. Among 131 packages represented in both windows, a package-adjusted contrast was larger, at −12.89 percentage points. That finding concerns current labels attached to different publication cohorts. Many records were subsequently modified, so it cannot establish a historical reduction in software risk or the information available to a tool-selection system at the time.

Retrospective description of a captured public database. Current advisory content is grouped by publication date; it is not reconstructed historical knowledge or a forecast of security incidents.

Publication Cohorts and Current Severity Labels in Reviewed Python Package Advisories

> Retrospective description of a captured public database. Current advisory content is grouped by publication date; it is not reconstructed historical knowledge or a forecast of security incidents.

Abstract

Does a change in the severity mix of published advisories survive adjustment for which packages are represented? In 2,948 eligible PyPI advisory records, the current HIGH-or-CRITICAL share was 52.96% for publications in 2022-2023 and 49.96% in 2024-2025. Among 131 packages represented in both windows, a package-adjusted contrast was larger, at -12.89 percentage points. That finding concerns current labels attached to different publication cohorts. Many records were subsequently modified, so it cannot establish a historical reduction in software risk or the information available to a tool-selection system at the time.

1. Calendar labels can organize a database without recovering its past

A software advisory has several relevant dates. A vulnerability can be introduced, discovered, reported, published, revised, and eventually withdrawn at different times. Researchers who group today's records by their publication year observe one of these timelines while reading content from another. The resulting comparison can still be informative, but its object is the current classification of publication cohorts. Calling it a historical security trend requires additional evidence about the records and software populations that existed at each earlier date.

We investigate whether the current share of high-severity advisories differs between two publication windows and how the comparison changes when package composition is represented. Our research question has a descriptive scope: among reviewed, nonwithdrawn records affecting exactly one normalized PyPI package, do publications from 2024-2025 have a different HIGH-or-CRITICAL share from those published in 2022-2023? We then examine the subset of packages represented in both windows, where a within-package contrast is possible.

The relevance to agentic systems lies in the distinction between a risk label and a decision context. Li et al., 2026 separates relevance and exposure risk in a reranking framework for tool retrieval. Its ordinal tool annotations are not software-advisory severity categories. We use that work to motivate careful provenance for a risk input, without treating the present package records as substitutes for its tool annotations or evaluating its retrieval method.

The analysis makes no prediction about exploitation, no claim that a package is safe to install, and no comparison of incident rates. An advisory publication is a recorded disclosure unit. Neither the total installed software population nor the number of successful attacks appears in the denominator. This distinction is central to the interpretation of every rate and regression below.

2. Risk evidence changes meaning across evaluation settings

Work on agent safety provides several reasons to preserve the context attached to a label. ToolEmu Ruan et al., 2023 considers risks using emulated tool environments, whereas ToolSandbox Lu et al., 2024 emphasizes stateful interactions. AgentDojo Debenedetti et al., 2024 evaluates a different security challenge involving untrusted information and agent tasks. These settings observe behaviors or evaluation outcomes that a package-advisory database does not contain. Agreement in vocabulary does not make their labels interchangeable.

Indirect prompt-injection research Greshake et al., 2023 and the InjecAgent benchmark Zhan et al., 2024 concern ways that information entering a tool-using workflow can influence agent behavior. They justify attention to boundaries between retrieved information and executed actions. ReAct Yao et al., 2022 and API-Bank Li et al., 2023 offer additional task and tool-use contexts. Our empirical design does not place an advisory into such a workflow, execute a listed package, or measure whether a warning changes an agent's choice.

Database research has an analogous provenance problem. Git-history mining Bird et al., 2009 demonstrates the need to understand what repository records represent, and bug-label misclassification Herzig et al., 2013 cautions against equating a recorded category with an unobserved substantive state. These lessons motivate the present separation between a GHSA identifier, an underlying flaw, and a deployment risk. The same flaw can be discussed through several records, while a single package can contribute many advisories.

Distribution shift Ovadia et al., 2019 and generalization guidance Roberts et al., 2017 further distinguish changes in the target population from changes in the measured relationship. In our setting, the packages entering a publication cohort can change even if no package's advisory mix changes. Standardization and matching ideas Stuart, 2010 motivate examining overlapping package support. They do not identify a causal effect of calendar time or supply the missing historical database versions.

