

## References

- J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL <https://arxiv.org/abs/2303.08774>
- E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=0g0X4H8yN4I>
- E. Akyürek, B. Wang, Y. Kim, and J. Andreas. In-context language learning: Architectures and algorithms. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=3Z9CRr5srL>
- L. Bottou and B. Schölkopf. The fiction machine, April 2025. URL <https://www.siam.org/publications/siam-news/articles/the-fiction-machine/>
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf)
- L. Charleux, E. Roux, T. Goyallon, G. Feverati, and P. Nagorny. Lotka-volterra equations, 2018. URL [https://scientific-python.readthedocs.io/en/latest/notebooks\\_rst/3\\_Ordinary\\_Differential\\_Equations/02\\_Examples/Lotka\\_Volterra\\_model.html](https://scientific-python.readthedocs.io/en/latest/notebooks_rst/3_Ordinary_Differential_Equations/02_Examples/Lotka_Volterra_model.html)
- K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL <https://aclanthology.org/D14-1179/>
- D. Dai, Y. Sun, L. Dong, Y. Hao, S. Ma, Z. Sui, and F. Wei. Why can GPT learn in-context? language models implicitly perform gradient descent as meta-optimizers. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023. URL <https://openreview.net/forum?id=fzbHRjAd8U>
- B. de Finetti. Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Accademia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Matematiche e Naturale*, 4:251–299, 1931. URL <http://www.brunodefinetti.it/Opere/funzioneCaratteristica.pdf>
- G. Deutch, N. Magar, T. Natan, and G. Dar. In-context learning and gradient descent revisited. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1017–1028, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.58. URL <https://aclanthology.org/2024.naacl-long.58/>
- B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. ISSN 00905364, 21688966. URL <http://www.jstor.org/stable/2958830>
- N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL <https://transformer-circuits.pub/2021/framework/index.html>
- J. L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. URL [https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402\\_1](https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402_1)

368 F. Falck, Z. Wang, and C. C. Holmes. Is in-context learning in large language models bayesian? A  
369 martingale perspective. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scar-  
370 lett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine*  
371 *Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12784–12805. PMLR,  
372 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/falck24a.html>.

373 S. Garg, D. Tsipras, P. Liang, and G. Valiant. What can transformers learn in-context? a case  
374 study of simple function classes. In *Proceedings of the 36th International Conference on Neural*  
375 *Information Processing Systems, NIPS '22*, Red Hook, NY, USA, 2022. Curran Associates Inc.  
376 ISBN 9781713871088. URL <https://openreview.net/pdf?id=f1NZJ2e0et>.

377 M. Grendar. Entropy and effective support size. *Entropy*, 8(3):169–174, 2006. ISSN 1099-4300. doi:  
378 10.3390/e8030169. URL <https://www.mdpi.com/1099-4300/8/3/169>.

379 S. Guo, J. B. Wildberger, and B. Schölkopf. Out-of-variable generalisation for discriminative  
380 models. In *The Twelfth International Conference on Learning Representations*, 2023. URL  
381 <https://openreview.net/pdf?id=zwMfg9PfpS>.

382 S. Guo, V. Tóth, B. Schölkopf, and F. Huszár. Causal de finetti: On the identification of invariant  
383 causal structure in exchangeable data, 2024a. URL <https://arxiv.org/abs/2203.15756>.

384 S. Guo, C. Zhang, K. Mohan, F. Huszár, and B. Schölkopf. Do finetti: On causal effects for  
385 exchangeable data, 2024b. URL <https://arxiv.org/abs/2405.18836>.

386 M. A. Hernan. *Causal Inference: What If*. Taylor & Francis, Boca Raton, 2024. URL <https://miguelhernan.org/whatifbook>.

387  
388 S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780,  
389 1997. URL <https://www.bioinf.jku.at/publications/older/2604.pdf>.

390 D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun,  
391 editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA,*  
392 *USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL [http://arxiv.org/abs/1412.](http://arxiv.org/abs/1412.6980)  
393 [6980](http://arxiv.org/abs/1412.6980).

394 A. Klenke. *Probability Theory: A Comprehensive Course*. Springer, 2008. URL [https://link.](https://link.springer.com/book/10.1007/978-1-84800-048-3)  
395 [springer.com/book/10.1007/978-1-84800-048-3](https://link.springer.com/book/10.1007/978-1-84800-048-3).

396 M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural*  
397 *information processing systems*, 30, 2017. URL [https://proceedings.neurips.cc/paper\\_](https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf)  
398 [files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf).

399 J. Li, L. Yu, and A. Ettinger. Counterfactual reasoning: Testing language models’ understanding of  
400 hypothetical scenarios. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the*  
401 *61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*,  
402 pages 804–815, Toronto, Canada, July 2023. Association for Computational Linguistics. doi:  
403 10.18653/v1/2023.acl-short.70. URL <https://aclanthology.org/2023.acl-short.70/>.

