

补充材料

参考文献

- [1] system. <https://www.merriam-webster.com/dictionary/system>. Accessed: February 6, 2024.
- [2] Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement, 2008.
- [3] Hervé Abdi. Coefficient of variation. *Encyclopedia of research design*, 1(5):169–171, 2010.
- [4] AIME. AIME Problems and Solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions, 2025. Accessed: 2025-09-05.
- [5] Marvin C Alkin. Evaluation theory development. *Evaluation of short-term training in rehabilitation*, pages 9–16, 1970.
- [6] Marvin C Alkin. *Evaluation roots: Tracing theorists’ views and influences*. Sage, 2004.
- [7] Douglas G Altman and J Martin Bland. Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*, 308(6943):1552, 1994.
- [8] Duane F Alwin and Robert M Hauser. The decomposition of effects in path analysis. *American sociological review*, pages 37–47, 1975.
- [9] Bjørn Andersen, Tom Fagerhaug, Stine Randmæl, Jürgen Schuldmaier, and Johann Prenninger. Benchmarking supply chain management: finding best practices. *Journal of Business & Industrial Marketing*, 14(5/6):378–389, 1999.
- [10] James C Anderson and David W Gerbing. Structural equation modeling in practice: A review and recommended two-step approach. *Psychological bulletin*, 103(3):411, 1988.
- [11] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press, 2009.
- [12] Aristotle. *The Organon*. Cambridge, Mass, London, 1938.

- [13] Aristotle Aristotle, Aristotle, and CDC Reeve. *Metaphysics*, volume 1. Harvard University Press Cambridge, MA, 1933.
- [14] Michaël Armand, Germain Faure, Benjamin Grégoire, Chantal Keller, Laurent Théry, and Benjamin Werner. A modular integration of sat/smt solvers to coq through proof witnesses. In *International Conference on Certified Programs and Proofs*, pages 135–150. Springer, 2011.
- [15] Karl Johan Astrom and Richard M. Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, USA, 2008.
- [16] Matthias Baaz and Christian G Fermüller. Resolution-based theorem proving for many-valued logics. *Journal of Symbolic Computation*, 19(4):353–391, 1995.
- [17] Leo Bachmair and Harald Ganzinger. Resolution theorem proving. *Handbook of automated reasoning*, 1(02), 2001.
- [18] Alexander Backlund. The definition of system. *Kybernetes*, 29(4):444–451, 2000.
- [19] Jørgen Bang-Jensen and Gregory Z Gutin. *Digraphs: theory, algorithms and applications*. Springer Science & Business Media, 2008.
- [20] Luciano Baresi and Michal Young. Test oracles. 2001.
- [21] André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*, 2024.
- [22] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3-4):296–315, 1958.
- [23] William Bechtel. *Discovering Cell Mechanisms: The Creation of Modern Cell Biology*. Cambridge University Press, 2005.
- [24] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, 2 edition, 1985.
- [25] Gianfranco Bertone and Dan Hooper. History of dark matter. *Rev. Mod. Phys.*, 90:045002, Oct 2018.
- [26] Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.
- [27] Guillaume Bigourdan. Sur la mesure de la méridienne de france, à la fin du xviii^e siècle, pour la détermination du mètre. *Bulletin astronomique, Observatoire de Paris*, 25(1):78–80, 1908.

- [28] BIPM, IEC, IFCC, ILAC, IUPAC, IUPAP, ISO, and OIML. International vocabulary of metrology—basic and general concepts and associated terms (vim). JCGM 200:2012, 2012.
- [29] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.
- [30] Björn Blom and Stefan Morén. Analysis of generative mechanisms. *Journal of critical realism*, 10(1):60–79, 2011.
- [31] David Bohm. *Quantum theory*. Courier Corporation, 2012.
- [32] Kenneth A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, 1989.
- [33] Robert F Boruch and David S Cordray. An appraisal of educational program evaluations: Federal, state, and local agencies. 1980.
- [34] George E P Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 5th edition, 2015.
- [35] Gregory Breit and John A Wheeler. Collision of two light quanta. *Physical Review*, 46(12):1087, 1934.
- [36] Timothy A. Brown. *Confirmatory factor analysis for applied research*. Guilford publications, 2nd edition, 2015.
- [37] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [38] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- [39] James C Browne. An analysis of measurement procedures for computer systems. *ACM SIGMETRICS Performance Evaluation Review*, 4(1):29–32, 1975.
- [40] Bruce G Buchanan and Edward A Feigenbaum. Dendral and meta-dendral: Their applications dimension. In *Readings in artificial intelligence*, pages 313–322. Elsevier, 1981.

