Bibliography

[1] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[2] N. Stiennon *et al.*, "Learning to summarize with human feedback," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3008–3021, 2020.

[3] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27730–27744, 2022.

[4] R. Nakano *et al.*, "Webgpt: Browser-assisted question-answering with human feedback," *arXiv preprint arXiv:2112.09332*, 2021.

[5] Y. Bai *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[6] N. Lambert *et al.*, "T\" ULU 3: Pushing frontiers in open language model post-training," *arXiv preprint arXiv:2411.15124*, 2024.

[7] R. Kirk *et al.*, "Understanding the effects of rlhf on llm generalisation and diversity," *arXiv preprint arXiv:2310.06452*, 2023.

[8] T. Chu *et al.*, "Sft memorizes, rl generalizes: A comparative study of foundation model post-training," *arXiv preprint arXiv:2501.17161*, 2025.

[9] P. Singhal, T. Goyal, J. Xu, and G. Durrett, "A long way to go: Investigating length correlations in rlhf," *arXiv preprint arXiv:2310.03716*, 2023.

[10] R. Park, R. Rafailov, S. Ermon, and C. Finn, "Disentangling length from quality in direct preference optimization," *arXiv preprint arXiv:2403.19159*, 2024.

[11] Allen Institute for Artificial Intelligence, "OLMoE, meet iOS." https://allenai.org/blog/olmoe-app, 2025.

[12] C. Zhou *et al.*, "Lima: Less is more for alignment," *Advances in Neural Information Processing Systems*, vol. 36, pp. 55006–55021, 2023.

[13] R. Taori *et al.*, "Stanford alpaca: An instruction-following LLaMA model," *GitHub repository*. https://github.com/tatsu-lab/stanford_alpaca; GitHub, 2023.

[14] W.-L. Chiang *et al.*, "Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality." 2023. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[15] X. Geng *et al.*, "Koala: A dialogue model for academic research." Blog post, 2023. Accessed: Apr. 03, 2023. [Online]. Available: https://bair.berkeley.edu/blog/2023/04/03/koala/

[16] M. Conover *et al.*, "Hello dolly: Democratizing the magic of ChatGPT with open models." Accessed: Jun. 30, 2023. [Online]. Available: https://www.databricks.com/blog/2023/03/24/hello-dolly-democratizing-magic-chatgpt-open-models.html

[17] A. Askell *et al.*, "A general language assistant as a laboratory for alignment," *arXiv preprint arXiv:2112.00861*, 2021.

[18] Y. Bai *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[19] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[20] L. Tunstall *et al.*, "Zephyr: Direct distillation of LM alignment," in *First conference on language modeling*, 2024. Available: https://openreview.net/forum?id=aKkAwZB6JV

[21] H. Ivison *et al.*, "Camels in a changing climate: Enhancing lm adaptation with tulu 2," *arXiv preprint arXiv:2311.10702*, 2023.

[22] G. Cui *et al.*, "Ultrafeedback: Boosting language models with high-quality feedback," 2023.

[23] A. Dubey *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[24] B. Adler *et al.*, "Nemotron-4 340B technical report," *arXiv preprint arXiv:2406.11704*, 2024.

[25] C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz, "A survey of preference-based reinforcement learning methods," *Journal of Machine Learning Research*, vol. 18, no. 136, pp. 1–46, 2017.

[26] T. Kaufmann, P. Weng, V. Bengs, and E. Hüllermeier, "A survey of reinforcement learning from human feedback," *arXiv preprint arXiv:2312.14925*, 2023.

[27] S. Casper *et al.*, "Open problems and fundamental limitations of reinforcement learning from human feedback," *arXiv preprint arXiv:2307.15217*, 2023.

[28] W. B. Knox and P. Stone, "Tamer: Training an agent manually via evaluative reinforcement," in *2008 7th IEEE international conference on development and learning*, IEEE, 2008, pp. 292–297.

[29] J. MacGlashan *et al.*, "Interactive learning from policy-dependent human feedback," in *International conference on machine learning*, PMLR, 2017, pp. 2285–2294.

[30] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei, "Reward learning from human preferences and demonstrations in atari," *Advances in neural information processing systems*, vol. 31, 2018.

[31] G. Warnell, N. Waytowich, V. Lawhern, and P. Stone, "Deep tamer: Interactive agent shaping in high-dimensional state spaces," in *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[32] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, "Scalable agent alignment via reward modeling: A research direction," *arXiv preprint arXiv:1811.07871*, 2018.

[33] D. M. Ziegler *et al.*, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, 2019.

[34] J. Wu *et al.*, "Recursively summarizing books with human feedback," *arXiv preprint arXiv:2109.10862*, 2021.

[35] J. Menick *et al.*, "Teaching language models to support answers with verified quotes," *arXiv preprint arXiv:2203.11147*, 2022.

[36] A. Glaese *et al.*, "Improving alignment of dialogue agents via targeted human judgements," *arXiv preprint arXiv:2209.14375*, 2022.

