M. D. Buhl, G. Sett, L. Koessler, J. Schuett, and M. Anderljung. Safety cases for frontier AI. *arXiv*, 2024. (Link).

L. Bürger, F. A. Hamprecht, and B. Nadler. Truth is universal: Robust detection of lies in LLMs. *arXiv*, 2024. (Link).

C. Burns. How "discovering latent knowledge in language models without supervision" fits into a broader alignment scheme. Alignment Forum, 2022. (Link).

C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv*, 2022. (Link).

C. Burns, P. Izmailov, J. H. Kirchner, B. Baker, L. Gao, L. Aschenbrenner, Y. Chen, A. Ecoffet, M. Joglekar, J. Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv*, 2023. (Link).

M. Burtell and T. Woodside. Artificial influence: An analysis of AI-driven persuasion. *arXiv*, 2023. (Link).

P. Bussmann, Bart ands Leask, J. Bloom, C. Tigges, and N. Nanda. Stitching SAEs of different sizes, 2024. (Link).

P. Butlin, R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. M. Fleming, C. Frith, X. Ji, et al. Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv*, 2023. (Link).

N. Cammarata, G. Goh, S. Carter, C. Voss, L. Schubert, and C. Olah. Curve circuits. *Distill*, 6(1): e00024–006, 2021. (Link).

M. Campbell, A. J. Hoane Jr, and F.-h. Hsu. Deep Blue. *Artificial intelligence*, 134(1-2):57–83, 2002. (Link).

S. K. Card. *The psychology of human-computer interaction*. CRC Press, 2018.

N. Carlini. Some lessons from adversarial machine learning. Talk, Jul 2024. Vienna Alignment Workshop.(Link).

N. Carlini, D. Paleka, K. D. Dvijotham, T. Steinke, J. Hayase, A. F. Cooper, K. Lee, M. Jagielski, M. Nasr, A. Conmy, et al. Stealing part of a production language model. In *International Conference on Machine Learning*, 2024. (Link).

J. Carlsmith. Is power-seeking AI an existential risk? *arXiv*, 2022. (Link).

J. Carlsmith. Scheming AIs: Will AIs fake alignment during training in order to get power? *arXiv*, 2023. (Link).

S. Carter, Z. Armstrong, L. Schubert, I. Johnson, and C. Olah. Activation atlas. *Distill*, 2019. (Link).

D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methodsandmetrics. *Electronics*,8(8):832,2019.(Link).

S. Casper, L. Schulze, O. Patel, and D. Hadfield-Menell. Defending against unforeseen failure modes withlatentadversarialtraining. *arXiv*,2024.(Link).

S. Cave and S. S. ÓhÉigeartaigh. An AI race for strategic advantage: rhetoric and risks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* , pages 36–40, 2018. (Link).

CCRL. Ccrl 40/15 rating list — all engines. Computer Chess Rating Lists (CCRL), 2024. Accessed: 2024-10-25. (Link).

CFI Team. Know your client (KYC), 2025. (Link).

D. Chalmers. The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17(9-10): 7–65, 2010. (Link).

L. Chan, A. Garriga-Alonso, N. Goldowsky-Dill, R. Greenblatt, J. Nitishinskaya, A. Radhakrishnan, B. Shlegeris, and N. Thomas. Causal scrubbing: a method for rigorously testing interpretability hypotheses, 2022. (Link).

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41 (3), Jul 2009. ISSN 0360-0300. (Link).

T. A. Chang, D. Rajagopal, T. Bolukbasi, L. Dixon, and I. Tenney. Scalable influence and fact tracing for large language model pretraining. *arXiv*, 2024. (Link).

P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Sehwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramer, et al. JailbreakBench: An open robustness benchmark for jailbreaking large language models. *arXiv*, 2024a. (Link).

P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong. Jailbreaking black box large language models in twenty queries. *arXiv*, 2024b. (Link).

Y. Chen, A. Wu, T. DePodesta, C. Yeh, K. Li, N. C. Marin, O. Patel, J. Riecke, S. Raval, O. Seow, et al. Designing a dashboard for transparency and control of conversational AI. *arXiv*, 2024. (Link).

J. Cheng, G. Novati, J. Pan, C. Bycroft, A. Žemgulytė ˙ , T. Applebaum, A. Pritzel, L. H. Wong, M. Zielinski, T. Sargeant, et al. Predictions for AlphaMissense. Zenodo, 2023. (Link).

ChessDB. All time top 100 ranklist by highest elo rating, 2016. Accessed: 2024-10-27. (Link).

D. Choi, V. Huang, K. Meng, D. D. Johnson, J. Steinhardt, and S. Schwettmann. Scaling automatic neuron description, 2024. (Link).

F. Chollet, M. Knoop, G. Kamradt, and B. Landers. Arc prize 2024: Technical report. *arXiv*, 2024. (Link).

P. Christensen, K. Gillingham, and W. Nordhaus. Uncertainty in forecasts of long-run economic growth. *Proceedings of the National Academy of Sciences*, 115(21):5409–5414, 2018. (Link).

P. Christiano. Approval directed agents, 2014. (Link).

P. Christiano. Corrigibility, 2017. (Link).

P. Christiano. Takeoff speeds, 2018. (Link).

P. Christiano. Worst-case guarantees (revisited), 2019. (Link).

P. Christiano. Learning the prior, 2020. (Link).

P. Christiano. Mundane solutions to exotic problems, may 2021. (Link).

P. Christiano. Comment on "Let's see you write that corrigibility tag", 2022. (Link).

P. Christiano and M. Xu. Mechanistic anomaly detection and ELK, Nov 2022. (Link).

P. Christiano, B. Shlegeris, and D. Amodei. Supervising strong learners by amplifying weak experts. *arXiv*, 2018. (Link).

P. Christiano, A. Cotra, and M. Xu. Eliciting latent knowledge: How to tell if your eyes deceive you, Dec 2021. (Link).

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Neural information processing systems*, 30, 2017. (Link).

J. Chua and O. Evans. Are DeepSeek R1 and other reasoning models more faithful? *arXiv*, 2025. (Link).

A. Cloud, J. Goldman-Wetzler, E. Wybitul, J. Miller, and A. M. Turner. Gradient routing: Masking gradients to localize computation in neural networks. *arXiv*, 2024. (Link).