404 X. Li, T.-K. L. Wong, R. T. Q. Chen, and D. Duvenaud. Scalable gradients for stochastic differential  
405 equations. *International Conference on Artificial Intelligence and Statistics*, 2020. URL [https:](https://proceedings.mlr.press/v108/li20i/li20i.pdf)  
406 [/proceedings.mlr.press/v108/li20i/li20i.pdf](https://proceedings.mlr.press/v108/li20i/li20i.pdf).

407 L. Lorch, A. Krause, and B. Schölkopf. Causal modeling with stationary diffusions. In S. Dasgupta,  
408 S. Mandt, and Y. Li, editors, *Proceedings of The 27th International Conference on Artificial*  
409 *Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1927–  
410 1935. PMLR, 02–04 May 2024. URL [https://proceedings.mlr.press/v238/lorch24a.](https://proceedings.mlr.press/v238/lorch24a.html)  
411 [html](https://proceedings.mlr.press/v238/lorch24a.html).

412 K. Lorenz. *Die Rückseite des Spiegels : Versuch einer Naturgeschichte menschlichen Erkennens*.  
413 Piper, München [u.a, 2. aufl. edition, 1973. ISBN 3492020305.

414 A. J. Lotka. Contribution to the theory of periodic reactions. *The Journal of Physical Chem-*  
415 *istry*, 14(3):271–274, 03 1910. doi: 10.1021/j150111a004. URL [https://doi.org/10.1021/](https://doi.org/10.1021/j150111a004)  
416 [j150111a004](https://doi.org/10.1021/j150111a004).

417 C. Lu, B. Huang, K. Wang, J. M. Hernández-Lobato, K. Zhang, and B. Schölkopf. Sample-efficient  
418 reinforcement learning via counterfactual-based data augmentation. In *Offline Reinforcement*  
419 *Learning - Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS)*,  
420 2020. URL <https://offline-rl-neurips.github.io/pdf/34.pdf>

421 G. Maruyama. Continuous markov processes and stochastic equations. *Rendiconti del Circolo*  
422 *Matematico di Palermo*, 4(1):48–90, 1955. doi: 10.1007/BF02846028. URL [https://doi.org/](https://doi.org/10.1007/BF02846028)  
423 [10.1007/BF02846028](https://doi.org/10.1007/BF02846028)

424 T. Mesnard, T. Weber, F. Viola, S. Thakoor, A. Saade, A. Harutyunyan, W. Dabney, T. S. Stepleton,  
425 N. Heess, A. Guez, E. Moulines, M. Hutter, L. Buesing, and R. Munos. Counterfactual credit  
426 assignment in model-free reinforcement learning. In M. Meila and T. Zhang, editors, *Proceedings*  
427 *of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine*  
428 *Learning Research*, pages 7654–7664. PMLR, 18–24 Jul 2021. URL [https://proceedings](https://proceedings.mlr.press/v139/mesnard21a.html)  
429 [mlr.press/v139/mesnard21a.html](https://proceedings.mlr.press/v139/mesnard21a.html)

430 J. Mooij, D. Janzing, and B. Schölkopf. From ordinary differential equations to structural causal  
431 models: the deterministic case. In A. Nicholson and P. Smyth, editors, *Proceedings of the Twenty-*  
432 *Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence*, pages 440–448,  
433 Corvallis, OR, 2013. AUAI Press. URL [http://www.is.tuebingen.mpg.de/fileadmin/](http://www.is.tuebingen.mpg.de/fileadmin/user_upload/files/publications/2013/MooijJS2013-uai.pdf)  
434 [user\\_upload/files/publications/2013/MooijJS2013-uai.pdf](http://www.is.tuebingen.mpg.de/fileadmin/user_upload/files/publications/2013/MooijJS2013-uai.pdf)

435 C. Olsson, N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai,  
436 A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones,  
437 J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and  
438 C. Olah. In-context learning and induction heads, 2022. URL [https://arxiv.org/abs/2209](https://arxiv.org/abs/2209.11895)  
439 [11895](https://arxiv.org/abs/2209.11895)

440 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin,  
441 N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Te-  
442 jani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. *PyTorch: an imperative*  
443 *style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY,  
444 USA, 2019. URL [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/](https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf)  
445 [bdbca288fee7f92f2bfa9f7012727740-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf)

446 J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition,  
447 2009. URL <https://bayes.cs.ucla.edu/B00K-2K/>

448 J. Peters, D. Janzing, and B. Schölkopf. Causal inference on discrete data using additive noise models.  
449 *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2436–2450, 2011. URL  
450 <https://ieeexplore.ieee.org/document/5740928>

451 J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive  
452 noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014. URL [https:](https://jmlr.org/papers/volume15/peters14a/peters14a.pdf)  
453 [//jmlr.org/papers/volume15/peters14a/peters14a.pdf](https://jmlr.org/papers/volume15/peters14a/peters14a.pdf)