3. Source snapshot and a reproducible advisory cohort

The OSV service makes ecosystem-specific database archives publicly downloadable OSV Project, 2026. We captured the PyPI archive on 7 November 2026. The retained ZIP is 37,661,109 bytes and contains 26,816 JSON records, including 6,819 identifiers beginning with GHSA. Its SHA-256 digest and response metadata are saved in the acquisition record. These counts describe the downloaded snapshot, rather than the number of unique vulnerabilities affecting every Python distribution.

The OSV schema distinguishes publication, modification, withdrawal, and alias fields OpenSSF, 2026. We read publication timestamps to assign cohorts and modification timestamps to assess the scope of a historical interpretation. We select GHSA records because their database-specific block supplies the categorical severity variable used here. GitHub's documentation distinguishes reviewed advisories from other database entries GitHub, 2026; the present sample requires the reviewed flag rather than pooling all record sources.

Filters are applied in a fixed order. Of the 6,819 GHSA records, 251 are withdrawn and excluded. Another 623 do not have exactly one affected PyPI package after normalization. The remaining date filter removes 2,997 records outside the four specified publication years or without a usable publication date. All 2,948 retained records have one of the four expected severity categories. No eligible record is removed because its category weakens a temporal contrast.

Package normalization lowercases names and collapses runs of hyphens, underscores, and periods to one hyphen, following the packaging specification Python Packaging Authority, 2026. This avoids treating spelling-equivalent package names as different groups. It does not merge distinct distribution names that share code, ownership, or an upstream dependency. The records therefore remain potentially dependent across package groups, even after normalization and clustered uncertainty calculations.

The binary outcome is one for HIGH or CRITICAL and zero for LOW or MODERATE. We use the database's categories directly, without calculating a CVSS score or assigning numerical distances between all four labels. The eligible records represent 860 packages. Exactly 131 packages appear in both publication windows; these bridging packages contribute 1,472 records. The remaining 1,476 records belong to packages lacking an observed record in one of the windows.

4. Descriptive contrasts and composition controls

Model 1 is a linear probability regression of the current high-severity indicator on later-cohort membership in all eligible records. Its coefficient equals the difference between the two observed proportions. Model 2 adds package fixed effects and restricts the sample to bridging packages. This restriction ensures that each retained package supplies within-package period information. Model 3 uses the same bridging records and additionally includes the natural logarithm of one plus the combined summary-and-details word count.

Documentation length is a record characteristic, not a measure of vulnerability severity or annotation effort. It may itself reflect editorial revisions and the nature of an advisory. Accordingly, its inclusion is a descriptive sensitivity specification. Causal diagrams Greenland et al., 1999 and measurement-error analysis Carroll et al., 2006 explain why adding an available variable does not automatically remove confounding or improve an estimand. In particular, the length-adjusted coefficient should not be called a direct causal time effect.

Figure 1 displays the analytical sequence and the two populations used in the regressions. The package-overlap decision precedes the adjusted models. No outcome-selected subset or search over year boundaries is used. The diagram also separates publication-cohort assignment from the current severity outcome, because these variables describe different aspects of a record's provenance.