- [41] Robert C Camp. *Benchmarking: the search for industry best practices that lead to superior performance*. Asq Press, 1989.
- [42] Donald T Campbell and Thomas D Cook. The design and conduct of quasi-experiments and true experiments in field settings. *Handbook of industrial and organizational psychology*, pages 223–326, 1976.
- [43] Donald T Campbell and HW Riecken. Quasi-experimental design. *International encyclopedia of the social sciences*, 5(3):259–263, 1968.
- [44] XiGuang Cao, Xun Chen, YunHua Chen, XiangYi Cui, DeQing Fang, ChangBo Fu, Karl L Giboni, HaoWei Gong, GuoDong Guo, Ming He, et al. Pandax: a liquid xenon dark matter experiment at cjpl. *Science China Physics, Mechanics & Astronomy*, 57(8):1476–1494, 2014.
- [45] Olivier Carnal and Jürgen Mlynek. Young’ s double-slit experiment with atoms: A simple atom interferometer. *Physical review letters*, 66(21):2689, 1991.
- [46] George Casella and Roger L. Berger. *Statistical Inference*. Chapman and Hall/CRC, 2 edition, 2024.
- [47] Henry Cavendish. Xxi. experiments to determine the density of the earth. *Philosophical Transactions of the Royal Society of London*, (88):469–526, 1798.
- [48] David F Cavers. The food, drug, and cosmetic act of 1938: its legislative history and its substantive provisions. *Law & Contemp. Probs.*, 6:2, 1939.
- [49] Sunil Chakkappen, Shreya Kunjibettu, Daniel McGreer, Masoomah Javidi Kishi, Hong Su, Mohamed Ziauddin, Mohamed Zait, Zhan Li, and Yuying Zhang. Automatic indexing in oracle. *Proceedings of the VLDB Endowment*, 18(12):4924–4937, 2025.
- [50] Hui Chen, Viet Luong, Lopamudra Mukherjee, and Vikas Singh. SimpleTM: A simple baseline for multivariate time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [51] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- [52] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.
- [53] China Computer Federation. Ccf recommended international conference list, 2022.

- [54] Chinese Academy of Sciences. Cas journal partition. <http://www.fenqubiao.com>, 2024.
- [55] LeeSun Choi. Introduction to logic. 1986.
- [56] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204. PMLR, 2015.
- [57] Dinesh Choudhary and Vijay Kumar. Software testing. *Journal of Computational Simulation and Modeling*, 1(1):1, 2011.
- [58] Clarivate Analytics. Journal citation reports. <https://clarivate.com/webometricsciencegroup/solutions/journal-citation-reports/>, 2024.
- [59] Jack Clark. Import ai 197: Facebook trains cyberpunk ai; chinese companies unite behind aibench evaluation system; how cloudflare uses ai. <https://jack-clark.net/2020/05/11/import-ai-197-facebook-trains-cyberpunk-ai-chinese-companies-unite-behind-aibench-evaluation-system-how-cloudflare-uses-ai/>, 2020. Accessed: 2025-10-26.
- [60] Edmund M Clarke. Model checking. In *International conference on foundations of software technology and theoretical computer science*, pages 54–56. Springer, 1997.
- [61] International Electrotechnical Commission. International electrotechnical vocabulary – part 192: Dependability. IEC 60050-192:2015, 2015.
- [62] Evaluation Research Society Standards Committee et al. Evaluation research society standards for program evaluation. *Standards for evaluation practice. New directions for program evaluation*, (15):7–19, 1982.
- [63] Computing Research and Education Association of Australasia. Core conference portal, 2024.
- [64] Conférence Générale des Poids et Mesures. 11th general conference on weights and measures, 1960.
- [65] Robert G Cowell, A Philip Dawid, Steffen L Lauritzen, and David J Spiegelhalter. *Probabilistic networks and expert systems*. Springer, 1999.
- [66] David Roxbee Cox. Planning of experiments. 1958.
- [67] Lee J Cronbach. Course improvement through evaluation. *Teachers college record*, 64(8):1–13, 1963.
- [68] Lee J Cronbach, Sueann Robinson Ambron, Sanford M Dornbusch, Robert D Hess, Robert C Hornik, Denis Charles Phillips, Decker F Walker, and Stephen S Weiner. *Toward reform of program evaluation*. JSTOR, 1980.

- [69] Randall Davis. Expert systems: Where are we? and where do we go from here? *AI magazine*, 3(2):3–3, 1982.
- [70] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–15, 09 1979.
- [71] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE.
- [72] Bureau International des Poids et Mesures. *The International System of Units (SI) —9th edition*. Bureau International des Poids and Measures (BIPM), Sèvres, France, 2019. Published May 2019; updated versions exist.
- [73] René Descartes and John Veitch. A discourse on method, 1977.
- [74] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
- [75] Jack J. Dongarra, Piotr Luszczek, and Antoine Petit. The linpack benchmark: past, present and future. *Concurrency & Computation Practice & Experience*, 15(9):803–820, 2010.
- [76] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- [77] Igor Douven. Abduction. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2025 edition, 2025.
- [78] Norman R Draper and Harry Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 1998.
- [79] Elliot W Eisner. On the uses of educational connoisseurship and criticism for evaluating classroom life. *Teachers College Record*, 78(3):1–11, 1977.
- [80] Junping Qiu et al. *Evaluation Science: Theory, Method and Practice*. Science Press, 1nd edition, 2010.
- [81] Matthew E Falagas, Vasilios D Kouranos, Ricardo Arencibia-Jorge, and Drosos E Karageorgopoulos. Comparison of scimago journal rank indicator with journal impact factor. *The FASEB journal*, 22(8):2623–2628, 2008.
- [82] Lawrence Fisher. Some new stock-market indexes. *The Journal of Business*, 39(1):191–225, 1966.