J. Clymer, N. Gabrieli, D. Krueger, and T. Larsen. Safety cases: Justifying the safety of advanced AI systems. *arXiv*, 2024. (Link).

A. D. Cobb, S. J. Roberts, and Y. Gal. Loss-calibrated approximate inference in bayesian neural networks. *arXiv*, 2018. (Link).

D. Cohn, Z. Ghahramani, and M. Jordan. Active learning with statistical models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Neural Information Processing Systems*, volume 7. MIT Press, 1994. (Link).

A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Neural information processing systems*, 36:16318–16352, 2023. (Link).

A. F. Cooper, C. A. Choquette-Choo, M. Bogen, M. Jagielski, K. Filippova, K. Z. Liu, A. Chouldechova, J. Hayes, Y. Huang, N. Mireshghallah, et al. Machine unlearning doesn't do what you think: Lessons for generative AI policy, research, and practice. *arXiv*, 2024. (Link).

A. Cotra. Forecasting transformative AI with biological anchors, 2020. (Link).

A. Cotra. Two-year update on my personal AI timelines, 2022a. (Link).

A. Cotra. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover, 2022b. (Link).

A. Cotra. LLMs could accelerate AI improvement and deployment. HLAI workshop 2023, 2023. (Link).

B. Cottier, R. Rahman, L. Fattorini, N. Maslej, and D. Owen. The rising costs of training frontier AI models. *arXiv*, 2024. (Link).

Coursera. The fastest-growing job skills of 2024 report. Coursera Blog, Dec 2023. (Link).

CPA Canada. Guide to comply with canada's anti-money laundering and anti-terrorist financing legislation, 2022. Accessed: 2025-01-10. (Link).

M.Crawford. *Why we drive: on freedom, risk and taking back control*RandomHouse,2020.

K. Cresswell and A. Sheikh. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. *International journal of medical informatics*, 82(5):e73–e86, 2013. (Link).

A. Critch. What multipolar failure looks like, and robust agent-agnostic processes (RAAPs), 2021. (Link).

A. Critch and D. Krueger. Ai research considerations for human existential safety (ARCHES). *arXiv*, 2020. (Link).

H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv*, 2023. (Link).

D. Dalrymple, J. Skalse, Y. Bengio, S. Russell, M. Tegmark, S. Seshia, S. Omohundro, C. Szegedy, B. Goldhaber, N. Ammann, et al. Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems. *arXiv*, 2024. (Link).

J. Danaher and S. Nyholm. Automation, work and the achievement gap. *AI and Ethics*, 1(3):227–237, 2021. (Link).

S. Dattani, L. Rodés-Guirao, H. Ritchie, E. Ortiz-Ospina, and M. Roser. Life expectancy. *Our World in Data*, 2023. (Link).

T. Davidson. Could advanced AI drive explosive economic growth? *Open Philanthropy*, 25, 2021a. (Link).

T. Davidson. Semi-informative priors over AI timelines. *Open Philanthropy*, Mar 2021b. (Link).

T. Davidson. What a compute-centric framework says about takeoff speeds. *Open Philanthropy*, 2023. (Link).

T. Davidson, J.-S. Denain, P. Villalobos, and G. Bas. AI capabilities can be significantly improved without expensive retraining. *arXiv*, 2023. (Link).

P. De Blanc. Ontological crises in artificial agents' value systems. *arXiv*, 2011. (Link).

A. Deeb and F. Roger. Do unlearning methods remove information from language model weights? *arXiv*, 2024. (Link).

DeepSeek. DeepSeek-R1-Lite-Preview is now live: unleashing supercharged reasoning power!, 2024. (Link).

G. Deng, Y. Liu, Y. Li, K. Wang, Y. Zhang, Z. Li, H. Wang, T. Zhang, and Y. Liu. MASTERKEY: automated jailbreaking of large language model chatbots. In *Proceedings 2024 Network and Distributed System Security Symposium*, NDSS 2024. Internet Society, 2024. (Link).

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. (Link).

C. Denison, M. MacDiarmid, F. Barez, D. Duvenaud, S. Kravec, S. Marks, N. Schiefer, R. Soklaski, A. Tamkin, J. Kaplan, et al. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv*, 2024. (Link).

M. Dennis, N. Jaques, E. Vinitsky, A. Bayen, S. Russell, A. Critch, and S. Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. *Advances in neural information processing systems*, 33:13049–13061, 2020. (Link).

Descartes Systems Group. Navigating OFAC compliance: The crucial role of IP address geolocation in sanctions screening, 2025. Retrieved January 10, 2025. (Link).

E. Dinan, S. Humeau, B. Chintagunta, and J. Weston. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China, nov 2019. Association for Computational Linguistics. (Link).

B.L.Docherty. *Losing humanity: The case against killer robots*.HumanRightsWatch,2012.

J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv*, 2021. (Link).

A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. (Link).

A. Douillard, Q. Feng, A. A. Rusu, R. Chhaparia, Y. Donchev, A. Kuncoro, M. Ranzato, A. Szlam, and J. Shen. DiLoCo: Distributed low-communication training of language models. *arXiv*, 2023. (Link).

A. Douillard, Q. Feng, A. A. Rusu, A. Kuncoro, Y. Donchev, R. Chhaparia, I. Gog, M. Ranzato, J. Shen, and A. Szlam. DiPaCo: Distributed path composition. *arXiv*, 2024. (Link).

M. K. B. Doumbouya, A. Nandi, G. Poesia, D. Ghilardi, A. Goldie, F. Bianchi, D. Jurafsky, and C. D. Manning. h4rm3l: A dynamic benchmark of composable jailbreak attacks for llm safety assessment. *arXiv*, 2024. (Link).

K. E. Drexler. Reframing superintelligence: Comprehensive AI services as general intelligence. *Future of Humanity Institute*, 2019. (Link).

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The Llama 3 herd of models. *arXiv*, 2024. (Link).