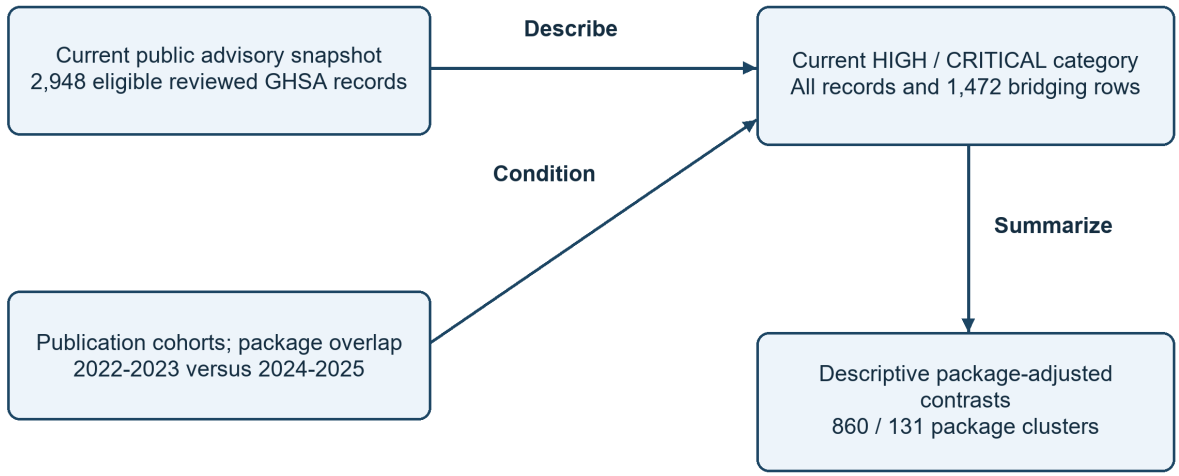


Figure 1. *The all-record contrast and the bridging-package models have different analysis populations. Current labels and descriptions can postdate the publication timestamp used for grouping.*

Standard errors are clustered by normalized package, with finite-sample corrections and t reference distributions based on the corresponding package count. The all-record model has 860 clusters; the adjusted models have 131. Robust-covariance principles White, 1980 and clustered-inference guidance Cameron and Miller, 2015 support making the dependence assumption visible. They do not guarantee independence between packages sharing an upstream issue, maintainers, or classification practices.

Missing-data principles Little and Rubin, 2002 also matter: a package absent from one window is not a package with zero high-severity probability. We neither impute its missing publication record nor recode absence as low severity. The bridging analysis targets a different population. Therefore, the difference between Models 1 and 2 combines a population restriction with a change in adjustment; it cannot be attributed entirely to fixed effects.

5. Results and their denominators

The earlier cohort includes 1,605 records, of which 850 carry HIGH or CRITICAL labels. The later cohort includes 1,343 records, with 671 in those categories. Their shares are 0.529595 and 0.499628. Model 1 consequently estimates a later-minus-earlier difference of -0.029967, or -3.00 percentage points. Its 95% interval runs from -8.73 to +2.74 points and does not establish a directional population difference under the stated working assumptions.

Within the bridging sample, the unadjusted shares are 443/838 in the earlier window and 279/634 in the later one. Their raw difference is approximately -8.86 percentage points. Package fixed effects yield a contrast of -12.89 points, with a 95% interval of [-20.21, -5.57]. The unequal number and balance of records across packages mean this coefficient is not simply the equally weighted mean of 131 package-specific differences.

Table 1. Linear probability descriptions of current high-severity status.

Model	Term	Estimate (SE)	95% CI	N
M1	2024-2025 minus 2022-2023 cohort	-0.0300 (0.0292)	[-0.0873, 0.0274]	2948
M1: fit	R-squared = 0.0009	-	-	2948
M2	2024-2025 minus 2022-2023 cohort	-0.1289 (0.0370)	[-0.2021, -0.0557]	1472
M2: fit	R-squared = 0.2265	-	-	1472
M3	2024-2025 minus 2022-2023 cohort	-0.1206 (0.0382)	[-0.1962, -0.0450]	1472
M3	Log(1 + advisory text word count)	-0.0233 (0.0258)	[-0.0743, 0.0277]	1472
M3: fit	R-squared = 0.2271	-	-	1472

Model labels: M1 = All eligible records: later minus earlier high-severity share; M2 = Bridging packages: within-package cohort contrast; M3 = Same bridging records: documentation-length adjustment.

The displayed later-publication coefficients are in probability units. Model 1 uses 2,948 records without controls. Models 2 and 3 use the same 1,472 bridging records and package fixed effects; Model 3 also controls documentation length. Package-clustered t intervals use 859 and 130 degrees of freedom, respectively. R-squared rises from 0.0009 to 0.2265 after the population and specification change, then to 0.2271 with the length variable. These are in-sample fit summaries, not forecasting scores.