[37] L. Gao, J. Schulman, and J. Hilton, "Scaling laws for reward model overoptimization," in *International conference on machine learning*, PMLR, 2023, pp. 10835–10866.

[38] D. Ganguli *et al.*, "Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned," *arXiv preprint arXiv:2209.07858*, 2022.

[39] R. Ramamurthy *et al.*, "Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization," *arXiv preprint arXiv:2210.01241*, 2022.

[40] A. Havrilla *et al.*, "TrlX: A framework for large scale reinforcement learning from human feedback," in *Proceedings of the 2023 conference on empirical methods in natural language processing*, Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8578–8595. doi: 10.18653/v1/2023.emnlp-main.530.

[41] L. von Werra *et al.*, "TRL: Transformer reinforcement learning," *GitHub repository*. https://github.com/huggingface/trl; GitHub, 2020.

[42] OpenAI, "ChatGPT: Optimizing language models for dialogue." https://openai.com/blog/chatgpt/, 2022.

[43] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[44] H. Lightman *et al.*, "Let's verify step by step," *arXiv preprint arXiv:2305.20050*, 2023.

[45] A. Kumar *et al.*, "Training language models to self-correct via reinforcement learning," *arXiv preprint arXiv:2409.12917*, 2024.

[46] A. Singh *et al.*, "Beyond human data: Scaling self-training for problem-solving with language models," *arXiv preprint arXiv:2312.06585*, 2023.

[47] OpenAI, "Introducing OpenAI o1-preview." Sep. 2024. Available: https://openai.com/index/introducing-openai-o1-preview/

[48] A. Vaswani *et al.*, "Attention is all you need," in *Neural information processing systems*, 2017. Available: https://api.semanticscholar.org/CorpusID:13756489

[49] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014, Available: https://api.semanticscholar.org/CorpusID:11212020

[50] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[51] G. Team *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.

[52] R. Agarwal *et al.*, "On-policy distillation of language models: Learning from self-generated mistakes," in *The twelfth international conference on learning representations*, 2024.

[53] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[54] R. S. Sutton, "Reinforcement learning: An introduction," *A Bradford Book*, 2018.

[55] N. Lambert, L. Castricato, L. von Werra, and A. Havrilla, "Illustrating reinforcement learning from human feedback (RLHF)," *Hugging Face Blog*, 2022.

[56] N. Lambert, T. K. Gilbert, and T. Zick, "Entangled preferences: The history and risks of reinforcement learning and human feedback," *arXiv preprint arXiv:2310.13595*, 2023.

[57] V. Conitzer *et al.*, "Social choice should guide AI alignment in dealing with diverse human feedback," *arXiv preprint arXiv:2404.10271*, 2024.

[58] A. Mishra, "Ai alignment and social choice: Fundamental limitations and policy implications," *arXiv preprint arXiv:2310.16048*, 2023.

[59] H. R. Kirk *et al.*, "The PRISM alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models," *arXiv preprint arXiv:2404.16019*, 2024.

[60] S. Poddar, Y. Wan, H. Ivison, A. Gupta, and N. Jaques, "Personalizing reinforcement learning from human feedback with variational preference learning," *arXiv preprint arXiv:2408.10075*, 2024.

[61] S. J. Russell and P. Norvig, *Artificial intelligence: A modern approach*. Pearson, 2016.

[62] B. Widrow and M. E. Hoff, "Adaptive switching circuits," Stanford Univ Ca Stanford Electronics Labs, 1960.

[63] B. F. Skinner, *The behavior of organisms: An experimental analysis*. BF Skinner Foundation, 2019.

[64] E. L. Thorndike, "The law of effect," *The American journal of psychology*, vol. 39, no. 1/4, pp. 212–222, 1927.

[65] A. Arnauld, *The port-royal logic*. 1662.

[66] J. Bentham, *An introduction to the principles of morals and legislation*. 1823.

[67] F. P. Ramsey, "Truth and probability," *Readings in Formal Epistemology: Sourcebook*, pp. 21–45, 2016.

[68] K. J. Arrow, "A difficulty in the concept of social welfare," *Journal of political economy*, vol. 58, no. 4, pp. 328–346, 1950.

[69] J. C. Harsanyi, "Rule utilitarianism and decision theory," *Erkenntnis*, vol. 11, no. 1, pp. 25–53, 1977.

[70] R. Pettigrew, *Choosing for changing selves*. Oxford University Press, 2019.

[71] N. Soares, B. Fallenstein, S. Armstrong, and E. Yudkowsky, "Corrigibility," in *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, 2015.

[72] W.-L. Chiang *et al.*, "Chatbot arena: An open platform for evaluating llms by human preference," *arXiv preprint arXiv:2403.04132*, 2024.

[73] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.

[74] J. Zhou *et al.*, "Instruction-following evaluation for large language models," *arXiv preprint arXiv:2311.07911*, 2023.

[75] K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela, "Kto: Model alignment as prospect theoretic optimization," *arXiv preprint arXiv:2402.01306*, 2024.

