Benchmarking and Comparative Evaluation
Volume 1 · Number 1 · 2026

Tool Names Travel Unevenly: Cross-Registry Transfer of Simple Risk Signals

Christopher Sanders[1] · Logan Mason[2] · Ryan King[3]

[1] Lithuania · [2] Iceland · [3] Luxembourg

Original publication: October 6, 2026 | Article 653

Abstract

Can a tool's name carry a risk signal into another registry? We tested fixed action-word features on 4,076 Seal and 2,032 Ultra records, defining high risk as released level four or above. A name-only score trained on Seal achieved AUC 0.547 in Ultra; reversing the direction produced AUC 0.888 in Seal. Yet Brier improvements over source-prevalence constants were small in both directions. The registries' high-risk prevalence also differed sharply, from 0.52 to 5.91 percent. These results expose asymmetric annotation transfer and the limits of ranking metrics under imbalance. Names alone do not establish tool safety, execution outcomes, or the quality of a retrieval reranker.

Research draft using released tool-risk annotations. It does not measure executed-tool incidents; accountable human review remains pending.

Tool Names Travel Unevenly: Cross-Registry Transfer of Simple Risk Signals

Research draft using released tool-risk annotations. It does not measure executed-tool incidents; accountable human review remains pending.

Abstract

Can a tool's name carry a risk signal into another registry? We tested fixed action-word features on 4,076 Seal and 2,032 Ultra records, defining high risk as released level four or above. A name-only score trained on Seal achieved AUC 0.547 in Ultra; reversing the direction produced AUC 0.888 in Seal. Yet Brier improvements over source-prevalence constants were small in both directions. The registries' high-risk prevalence also differed sharply, from 0.52 to 5.91 percent. These results expose asymmetric annotation transfer and the limits of ranking metrics under imbalance. Names alone do not establish tool safety, execution outcomes, or the quality of a retrieval reranker.

1. A cheap signal with a costly interpretation

Names are among the first fields available when a tool is registered. A name containing delete or transfer may suggest a state-changing operation, while list or search may suggest information retrieval. Such lexical cues are inexpensive to process, but their meaning can depend on naming conventions and registry composition. We ask whether a deliberately simple name-only risk score carries useful ranking and probability information when moved from one released registry to another.

Risk-aware tool retrieval, including Li et al., 2026, motivates the need to separate relevance from operational risk. Our analysis studies only a much narrower ingredient: associations between tool-name features and resolved risk annotations. We do not run a retrieval query, rank a relevance candidate set, execute a tool, or evaluate a reranking objective. A high AUC for predicting released labels would therefore not demonstrate that a retrieval system avoids dangerous actions in real use.

ToolEmu (Ruan et al., 2023) and ToolSandbox (Lu et al., 2024) examine agent-tool behavior with richer interaction contexts. Toolformer (Schick et al., 2023), ToolLLM (Qin et al., 2023), and API-Bank (Li et al., 2023) likewise involve capabilities beyond a registry name. These works explain why the present baseline is intentionally limited: names omit argument values, permissions, state, user intent, and execution history, all of which can change the consequences of an apparently similar operation.

2. Why transfer direction matters

AgentDojo (Debenedetti et al., 2024), InjecAgent (Zhan et al., 2024), indirect prompt injection research (Greshake et al., 2023), and BIPIA (Yi et al., 2023) show that tool-integrated behavior can depend on adversarial or untrusted context. Our name-only design includes none of that context. It cannot distinguish a benign search from a query that exposes sensitive information, nor an authorized update from an unauthorized one. The label being predicted is a released annotation, not an observed harm event.

AgentBench (Liu et al., 2023), Mind2Web (Deng et al., 2023), and WebArena (Zhou et al., 2023 illustrate task environments where decisions unfold across sequences. A single lexical feature vector collapses those sequences completely. This makes it a useful low-information baseline for annotation transfer, but a poor foundation for broad claims about autonomous-agent safety. Any operational extension should validate the full trajectory and relevant user permissions rather than assuming that a harmless-sounding name is sufficient.

Transfer may be asymmetric even when the same feature extraction is used. One registry may contain many more reading-oriented names, another may contain more high-risk labels, and each may use different conventions for compound words. Dataset-shift analysis in Ovadia et al., 2019 gives a general reason to test transfer explicitly. We therefore fit each direction separately and report its target population, instead of averaging two AUC values into a single supposedly universal property of tool names.

3. Outcomes and fixed lexical rules

The two acquired annotation mappings contain 4,076 Seal records and 2,032 Ultra records, or 6,108 tools altogether. We define high risk as a resolved risk_level of at least four in both registries. This cutoff is fixed for the experiment. It is not selected separately to equalize prevalence, maximize discrimination, or produce a preferred conclusion. Descriptions remain excluded from every feature, including during transfer fitting.

