This document is a summary of all the papers discussed in the course AI for Business Research offered at CUHK Business School in Spring 2025.

0. What's News in AI

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... & He, Y. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*. [Link to paper] [Link to GitHub]

Kevin A. Bryan. (2025). Revisiting the McKinley Tariff of 1890 through the Lens of Modern Trade Theory. [Link to paper]

Kevin A. Bryan. (2025). Information Frictions and Innovation: A Formal Theory. [Link to paper]

Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., ... & Hashimoto, T. (2025). s1: Simple test-time scaling. arXiv preprint arXiv:2501.19393. [Link to paper] [Link to GitHub]

El-Kishky, A., Wei, A., Saraiva, A., Minaev, B., Selsam, D., Dohan, D., ... & Zhou, W. (2025). Competitive Programming with Large Reasoning Models. arXiv preprint arXiv:2502.06807. [Link to paper]

Starace, G., Jaffe, O., Sherburn, D., Aung, J., Chan, J. S., Maksin, L., ... & Patwardhan, T. (2025). PaperBench: Evaluating AI's Ability to Replicate AI Research. *arXiv preprint arXiv:2504.01848*. [Link to paper] [Link to GitHub]

1. Introduction

Cohen M, Zhang R, Jiao K. (2022) Data aggregation and demand prediction. *Operations Research*, 70(5): 2597-2618. [Link to paper] [Link to GitHub]

Ye, Z., Zhang, Z, Zhang, D. J., Zhang, H., Zhang, R. (2023) Deep Learning Based Causal Inference for Large-Scale Combinatorial Experiments: Theory and Empirical Evidence, working paper and EC'23. [Link to paper] [Link to GitHub]

Ye, Z., Zhang, D. J., Zhang, H., Zhang, R., Chen, X., and Xu, Z. (2023) Cold start to improve market thickness on online advertising platforms: Data-driven algorithms and field experiments. Management Science, 69(7), 3838-3860. [Link to paper] [Link to GitHub]

Zhang, X., Sun, C., Zhang, R., and Goh, K-Y (2024) The Value of AI-Generated Metadata for UGC Platforms: Evidence from a Large-scale Field Experiment, *working paper* and *CIST 2024*. [Link to paper]

Jeremy Yang; , Juanjuan Zhang; , Yuhan Zhang (2024) Engagement That Sells: Influencer Video Advertising on TikTok. *Marketing Science* 0(0). [Link to paper]

Kelly, Bryan T. and Xiu, Dacheng, Financial Machine Learning (July 1, 2023). Available at SSRN: [Link to the paper]

Allon, G., Chen, D., Jiang, Z., & Zhang, D. (2023). Machine learning and prediction errors in causal inference. The Wharton School Research Paper Forthcoming. [Link to the paper]

Battaglia, L., Christensen, T., Hansen, S., & Sacher, S. (2024). Inference for regression with variables generated from unstructured data. [Link to the paper]

Tian, Z., Dew, R., & Iyengar, R. (2024). Mega or Micro? Influencer Selection Using Follower Elasticity. *Journal of Marketing Research*, *61*(3), 472-495. [Link to paper]

Jeremy Yang, Dean Eckles, Paramveer Dhillon, Sinan Aral (2023) Targeting for Long-Term Outcomes. *Management Science* 70(6):3841-3855. [Link to paper]

Jacob Feldman, Dennis J. Zhang, Xiaofei Liu, Nannan Zhang (2021) Customer Choice Models vs. Machine Learning: Finding Optimal Product Displays on Alibaba. *Operations Research* 70(1):309-328. [Link to paper] [Link to GitHub]

Xiao Liu (2022) Dynamic Coupon Targeting Using Batch Deep Reinforcement Learning: An Application to Livestream Shopping. *Marketing Science* 42(4):637-658. [Link to paper]

Shakked Noy, Whitney Zhang ,Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381,187-192(2023). [Link to paper]