- [83] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, 222(594–604):309–368, 1922.
- [84] Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, 222:309–368, 1922.
- [85] Ronald A. Fisher. Statistical methods for research workers. 1925.
- [86] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.
- [87] Ronald Aylmer Fisher. *The design of experiments*. Springer, 1971.
- [88] Jay Wright Forrester. Industrial dynamics. *Journal of the Operational Research Society*, 48(10):1037–1041, 1997.
- [89] James Franck and Gustav Hertz. Über zusammenstöße zwischen elektronen und den molekülen des quecksilberdampfes und die ionisierungsspannung desselben. *Physikalische Blätter*, 23(7):294–301, 1967.
- [90] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.
- [91] Gordon Fraser, Franz Wotawa, and Paul E Ammann. Testing with model checkers: a survey. *Software Testing, Verification and Reliability*, 19(3):215–261, 2009.
- [92] Gottlob Frege. *Begriffsschrift: A Formula Language, Modeled upon that of Arithmetic, for Pure Thought*. Halle: Louis Nebert, 1879.
- [93] Mike Furr. Scale construction and psychometrics for social and personality psychology. *Scale Construction and Psychometrics for Social and Personality Psychology*, pages 1–160, 2011.
- [94] Eugene Garfield et al. The impact factor. *Current contents*, 25(20):3–7, 1994.
- [95] Michael R Genesereth and Nils J Nilsson. *Logical foundations of artificial intelligence*. Morgan Kaufmann, 2012.
- [96] Gene V Glass. *The growth of evaluation methodology*. Number 27. Laboratory of Educational Research, University of Colorado, 1969.
- [97] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.
- [98] Kurt Gödel. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für mathematik und physik*, 38(1):173–198, 1931.

- [99] Mario Gómez-Torrente. Tarski on logical consequence. *Notre Dame Journal of Formal Logic*, 37(1):125–151, 1996.
- [100] Weiwei Gong and Xu Zhou. A survey of sat solver. In *AIP Conference Proceedings*, volume 1836, page 020059. AIP Publishing LLC, 2017.
- [101] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
- [102] Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of econometrics*, 225(2):254–277, 2021.
- [103] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.
- [104] David J. Griffiths. *Introduction to Quantum Mechanics*. Pearson, 2 edition, 2005.
- [105] Egon G Guba and Yvonna S Lincoln. *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches*. Jossey-Bass, 1981.
- [106] Egon G Guba and Yvonna S Lincoln. *Fourth generation evaluation*. Sage, 1989.
- [107] BI Guo-Qiang. Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic type. *The Journal Neuroscience*, 18(24):10464–10472, 1988.
- [108] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.
- [109] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE intelligent systems*, 24(2):8–12, 2009.
- [110] James D Hamilton. *Time series analysis*. Princeton university press, 2020.
- [111] Gilbert H. Harman. The inference to the best explanation. *Philosophical Review*, 74(1):88–95, 1965.
- [112] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009.
- [113] Daniel M Hausman and James Woodward. Independence, invariance and the causal markov condition. *The British journal for the philosophy of science*, 50(4):521–583, 1999.
- [114] Thomas L Heath et al. *The Thirteen Books of the Elements, Vol. 3*, volume 3. Courier Corporation, 2013.

- [115] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Psychology press, 2005.
- [116] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.
- [117] John L Hennessy and David A Patterson. *Computer architecture: a quantitative approach*. Elsevier, 2011.
- [118] Boris Hennig. *Aristotle's Four Causes*. Peter Lang Verlag, New York, United States of America, 2019.
- [119] Miguel A. Hernan. *Causal Inference: What If*. Taylor & Francis, Boca Raton, 2024.
- [120] John A Hertz. *Introduction to the theory of neural computation*. Crc Press, 2018.
- [121] Björn Hibell, Barbro Andersson, Salme Ahlström, Olga Balakireva, Thoroddur Bjarnason, Anna Kokkevi, Mark Morgan, et al. The 1999 espad report. *Alcohol and other drug use among students in*, 26, 1997.
- [122] David Hilbert. *Grundlagen der geometrie*, volume 7. , 1913.
- [123] David Hilbert and Wilhelm Ackermann. *Grundzüge der theoretischen Logik*. Springer, 1934.
- [124] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [125] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
- [126] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.
- [127] Ernest R House. *Evaluating With Validity. Beverly Hills*. California: Sage Publications, 1980.
- [128] Colin Howson. Popper, prior probabilities, and inductive inference. *The British journal for the philosophy of science*, 38(2):207–224, 1987.
- [129] Rick H Hoyle. *Structural equation modeling: Concepts, issues, and applications*. Sage, 1995.