J. Dunefsky, P. Chlenski, and N. Nanda. Transcoders find interpretable LLM feature circuits. *arXiv*, 2024. (Link).

Y. Elazar, S. Ravfogel, A. Jacovi, and Y. Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. (Link).

R. Eldan and M. Russinovich. Who's Harry Potter? approximate unlearning in LLMs. *arXiv*, 2023. (Link).

N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. (Link).

N. Elhage, T. Hume, C. Olsson, N. Nanda, T. Henighan, S. Johnston, S. ElShowk, N. Joseph, N. DasSarma, B. Mann, et al. Softmax linear units. *Transformer Circuits Thread*, 2022a. (Link).

N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022b. (Link).

L. Engstrom, A. Feldmann, and A. Madry. DsDm: Model-aware dataset selection with datamodels. *arXiv*, 2024. (Link).

Epoch AI. Data on notable AI models, 2024. Accessed: 2024-10-29. (Link).

E.ErdilandT.Besiroglu.Algorithmicprogressincomputervision. *arXiv*,2022.(Link).

E. Erdil and T. Besiroglu. Explosive growth from AI automation: A review of the arguments. *arXiv*, 2023. (Link).

E. Erdil, T. Besiroglu, and A. Ho. Estimating idea production: A methodological survey. *arXiv*, 2024. (Link).

E. Erdil, A. Potlogea, T. Besiroglu, E. Roldan, A. Ho, J. Sevilla, M. Barnett, M. Vrzla, and R. Sandler. Gate: An integrated assessment model for ai automation, 2025. (Link).

K. M. Esvelt. Inoculating science against potential pandemics and information hazards. *PLoS Pathogens*, 14(10):e1007286, 2018. (Link).

J. S. B. Evans. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10): 454–459, 2003. (Link).

O. Evans, A. Stuhlmüller, and N. Goodman. Learning the preferences of ignorant, inconsistent agents. In *AAAI Conference on Artificial Intelligence*, volume 30, 2016. (Link).

T. Everitt, G. Lea, and M. Hutter. AGI safety literature review. *arXiv*, 2018. (Link).

R. Fang, R. Bindu, A. Gupta, Q. Zhan, and D. Kang. Teams of LLM agents can exploit zero-day vulnerabilities. *arXiv*, 2024. (Link).

S. Farquhar, V. Varma, Z. Kenton, J. Gasteiger, V. Mikulik, and R. Shah. Challenges with unsupervised LLM knowledge discovery. *arXiv*, 2023. (Link).

S. Farquhar, V. Varma, D. Lindner, D. Elson, C. Biddulph, I. Goodfellow, and R. Shah. MONA: myopic optimization with non-myopic approval can mitigate multi-step reward hacking. *arXiv*, 2025. (Link).

E. Farrell, Y.-T. Lau, and A. Conmy. Applying sparse autoencoders to unlearn knowledge in language models. *arXiv*, 2024. (Link).

Federal Deposit Insurance Corporation. Bank secrecy act, anti-money laundering, and office of foreign assets control. DSC Risk Management Manual of Examination Policies Section 8.1, Federal Deposit Insurance Corporation, 12 2004. (Link).

M. Feffer, A. Sinha, W. H. Deng, Z. C. Lipton, and H. Heidari. Red-teaming for generative AI: Silver bullet or security theater? *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 421–437, Oct 2024. (Link).

J. Ferrando and M. R. Costa-jussà. On the similarity of circuits across languages: a case study on the subject-verb agreement task. *arXiv*, 2024. (Link).

J. Ferrando, G. Sarti, A. Bisazza, and M. R. Costa-jussà. A primer on the inner workings of transformer-based language models. *arXiv*, 2024. (Link).

D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, et al. Building Watson: An overview of the DeepQA project. *AI magazine*, 31 (3):59–79, 2010. (Link).

K. Fiedler, J. Prager, and L. McCaughey. Metacognitive myopia: A major obstacle on the way to rationality. *Current Directions in Psychological Science*, 32(1):49–56, 2023. (Link).

Financial Action Task Force. High-risk jurisdictions subject to a call for action. Public statement, Financial Action Task Force, 6 2023. (Link).

Financial Crimes Enforcement Network. 31 c.f.r. § 1020.315 transactions of exempt persons. Electronic Code of Federal Regulations, 2025. (Link).

L. Fluri, D. Paleka, and F. Tramèr. Evaluating superhuman models with consistency checks. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 194–232. IEEE, 2024. (Link).

R. M. French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4): 128–135, 1999. (Link).

D. Friedman, A. Lampinen, L. Dixon, D. Chen, and A. Ghandeharioun. Interpretability illusions in the generalization of simplified models. *arXiv*, 2023. (Link).

A. Fügener, J. Grahl, A. Gupta, and W. Ketter. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, 33(2):678–696, 2022. (Link).

I. Gabriel, A. Manzini, G. Keeling, L. A. Hendricks, V. Rieser, H. Iqbal, N. Tomašev, I. Ktena, Z. Kenton, M. Rodriguez, et al. The ethics of advanced AI assistants. *arXiv*, 2024. (Link).

S. Y. Gadre, G. Smyrnis, V. Shankar, S. Gururangan, M. Wortsman, R. Shao, J. Mercat, A. Fang, J. Li, S. Keh, et al. Language models scale reliably with over-training and on downstream tasks. *arXiv*, 2024. (Link).

Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning*, pages 1050–1059. PMLR, 2016. (Link).

I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3): 1097–1179, 2024. (Link).

C. Gamble and J. Gao. Safety-first AI for autonomous data centre cooling and industrial control. *Google DeepMind Blog*, Aug 2018. (Link).

R. Gandikota, S. Feucht, S. Marks, and D. Bau. Erasing conceptual knowledge from language models. *arXiv*, 2024. (Link).

D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv*, 2022. (Link).

L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, and J. Wu. Scaling and evaluating sparse autoencoders. *arXiv*, 2024. (Link).

A. C. B. Garcia, M. G. P. Garcia, and R. Rigobon. Algorithmic discrimination in the credit domain: what do we know about it? *AI & SOCIETY*, 39(4):2059–2098, 2024. (Link).

S. Garcin, J. Doran, S. Guo, C. G. Lucas, and S. V. Albrecht. Dred: Zero-shot transfer in reinforcement learning via data-regularised environment design. *arXiv*, 2024. (Link).