Anja Lambrecht, Catherine Tucker (2019) Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science* 65(7):2966-2981. [Link to paper]

Tetsuya Kaji, Elena Manresa, Guillaume Pouliot, An Adversarial Approach to Structural Estimation, *Econometrica*, 10.3982/ECTA18707, 91, 6, (2041-2063), (2023). [Link to paper] [Link to GitHub]

Yanhao (Max) Wei; , Zhenling Jiang (2024) Estimating Parameters of Structural Models Using Neural Networks. Marketing Science 44(1):102-128. [Link to paper]

Ishai Menache, Konstantina Mellou, Marco Molinaro, Zijie Zhou, "Foundational Model for LLM Inference: Online Batching and Scheduling". working paper

Patrick Jaillet, Jiashio Jiang, Chara Podimata, Zijie Zhou, Fundamental Modeling for LLM Inference withExploding KV Cache Demands.

2. Prediction

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review*, 105 (5): 491–95. [Link to paper]

Neal Jean et al. ,Combining satellite imagery and machine learning to predict poverty. *Science* 353,790-794(2016). [Link to paper] [Link to GitHub]

Binsbergen, J. H., Bryzgalova, S., Mukhopadhyay, M., & Sharma, V. (2024). (Almost) 200 years of news-based economic sentiment (NBER Working Paper No. 32026). *National Bureau of Economic Research*. [Link to paper]

Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. 2015. "Machine Learning Methods for Demand Estimation." *American Economic Review*, 105 (5): 481–85. [Link to paper]

Cohen M, Zhang R, Jiao K. (2022) Data aggregation and demand prediction. *Operations Research*, 70(5): 2597-2618. [Link to paper] [Link to GitHub]

Cui, R., Gallino, S., Moreno, A., & Zhang, D. J. (2018). The operational value of social media information. *Production and operations management*, 27(10), 1749-1769. [Link to paper]

Farias, V. F., & Li, A. A. (2019). Learning preferences with side information. *Management Science*, 65(7), 3131-3149. [Link to paper]

Peng, Jing and Liang, Chen. 2023. "On the Differences Between View-Based and Purchase-Based Recommender Systems," *MIS Quarterly*, (47: 2) pp.875-900. [Link to paper]

Zhan, R., Pei, C., Su, Q., Wen, J., Wang, X., Mu, G., ... & Gai, K. (2022, August). Deconfounding duration bias in watch-time prediction for video recommendation. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 4472-4481). [Link to paper]

Covington, P., Adams, J., & Sargin, E. (2016, September). Deep neural networks for youtube recommendations. In Proceedings of the 10th ACM conference on recommender systems (pp. 191-198). [Link to paper]

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5), 2223-2273. [Link to paper]

Cao, K., Xia, Y., Yao, J., Han, X., Lambert, L., Zhang, T., ... & Lu, J. (2023). Large-scale pancreatic cancer detection via non-contrast CT and deep learning. *Nature medicine*, 29(12), 3033-3043. [Link to paper]

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan, Human Decisions and Machine Predictions, *The Quarterly Journal of Economics*, Volume 133, Issue 1, February 2018, Pages 237–293. [Link to paper]

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., ... & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, 595(7866), 181-188. [Link to paper]

3. Intro2ML

Demirkaya, E., Fan, Y., Gao, L., Lv, J., Vossler, P., & Wang, J. (2024). Optimal nonparametric inference with two-scale distributional nearest neighbors. *Journal of the American Statistical Association*, 119(545), 297-307. [Link to paper]

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353-7360. [Link to paper]

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In Breakthroughs in statistics: Methodology and distribution (pp. 569-593). New York, NY: Springer New York. [Link to paper]

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24, 123-140. [Link to paper]

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. [Link to paper]

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228-1242. [Link to paper]

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). [Link to paper] [Link to GitHub]

4. Intro2DL

Kingma, D. P. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. [Link to paper]

Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747. [Link to paper]

Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., ... & Kohli, P. (2021). Advancing mathematics by guiding human intuition with AI. *Nature*, 600(7887), 70-74. [Link to paper] [Link to GitHub]

Ye, Z., Zhang, Z., Zhang, D., Zhang, H., & Zhang, R. P. (2023). Deep-Learning-Based Causal Inference for Large-Scale Combinatorial Experiments: Theory and Empirical Evidence. Available at SSRN 4375327. [Link to paper] [Link to GitHub]

Wang, Zhi and Gao, Rui and Li, Shuang（2024）, Neural-Network Mixed Logit Choice Model: Statistical and Optimality Guarantees. [Link to paper]

Luyang Chen, Markus Pelger, Jason Zhu (2023) Deep Learning in Asset Pricing. *Management Science* 70(2):714-750. [Link to paper] [Link to GitHub]

Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. (2018). Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31. [Link to paper] [Link to GitHub]

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778). [Link to paper] [Link to GitHub]

Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference

on artificial intelligence and statistics (pp. 249-256). JMLR Workshop and Conference Proceedings. [Link to paper]

Ioffe, S. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167. [Link to paper]

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034). [Link to paper]

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929-1958. [Link to paper] [Link to GitHub]

5. Transformer

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901. [Link to paper] [Link to GitHub]

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. [Link to paper] [Link to GitHub]

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol. 1, No. 2). [Link to paper] [Link to GitHub]

Mikolov, T. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 3781. [Link to paper]

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9. [Link to paper] [Link to GitHub]

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543). [Link to paper] [Link to GitHub]

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744. [Link to paper]

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144. [Link to paper]

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. (2014). Sequence to sequence learning with neural networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14), Vol. 2. MIT Press, Cambridge, MA, USA, 3104–3112. [Link to paper]

Bahdanau, D. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. [Link to paper]

Lei Ba, J., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. ArXiv e-prints, arXiv-1607. [Link to paper]

Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research, 304, 114135. [Link to paper]

Card, D., Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., ... & Jurafsky, D. (2022). Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. Proceedings of the National Academy of Sciences, 119(31), e2120510119. [Link to paper] [Link to GitHub]

Hansen, S., Lambert, P. J., Bloom, N., Davis, S. J., Sadun, R., & Taska, B. (2023). Remote work across jobs, companies, and space (No. w31007). National Bureau of Economic Research. [Link to paper]

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535-574. [Link to paper]

Ash, E., & Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, *15*(1), 659-688. [Link to paper] [Link to GitHub]

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. [Link to paper] [Link to GitHub]

Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4195-4205). [Link to paper] [Link to GitHub]

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., ... & Mordatch, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, *34*, 15084-15097. [Link to paper] [Link to GitHub]

Qi, M., Shi, Y., Qi, Y., Ma, C., Yuan, R., Wu, D., & Shen, Z. J. (2023). A practical end-to-end inventory management model with deep learning. *Management Science*, *69*(2), 759-773. [Link to paper] [Link to GitHub]

Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*. [Link to paper] [Link to GitHub]

Wang, S., Li, B. Z., Khabsa, M., Fang, H., & Ma, H. (2020). Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*. [Link to paper] [Link to GitHub]

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., ... & Weller, A. (2020). Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*. [Link to paper] [Link to GitHub]

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*. [Link to paper] [Link to GitHub]

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, *23*(120), 1-39. [Link to paper] [Link to GitHub]

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*. [Link to paper] [Link to GitHub]

6. Pretraining

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901. [Link to paper] [Link to GitHub]

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. [Link to paper] [Link to GitHub]

Mikolov, T. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 3781. [Link to paper]

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744. [Link to paper]

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543). [Link to paper] [Link to GitHub]

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361. [Link to paper]

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556. [Link to paper]

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*(2), 806-841. [Link to paper] [Link to GitHub]

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol. 1, No. 2). [Link to paper] [Link to GitHub]

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9. [Link to paper] [Link to GitHub]

Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 364. [Link to paper] [Link to GitHub]