[76] Z. Wu *et al.*, "Fine-grained human feedback gives better rewards for language model training," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[77] A. Chen *et al.*, "Learning from natural language feedback," *Transactions on Machine Learning Research*, 2024.

[78] OpenAI, "Introducing the model spec." May 2024. Available: https://openai.com/index/introducing-the-model-spec/

[79] A. Y. Ng, S. Russell, *et al.*, "Algorithms for inverse reinforcement learning." in *Proceedings of the seventeenth international conference on machine learning*, in ICML '00. 2000, pp. 663--670.

[80] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952, Accessed: Feb. 13, 2023. [Online]. Available: http://www.jstor.org/stable/2334029

[81] B. Zhu *et al.*, "Starling-7b: Improving helpfulness and harmlessness with rlaif," in *First conference on language modeling*, 2024.

[82] A. Liu, Z. Zhao, C. Liao, P. Lu, and L. Xia, "Learning plackett-luce mixtures from partial preferences," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, pp. 4328–4335.

[83] B. Zhu, M. Jordan, and J. Jiao, "Principled reinforcement learning with human feedback from pairwise or k-wise comparisons," in *International conference on machine learning*, PMLR, 2023, pp. 43037–43067.

[84] K. Cobbe *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[85] C. Lyu *et al.*, "Exploring the limit of outcome reward for learning mathematical reasoning," *arXiv preprint arXiv:2502.06781*, 2025.

[86] L. Zheng *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46595–46623, 2023.

[87] Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto, "Length-controlled alpacaeval: A simple way to debias automatic evaluators," *arXiv preprint arXiv:2404.04475*, 2024.

[88] T. Li *et al.*, "From crowdsourced data to high-quality benchmarks: Arena-hard and BenchBuilder pipeline," *arXiv preprint arXiv:2406.11939*, 2024.

[89] B. Y. Lin *et al.*, "WILDBENCH: Benchmarking LLMs with challenging tasks from real users in the wild," *arXiv preprint arXiv:2406.04770*, 2024.

[90] D. Mahan *et al.*, "Generative reward models," 2024, Available: https://www.synthlabs.ai/pdf/Generative_Reward_Models.pdf

[91] L. Zhang, A. Hosseini, H. Bansal, M. Kazemi, A. Kumar, and R. Agarwal, "Generative verifiers: Reward modeling as next-token prediction," *arXiv preprint arXiv:2408.15240*, 2024.

[92] Z. Ankner, M. Paul, B. Cui, J. D. Chang, and P. Ammanabrolu, "Critique-out-loud reward models," *arXiv preprint arXiv:2408.11791*, 2024.

[93] S. Kim *et al.*, "Prometheus: Inducing fine-grained evaluation capability in language models," in *The twelfth international conference on learning representations*, 2023.

[94] N. Lambert *et al.*, "Rewardbench: Evaluating reward models for language modeling," *arXiv preprint arXiv:2403.13787*, 2024.

[95] X. Wen *et al.*, "Rethinking reward model evaluation: Are we barking up the wrong tree?" *arXiv preprint arXiv:2410.05584*, 2024.

[96] S. Gureja *et al.*, "M-RewardBench: Evaluating reward models in multilingual settings," *arXiv preprint arXiv:2410.15522*, 2024.

[97] Z. Jin *et al.*, "RAG-RewardBench: Benchmarking reward models in retrieval augmented generation for preference alignment," *arXiv preprint arXiv:2412.13746*, 2024.

[98] E. Zhou *et al.*, "RMB: Comprehensively benchmarking reward models in LLM alignment," *arXiv preprint arXiv:2410.09893*, 2024.

[99] Y. Liu, Z. Yao, R. Min, Y. Cao, L. Hou, and J. Li, "RM-bench: Benchmarking reward models of language models with subtlety and style," *arXiv preprint arXiv:2410.16184*, 2024.

[100] Z. Wu, M. Yasunaga, A. Cohen, Y. Kim, A. Celikyilmaz, and M. Ghazvininejad, "reWordBench: Benchmarking and improving the robustness of reward models with transformed inputs," *arXiv preprint arXiv:2503.11751*, 2025.

[101] Z. Chen *et al.*, "MJ-bench: Is your multimodal reward model really a good judge for text-to-image generation?" *arXiv preprint arXiv:2407.04842*, 2024.

[102] M. Yasunaga, L. Zettlemoyer, and M. Ghazvininejad, "Multimodal rewardbench: Holistic evaluation of reward models for vision language models," *arXiv preprint arXiv:2502.14191*, 2025.

[103] L. Li *et al.*, "VLRewardBench: A challenging benchmark for vision-language generative reward models," *arXiv preprint arXiv:2411.17451*, 2024.

[104] J. Ruan *et al.*, "Vlrmbench: A comprehensive and challenging benchmark for vision-language reward models," *arXiv preprint arXiv:2503.07478*, 2025.

[105] E. Frick *et al.*, "How to evaluate reward models for RLHF," *arXiv preprint arXiv:2410.14872*, 2024.