A. Geiger, H. Lu, T. Icard, and C. Potts. Causal abstractions of neural networks. *Neural information processing systems*34:9574–9586, 2021. (Link).

A. Geiger, C. Potts, and T. Icard. Causal abstraction for faithful model interpretation. *arXiv*, 2023. (Link).

Gemini Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv*, 2023. (Link).

Gemini Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv*, 2024. (Link).

T. Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, 2021.

J. Gilmer, L. Metz, F. Faghri, S. S. Schoenholz, M. Raghu, M. Wattenberg, and I. Goodfellow. Adversarial spheres. *arXiv*,2018.(Link).

A. Glaese, N. McAleese, M. Trębacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv*,2022.(Link).

S. Glazunov and M. Brand. Project naptime: Evaluating offensive security capabilities of large language models. Google Project Zero Blog, 2024. (Link).

A. Gleave and G. Irving. Uncertainty estimation for language reward models. *arXiv*, 2022. (Link).

A. Gleave, M. Dennis, C. Wild, N. Kant, S. Levine, and S. Russell. Adversarial policies: Attacking deep reinforcement learning. *arXiv*, 2019. (Link).

G. Goh, N. Cammarata, C. Voss, S. Carter, M. Petrov, L. Schubert, A. Radford, and C. Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. (Link).

N. Goldowsky-Dill, C. MacLeod, L. Sato, and A. Arora. Localizing model behavior with path patching. *arXiv*, 2023. (Link).

N. Goldowsky-Dill, B. Chughtai, S. Heimersheim, and M. Hobbhahn. Detecting strategic deception using linear probes, 2025. (Link).

J. M. Gómez, M. Verdú, A. González-Megías, and M. Méndez. The phylogenetic roots of human lethal violence. *Nature*, 538(7624):233–237, 2016. (Link).

I. J. Good. Speculations on perceptrons and other automata. International Business Machines Corporation, 1959. (Link).

I. J. Good. Some future social repercussions of computers. *International Journal of Environmental Studies*, 1(1-4):67–79, 1970. (Link).

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv*, 2014. (Link).

Google. Gemini 2.0 Flash Thinking mode, 2024. (Link).

Google Cloud. How google protects its production services, 2024. (Link).

GoogleDeepMind.UpdatingtheFrontierSafetyFramework. *Google DeepMind Blog*2025.(Link).

A. Gopal, N. Helm-Burger, L. Justen, E. H. Soice, T. Tzeng, G. Jeyapragasan, S. Grimm, B. Mueller, and K. M. Esvelt. Will releasing the weights of future large language models grant widespread access to pandemic agents? *arXiv*, 2023. (Link).

K. Grace. Algorithmic progress in six domains. *Machine Intelligence Research Institute*, 2013. (Link).

K. Grace. Likelihood of discontinuous progress around the development of AGI, 2020. Published Feb 2018; last updated 26 May 2020. (Link).

K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans. When will AI exceed human performance? evidence from AI experts. *Journal of Artificial Intelligence Research*, 62:729–754, 2018. (Link).

K. Grace, R. Korzekwa, A. Bergal, and D. Kokotajlo. Discontinuous progress investigation. Technical report, AI Impacts, 2021. (Link).

K. Grace, Z. Stein-Perlman, B. Weinstein-Raun, and J. Salvatier. Expert survey on progress in AI. AI Impacts, 2022. (Link).

K. Grace, H. Stewart, J. F. Sandkühler, S. Thomas, B. Weinstein-Raun, and J. Brauner. Thousands of AI authors on the future of AI. *arXiv*, 2024. (Link).

R. Greenblatt and B. Shlegeris. The case for ensuring that powerful AIs are controlled, 2024a. (Link).

R. Greenblatt and B. Shlegeris. Managing catastrophic misuse without robust AIs, 2024b. (Link).

R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, et al. Alignment faking in large language models. *arXiv*, 2024a. (Link).

R. Greenblatt, F. Roger, D. Krasheninnikov, and D. Krueger. Stress-testing capability elicitation with password-locked models. *arXiv*, 2024b. (Link).

R. Greenblatt, B. Shlegeris, K. Sachan, and F. Roger. AI control: Improving safety despite intentional subversion. In *International Conference on Machine Learning*, 2024c. (Link).

C. Griffin, D. Wallace, J. Mateos-Garcia, H. Schieve, and P. Kohli. Seizing the AI for Science opportunity, 2024. (Link).

Z. Griliches. Issues in assessing the contribution of research and development to productivity growth. *The bell journal of economics*, pages 92–116, 1979. (Link).

J. Gross, R. Agrawal, T. Kwa, E. Ong, C. H. Yip, A. Gibson, S. Noubir, and L. Chan. Compact proofs of model performance via mechanistic interpretability. In *Neural Information Processing Systems*, 2024.

R. Grosse. Three sketches of ASL-4 safety case components. Anthropic Alignment Science Blog, 2024. (Link).

R. Grosse, J. Bae, C. Anil, N. Elhage, A. Tamkin, A. Tajdini, B. Steiner, D. Li, E. Durmus, E. Perez, et al. Studying large language model generalization with influence functions. *arXiv*, 2023. (Link).

M. Y. Guan, M. Joglekar, E. Wallace, S. Jain, B. Barak, A. Heylar, R. Dias, A. Vallone, H. Ren, J. Wei, et al. Deliberative alignment: Reasoning enables safer language models. *arXiv*, 2024. (Link).

P. Guo, A. Syed, A. Sheshadri, A. Ewart, and G. K. Dziugaite. Mechanistic unlearning: Robust knowledgeunlearningandeditingviamechanisticlocalization. *arXiv*,2024.(Link).

W. Gurnee, N. Nanda, M. Pauly, K. Harvey, D. Troitskii, and D. Bertsimas. Finding neurons in a haystack:Casestudieswithsparseprobing. *arXiv*,2023.(Link).

S. Gust, E. A. Hanushek, and L. Woessmann. Global universal basic skills: Current deficits and implicationsforworlddevelopment. *Journal of Development Economics* ,166:103205,2024.(Link).