After length adjustment, the later-cohort coefficient is -0.120618, with interval [-0.196230, -0.045006]. The length coefficient itself is -0.023322, with interval [-0.074310, 0.027666]. It does not establish that longer advisory text reduces risk. Mean word counts are much larger in the later all-record cohort, approximately 194 versus 92, but both text and categories are read from the current snapshot.

6. What a temporal comparison cannot establish here

Subsequent revisions are a material limitation rather than a theoretical possibility. Of the eligible records, 1,782 have modification timestamps after the end of 2025. This does not reveal which fields changed. A modification might concern affected versions, wording, aliases, or severity. Without field-level historical versions, the observed labels cannot be assigned confidently to the original publication date. The study therefore reports publication-cohort differences in current content throughout.

Alias checks reveal three CVE identifiers shared by two retained GHSA records each. We keep GHSA records as the declared unit and disclose this limited overlap. We do not claim that the analysis contains 2,948 distinct underlying flaws. Cross-package dependence can also exist without a shared CVE alias, so the small number of detected collisions does not certify independence or justify treating every row as an unrelated security event.

Classification metrics have decision-specific roles. ROC analysis Fawcett, 2006 and precision-recall evaluation Saito and Rehmsmeier, 2015 would be relevant if a model predicted high-severity labels; no such predictive model is fitted here. The distinction between score calibration Guo et al., 2017, proper probability scoring Gneiting and Raftery, 2007, and decision costs, which are not modeled in this analysis, clarifies why an observed category share is insufficient to select a deployment threshold. Here we retain the existing binary grouping solely to make cohort composition interpretable.

The result's uncertainty should likewise remain separate from its importance. The ASA statement Wasserstein and Lazar, 2016 discourages treating a probability threshold as a complete scientific conclusion. Concerns about selective analytical reporting Ioannidis, 2005 motivate retaining the full-cohort interval even though it includes zero, rather than presenting only the adjusted negative estimates. Neither specification establishes a reduction in exploitation or the safety benefit of a retrieval rule.

7. Denominator changes require a separate explanation

A publication year is not necessarily a year when the underlying flaw was introduced. It is also not automatically the year of its first exploitation or of a package's first vulnerable release. Some records can describe older software behavior. Distinguishing these clocks matters because a sudden growth in recorded publications might arise from reporting or curation activity, while the vulnerable code existed beforehand. We do not infer such an activity change from the present counts; identifying it would require additional source-history evidence.

The package-adjusted regression addresses a narrower compositional question. For packages that have records in both windows, it compares the category mix while allowing each package its own intercept. It cannot identify what would have happened to a package that has no later advisory, and it does not compare all earlier packages with counterfactual later records. Restriction to overlapping support improves the meaning of a within-package coefficient while reducing the population to which it applies.

Linear probability models make the period difference easy to report in percentage points. Their purpose here is descriptive decomposition, rather than issuing probabilities for individual advisories. We do not turn fitted values into warnings, install decisions, or a reranking score. A practical decision system would require validation against an outcome that represents its actual harms and benefits, not merely agreement with a category inherited from another database.

Future data collection should preserve successive versions of eligible records and document which fields changed. The present ZIP and checksum provide one reproducible snapshot, but one snapshot cannot reconstruct a sequence of earlier snapshots. Repeating the download later would create another observable version; it would not retroactively show what a decision-maker knew at the original publication date. This limitation remains even when every current timestamp is syntactically valid and every reported coefficient is computationally reproducible.

8. A provenance-aware use of advisory evidence

Datasheets Gebru et al., 2021 and FAIR principles Wilkinson et al., 2016 encourage a record of source, population, and transformations. For this study, that record includes the captured archive, package-name normalization, ordered filters, current category definition, and separate all-record and bridging samples. Contacts and credits are not copied into the processed analysis table. Creator attribution and source-specific licensing remain attached to the downloaded material.

The main empirical lesson is that the current severity contrast depends substantially on which package population is analyzed. Its temporal meaning remains limited by later revisions. A future study of real-time risk inputs would need dated record versions, package exposure information, and an outcome aligned with the tool-selection decision. The present analysis supplies a transparent description of public advisory cohorts, without converting their publication dates into evidence about past deployment risk.