[106] S. Kim *et al.*, "Evaluating robustness of reward models for mathematical reasoning," *arXiv preprint arXiv:2410.01729*, 2024.

[107] M. Song, Z. Su, X. Qu, J. Zhou, and Y. Cheng, "PRMBench: A fine-grained and challenging benchmark for process-level reward models," *arXiv preprint arXiv:2501.03124*, 2025.

[108] W. Wang *et al.*, "VisualPRM: An effective process reward model for multimodal reasoning," *arXiv preprint arXiv:2503.10291*, 2025.

[109] H. Tu, W. Feng, H. Chen, H. Liu, X. Tang, and C. Xie, "ViLBench: A suite for vision-language process reward modeling." Mar. 2025. Available: https://arxiv.org/abs/2503.20271

[110] H. Wang, W. Xiong, T. Xie, H. Zhao, and T. Zhang, "Interpretable preferences via multi-objective reward modeling and mixture-of-experts," *arXiv preprint arXiv:2406.12845*, 2024.

[111] Z. Wang *et al.*, "HelpSteer2: Open-source dataset for training top-performing reward models," *arXiv preprint arXiv:2406.08673*, 2024.

[112] Z. Wang *et al.*, "HelpSteer2-preference: Complementing ratings with preferences," *arXiv preprint arXiv:2410.01257*, 2024.

[113] J. Park, S. Jwa, M. Ren, D. Kim, and S. Choi, "Offsetbias: Leveraging debiased data for tuning evaluators," *arXiv preprint arXiv:2407.06551*, 2024.

[114] N. Jaques, S. Gu, D. Bahdanau, J. M. Hernández-Lobato, R. E. Turner, and D. Eck, "Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control," in *International conference on machine learning*, PMLR, 2017, pp. 1645–1654.

[115] N. Jaques *et al.*, "Human-centric dialog training via offline reinforcement learning," *arXiv preprint arXiv:2010.05848*, 2020.

[116] J. Schulman, "Approximating KL-divergence." http://joschu.net/blog/kl-approx.html, 2016.

[117] R. Y. Pang, W. Yuan, K. Cho, H. He, S. Sukhbaatar, and J. Weston, "Iterative reasoning preference optimization," *arXiv preprint arXiv:2404.19733*, 2024.

[118] Z. Gao *et al.*, "Rebel: Reinforcement learning via regressing relative rewards," *arXiv preprint arXiv:2404.16767*, 2024.

[119] T. B. Brown *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[120] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[121] J. Wei *et al.*, "Finetuned language models are zero-shot learners," in *International conference on learning representations*, 2022. Available: https://openreview.net/forum?id=gEZrGCozdqR

[122] V. Sanh *et al.*, "Multitask prompted training enables zero-shot task generalization," in *International conference on learning representations*, 2022. Available: https://openreview.net/forum?id=9Vrb9D0WI4

[123] S. Mishra, D. Khashabi, C. Baral, and H. Hajishirzi, "Cross-task generalization via natural language crowdsourcing instructions," in *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, Association for Computational Linguistics, May 2022, pp. 3470–3487. doi: 10.18653/v1/2022.acl-long.244.

[124] E. Wallace, K. Xiao, R. Leike, L. Weng, J. Heidecke, and A. Beutel, "The instruction hierarchy: Training llms to prioritize privileged instructions," *arXiv preprint arXiv:2404.13208*, 2024.

[125] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in neural information processing systems*, vol. 36, pp. 10088–10115, 2023.

[126] N. Rajani, L. Tunstall, E. Beeching, N. Lambert, A. M. Rush, and T. Wolf, "No robots," *Hugging Face repository*. https://huggingface.co/datasets/HuggingFaceH4/no_robots; Hugging Face, 2023.

[127] W. R. Gilks and P. Wild, "Adaptive rejection sampling for gibbs sampling," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 41, no. 2, pp. 337–348, 1992.

[128] A. Ahmadian *et al.*, "Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms," *arXiv preprint arXiv:2402.14740*, 2024.

[129] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *Proceedings of the international conference on learning representations (ICLR)*, 2016.

[130] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, pp. 229–256, 1992.

[131] S. C. Huang, A. Ahmadian, and C. F. AI, "Putting RL back in RLHF." https://huggingface.co/blog/putting_rl_back_in_rlhf_with_rloo, 2024.

[132] W. Kool, H. van Hoof, and M. Welling, "Buy 4 reinforce samples, get a baseline for free!" 2019.

[133] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[134] C. Berner *et al.*, "Dota 2 with large scale deep reinforcement learning," *arXiv preprint arXiv:1912.06680*, 2019.

[135] Z. Liu *et al.*, "Understanding R1-zero-like training: A critical perspective," *arXiv preprint arXiv:2503.20783*, Mar. 2025, Available: https://arxiv.org/abs/2503.20783

[136] Z. Shao *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[137] A. Liu *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[138] D. Guo *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[139] H. Ivison *et al.*, "Unpacking DPO and PPO: Disentangling best practices for learning from preference feedback," *arXiv preprint arXiv:2406.09279*, 2024.