- [130] D. Hume. *An Enquiry Concerning Human Understanding*. Dover philosophical classics. Dover Publications, 2004.
- [131] Rob J Hyndman, George Athanasopoulos, et al. *Forecasting: principles and practice*, volume 2. OTexts Melbourne, Australia, 2014.
- [132] Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010.
- [133] Guido W Imbens and Donald B Rubin. *Causal inference for statistics, social, and biomedical sciences*. Cambridge university press, 2015.
- [134] OT Inan, P Tenaerts, SA Prindiville, HR Reynolds, DS Dizon, K Cooper-Arnold, M Turakhia, MJ Pletcher, KL Preston, HM Krumholz, et al. Digitizing clinical trials. *NPJ digital medicine*, 3(1):101, 2020.
- [135] International Software Testing Qualifications Board. Istqb glossary of terms used in software testing.
- [136] BSEN ISO and BRITISH STANDARD. Ergonomics of human-system interaction. *British Standards Institution*, 3, 2010.
- [137] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [138] Raj Jain. *The art of computer systems performance analysis*, volume 182. John Wiley & Sons Chichester, 1991.
- [139] Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.
- [140] JCGM Jcgm et al. Evaluation of measurement data—guide to the expression of uncertainty in measurement. *Int. Organ. Stand. Geneva ISBN*, 50:134, 2008.
- [141] Jean Jenkins and Susan Hubbard. History of clinical trials. In *Seminars in oncology nursing*, volume 7, pages 228–234, 1991.
- [142] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024.
- [143] Lizy Kurian John and Lieven Eeckhout. *Performance evaluation and benchmarking*. CRC Press, 2018.

- [144] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Upper Saddle River, N.J., 6 edition, 2007.
- [145] Karl G. Jöreskog. A general method for estimating a linear structural equation system. *ETS Research Bulletin Series*, 1970(2):i–41, 1970.
- [146] Raghu N Kacker. On quantity, value, unit, and other terms in the jcgim international vocabulary of metrology. *Measurement Science and Technology*, 32(12):125015, 2021.
- [147] Guoxin Kang, Wanling Gao, and Jianfeng Zhan. Evaluatology-driven artificial intelligence. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, page 100245, 2025.
- [148] Immanuel Kant. *Critique of Pure Reason*. Cambridge University Press, Cambridge, 1929.
- [149] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [150] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.
- [151] Johannes Kepler. *Astronomia Nova*. 1609.
- [152] Johannes Kepler. *Harmonices Mundi*. 1619.
- [153] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772):700–721, 1927.
- [154] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*, 2021.
- [155] Roger E Kirk. Experimental design. *Sage handbook of quantitative methods in psychology*, pages 23–45, 2009.
- [156] Les Kirkup and Robert B Frenkel. *An introduction to uncertainty in measurement: using the GUM (guide to the expression of uncertainty in measurement)*. Cambridge University Press, 2006.
- [157] Rex B Kline. *Principles and practice of structural equation modeling*. Guilford publications, 2023.

- [158] Michael E Knudson. A performance measurement and system evaluation project plan proposal. *ACM SIGMETRICS Performance Evaluation Review*, 13(1):20–31, 1985.
- [159] Andrey Nikolaevich Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933. Cited for the axiomatic definition of Probability, Random Variable, and Distribution Function.
- [160] Samuel Kounev, Klaus-Dieter Lange, and Joakim Von Kistowski. *Systems Benchmarking*. Springer, 2020.
- [161] Robert Kowalski. Logic programming. In *Handbook of the History of Logic*, volume 9, pages 523–569. Elsevier, 2014.
- [162] Ludwig Kraus, Ulf Guttormsson, Håkan Leifman, Sharon Arpa, Sabrina Molinaro, Karin Monshouwer, Marcis Trapencieris, Julian Vicente, Ársæll Már Arnarsson, Olga Balakireva, et al. *ESPAD report 2015: results from the European school survey project on alcohol and other drugs: results from the European school survey project on alcohol and other drugs*. Luxembourg: Publications Office of the European Union, 2016.
- [163] Alfgeir L. Kristjansson, Michael J. Mann, Jon Sigfusson, Ingibjorg E. Thorisdottir, John P. Allegrante, and Inga Dora Sigfusdottir. Development and Guiding Principles of the Icelandic Model for Preventing Adolescent Substance Use. *Health Promotion Practice*, 21(1):62–69, January 2020.
- [164] Alfgeir L. Kristjansson, Michael J. Mann, Jon Sigfusson, Ingibjorg E. Thorisdottir, John P. Allegrante, and Inga Dora Sigfusdottir. Implementing the Icelandic Model for Preventing Adolescent Substance Use. *Health Promotion Practice*, 21(1):70–79, January 2020.
- [165] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [166] Christoph Kröger. Evaluation: Definitions and concept. 1998). *Evaluation drug prevention in the European union. Scientific monograph series*, (2):61–66, 1998.
- [167] Sean P Lally. Henry cavendish and the density of the earth. *The Physics Teacher*, 37(1):34–37, 1999.
- [168] Antoine Lavoisier. *Traité Élémentaire de Chimie*. Cuchet, Paris, 1789.
- [169] John Law. Robust statistics—the approach based on influence functions, 1986.
- [170] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