454 J. Peters, S. Bauer, and N. Pfister. *Causal Models for Dynamical Systems*, page 671–690. Association  
455 for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL  
456 <https://doi.org/10.1145/3501714.3501752>

457 L. Qin, A. Bosselut, A. Holtzman, C. Bhagavatula, E. Clark, and Y. Choi. Counterfactual story  
458 reasoning and generation. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019*  
459 *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint*  
460 *Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053, Hong Kong,  
461 China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1509. URL  
462 <https://aclanthology.org/D19-1509/>

463 A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsuper-  
464 vised multitask learners. 2019. URL [https://cdn.openai.com/better-language-models/](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf)  
465 [language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf) OpenAI Technical Re-  
466 port.

- 467 S. Ravfogel, A. Svete, V. Snæbjarnarson, and R. Cotterell. Gumbel counterfactual generation from  
468 language models. In *The Thirteenth International Conference on Learning Representations*, 2025.  
469 URL <https://openreview.net/forum?id=TUC0ZT2zIQ>.
- 470 P. Reizinger, S. Guo, F. Huszár, B. Schölkopf, and W. Brendel. Identifiable exchangeable mechanisms  
471 for causal structure and representation learning. In *The Thirteenth International Conference on*  
472 *Learning Representations*, 2025. URL <https://openreview.net/forum?id=k03mB41vyM>.
- 473 A. Sauer and A. Geiger. Counterfactual generative networks. In *International Conference on Learning*  
474 *Representations*, 2021. URL <https://openreview.net/forum?id=BXewfAYMmJw>.
- 475 B. Schölkopf. Causality for machine learning. In *Probabilistic and causal inference: The works*  
476 *of Judea Pearl*, pages 765–804. 2022. URL [https://dl.acm.org/doi/10.1145/3501714](https://dl.acm.org/doi/10.1145/3501714.3501755)  
477 [3501755](https://dl.acm.org/doi/10.1145/3501714.3501755).
- 478 C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27  
479 (3):379–423, 1948. URL [https://people.math.harvard.edu/~ctm/home/text/others/](https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf)  
480 [shannon/entropy/entropy.pdf](https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf).
- 481 V. Southgate and A. Vernetti. Belief-based action prediction in preverbal infants. *Cogni-*  
482 *tion*, 130(1):1–10, 2014. URL [https://www.sciencedirect.com/science/article/pii/](https://www.sciencedirect.com/science/article/pii/S0010027713001650?via%3Dihub)  
483 [S0010027713001650?via%3Dihub](https://www.sciencedirect.com/science/article/pii/S0010027713001650?via%3Dihub).
- 484 N. Tandon, B. Dalvi, K. Sakaguchi, P. Clark, and A. Bosselut. WIQA: A dataset for “what if...”  
485 reasoning over procedural text. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the*  
486 *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*  
487 *Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6076–6085, Hong  
488 Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1629.  
489 URL <https://aclanthology.org/D19-1629/>.
- 490 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polo-  
491 sukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural*  
492 *Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA, 2017. Curran  
493 Associates Inc. ISBN 9781510860964. URL [https://proceedings.neurips.cc/paper\\_](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)  
494 [files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf).
- 495 J. von Oswald, E. Niklasson, E. Randazzo, J. a. Sacramento, A. Mordvintsev, A. Zhmoginov,  
496 and M. Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of*  
497 *the 40th International Conference on Machine Learning, ICML’23. JMLR.org*, 2023. URL  
498 <https://proceedings.mlr.press/v202/von-oswald23a/von-oswald23a.pdf>.
- 499 S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black  
500 box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017. URL [https://arxiv](https://arxiv.org/pdf/1711.00399)  
501 [org/pdf/1711.00399](https://arxiv.org/pdf/1711.00399).
- 502 S. M. Xie, A. Raghunathan, P. Liang, and T. Ma. An explanation of in-context learning as implicit  
503 bayesian inference, 2022. URL <https://arxiv.org/abs/2111.02080>.
- 504 H. Yan, L. Kong, L. Gui, Y. Chi, E. Xing, Y. He, and K. Zhang. Counterfactual generation with  
505 identifiability guarantees. In *Thirty-seventh Conference on Neural Information Processing Systems*,  
506 2023. URL <https://openreview.net/forum?id=cslnCXE9XA>.
- 507 N. Ye and H. Namkoong. Exchangeable sequence models quantify uncertainty over latent concepts,  
508 2024. URL <https://arxiv.org/abs/2408.03307>.
- 509 L. Zhang, R. T. McCoy, T. R. Sumers, J.-Q. Zhu, and T. L. Griffiths. Deep de finetti: Recovering topic  
510 distributions from large language models, 2023. URL <https://arxiv.org/abs/2312.14226>.