[140] S. Huang, M. Noukhovitch, A. Hosseini, K. Rasul, W. Wang, and L. Tunstall, "The n+ implementation details of RLHF with PPO: A case study on TL;DR summarization," in *First conference on language modeling*, 2024. Available: https://openreview.net/forum?id=kHO2ZTa8e3

[141] L. Weng, "Policy gradient algorithms," *lilianweng.github.io*, 2018, Available: https://lilianweng.github.io/posts/2018-04-08-policy-gradient/

[142] A. Baheti, X. Lu, F. Brahman, R. L. Bras, M. Sap, and M. Riedl, "Leftover lunch: Advantage-based offline reinforcement learning for language models," *arXiv preprint arXiv:2305.14718*, 2023.

[143] Q. Yu *et al.*, "DAPO: An open-source LLM reinforcement learning system at scale." 2025.

[144] D. Seita, "Notes on the generalized advantage estimation paper." 2017. Available: https://danieltakeshi.github.io/2017/04/02/notes-on-the-generalized-advantage-estimation-paper/

[145] T. Wu, B. Zhu, R. Zhang, Z. Wen, K. Ramchandran, and J. Jiao, "Pairwise proximal policy optimization: Harnessing relative feedback for llm alignment," *arXiv preprint arXiv:2310.00212*, 2023.

[146] Y. Flet-Berliac *et al.*, "Contrastive policy gradient: Aligning LLMs on sequence-level scores in a supervised-friendly fashion," *arXiv preprint arXiv:2406.19185*, 2024.

[147] T. Cohere *et al.*, "Command a: An enterprise-ready large language model," *arXiv preprint arXiv:2504.00698*, 2025.

[148] Z. Li *et al.*, "Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models," in *Forty-first international conference on machine learning*, 2023.

[149] T. Gunter *et al.*, "Apple intelligence foundation language models," *arXiv preprint arXiv:2407.21075*, 2024.

[150] K. Team *et al.*, "Kimi k1. 5: Scaling reinforcement learning with llms," *arXiv preprint arXiv:2501.12599*, 2025.

[151] M. Tomar, L. Shani, Y. Efroni, and M. Ghavamzadeh, "Mirror descent policy optimization," *arXiv preprint arXiv:2005.09814*, 2020.

[152] Y. Zhang *et al.*, "Improving LLM general preference alignment via optimistic online mirror descent," *arXiv preprint arXiv:2502.16852*, 2025.

[153] Y. Yuan *et al.*, "VAPO: Efficient and reliable reinforcement learning for advanced reasoning tasks," *arXiv preprint arXiv:2504.05118*, 2025.

[154] Y. Yuan, Y. Yue, R. Zhu, T. Fan, and L. Yan, "What's behind PPO's collapse in long-CoT? Value optimization holds the secret," *arXiv preprint arXiv:2503.01491*, 2025.

[155] Y. Zhao, R. Joshi, T. Liu, M. Khalman, M. Saleh, and P. J. Liu, "Slic-hf: Sequence likelihood calibration with human feedback," *arXiv preprint arXiv:2305.10425*, 2023.

[156] M. G. Azar *et al.*, "A general theoretical paradigm to understand learning from human preferences," in *International conference on artificial intelligence and statistics*, PMLR, 2024, pp. 4447–4455.

[157] A. Amini, T. Vieira, and R. Cotterell, "Direct preference optimization with an offset," *arXiv preprint arXiv:2402.10571*, 2024.

[158] J. Hong, N. Lee, and J. Thorne, "Reference-free monolithic preference optimization with odds ratio," *arXiv e-prints*, pp. arXiv–2403, 2024.

[159] Y. Meng, M. Xia, and D. Chen, "Simpo: Simple preference optimization with a reference-free reward," *Advances in Neural Information Processing Systems*, vol. 37, pp. 124198–124235, 2025.

[160] N. Razin, S. Malladi, A. Bhaskar, D. Chen, S. Arora, and B. Hanin, "Unintentional unalignment: Likelihood displacement in direct preference optimization," *arXiv preprint arXiv:2410.08847*, 2024.

[161] Y. Ren and D. J. Sutherland, "Learning dynamics of llm finetuning," *arXiv preprint arXiv:2407.10490*, 2024.

[162] T. Xiao, Y. Yuan, H. Zhu, M. Li, and V. G. Honavar, "Cal-dpo: Calibrated direct preference optimization for language model alignment," *arXiv preprint arXiv:2412.14516*, 2024.

[163] A. Gupta *et al.*, "AlphaPO–reward shape matters for LLM alignment," *arXiv preprint arXiv:2501.03884*, 2025.

[164] S. Guo *et al.*, "Direct language model alignment from online ai feedback," *arXiv preprint arXiv:2402.04792*, 2024.

[165] P. Singhal, N. Lambert, S. Niekum, T. Goyal, and G. Durrett, "D2po: Discriminator-guided dpo with response evaluation models," *arXiv preprint arXiv:2405.01511*, 2024.