- [171] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.
- [172] Adam J Lee and Marianne Winslett. Enforcing safety and consistency constraints in policy-based authorization systems. *ACM Transactions on Information and System Security (TISSEC)*, 12(2):1–33, 2008.
- [173] Tsung-Dao Lee and Chen Ning Yang. Question of parity conservation in weak interactions. *Physical Review*, 104(1):254, 1956.
- [174] E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, 2 edition, 1998.
- [175] David Lewis. *Counterfactuals*. Blackwell, Malden, Mass., 1973.
- [176] Loet Leydesdorff and Tobias Opthof. Scopus’s source normalized impact per paper (snip) versus a journal impact factor based on fractional counting of citations. *Journal of the American society for information science and technology*, 61(11):2365–2369, 2010.
- [177] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Forty-second International Conference on Machine Learning*, 2025.
- [178] Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. In *Thirty-eighth Conference on Neural Information Processing Systems*, 2024.
- [179] Peter Lipton. *Inference to the Best Explanation*. Routledge, 2nd edition, 2004.
- [180] Peter Lipton. Inference to the best explanation. *A Companion to the Philosophy of Science*, pages 184–193, 2017.
- [181] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.
- [182] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [183] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pages 11116–11140, 2024.

- [184] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.
- [185] Peter JF Lucas and Linda C Van Der Gaag. *Principles of expert systems*. Addison Wesley Longman, 1991.
- [186] Peter Machamer, Lindley Darden, and Carl F. Craver. Thinking about mechanisms. *Philosophy of Science*, 67(1):1–25, 2000.
- [187] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK, 2003.
- [188] John Leslie Mackie. *The Cement of the Universe: A Study of Causation*. Clarendon Press, Oxford, GB, 1974.
- [189] Melvin H Marx. Philosophy of natural science. *PsycCRITIQUES*, 12(2):95, 1967.
- [190] Sandra Mathison. *Encyclopedia of evaluation*. Sage publications, 2004.
- [191] William Anderson McCall. *How to experiment in education*. Macmillan, 1926.
- [192] Kenneth O McGraw and Seok P Wong. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30, 1996.
- [193] Gregor Mendel. Versuche uber pflanzen-hybriden. *Vorgelegt in den Sitzungen*, 1865.
- [194] Tim Menzies. Applications of abduction: knowledge-level modelling. *International journal of human-computer studies*, 45(3):305–335, 1996.
- [195] Albert Messiah. *Quantum mechanics*. Courier Corporation, 2014.
- [196] David Moher, Kenneth F Schulz, Douglas G Altman, and CONSORT Group*. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Annals of internal medicine*, 134(8):657–662, 2001.
- [197] Douglas C Montgomery. *Design and analysis of experiments*. John wiley & sons, 2017.
- [198] Sara Monti, Vittorio Grosso, Monica Todoerti, and Roberto Caporali. Randomized controlled trials and real-world data: differences and similarities to untangle literature data. *Rheumatology*, 57(Supplement_7):vii54–vii58, 2018.
- [199] Alhassan Mumuni, Fuseini Mumuni, and Nana Kobina Gerrar. A survey of synthetic data augmentation methods in computer vision. *arXiv preprint arXiv:2403.10075*, 2024.

- [200] Vikas Nagaraj. Automating test vector validation for silicon verification at scale. *International Journal of Engineering and Architecture (IJEa)*, 2(1):76–113, 2025.
- [201] Leland Gerson Neuberger. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685, 2003.
- [202] D Nevo, A Lewy, S Kugelmass, G Ben-Shakar, N Blass, RF Boruch, DJ Davis, B Nevo, D Nevo, P Tamir, et al. The evaluation of a multi-dimensional project. *Lewy, A.(et al.) Decision Oriented Evaluation In Education, International Science Services*, 1981.
- [203] David Nevo. The conceptualization of educational evaluation: An analytical review of the literature. *Review of Educational Research*, 53(1):117–128, Spring, 1983.
- [204] Allen Newell, John C Shaw, and Herbert A Simon. Report on a general problem solving program. In *IFIP congress*, volume 256, page 1959. Pittsburgh, PA, 1959.
- [205] Allen Newell, Herbert Alexander Simon, et al. *Human problem solving*, volume 104. Prentice-hall Englewood Cliffs, NJ, 1972.
- [206] Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A*, 236(767):333–380, 1937.
- [207] Jerzy Neyman and Egon S. Pearson. The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, 29:492–510, 1933.
- [208] Jerzy Neyman and Egon Sharpe Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.
- [209] Ngoc-Hai Nguyen, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy Chen, Min-Yen Kan, et al. Llms are biased towards output formats! systematically evaluating and mitigating output format bias of llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 299–330, 2025.
- [210] Trang Quynh Nguyen, Elizabeth L Ogburn, Ian Schmid, Elizabeth B Sarker, Noah Greifer, Ina M Koning, and Elizabeth A Stuart. Causal mediation analysis: From simple to more robust strategies for estimation of marginal natural (in) direct effects. *Statistics surveys*, 17:1, 2023.
- [211] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