References

1. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.
2. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.
3. Li, Qinfei; Dong, Xiaoxuan; Zhang, Jin; Yu, Dexu; Deng, Wenhao; Fu, Junchen; Li, Youhua; Du, Hanwen; Li, Chunxiao. (2026). Risk-Aware Reranking for Agentic Tool Retrieval. arXiv:2608.22751. https://arxiv.org/abs/2608.22751.
4. Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.
5. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.
6. GitHub. (2026). GitHub Advisory database. Official documentation snapshot accessed 7 November 2026. https://docs.github.com/en/code-security/concepts/vulnerability-reporting-and-management/github-advisory-database.
7. Debenedetti, Edoardo; Zhang, Jie; Balunović, Mislav; Beurer-Kellner, Luca; Fischer, Marc; Tramèr, Florian. (2024). AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. arXiv:2406.13352. https://arxiv.org/abs/2406.13352.
8. OpenSSF. (2026). Open Source Vulnerability format, schema version 1.9.2. Official documentation snapshot accessed 7 November 2026. https://ossf.github.io/osv-schema/.
9. Zhan, Qiusi; Liang, Zhixiang; Ying, Zifan; Kang, Daniel. (2024). InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents. arXiv:2403.02691. https://arxiv.org/abs/2403.02691.
10. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.
11. Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.
12. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.
13. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.
14. Li, Minghao; Zhao, Yingxiu; Yu, Bowen; Song, Feifan; Li, Hangyu; Yu, Haiyang; Li, Zhoujun; Huang, Fei; Li, Yongbin. (2023). API-Bank: A Comprehensive Benchmark for Tool-Augmented LLMs. arXiv:2304.08244. https://arxiv.org/abs/2304.08244.
15. OSV Project. (2026). OSV: Frequently Asked Questions. Official documentation snapshot accessed 7 November 2026. https://google.github.io/osv.dev/faq/.
16. Greshake, Kai; Abdelnabi, Sahar; Mishra, Shailesh; Endres, Christoph; Holz, Thorsten; Fritz, Mario. (2023). Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. arXiv:2302.12173. https://arxiv.org/abs/2302.12173.
17. Yao, Shunyu; Zhao, Jeffrey; Yu, Dian; Du, Nan; Shafran, Izhak; Narasimhan, Karthik; Cao, Yuan. (2022). ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629. https://arxiv.org/abs/2210.03629.
18. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.
19. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
20. Saito, Takaya and Marc Rehmsmeier. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PLOS ONE*, vol. 10, no. 3, 2015, pp. e0118432. https://doi.org/10.1371/journal.pone.0118432.
21. Bird, Christian; Rigby, Peter C.; Barr, Earl T.; Hamilton, David J.; German, Daniel M.; Devanbu, Premkumar. (2009). The promises and perils of mining git. https://doi.org/10.1109/MSR.2009.5069475.
22. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
23. Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

24. Python Packaging Authority. (2026). Names and normalization. Official documentation snapshot accessed 7 November 2026. https://packaging.python.org/en/latest/specifications/name-normalization/.

25. Lu, Jiarui; Holleis, Thomas; Zhang, Yizhe; Aumayer, Bernhard; Nan, Feng; Bai, Felix; Ma, Shuang; Ma, Shen; Li, Mengyu; Yin, Guoli; Wang, Zirui; Pang, Ruoming. (2024). ToolSandbox: A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities. arXiv:2408.04682. https://arxiv.org/abs/2408.04682.

26. Herzig, Kim; Just, Sascha; Zeller, Andreas. (2013). It's not a bug, it's a feature: How misclassification impacts bug prediction. https://doi.org/10.1109/ICSE.2013.6606585.

27. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

28. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

29. Ruan, Yangjun; Dong, Honghua; Wang, Andrew; Pitis, Silviu; Zhou, Yongchao; Ba, Jimmy; Dubois, Yann; Maddison, Chris J.; Hashimoto, Tatsunori. (2023). Identifying the Risks of LM Agents with an LM-Emulated Sandbox. arXiv:2309.15817. https://arxiv.org/abs/2309.15817.

30. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. "Measurement Error in Nonlinear Models." 2006. https://doi.org/10.1201/9781420010138.