D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan. Cooperative inverse reinforcement learning. *Neural information processing systems*,29, 2016. (Link).

D. Halawi, A. Wei, E. Wallace, T. T. Wang, N. Haghtalab, and J. Steinhardt. Covert malicious finetuning: Challengesinsafeguardingllmadaptation. *arXiv*,2024.(Link).

B.H.Hall.Theprivateandsocialreturnstoresearchanddevelopment. *NBER Working Paper*,1996. (Link).

J.L.Hall.Columbiaandchallenger:organizationalfailureatnasa. *Space Policy*,19(4):239–247, 2003. (Link).

M. Hanna, O. Liu, and A. Variengien. How does GPT-2 compute greater-than?: Interpreting mathematicalabilitiesinapre-trainedlanguagemodel. *Neural information processing systems*,36,2024. (Link).

R. Hanson. Long-term growth as a sequence of exponential modes, 2000. (Link).

R. Hanson. Economic growth given machine intelligence, 2001. (Link).

R. Hanson. *The age of Em: Work, love, and life when robots rule the earth*. Oxford University Press, 2016.

L. Hanu and Unitary team. Detoxify. Github, 2020. (Link).

S. Hao, S. Sukhbaatar, D. Su, X. Li, Z. Hu, J. Weston, and Y. Tian. Training large language models to reason in a continuous latent space. *arXiv*, 2024. (Link).

Y. N. Harari. *21 Lessons for the 21st Century*. Random House, 2018.

Y. N. Harari. *Nexus: A brief history of information networks from the Stone Age to AI*. Signal, 2024.

P. Hase, M. Diab, A. Celikyilmaz, X. Li, Z. Kozareva, V. Stoyanov, M. Bansal, and S. Iyer. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2714–2731, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. (Link).

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. (Link).

S. Heimersheim and N. Nanda. How to use and interpret activation patching. *arXiv*, 2024. (Link).

P. Hemmer, M. Westphal, M. Schemmer, S. Vetter, M. Vössing, and G. Satzger. Human-AI collaboration: the effect of AI delegation on human task performance and task satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 453–463, 2023.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv*, 2020. (Link).

D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt. Unsolved problems in ML safety. *arXiv*, 2021. (Link).

D. Hendrycks, M. Mazeika, and T. Woodside. An overview of catastrophic AI risks. *arXiv*, 2023. (Link).

O. Henkel, L. Hills, A. Boxer, B. Roberts, and Z. Levonian. Can large language models make the grade? an empirical study evaluating LLMs ability to mark short answer questions in K-12 education. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, pages 300–304, 2024. (Link).

D. Hernandez and T. B. Brown. Measuring the algorithmic efficiency of neural networks. *arXiv*, 2020. (Link).

J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y.Zhou.Deeplearningscalingispredictable,empirically. *arXiv*,2017.(Link).

G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network, 2015. (Link).

A. Ho, T. Besiroglu, E. Erdil, D. Owen, R. Rahman, Z. C. Guo, D. Atkinson, N. Thompson, and J. Sevilla. Algorithmicprogressinlanguagemodels. *arXiv*,2024.(Link).

T.Hobbes. *Leviathan: Or, the Matter, Forme, and Power of a Common-Wealth Ecclesiasticall and Civill* Printed for Andrew Crooke, London, 1651. Often referred to as the "First Edition" despite not being explicitly numbered.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv*, 2022. (Link).

J. Hong, S. Levine, and A. Dragan. Zero-shot goal-directed dialogue via RL on imagined conversations. *arXiv*, 2023. (Link).

J. Hoogland, G. Wang, M. Farrugia-Roberts, L. Carroll, S. Wei, and D. Murfet. The developmental landscape of in-context learning. *arXiv*, 2024. (Link).

N. Howe, I. McKenzie, O. Hollinsworth, M. Zajac, T. Tseng, A. Tucker, P.-L. Bacon, and A. Gleave. Effects of scale on language model robustness. *arXiv*, 2024. (Link).

C.-Y. Hsu, Y.-L. Tsai, C.-H. Lin, P.-Y. Chen, C.-M. Yu, and C.-Y. Huang. Safe LoRA: the silver lining of reducing safety risks when fine-tuning large language models. *arXiv*, 2025. (Link).

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. (Link).

J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 42(08):2011–2023, 2020. (Link).

N. Hu, B. Wright, C. Denison, S. Marks, J. Treutlein, J. Uesato, and E. Hubinger. Training on documents about reward hacking induces reward hacking. *Anthropic Blog*, January 2025. (Link).

S. Hu, X. Liu, X. Han, X. Zhang, C. He, W. Zhao, Y. Lin, N. Ding, Z. Ou, G. Zeng, et al. Predicting emergent abilities with infinite resolution evaluation. In *International Conference on Learning Representations*, 2023. (Link).

J. Huang, A. Geiger, K. D'Oosterlinck, Z. Wu, and C. Potts. Rigorously assessing natural language explanations of neurons. *arXiv*, 2023. (Link).

E.Hubinger.Gradienthacking.AlignmentForum,2019.(Link).

E. Hubinger. When can we trust model evaluations? Alignment Forum, 2023. (Link).

E. Hubinger, C. Denison, J. Mu, M. Lambert, M. Tong, M. MacDiarmid, T. Lanham, D. M. Ziegler, T. Maxwell, N. Cheng, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv*, 2024. (Link).

M. Hutter, D. Quarel, and E. Catt. *An Introduction to Universal Artificial Intelligence*. CRC Press, 2024.

H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, D. Testuggine, and M. Khabsa. Llama guard: LLM-based input-output safeguard for human-AI conversations. *arXiv*, 2023. (Link).

International Labour Organisation. Labour share of GDP, 2022. Original data accessed via Our World in Data. (Link).

G. Irving. Safety cases at AISI. Association of International Safety-Critical-System Institutions Website, aug 2024. (Link).

G. Irving and A. Askell. AI safety needs social scientists. *Distill*, 4(2):e14, 2019. (Link).

G. Irving, P. Christiano, and D. Amodei. AI safety via debate. *arXiv*, 2018. (Link).

G. Irving, T. Korbak, and B. Hilton. Automation collapse, 2024. (Link).

A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, et al. OpenAI o1 system card. *arXiv*, 2024. (Link).