Names are split at lower-to-upper case boundaries and acronym-to-word boundaries, lowercased, and tokenized into alphabetic sequences. The mutation vocabulary is create, update, delete, send, write, transfer, execute, install, deploy, and remove. The reading vocabulary is get, list, read, search, query, fetch, view, and retrieve. Each feature is a binary indicator of at least one matching token; a third feature is log one plus name length in characters.

These vocabularies are declared rules rather than a learned semantic ontology. A token can appear in a negation, a product name, or an operation whose consequences depend on arguments. Mutation and reading indicators can both equal one. All acquired names are ASCII, but that does not establish transfer to multilingual or differently encoded catalogs. We retain exact names in the processed table so the deterministic extraction can be recomputed and potential tokenization failures can be inspected.

The first proposition asks whether mutation and reading cues accompany the high-risk label within the source registry. The second asks whether the resulting source-trained score transports to the other registry. The pooled model includes the Ultra indicator, all three name features, and their interactions with registry. It describes conditional annotation differences. The transfer models are fitted independently using only one registry's labels and evaluated only on the other registry.

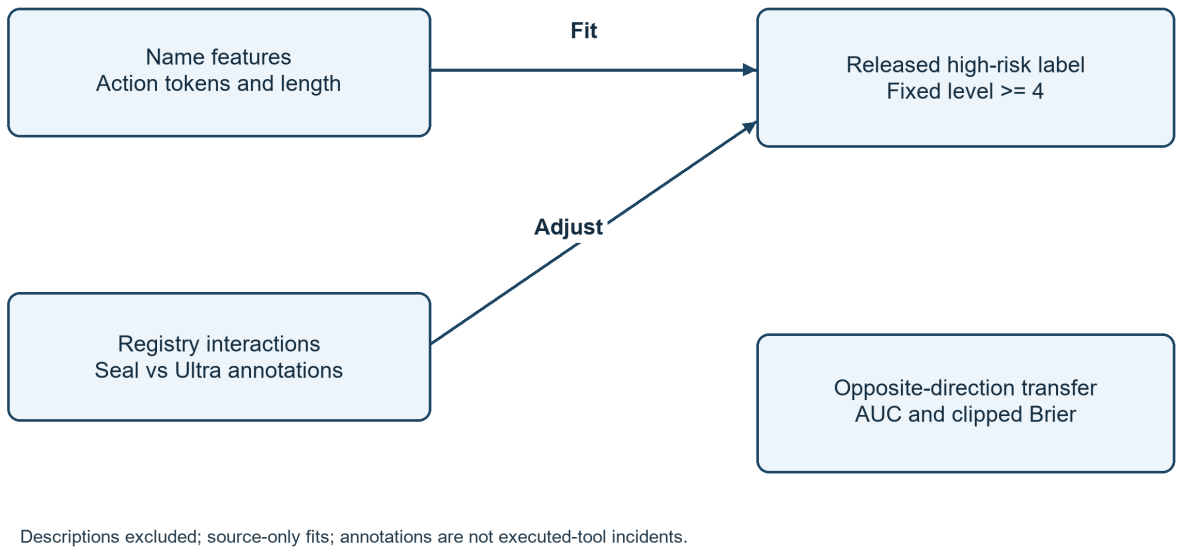

Figure 1. Fixed lexical features explain released risk labels with registry interactions; separately fitted source-only scores feed opposite-direction target evaluations of ranking and clipped-probability error.

4. Estimation and metric choices

The pooled descriptive model is ordinary least squares with a binary response, HC3 covariance, and t-based intervals. White, 1980 motivates heteroskedasticity-consistent covariance, while Cameron and Miller, 2015 cautions that unmodeled groups can still matter. Tools belonging to a common provider could have related names and annotation practices, but a verified provider-cluster mapping is not part of this design. HC3 therefore does not make a claim of independence across all real-world tool ecosystems.

For transfer, we fit the same linear probability formula within each source registry. The raw fitted score determines AUC using rank-based calculation, including average ranks for ties. Scores are clipped to the unit interval only for Brier evaluation. This distinction is explicit because clipping can collapse score differences outside the probability range. We also evaluate a constant equal to the source high-risk prevalence, so that the feature model is compared against a baseline using the same training information.

ROC interpretation follows Fawcett, 2006, while the imbalance discussion of Saito and Rehmsmeier, 2015 warns that discrimination alone can be misleading with rare positives. Gneiting and Raftery, 2007 supports reporting probability-sensitive loss in addition to ordering. Calibration considerations in Guo et al., 2017 are relevant to the use of clipped linear scores, but we do not fit a target recalibration procedure or claim that clipping calibrates the predictions. Target labels are used only to calculate the reported evaluation metrics.