Gorodnichenko, Y., Pham, T., & Talavera, O. (2023). The voice of monetary policy. American Economic Review, 113(2), 548-584. [Link to paper] [Link to GitHub]

Radford, A. (2018). Improving language understanding by generative pre-training. [Link to paper] [Link to GitHub]

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901. [Link to paper] [Link to GitHub]

Rajpurkar, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*. [Link to paper] [Link to GitHub]

Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*. [Link to paper]

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., ... & Lo, K. (2024). Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*. [Link to paper] [Link to GitHub]

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., ... & Piao, Y. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*. [Link to paper] [Link to GitHub]

Reisenbichler, M., Reutterer, T., Schweidel, D. A., & Dan, D. (2022). Frontiers: Supporting content marketing with natural language generation. *Marketing Science*, *41*(3), 441-452. [Link to paper]

Dao, T., Fu, D., Ermon, S., Rudra, A., & Ré, C. (2022). Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, *35*, 16344-16359. [Link to paper] [Link to GitHub]

Dai, D., Deng, C., Zhao, C., Xu, R. X., Gao, H., Chen, D., ... & Liang, W. (2024). Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*. [Link to paper] [Link to GitHub]

Yuan, J., Gao, H., Dai, D., Luo, J., Zhao, L., Zhang, Z., ... & Zeng, W. (2025). Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention. *arXiv preprint arXiv:2502.11089*. [Link to paper]

Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*. [Link to paper] [Link to GitHub]

7. Posttraining

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901. [Link to paper] [Link to GitHub]

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. [Link to paper] [Link to GitHub]

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol. 1, No. 2). [Link to paper] [Link to GitHub]

Mikolov, T. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 3781. [Link to paper]

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543). [Link to paper] [Link to GitHub]

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9. [Link to paper] [Link to GitHub]

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744. [Link to paper]

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*. [Link to paper] [Link to GitHub]

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019, May). Parameter-efficient transfer learning for NLP. In the International *conference on machine learning* (pp. 2790-2799). PMLR. [Link to paper] [Link to GitHub]

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3. [Link to paper] [Link to GitHub]

Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., ... & Ma, Y. (2025). Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*. [Link to paper] [Link to GitHub]

Silver, D., Schrittwieser, J., Simonyan, K. *et al.* Mastering the game of Go without human knowledge. *Nature* 550, 354–359 (2017). [Link to paper] [Link to GitHub]

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*. [Link to paper]

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*. [Link to paper] [Link to GitHub]

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... & Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in neural information processing systems*, *33*, 3008-3021. [Link to paper]

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744. [Link to paper]

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, *36*, 53728-53741. [Link to paper] [Link to GitHub]

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837. [Link to paper] [Link to GitHub]

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*. [Link to paper]

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage?. *Advances in Neural Information Processing Systems*, *36*, 55565-55581. [Link to paper]

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, *36*, 11809-11822. [Link to paper] [Link to GitHub]

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., & Han, J. (2022). Large language models can self-improve. *arXiv preprint arXiv:2210.11610*. [Link to paper] [Link to GitHub]

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*. [Link to paper] [Link to GitHub]

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., ... & Cobbe, K. (2023, May). Let's verify step by step. In *The Twelfth International Conference on Learning Representations*. [Link to paper] [Link to GitHub]

Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., ... & Higgins, I. (2022). Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*. [Link to paper]

Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*. [Link to paper]

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... & He, Y. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*. [Link to paper] [Link to GitHub]

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., ... & Guo, D. (2024). Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*. [Link to paper] [Link to GitHub]

Team, K., Du, A., Gao, B., Xing, B., Jiang, C., Chen, C., ... & Yang, Z. (2025). Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*. [Link to paper] [Link to GitHub]

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, qjae044. [Link to paper]

Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., ... & Hashimoto, T. (2025). s1: Simple test-time scaling. arXiv preprint arXiv:2501.19393. [Link to paper] [Link to GitHub]

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*. [Link to paper]