S. Jain, R. Kirk, E. S. Lubana, R. P. Dick, H. Tanaka, E. Grefenstette, T. Rocktäschel, and D. S. Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv*, 2023. (Link).

M. Jakesch, J. T. Hancock, and M. Naaman. Human heuristics for AI-generated language are flawed. *PNAS*, 2023. (Link).

J. Jang, D. Yoon, S. Yang, S. Cha, M. Lee, L. Logeswaran, and M. Seo. Knowledge unlearning for mitigating privacy risks in language models. *arXiv*, 2022. (Link).

H. J. Jeon, S. Milli, and A. Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426, 2020. (Link).

J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang, et al. AI alignment: A comprehensive survey. *arXiv*, 2023. (Link).

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. *arXiv*, 2023. (Link).

M. Jiang, M. Dennis, J. Parker-Holder, A. Lupu, H. Küttler, E. Grefenstette, T. Rocktäschel, and J. Foerster. Grounding aleatoric uncertainty for unsupervised environment design. *Advances in Neural Information Processing Systems*, 35:32868–32881, 2022. (Link).

C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan. SWE-bench: Can language models resolve real-world Github issues? *arXiv*, 2023. (Link).

A. L. Jones. Scaling scaling laws with board games. *arXiv*, 2021. (Link).

B. F. Jones and L. H. Summers. A calculation of the social returns to innovation. *National Bureau of Economic Research*, 27863, 2020. (Link).

C. I. Jones. R&D-based models of economic growth. *Journal of Political Economy*, 103(4):759–784, 1995. (Link).

E. Jones, A. Dragan, A. Raghunathan, and J. Steinhardt. Automatically auditing large language models via discrete optimization. *International Conference on Machine Learning*, 202:1934–1954, 2023. (Link).

E. Jones, A. Dragan, and J. Steinhardt. Adversaries can misuse combinations of safe models. *arXiv*, 2024. (Link).

L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2): 29–48, 2022. (Link).

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. (Link).

S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv*, 2022. (Link).

D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv*, 2020. (Link).

B. Karlan. Human achievement and artificial intelligence. *Ethics and Information Technology*, 25(3): 40, 2023. (Link).

A. Karpathy. Deep neural nets: 33 years ago and 33 years from now, 2022. (Link).

A. Karvonen, D. Pai, M. Wang, and B. Keigwin. Sieve: SAEs beat baselines on a real-world task (a code generation case study). *Tilde Research Blog*, dec 2024. (Link).

K. Kelly. The Maes-Garreau point, 2007. Accessed: 2024-06. (Link).

Z. Kenton, T. Everitt, L. Weidinger, I. Gabriel, V. Mikulik, and G. Irving. Alignment of language agents. *arXiv*, 2021. (Link).

Z. Kenton, N. Y. Siegel, J. Kramár, J. Brown-Cohen, S. Albanie, J. Bulian, R. Agarwal, D. Lindner, Y. Tang, N. D. Goodman, et al. On scalable oversight with weak LLMs judging strong LLMs. *arXiv*, 2024. (Link).

J. Keynes. *Economic possibilities for our grandchildren - Essays in persuasion*. Springer, 1931.

A. Khan, J. Hughes, D. Valentine, L. Ruis, K. Sachan, A. Radhakrishnan, E. Grefenstette, S. R. Bowman, T. Rocktäschel, and E. Perez. Debating with more persuasive LLMs leads to more truthful answers. *arXiv*, 2024. (Link).

D. Kharlapenko, S. Shabalin, F. Barez, N. Nanda, and A. Conmy. Scaling sparse feature circuits for studying in-context learning, 2025. (Link).

H. Kim, X. Yi, J. Yao, J. Lian, M. Huang, S. Duan, J. Bak, and X. Xie. The road to artificial superintelligence:Acomprehensivesurveyofsuperalignment. *arXiv*,2024.(Link).

M. Kinniment, L. J. K. Sato, H. Du, B. Goodrich, M. Hasin, L. Chan, L. H. Miles, T. R. Lin, H. Wijk, J. Burget, et al. Evaluating language-model agents on realistic autonomous tasks. *arXiv*, 2023. (Link).

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. (Link).

H. Kitano. Nobel Turing Challenge: creating the engine for scientific discovery. *NPJ systems biology and applications*, 7(1):29, 2021. (Link).

P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017. (Link).

R. Koncel-Kedziorski, S. Roy, A. Amini, N. Kushman, and H. Hajishirzi. MAWPS: A math word problemrepository.In*Associationforcomputationallinguistics:humanlanguagetechnologies*,pages 1152–1157, 2016. (Link).

E. Konopinski, C. Marvin, and E. Teller. Ignition of the atmosphere with nuclear bombs. *Report LA-602. Los Alamos, NM: Los Alamos Laboratory*, 1946. (Link).

T. Korbak, K. Shi, A. Chen, R. Bhalerao, C. L. Buckley, J. Phang, S. R. Bowman, and E. Perez. Pretraining language models with human preferences. *arXiv*, 2023. (Link).

T. Korbak, J. Clymer, B. Hilton, B. Shlegeris, and G. Irving. A sketch of an AI control safety case. *arXiv*, 2025. (Link).

A. Korinek. Economic policy challenges for the age of AI. Technical report, National Bureau of Economic Research, 2024. (Link).

R. Korzekwa and H. Stewart. Views of prominent AI developers on risk from AI, 2023. (Link).

J. Kossen, J. Han, M. Razzak, L. Schut, S. Malik, and Y. Gal. Semantic entropy probes: Robust and cheaphallucinationdetectioninLLMs. *arXiv*,2024.(Link).

V. Krakovna, J. Uesato, V. Mikulik, M. Rahtz, T. Everitt, R. Kumar, Z. Kenton, J. Leike, and S. Legg. Specification gaming: the flip side of AI ingenuity. GDM blog, 2020. (Link).

R. M. Kramer, E. Newton, and P. L. Pommerenke. Self-enhancement biases and negotiator judgment: Effectsofself-esteemandmood. *Organizational Behavior and Human Decision Processes*,56(1): 110–133, 1993. (Link).