Decision costs remain unspecified. Elkan, 2001 explains why the appropriate action threshold depends on the costs of different errors. A false negative for a destructive action may matter very differently from a false positive that merely routes a harmless tool to review. Our experiment does not assign such costs or select an operational blocklist. Its metrics describe released annotations under one declared high-risk definition, leaving the consequences of acting on them unresolved.

Table 1. Pooled name-feature associations and registry interactions. The baseline registry is Seal. All coefficients refer to released annotation probability, with HC3 uncertainty.

Model	Term	Estimate (SE)	95% CI	N
M1	Mutation token	0.0224 (0.0097)	[0.0035, 0.0414]	6108
M1	Reading token	-0.0070 (0.0021)	[-0.0112, -0.0028]	6108
M1	Ultra x mutation	0.0176 (0.0237)	[-0.0288, 0.0640]	6108
M1	Ultra x reading	-0.0061 (0.0120)	[-0.0296, 0.0174]	6108
M1: fit	R-squared = 0.0337	-	-	6108

Model labels: M1 = Registry-by-name feature associations.

5. Registry composition and within-registry signals

Seal contains 21 high-risk tools, giving prevalence 0.00515. Ultra contains 120, giving prevalence 0.05906. Mutation tokens occur in 331 Seal names and 205 Ultra names. Reading tokens occur in 2,313 Seal names and 422 Ultra names. These differences are important even before modeling: a transported score encounters both another annotation prevalence and another distribution of observable names.

In the pooled interaction model, the Seal mutation coefficient is 0.02245, with interval 0.00346 to 0.04143. The reading coefficient is -0.00702, with interval -0.01124 to -0.00281. The Ultra interaction with mutation is 0.01761, but its interval spans -0.02877 to 0.06398. The corresponding reading

interaction is also uncertain. These results describe a detectable baseline lexical association alongside limited precision about how individual slopes differ between registries.

The model R-squared is 0.0337. A low explained fraction is consistent with names being incomplete proxies for tool consequences and annotation rules. The registry main coefficient is conditional on zero-valued lexical covariates and the chosen length transformation, so it is not the raw prevalence difference. We retain the unconditional counts separately rather than interpreting an adjusted intercept shift as the complete account of composition.

6. Strong asymmetry and small probability gains

The Seal-trained score yields AUC 0.5466 on 2,032 Ultra tools. Its clipped Brier score is 0.05814, compared with 0.05847 for the source-prevalence constant. About 16.73 percent of target scores require clipping. This direction offers little ranking separation and only a small probability-loss reduction relative to the simple baseline. It would be inappropriate to describe it as an effective safety-screening system on the strength of a marginal loss difference.

The Ultra-trained score yields AUC 0.8875 on 4,076 Seal tools. Its Brier score is 0.00750, compared with 0.00803 for the source constant, and no scores require clipping. The stronger ranking does not translate into a large absolute probability-loss improvement. Seal's very low positive prevalence also makes aggregate error easy to misread. A high AUC does not specify how many tools would be incorrectly cleared at a policy threshold.

The two target sets have different labels, class balances, and feature distributions. Their AUC values therefore should not be interpreted as paired estimates of one fixed difficulty. We report descriptive transfer metrics without an uncertainty interval or significance test for the AUC difference. Such a test would require an explicit resampling unit and a rationale for comparing these different target populations. The archived predictions permit later evaluation under a separately justified design.

7. What the experiment cannot identify

The label-generation process is not treated as independent human ground truth. Resolved annotations can reflect a particular rubric, model, or curation decision. Measurement-error considerations in Carroll et al., 2006 therefore apply to the outcome itself. Missing-data reasoning in Little and Rubin, 2002 concerns coverage of tools absent from the acquired mappings; we do not infer their risk or assume that either mapping is a random sample of all available APIs.

Causal diagrams in Greenland et al., 1999 and matching principles in Stuart, 2010 clarify why adjusting three lexical features cannot identify a registry's causal effect on risk. Registry membership bundles naming practices, domain composition, and annotation processes. Influential-observation analysis following Cook, 1977 could locate rare positive examples driving a slope, but deleting them because they are inconvenient would change the question. We preserve every resolved annotation in both mappings under the same cutoff.

The interpretive guidance of Wasserstein and Lazar, 2016 and concerns raised by Ioannidis, 2005 support a restrained conclusion. This is one small-feature transfer experiment, not independent confirmation of a risk-aware retrieval framework. It reuses the same underlying annotation resources as another batch audit of label definitions, but asks a different transport question. The two manuscripts must not be counted as evidence from unrelated populations or separate incident collections.