[166] C. Rosset, C.-A. Cheng, A. Mitra, M. Santacroce, A. Awadallah, and T. Xie, "Direct nash optimization: Teaching language models to self-improve with general preferences," *arXiv preprint arXiv:2404.03715*, 2024.

[167] S. Jung, G. Han, D. W. Nam, and K.-W. On, "Binary classifier optimization for large language model alignment," *arXiv preprint arXiv:2404.04656*, 2024.

[168] H. Zhao *et al.*, "Rainbowpo: A unified framework for combining improvements in preference optimization," *arXiv preprint arXiv:2410.04203*, 2024.

[169] A. Gorbatovski, B. Shaposhnikov, V. Sinii, A. Malakhov, and D. Gavrilov, "The differences between direct alignment algorithms are a blur," *arXiv preprint arXiv:2502.01237*, 2025.

[170] S. Xu *et al.*, "Is dpo superior to ppo for llm alignment? A comprehensive study," *arXiv preprint arXiv:2404.10719*, 2024.

[171] F. Tajwar *et al.*, "Preference fine-tuning of llms should leverage suboptimal, on-policy data," *arXiv preprint arXiv:2404.14367*, 2024.

[172] H. Lee *et al.*, "Rlaif: Scaling reinforcement learning from human feedback with ai feedback," 2023.

[173] A. Sharma, S. Keh, E. Mitchell, C. Finn, K. Arora, and T. Kollar, "A critical evaluation of AI feedback for aligning large language models." 2024. Available: https://arxiv.org/abs/2402.12366

[174] L. Castricato, N. Lile, S. Anand, H. Schoelkopf, S. Verma, and S. Biderman, "Suppressing pink elephants with direct principle feedback." 2024. Available: https://arxiv.org/abs/2402.07896

[175] L. J. V. Miranda *et al.*, "Hybrid preferences: Learning to route instances for human vs. AI feedback," *arXiv preprint arXiv:2410.19133*, 2024.

[176] T. Wang *et al.*, "Shepherd: A critic for language model generation," *arXiv preprint arXiv:2308.04592*, 2023.

[177] P. Ke *et al.*, "CritiqueLLM: Towards an informative critique generation model for evaluation of large language model generation," *arXiv preprint arXiv:2311.18702*, 2023.

[178] J. Li, S. Sun, W. Yuan, R.-Z. Fan, H. Zhao, and P. Liu, "Generative judge for evaluating alignment," *arXiv preprint arXiv:2310.05470*, 2023.

[179] S. Kim *et al.*, "Prometheus 2: An open source language model specialized in evaluating other language models," *arXiv preprint arXiv:2405.01535*, 2024.

[180] S. Lee, S. Kim, S. Park, G. Kim, and M. Seo, "Prometheus-vision: Vision-language model as a judge for fine-grained evaluation," in *Findings of the association for computational linguistics ACL 2024*, 2024, pp. 11286–11315.

[181] M. Y. Guan *et al.*, "Deliberative alignment: Reasoning enables safer language models," *arXiv preprint arXiv:2412.16339*, 2024.

[182] Anthropic, "Claude's constitution." Accessed: Feb. 07, 2024. [Online]. Available: https://www.anthropic.com/news/claudes-constitution

[183] D. Ganguli *et al.*, "Collective constitutional AI: Aligning a language model with public input." Anthropic, 2023.

[184] S. Huang *et al.*, "Constitutional AI recipe," *Hugging Face Blog*, 2024.

[185] N. Lambert, H. Schoelkopf, A. Gokaslan, L. Soldaini, V. Pyatkin, and L. Castricato, "Self-directed synthetic dialogues and revisions technical report," *arXiv preprint arXiv:2407.18421*, 2024.

[186] Z. Sun *et al.*, "Principle-driven self-alignment of language models from scratch with minimal human supervision," in *Thirty-seventh conference on neural information processing systems*, 2023. Available: https://openreview.net/forum?id=p40XRfBX96

[187] Z. Sun *et al.*, "SALMON: Self-alignment with principle-following reward models," in *The twelfth international conference on learning representations*, 2024. Available: https://openreview.net/forum?id=xJbsmB8UMx

[188] A. Irpan, "Deep reinforcement learning doesn't work yet." 2018. Available: https://www.alexirpan.com/2018/02/14/rl-hard.html

[189] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Proceedings of the AAAI conference on artificial intelligence*, 2018. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11694

[190] G. Sheng *et al.*, "HybridFlow: A flexible and efficient RLHF framework," *arXiv preprint arXiv: 2409.19256*, 2024.

[191] J. Hu *et al.*, "OpenRLHF: An easy-to-use, scalable and high-performance RLHF framework," *arXiv preprint arXiv:2405.11143*, 2024.

[192] J. Liu, A. Cohen, R. Pasunuru, Y. Choi, H. Hajishirzi, and A. Celikyilmaz, "Don't throw away your value model! Generating more preferable text with value-guided monte-carlo tree search decoding," *arXiv preprint arXiv:2309.15028*, 2023.