J. Kramár, T. Lieberum, R. Shah, and N. Nanda. AtP*: An efficient and scalable method for localizing LLMbehaviourtocomponents. *arXiv*,2024.(Link).

D.Krasheninnikov,E.Krasheninnikov,andD.Krueger.Out-of-contextmeta-learninginlargelanguage models.In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023. (Link).

L. M. Krauss and G. D. Starkman. Universal limits on computation. *arXiv*, 2004. (Link).

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Neural information processing systems*,25,2012.

J. A. Krosnick. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3):213–236, 1991.

D. Krueger, T. Maharaj, and J. Leike. Hidden incentives for auto-induced distributional shift. In *Workshop on Safe Machine Learning at the 7th International Conference on Learning Representations (ICLR 2019)*, pages 1–7, 2019. (Link).

L. Kuhn, Y. Gal, and S. Farquhar. Semantic uncertainty: Linguistic invariants for uncertainty estimation innaturallanguagegeneration.In *International Conference on Learning Representations*,2023. (Link).

T.S.Kuhn. *The Structure of Scientific Revolutions*.UniversityofChicagoPress,1962.

J. Kulveit, R. Douglas, N. Ammann, D. Turan, D. Krueger, and D. Duvenaud. Gradual disempowerment: SystemicexistentialrisksfromincrementalAIdevelopment. *arXiv*,2025.(Link).

H. Kumar, D. M. Rothschild, D. G. Goldstein, and J. M. Hofman. Math education with large language models:Perilorpromise? *SSRN 4641653* ,2023.(Link).

F. Kunstner, P. Hennig, and L. Balles. Limitations of the empirical fisher approximation for natural gradientdescent. *Neural information processing systems*,32,2019.(Link).

R. Kurzweil. By 2029 no computer - or "machine intelligence" - will have passed the Turing Test, 2002. Predictor: Mitchell Kapor, Challenger: Ray Kurzweil. (Link).

R.Kurzweil.Thesingularityisnear.In *Ethics and emerging technologies*,pages393–406.Springer, 2005.

V. Lai and C. Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 29–38, 2019. (Link).

R. Laine, B. Chughtai, J. Betley, K. Hariharan, M. Balesni, J. Scheurer, M. Hobbhahn, A. Meinke, and O. Evans. Me, myself, and AI: The situational awareness dataset (SAD) for LLMs. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. (Link).

B. Lake and M. Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequencerecurrentnetworks.In *International conference on machine learning*,pages2873–2882. PMLR, 2018. (Link).

Y. Lakretz, G. Kruszewski, T. Desbordes, D. Hupkes, S. Dehaene, and M. Baroni. The emergence of numberandsyntaxunitsinLSTMlanguagemodels. *arXiv*,2019.(Link).

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimationusingdeepensembles. *Advances in neural information processing systems* ,30,2017. (Link).

L. Langosco, J. Koch, L. Sharkey, J. Pfau, L. Orseau, and D. Krueger. Goal misgeneralization in deep reinforcement learning. *arXiv*, 2023. (Link).

T. Lanham, A. Chen, A. Radhakrishnan, B. Steiner, C. Denison, D. Hernandez, D. Li, E. Durmus, E. Hubinger, J. Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv*, 2023. (Link).

A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learnedsimilaritymetric.In *International conference on machine learning*pages1558–1566.PMLR, 2016. (Link).

W. Laurito, S. Maiya, G. Dhimoïla, K. Hänni, et al. Cluster-norm for unsupervised probing of knowledge. *arXiv*, 2024. (Link).

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Back-propagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. (Link).

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. (Link).

B. W. Lee, I. Padhi, K. N. Ramamurthy, E. Miehling, P. Dognin, M. Nagireddy, and A. Dhurandhar. Programming refusal with conditional activation steering. *arXiv*, 2024. (Link).

J. D. Lee and N. Moray. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, 40, 1994. (Link).

J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46, 2004. (Link).

S. Legg and M. Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17:391–444, 2007. (Link).

J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv*, 2018. (Link).

J. Leike, L. Weng, and J. Wu. Our approach to alignment research. OpenAI Blog, 2022. (Link).

N. G. Leveson. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press, 2016.

J. Li. A comparative study on annotation quality of crowdsourcing and LLM via label aggregation. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. (Link).

K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Neural Information Processing Systems*, 36, 2024a. (Link).

K. Li, S. Zhuang, Y. Zhang, M. Xu, R. Wang, K. Xu, X. Fu, and X. Cheng. I'm spartacus, no, i'm spartacus: Measuring and understanding LLM identity confusion. *arXiv*, 2024b. (Link).

N. Li, Z. Han, I. Steneker, W. Primack, R. Goodside, H. Zhang, Z. Wang, C. Menghini, and S. Yue. LLM defenses are not robust to multi-turn human jailbreaks yet. *arXiv*, 2024c. (Link).

N. Li, A. Pan, A. Gopal, S. Yue, A. Gatti, J. D. Li, A.-K. Dombrowski, S. Goel, L. Phan, G. Mukobi, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv*, 2024d. (Link).

Y. Li, D. Choi, J. Chung, N. Kushman, J. Schrittwieser, R. Leblond, T. Eccles, J. Keeling, F. Gimeno, A. Dal Lago, et al. Competition-level code generation with AlphaCode. *Science*, 378(6624): 1092–1097, 2022. (Link).

Y. Li, F. Wei, J. Zhao, C. Zhang, and H. Zhang. RAIN: Your language models can align themselves withoutfinetuning. *arXiv*,2023.(Link).

Z. Li and D. Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. (Link).

T. Lieberum, M. Rahtz, J. Kramár, N. Nanda, G. Irving, R. Shah, and V. Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in Chinchilla. *arXiv*, 2023. (Link).

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step. *arXiv*, 2023. (Link).

S. Lin, J. Hilton, and O. Evans. Teaching models to express their uncertainty in words. *arXiv*, 2022. (Link).

P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable AI: A review of machine learning interpretabilitymethods. *Entropy*,23(1):18,2020.(Link).