Wang, M., Zhang, D. J., & Zhang, H. (2024). Large Language Models for Market Research: A Data-augmentation Approach. *arXiv preprint arXiv:2412.19363*. [Link to paper]

8. Inference

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., ... & Piao, Y. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*. [Link to paper] [Link to GitHub]

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., & Liu, Y. (2024). Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, *568*, 127063. [Link to paper] [Link to GitHub]

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, *36*, 10088-10115. [Link to paper] [Link to GitHub]

Patrick Jaillet, Jiashuo Jiang, Chara Podimata, Zijie Zhou, "LLM Inference with Exploding KV Cache Demands", Accepted by NeurIPS 2024 Workshop

9. LLM as Research Tools

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., ... & Piao, Y. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*. [Link to paper] [Link to GitHub]

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). ReAct: Synergizing Reasoning and Acting in Language Models. *ArXiv, abs/2210.03629*. [Link to paper] [Link to GitHub]

Shinn, N., Cassano, F., Labash, B., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: language agents with verbal reinforcement learning. *Neural Information Processing Systems*. [Link to paper] [Link to GitHub]

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023, October). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th annual acm symposium on user interface software and technology (pp. 1-22). [Link to paper] [Link to GitHub]

Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C.J., Morris, M.R., Willer, R., Liang, P., & Bernstein, M.S. (2024). Generative Agent Simulations of 1,000 People. *ArXiv, abs/2411.10109*. [Link to paper] [Link to GitHub]

Yang, Z., Zhang, Z., Zheng, Z., Jiang, Y., Gan, Z., Wang, Z., Ling, Z., Chen, J., Ma, M., Dong, B., Gupta, P., Hu, S., Yin, Z., Li, G., Jia, X., Wang, L., Ghanem, B., Lu, H., Ouyang, W., Qiao, Y., Torr, P., & Shao, J. (2024). OASIS: Open Agent Social Interaction Simulations with One Million Agents. *ArXiv, abs/2411.11581*. [Link to paper] [Link to GitHub]

Ludwig, J., Mullainathan, S., & Rambachan, A. (2025). Large language models: An applied econometric framework (No. w33344). National Bureau of Economic Research. [Link to paper]

Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2024). Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina. arXiv preprint arXiv:2410.19599. [Link to paper]

Ties de Kok (2025) ChatGPT for Textual Analysis? How to Use Generative LLMs in Accounting Research. *Management Science* forthcoming. [Link to paper] [Link to GitHub]

10. Causal Inference Fundamentals

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, *10*(1), 1-51. [Link to paper]

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688. [Link to paper]

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science*, *38*(2), 193-225. [Link to paper]

Imbens, G., & Xu, Y. (2024). Lalonde (1986) after nearly four decades: Lessons learned. *arXiv preprint arXiv:2406.00827*. [Link to paper] [Link to GitHub]

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604-620. [Link to paper]

Gordon, B. R., Zettelmeyer, F., Bhargava, N., & Chapsky, D. (2019). A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science*, *38*(2), 193-225. [Link to paper]

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, *89*(427), 846-866. [Link to paper]

Dudík, M., Erhan, D., Langford, J., & Li, L. (2014). Doubly robust policy evaluation and optimization. [Link to paper]

Liu, X. (2023). Dynamic coupon targeting using batch deep reinforcement learning: An application to livestream shopping. *Marketing Science*, 42(4), 637-658. [Link to paper]

Yang, J., Eckles, D., Dhillon, P., & Aral, S. (2024). Targeting for long-term outcomes. *Management Science*, 70(6), 3841-3855. [Link to paper] [Link to GitHub]

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press. [Link to book]

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55. [Link to paper]

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, *113*(521), 390-400. [Link to paper]

Paul B. Ellickson, Wreetabrata Kar, James C. Reeder, III (2022) Estimating Marketing Component Effects: Double Machine Learning from Targeted Digital Promotions. *Marketing Science 42*(4):704-728. [Link to paper]