- [212] Nils J. Nilsson. *Logic and Artificial Intelligence*. Morgan Kaufmann, 1991.
- [213] Nils J Nilsson. *Principles of artificial intelligence*. Morgan Kaufmann, 2014.
- [214] National Institute of Standards and Technology. Meter bar 27, n.d. Accessed: 2025-10-26.
- [215] US General Accounting Office. Assessing social program impact evaluations: A checklist approach, 1978.
- [216] OpenAI. Introducing GPT-4.1 in the API. <https://openai.com/index/gpt-4-1/>, 2025. Accessed: 2025-09-05.
- [217] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.
- [218] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab, 1999.
- [219] Michael Quinn Patton. *Utilization-focused evaluation*. Beverly Hills. Ca: Sage, 1978.
- [220] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
- [221] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 1 edition, 2000. Cited for the formal definition of Causality and the do-operator.
- [222] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.
- [223] Judea Pearl. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392. 2022.
- [224] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic books, 2018.
- [225] Karl Pearson. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London, Series A: Containing Papers of a Mathematical or Physical Character*, (187):253–318, 12 1896.
- [226] Charles S. Peirce. *Collected Papers of Charles Sanders Peirce, Vol. 1–8*. Harvard University Press, 1931.

- [227] Charles Sanders Peirce. *Collected papers of charles sanders peirce*, volume 5. Harvard University Press, 1934.
- [228] Rupert Pennick. The history of the metric system. *Journal of Geomancy*, 2(3):61–65, April 1978.
- [229] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017.
- [230] Ahti-Veikko Pietarinen. Abduction and diagrams. *Logic Journal of the IGPL*, 29(4):447–468, 2021.
- [231] David Poole. A methodology for using a default and abductive reasoning system. *Artificial Intelligence*, 31(1):125–157, 1990.
- [232] David Poole. Probabilistic horn abduction and bayesian networks. *Artificial intelligence*, 64(1):81–129, 1993.
- [233] M Provus. Evaluation as public policy. *Curriculum Theory Network*, 3(8-9):33–44, 1972.
- [234] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392, 2016.
- [235] Sreeram V Ramagopalan, Alex Simpson, and Cormac Sammon. Can real-world data really replace randomised clinical trials? *BMC medicine*, 18(1):1–2, 2020.
- [236] Charles S Reichardt. Experimental and quasi-experimental designs for generalized causal inference, 2002.
- [237] Olav Reiersøl. *Confluence analysis by means of instrumental sets of variables*. PhD thesis, Almqvist & Wiksell, 1945.
- [238] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.
- [239] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.
- [240] LS Robson, HS Shannon, LM Goldenhar, and AR Hale. Quasi-experimental and experimental designs: more powerful evaluation designs. *Guide to Evaluating the Effectiveness of Strategies for Preventing Work Injuries*. Department of Health and Human Services: Cincinnati, OH, pages 29–42, 2001.

- [241] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- [242] Peter H Rossi, Mark W Lipsey, and Gary T Henry. *Evaluation: A systematic approach*. Sage publications, 2018.
- [243] PH Rossi and HE Freeman. *Evaluation: A systematic approach (pp. 375-415)*. Newbury Park, CA: Sage, 1989.
- [244] Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*. John wiley & sons, 2003.
- [245] Donald Rubin. Estimating causal effects of treatments in experimental and observational studies. *ETS Research Bulletin Series*, 1972(2):i–31, 1972.
- [246] Donald Rubin. Estimating causal effects of treatments in experimental and observational studies. *Ets research bulletin series*, 1972(2):i–31, 1972.
- [247] Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.
- [248] Donald B Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of educational Psychology*, 66(5):688, 1974.
- [249] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.
- [250] Donald B Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328, 1979.
- [251] Donald B Rubin. [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.
- [252] Donald B Rubin. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31(2):161–170, 2004.
- [253] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American statistical Association*, 100(469):322–331, 2005.
- [254] Donald B Rubin. Causal inference through potential outcomes and principal stratification: application to studies with “censoring” due to death. *Statistical Science*, pages 299–309, 2006.
- [255] Donald B Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, 26(1):20–36, 2007.

- [256] Vera C. Rubin, W. Kent Ford, and N. Thonnard. Extended rotation curves of high-luminosity spiral galaxies. iv. systematic dynamical properties, sa to sc. *The Astrophysical Journal Letters*, 225:L107–L111, 1978.
- [257] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
- [258] Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7), 2018.
- [259] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting causal associations in large nonlinear time series datasets. *arXiv preprint arXiv:1702.07007*, 2017.
- [260] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.
- [261] Bertrand Russell. *History of Western Philosophy*. Routledge, 2004.
- [262] Bertrand Russell and Alfred North Whitehead. *Principia Mathematica*. Cambridge University Press, 1910.
- [263] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 4th edition, 2021.
- [264] Wesley C. Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, 1984.
- [265] Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.
- [266] John D Sargan. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the econometric society*, pages 393–415, 1958.
- [267] Henry Scheffe. *The analysis of variance*. John Wiley & Sons, 1999.
- [268] Daniel P Scheitrum, Colin A Carter, and Cesar Revoredo-Giha. Wti and brent futures pricing structure. *Energy Economics*, 72:462–469, 2018.
- [269] Erwin Schrödinger. Die gegenwärtige situation in der quantenmechanik. *Naturwissenschaften*, 23(50):844–849, 1935.
- [270] M Scriven. The pathway comparison model of evaluation, 1972.