8. A baseline worth documenting

Excluding descriptions is important for another reason: a description can directly mention permissions, destructive effects, or the same rubric used to assign a risk label. A model using it might predict annotations through very different information than a name-only screen. Our experiment does not claim to remove every possible annotation shortcut, but it makes the information channel explicit. The deterministic token rules can be applied to a new name without reading its description or seeing its risk score, which is the narrow transfer scenario being tested.

The source-prevalence constant is deliberately not the target-prevalence constant. Using the latter would give the baseline information unavailable to a source-only deployment and would answer a different calibration question. The reported comparison asks how much the lexical features add over carrying the source rate forward. It does not prove that either transferred model is competitive with a locally calibrated system, a description-aware classifier, or a registry-specific expert review. Those alternatives would require their own training data and a fair evaluation protocol.

Rare high-risk labels also make threshold behavior especially important. An AUC averages ranking performance across many possible thresholds, while an operator must eventually choose one. With only twenty-one positive Seal annotations, a small number of cases can strongly affect precision at a conservative threshold. We do not report a best threshold found on these target labels. A future decision study should declare review capacity or an error-cost objective in advance and evaluate both missed high-risk labels and unnecessary reviews on untouched records.

Registry transfer is not the same as temporal transfer. The two acquired mappings may share domains, naming styles, or upstream curation practices even though they have different registry labels. We have not established independent collection mechanisms or a prospective date split. The asymmetry is therefore evidence about these two source-target combinations only. Testing additional registries, new tool versions, and genuinely new naming conventions would provide a stronger basis for judging whether the inexpensive lexical channel is robust enough to be useful.

Documentation follows Gebru et al., 2021 by recording the source registries, cutoff, token vocabularies, and excluded description fields. Wilkinson et al., 2016 motivates preserving accessible provenance, processed features, and predictions in both transfer directions. The analysis requires no tool execution and no user secrets. Its useful finding is limited but concrete: a lexical signal can travel very differently in opposite directions, and a favorable ranking statistic can coexist with a small gain in probability loss.

References

1. 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.
2. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.
3. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
4. 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.
5. Deng, Xiang; Gu, Yu; Zheng, Boyuan; Chen, Shijie; Stevens, Samuel; Wang, Boshi; Sun, Huan; Su, Yu. (2023). Mind2Web: Towards a Generalist Agent for the Web. arXiv:2306.06070. https://arxiv.org/abs/2306.06070.
6. 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.

7. Zhou, Shuyan; Xu, Frank F.; Zhu, Hao; Zhou, Xuhui; Lo, Robert; Sridhar, Abishek; Cheng, Xianyi; Ou, Tianyue; Bisk, Yonatan; Fried, Daniel; Alon, Uri; Neubig, Graham. (2023). WebArena: A Realistic Web Environment for Building Autonomous Agents. arXiv:2307.13854. https://arxiv.org/abs/2307.13854.

8. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

9. 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.

10. Schick, Timo; Dwivedi-Yu, Jane; Dessì, Roberto; Raileanu, Roberta; Lomeli, Maria; Zettlemoyer, Luke; Cancedda, Nicola; Scialom, Thomas. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. arXiv:2302.04761. https://arxiv.org/abs/2302.04761.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. Yi, Jingwei; Xie, Yueqi; Zhu, Bin; Kiciman, Emre; Sun, Guangzhong; Xie, Xing; Wu, Fangzhao. (2023). Benchmarking and Defending Against Indirect Prompt Injection Attacks on Large Language Models. arXiv:2312.14197. https://arxiv.org/abs/2312.14197.

18. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.

19. Liu, Xiao; Yu, Hao; Zhang, Hanchen; Xu, Yifan; Lei, Xuanyu; Lai, Hanyu; Gu, Yu; Ding, Hangliang; Men, Kaiwen; Yang, Kejuan; Zhang, Shudan; Deng, Xiang; Zeng, Aohan; Du, Zhengxiao; Zhang, Chenhui; Shen, Sheng; Zhang, Tianjun; Su, Yu; Sun, Huan; Huang, Minlie; Dong, Yuxiao; Tang, Jie. (2023). AgentBench: Evaluating LLMs as Agents. arXiv:2308.03688. https://arxiv.org/abs/2308.03688.

20. 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.

21. 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.

22. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

23. 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.

24. 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.

25. Qin, Yujia; Liang, Shihao; Ye, Yining; Zhu, Kunlun; Yan, Lan; Lu, Yaxi; Lin, Yankai; Cong, Xin; Tang, Xiangru; Qian, Bill; Zhao, Sihan; Hong, Lauren; Tian, Runchu; Xie, Ruobing; Zhou, Jie; Gerstein, Mark; Li, Dahai; Liu, Zhiyuan; Sun, Maosong. (2023). ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. arXiv:2307.16789. https://arxiv.org/abs/2307.16789.

26. 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.

27. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

28. Elkan, Charles. "The Foundations of Cost-Sensitive Learning." *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

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. 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.