11. Double Machine Learning

Fuhr, J., Berens, P., & Papies, D. (2024). Estimating Causal Effects with Double Machine Learning--A Method Evaluation. arXiv preprint arXiv:2403.14385. [Link to paper]

Gordon, B. R., Moakler, R., & Zettelmeyer, F. (2023). Close enough? a large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Science*, 42(4), 768-793. [Link to paper]

Manzoor, E., Chen, G. H., Lee, D., & Smith, M. D. (2024). Influence via ethos: On the persuasive power of reputation in deliberation online. *Management Science*, 70(3), 1613-1634. [Link to paper]

Shi, B., Mao, X., Yang, M., & Li, B. (2023). What, Why, and How: An Empiricist's Guide to Double/Debiased Machine Learning. Debiased Machine Learning (December 27, 2023). [Link to paper]

Ye, Z., Zhang, Z., Zhang, D., Zhang, H., & Zhang, R. P. (2023). Deep-Learning-Based Causal Inference for Large-Scale Combinatorial Experiments: Theory and Empirical Evidence. Available at SSRN 4375327. [Link to paper] [Link to GitHub]

Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, 387-401. [Link to paper]

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2), 608-650. [Link to paper]

Dube, A., Jacobs, J., Naidu, S., & Suri, S. (2020). Monopsony in online labor markets. *American Economic Review: Insights*, *2*(1), 33-46. [Link to paper]

Xu, Y., Ghose, A., & Xiao, B. (2024). Mobile payment adoption: An empirical investigation of Alipay. *Information Systems Research*, *35*(2), 807-828. [Link to paper]

Manzoor, E., Chen, G. H., Lee, D., & Smith, M. D. (2024). Influence via ethos: On the persuasive power of reputation in deliberation online. *Management Science*, *70*(3), 1613-1634. [Link to paper]

Fuhr, J., Berens, P., & Papies, D. (2024). Estimating Causal Effects with Double Machine Learning--A Method Evaluation. *arXiv preprint arXiv:2403.14385*. [Link to paper]

Harrison Jr, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, *5*(1), 81-102. [Link to paper]

Farrell, M. H., Liang, T., & Misra, S. (2020). Deep learning for individual heterogeneity: An automatic inference framework. *arXiv preprint arXiv:2010.14694*. [Link to paper]

Farrell, M. H., Liang, T., & Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, *89*(1), 181-213. [Link to paper]

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931-954. [Link to paper]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68.[Link to paper]

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353-7360. [Link to paper]

Belloni, A., Chernozhukov, V., & Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4), 606-619. [Link to paper]

Manzoor, E., Chen, G. H., Lee, D., & Smith, M. D. (2024). Influence via ethos: On the persuasive power of reputation in deliberation online. *Management Science*, *70*(3), 1613-1634. [Link to paper]

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The review of economic studies*, *72*(1), 1-19. [Link to paper]

Gordon, B. R., Moakler, R., & Zettelmeyer, F. (2023). Close enough? a large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Science*, *42*(4), 768-793. [Link to paper]

Chang, N. C. (2020). Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal*, *23*(2), 177-191. [Link to paper]

Chernozhukov, V., Newey, W. K., & Singh, R. (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica*, *90*(3), 967-1027. [Link to paper]

Fuhr, Jonathan, Philipp Berens, and Dominik Papies. "Estimating Causal Effects with Double Machine Learning--A Method Evaluation." *arXiv preprint arXiv:2403.14385* (2024). [Link to paper]

Harrison Jr, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, *5*(1), 81-102. [Link to paper]

12. HTE

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353-7360. [Link to paper]

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299-319. [Link to paper] [Link to GitHub]

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, *116*(10), 4156-4165. [Link to paper]

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228-1242. [Link to paper]

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. [Link to paper]

Oprescu, M., Syrgkanis, V., & Wu, Z. S. (2019, May). Orthogonal random forest for causal inference. In *International Conference on Machine Learning* (pp. 4932-4941). PMLR. [Link to paper]