- [271] Michael Scriven. Prose and cons about goal-free evaluation. *Evaluation Comment*, 3(4).
- [272] Michael Scriven. The methodology of evaluation, social science education consortium. publication 110, . 1966.
- [273] Michael Scriven. Maximizing the power of causal investigations: The modus operandi method. *Evaluation studies review annual*, 1:101–118, 1976.
- [274] Michael Scriven. *Evaluation thesaurus*. Sage, 1991.
- [275] Michael Scriven. Evaluation as a discipline. *Studies in Educational Evaluation*, 20(1):147–166, 1994.
- [276] Michael Scriven. The logic of evaluation. *Dissensus and the search for common ground*, 1:1–16, 2007.
- [277] Michael Scriven. Roadblocks to recognition and revolution. *American Journal of Evaluation*, 37(1):27–44, 2016.
- [278] Jenn W Sellers, Camelia M Mihaescu, Kassa Ayalew, Phillip D Kronstein, Bei Yu, Yang-Min Ning, Miguel Rodriguez, LaKisha Williams, and Ni A Khin. Descriptive analysis of good clinical practice inspection findings from us food and drug administration and european medicines agency. *Therapeutic Innovation & Regulatory Science*, 56(5):753–764, 2022.
- [279] William R Shadish, Thomas D Cook, and Laura C Leviton. *Foundations of program evaluation: Theories of practice*. Sage, 1991.
- [280] Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts, 2024.
- [281] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.
- [282] Michael Sipser et al. *Introduction to the Theory of Computation*, volume 2. Thomson course technology Boston, 2006.
- [283] Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4195–4206, 2025.
- [284] SPEC. SPEC CPU Benchmark Suite. <https://www.spec.org/benchmarks.html#cpu>.

- [285] SPEC. SPEC CPU2017, 2017. Available at <https://www.spec.org/cpu2017>.
- [286] SPEC. SPEC CPU2026, 2026. Available at <https://www.spec.org/cpu2026>.
- [287] Jerzy Splawa-Neyman, Dorota M Dabrowska, and Terrence P Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.
- [288] Robert E Stake. The countenance of educational evaluation. *Teachers college record*, 68(7):1–15, 1967.
- [289] Robert E Stake. *Evaluating the arts in education: A responsiveness approach*. Merrill Publishing Co, Columbus, Ohio, 1975.
- [290] Robert E Stake. Evaluating educational programmes: The need and the response. 1976.
- [291] Robert E Stake. Setting standards for educational evaluators. *Evaluation News*, 2(2):148–152, 1981.
- [292] Daren S Starnes, Dan Yates, and David S Moore. *The practice of statistics*. Macmillan, 2010.
- [293] Stanley Smith Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.
- [294] James Stewart. *Single variable calculus: Concepts and contexts*. Cengage Learning, 2018.
- [295] Harald O Stolberg, Geoffrey Norman, and Isabelle Trop. Randomized controlled trials. *American Journal of Roentgenology*, 183(6):1539–1544, 2004.
- [296] Daniel L Stufflebeam. Evaluation as enlightenment for decision-making. 1968.
- [297] Daniel L Stufflebeam. The relevance of the cipp evaluation model for educational accountability. 1971.
- [298] Daniel L Stufflebeam, Phi Delta Kappa, and Bloomington Ind. *Educational evaluation [and] decision making*. FE Peacock Itasca, IL, 1971.
- [299] Daniel L Stufflebeam and George F Madaus. The standards for evaluation of educational programs, projects, and materials: A description and summary. In *Evaluation models: Viewpoints on educational and human services evaluation*, pages 395–404. Springer, 1983.
- [300] DL Stufflebeam. Meta-evaluation (occasional paper no. 3). *Kalamazoo: Western Michigan University, December*, 1974.
- [301] William Stukeley. Memoirs of sir isaac newton’s life, 1752. Folio 15r.