[193] B. Brown *et al.*, "Large language monkeys: Scaling inference compute with repeated sampling," *arXiv preprint arXiv:2407.21787*, 2024.

[194] Z. Liu *et al.*, "Inference-time scaling for generalist reward modeling," *arXiv preprint arXiv:2504.02495*, 2025.

[195] N. Muennighoff *et al.*, "s1: Simple test-time scaling," *arXiv preprint arXiv:2501.19393*, 2025.

[196] L. Chen *et al.*, "Are more llm calls all you need? Towards scaling laws of compound inference systems," *arXiv preprint arXiv:2403.02419*, 2024.

[197] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "AI models collapse when trained on recursively generated data," *Nature*, vol. 631, no. 8022, pp. 755–759, 2024.

[198] M. Gerstgrasser *et al.*, "Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data," *arXiv preprint arXiv:2404.01413*, 2024.

[199] Y. Feng, E. Dohmatob, P. Yang, F. Charton, and J. Kempe, "Beyond model collapse: Scaling up with synthesized data requires reinforcement," in *ICML 2024 workshop on theoretical foundations of foundation models*, 2024.

[200] Y. Wang *et al.*, "Self-instruct: Aligning language models with self-generated instructions," *arXiv preprint arXiv:2212.10560*, 2022.

[201] E. Beeching *et al.*, "NuminaMath 7B TIR," *Hugging Face repository*. https://huggingface.co/AI-MO/NuminaMath-7B-TIR; Numina & Hugging Face, 2024.

[202] M. Li *et al.*, "Superfiltering: Weak-to-strong data filtering for fast instruction-tuning," *arXiv preprint arXiv:2402.00530*, 2024.

[203] K. Shridhar, A. Stolfo, and M. Sachan, "Distilling reasoning capabilities into smaller language models," *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7059–7073, 2023.

[204] C.-Y. Hsieh *et al.*, "Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes," *arXiv preprint arXiv:2305.02301*, 2023.

[205] D. Hendrycks *et al.*, "Measuring massive multitask language understanding," *arXiv preprint arXiv:2009.03300*, 2020.

[206] A. Mallen, A. Asai, V. Zhong, R. Das, H. Hajishirzi, and D. Khashabi, "When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories," *arXiv preprint*, 2022.

[207] S. Lin, J. Hilton, and O. Evans, "Truthfulqa: Measuring how models mimic human falsehoods," *arXiv preprint arXiv:2109.07958*, 2021.

[208] M. Suzgun *et al.*, "Challenging BIG-bench tasks and whether chain-of-thought can solve them," *arXiv preprint arXiv:2210.09261*, 2022.

[209] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, "DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs," *arXiv preprint arXiv:1903.00161*, 2019.

[210] D. Hendrycks *et al.*, "Measuring mathematical problem solving with the MATH dataset," *NeurIPS*, 2021.

[211] K. Cobbe *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[212] M. Chen *et al.*, "Evaluating large language models trained on code," 2021, Available: https://arxiv.org/abs/2107.03374

[213] J. Liu, C. S. Xia, Y. Wang, and L. Zhang, "Is your code generated by chatGPT really correct? Rigorous evaluation of large language models for code generation," in *Thirty-seventh conference on neural information processing systems*, 2023. Available: https://openreview.net/forum?id=1qvx610Cu7

[214] J. Zhou *et al.*, "Instruction-following evaluation for large language models." 2023. Available: https://arxiv.org/abs/2311.07911

[215] D. Rein *et al.*, "GPQA: A graduate-level google-proof q&a benchmark," *arXiv preprint arXiv:2311.12022*, 2023.

[216] L. Phan, A. Gatti, Z. Han, N. Li, and H. et al. Zhang, "Humanity's last exam," *arXiv preprint arXiv:2501.14249*, 2025.

[217] R. Aleithan, H. Xue, M. M. Mohajer, E. Nnorom, G. Uddin, and S. Wang, "SWE-Bench+: Enhanced coding benchmark for LLMs," *arXiv preprint arXiv:2410.06992*, 2024.

[218] N. Jain *et al.*, "LiveCodeBench: Holistic and contamination-free evaluation of large language models for code," *arXiv preprint arXiv:2403.07974*, 2024.

[219] S. AI, "SEAL LLM leaderboards: Expert-driven private evaluations." 2024. Available: https://scale.com/leaderboard

[220] S. Schulhoff *et al.*, "The prompt report: A systematic survey of prompting techniques," *arXiv preprint arXiv:2406.06608*, 2024.

[221] J. Robinson, C. M. Rytting, and D. Wingate, "Leveraging large language models for multiple choice question answering," in *International conference on learning representations*, 2023. Available: https://openreview.net/forum?id=upQ4o-ygvJ

[222] J. Wei *et al.*, "Finetuned language models are zero-shot learners," in *International conference on learning representations*, 2022.

[223] V. Sanh *et al.*, "Multitask prompted training enables zero-shot task generalization," in *International conference on learning representations*, 2022.

[224] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22199–22213, 2022.