- [302] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.
- [303] G Kasten Tallmadge. The joint dissemination review panel ideabook. 1977.
- [304] Peiwang Tang and Weitai Zhang. Unlocking the power of patch: Patch-based mlp for long-term time series forecasting, 2024.
- [305] Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.
- [306] Jaime A Teixeira da Silva. Citescor: Advances, evolution, applications, and limitations. *Publishing Research Quarterly*, 36(3):459–468, 2020.
- [307] Jacqueline K Telford. A brief introduction to design of experiments. *Johns Hopkins apl technical digest*, 27(3):224–232, 2007.
- [308] Ambler Thompson and Barry N Taylor. Use of the international system of units (si). *NIST Special Publication, Gaithersburg*, 2008.
- [309] Krishnaiyan Thulasiraman and Madisetti NS Swamy. *Graphs: theory and algorithms*. John Wiley & Sons, 2011.
- [310] Zhengshuyuan Tian, Wanling Gao, Chuanxin Lan, Chenxi Wang, Lei Wang, Guoxin Kang, Zhengxin Yang, Yunyou Huang, Xuehai Hong, and Jianfeng Zhan. Beyond benchmarks: Toward causally faithful evaluation of large language models. *Forty-third International Conference on Machine Learning (ICML)*, 2026.
- [311] Stephen E Toulmin. *The uses of argument*. Cambridge university press, 2003.
- [312] Dmitry Tsarkov and Ian Horrocks. Fact++ description logic reasoner: System description. In *International joint conference on automated reasoning*, pages 292–297. Springer, 2006.
- [313] Alan M Turing. Computing machinery and intelligence. In *Parsing the Turing test: Philosophical and methodological issues in the quest for the thinking computer*, pages 23–65. Springer, 2007.
- [314] Ralph W Tyler. *Basic principles of curriculum and instruction*. University of Chicago Pres, 1950.
- [315] Jodie B Ullman and Peter M Bentler. Structural equation modeling. *Handbook of psychology, second edition*, 2, 2012.
- [316] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

- [317] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.
- [318] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [319] Juan D Velásquez and Vasile Palade. A knowledge base for the maintenance of knowledge extracted from web data. *Knowledge-Based Systems*, 20(3):238–248, 2007.
- [320] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.
- [321] Chenxi Wang, Lei Wang, Wanling Gao, Fanda Fan, Guoxin Kang, Hongxiao Li, Yuchen Su, and Jianfeng Zhan. Inference of component effect on system performance. *arXiv preprint arXiv:2605.26643*, 2026.
- [322] Chenxi Wang, Lei Wang, Wanling Gao, Yikang Yang, Yutong Zhou, and Jianfeng Zhan. Achieving consistent and comparable cpu evaluation outcomes. *arXiv preprint arXiv:2411.08494*, 2024.
- [323] Chenxi Wang, Lei Wang, Wanling Gao, Yikang Yang, Yutong Zhou, and Jianfeng Zhan. Achieving consistent and comparable cpu evaluation outcomes. *Technical Report, International Open Benchmark Council*, 2024.
- [324] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International conference on learning representations*, volume 2024, pages 38626–38652, 2024.
- [325] Xiaorui Wang, Fanda Fan, Chenxi Wang, Yuxuan Yang, Rui Tang, Kuoyu Gao, simiao pang, Yuanfeng Shang, Zhipeng Liu, Wanling Gao, Lei Wang, and Jianfeng Zhan. CombinationTS: A modular framework for understanding time-series forecasting models. In *Forty-third International Conference on Machine Learning*, 2026.
- [326] Alfred North Whitehead and Bertrand Arthur Russell. *Principia Mathematica*. Cambridge University Press, 2 edition, 1927. Cited for formal definitions of Variable and Function.
- [327] James A Whittaker. What is software testing? And why is it so hard? *IEEE software*, 17(1):70–79, 2000.
- [328] Robert Wild, Markus Nötzold, Malcolm Simpson, Thuy Dung Tran, and Roland Wester. Tunnelling measured in a very slow ion–molecule reaction. *Nature*, 615(7952):425–429, 2023.

- [329] Philip Green Wright. *The tariff on animal and vegetable oils*. Number 26. Macmillan, 1928.
- [330] Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20(7):557–585, 1921.
- [331] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson. Experimental test of parity conservation in beta decay. *Phys. Rev.*, 105:1413–1415, Feb 1957.
- [332] Chien-Shiung Wu and Irving Shakhnov. The angular correlation of scattered annihilation radiation. *Physical Review*, 77(1):136, 1950.
- [333] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [334] Hugh D Young, Roger A Freedman, and Lewis A Ford. *University physics with modern physics*. 2020.
- [335] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? 2023.
- [336] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.
- [337] Jianfeng Zhan. Call for establishing benchmark science and engineering. *Bench-Council Transactions on Benchmarks, Standards and Evaluations*, 1(1):100012, 2021.
- [338] Jianfeng Zhan. Five axioms of things. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, page 100184, 2024.
- [339] Jianfeng Zhan, Lei Wang, Wanling Gao, Hongxiao Li, Chenxi Wang, Yunyou Huang, Yatao Li, Zhengxin Yang, Guoxin Kang, Chunjie Luo, Hainan Ye, Shaopeng Dai, and Zhifei Zhang. Evaluatology: The science and engineering of evaluation. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 4(1):100162, 2024.
- [340] Jianfeng Zhan, Lei Wang, Wanling Gao, Chenxi Wang, Hongxiao Li, Fanda Fan, and Guoxin Kang. *Evaluatology: The Science of Uncovering the Effects*. Bench-Council Press, Hong Kong, version 1.0, first edition, 2025.
- [341] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

- [342] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pages 27268–27286. PMLR, 2022.