[225] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[226] OpenAI, "Introducing SWE-bench verified." Aug. 2024. Available: https://openai.com/index/introducing-swe-bench-verified/

[227] J. Li *et al.*, "Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions," *Hugging Face repository*, vol. 13, p. 9, 2024.

[228] L. Yu *et al.*, "Metamath: Bootstrap your own mathematical questions for large language models," *arXiv preprint arXiv:2309.12284*, 2023.

[229] A. K. Singh *et al.*, "Evaluation data contamination in LLMs: How do we measure it and (when) does it matter?" *arXiv preprint arXiv:2411.03923*, 2024.

[230] K. Huang *et al.*, "MATH-perturb: Benchmarking LLMs' math reasoning abilities against hard perturbations," *arXiv preprint arXiv:2502.06453*, 2025.

[231] UK AI Safety Institute, "Inspect AI: Framework for Large Language Model Evaluations." https://github.com/UKGovernmentBEIS/inspect_ai, 2024.

[232] C. Fourrier, N. Habib, H. Kydlicek, T. Wolf, and L. Tunstall, "LightEval: A lightweight framework for LLM evaluation." https://github.com/huggingface/lighteval, 2023.

[233] C. Fourrier, N. Habib, A. Lozovskaya, K. Szafer, and T. Wolf, "Open LLM leaderboard v2." https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard; Hugging Face, 2024.

[234] L. Gao *et al.*, "A Framework for Few-Shot Language Model Evaluation." Zenodo, 2023. doi: 10.5281/zenodo.10256836.

[235] S. Black *et al.*, "GPT-NeoX-20B: An open-source autoregressive language model," in *Proceedings of the ACL workshop on challenges & perspectives in creating large language models*, 2022. Available: https://arxiv.org/abs/2204.06745

[236] Y. Gu, O. Tafjord, B. Kuehl, D. Haddad, J. Dodge, and H. Hajishirzi, "OLMES: A Standard for Language Model Evaluations," *arXiv preprint arXiv:2406.08446*, 2024.

[237] P. Liang *et al.*, "Holistic Evaluation of Language Models," *Transactions on Machine Learning Research*, 2023, doi: 10.1111/nyas.15007.

[238] MosaicML, "Mosaic Eval Gauntlet v0.3.0 — Evaluation Suite." https://github.com/mosaicml/llm-foundry/blob/main/scripts/eval/local_data/EVAL_GAUNTLET.md, 2024.

[239] J. Schulman, "Proxy objectives in reinforcement learning from human feedback." Invited talk at the International Conference on Machine Learning (ICML), 2023. Available: https://icml.cc/virtual/2023/invited-talk/21549

[240] C. Zhang, O. Vinyals, R. Munos, and S. Bengio, "A study on overfitting in deep reinforcement learning," *arXiv preprint arXiv:1804.06893*, 2018.

[241] C. A. Goodhart and C. Goodhart, *Problems of monetary management: The UK experience*. Springer, 1984.

[242] K. Hoskin, "The 'awful idea of accountability': Inscribing people into the measurement of objects," *Accountability: Power, ethos and the technologies of managing*, vol. 265, 1996.

[243] M. Sharma *et al.*, "Towards understanding sycophancy in language models," *arXiv preprint arXiv:2310.13548*, 2023.

[244] T. Lu and C. Boutilier, "Learning mallows models with pairwise preferences," in *Proceedings of the 28th international conference on machine learning (icml-11)*, 2011, pp. 145–152.

[245] S. Han *et al.*, "Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms," *arXiv preprint arXiv:2406.18495*, 2024.

[246] H. Inan *et al.*, "Llama guard: Llm-based input-output safeguard for human-ai conversations," *arXiv preprint arXiv:2312.06674*, 2023.

[247] P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy, "Xstest: A test suite for identifying exaggerated safety behaviours in large language models," *arXiv preprint arXiv:2308.01263*, 2023.

[248] T. Coste, U. Anwar, R. Kirk, and D. Krueger, "Reward model ensembles help mitigate overoptimization," *arXiv preprint arXiv:2310.02743*, 2023.

[249] T. Moskovitz *et al.*, "Confronting reward model overoptimization with constrained RLHF," *arXiv preprint arXiv:2310.04373*, 2023.

[250] R. Rafailov *et al.*, "Scaling laws for reward model overoptimization in direct alignment algorithms," *Advances in Neural Information Processing Systems*, vol. 37, pp. 126207–126242, 2024.

[251] S. Zhuang and D. Hadfield-Menell, "Consequences of misaligned AI," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15763–15773, 2020.

[252] W. Yuan *et al.*, "Self-rewarding language models." 2025. Available: https://arxiv.org/abs/2401.10020

[253] J. Bai *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[254] G. Wang, S. Cheng, X. Zhan, X. Li, S. Song, and Y. Liu, "Openchat: Advancing open-source language models with mixed-quality data," *arXiv preprint arXiv:2309.11235*, 2023.

[255] Anthropic, "Claude's character." 2024. Available: https://www.anthropic.com/research/claude-character