C. Liu, T. Arnon, C. Lazarus, C. Strong, C. Barrett, M. J. Kochenderfer, et al. Algorithms for verifying deepneuralnetworks. *Foundations and Trends® in Optimization* ,4(3-4):244–404,2021a.(Link).

J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen. What makes good in-context examples for GPT-3? *arXiv*,2021b.(Link).

Y. Liu, G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, L. Zhao, T. Zhang, and K. Wang. A hitchhiker's guide to jailbreaking chatgpt via prompt engineering. In *Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things (SEAIoT)*, *Proceedings of the 4th International Workshop on*, pages 12–21. ACM, 2024. (Link).

S.Lloyd.Ultimatephysicallimitstocomputation. *Nature*,406(6799):1047–1054,2000.(Link).

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. (Link).

L. Luo, Y. Liu, R. Liu, S. Phatale, M. Guo, H. Lara, Y. Li, L. Shu, Y. Zhu, L. Meng, et al. Improve mathematicalreasoninginlanguagemodelsbyautomatedprocesssupervision. *arXiv* ,2024.(Link).

A. Ly, M. Marsman, J. Verhagen, R. P. Grasman, and E.-J. Wagenmakers. A tutorial on Fisher information. *Journal of Mathematical Psychology*,80:40–55,2017.(Link).

C. Lyle and R. Pascanu. Switching between tasks can cause AI to lose the ability to learn. *Nature*, 632:745–747, aug 2024. (Link).

A. Lynch, P. Guo, A. Ewart, S. Casper, and D. Hadfield-Menell. Eight methods to evaluate robust unlearninginLLMs. *arXiv*,2024.(Link).

C. Ma, Z. Yang, H. Ci, J. Gao, M. Gao, X. Pan, and Y. Yang. Evolving diverse red-team language models inmulti-roundmulti-agentgames. *arXiv*,2024.(Link).

S. Ma, Y. Lei, X. Wang, C. Zheng, C. Shi, M. Yin, and X. Ma. Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*,2023. (Link).

D.J.C.MacKay.Information-basedobjectivefunctionsforactivedataselection. *Neural Computation*, 4(4):590–604, jul 1992. ISSN 0899-7667. (Link).

A. Mądry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant toadversarialattacks. *International Conference on Learning Representations*,2018.(Link).

A. Makelov, G. Lange, and N. Nanda. Is this the subspace you are looking for? an interpretability illusionforsubspaceactivationpatching. *arXiv*,2023.(Link).

A. Mallen, M. Brumley, J. Kharchenko, and N. Belrose. Eliciting latent knowledge from quirky language models. *arXiv*, 2023. (Link).

A. Mallen, C. Griffin, A. Abate, and B. Shlegeris. Subversion strategy eval: Evaluating ai's stateless strategiccapabilitiesagainstcontrolprotocols. *arXiv*,2024.(Link).

A. Manzini, G. Keeling, N. Marchal, K. R. McKee, V. Rieser, and I. Gabriel. Should users trust advanced AI assistants? Justified trust as a function of competence and alignment. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*2024. (Link).

B. Maples, M. Cerit, A. Vishwanath, and R. Pea. Loneliness and suicide mitigation for students using GPT3-enabledchatbots. *NPJ mental health research*,3(1):4,2024.(Link).

N. Marchal, R. Xu, R. Elasmar, I. Gabriel, B. Goldberg, and W. Isaac. Generative AI misuse: A taxonomy oftacticsandinsightsfromreal-worlddata. *arXiv*,abs/2406.13843,2024.(Link).

S. Marks and M. Tegmark. The geometry of truth: Emergent linear structure in large language model representationsoftrue/falsedatasets. *arXiv*,Oct2023.(Link).

S. Marks, C. Rager, E. J. Michaud, Y. Belinkov, D. Bau, and A. Mueller. Sparse feature circuits: Discoveringandeditinginterpretablecausalgraphsinlanguagemodels. *arXiv*,2024.(Link).

S. Marks, J. Treutlein, T. Bricken, J. Lindsey, J. Marcus, S. Mishra-Sharma, D. Ziegler, et al. Auditing languagemodelsforhiddenobjectives. *Anthropic*,2025.(Link).

Y. Mathew, O. Matthews, R. McCarthy, J. Velja, C. S. de Witt, D. Cope, and N. Schoots. Hidden in plain text: Emergence & mitigation of steganographic collusion in LLMs. *arXiv*, 2024. (Link).

M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv*, 2024. (Link).

N. McAleese, R. M. Pokorny, J. F. C. Uribe, E. Nitishinskaya, M. Trebacz, and J. Leike. LLM critics help catch LLM bugs. *arXiv*, 2024. (Link).

C. McDougall, A. Conmy, C. Rushing, T. McGrath, and N. Nanda. Copy suppression: Comprehensively understanding an attention head. *arXiv*, 2023. (Link).

T. McGrath, M. Rahtz, J. Kramár, V. Mikulik, and S. Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv*, 2023. (Link).

T. McGrath, D. Balsam, M. Deng, and E. Ho. Understanding and steering Llama 3 with sparse autoencoders, 2024. (Link).

I. R. McKenzie, A. Lyzhov, M. Pieler, A. Parrish, A. Mueller, A. Prabhu, E. McLean, A. Kirtland, A. Ross, A. Liu, et al. Inverse scaling: When bigger isn't better. *arXiv*, 2023. (Link).

A. Mehrotra, M. Zampetakis, P. Kassianik, B. Nelson, H. Anderson, Y. Singer, and A. Karbasi. Tree of attacks:Jailbreakingblack-boxLLMsautomatically. *arXiv*,2023.(Link).

A. Meinke, B. Schoen, J. Scheurer, M. Balesni, R. Shah, and M. Hobbhahn. Frontier models are capable ofin-contextscheming. *arXiv*,2024.(Link).

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. *Neural information processing systems*,35:17359–17372, 2022. (Link).

K. Meng, V. Huang, N. Chowdhury, D. Choi, J. Steinhardt, and S. Schwettmann. Monitor: An AI-driven observability interface, 2024. (Link).

A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk. Scaling deep learning formaterialsdiscovery. *Nature* ,624(7990):80–85,2023.(Link).

W. Merrill and A. Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*,11:531–545, 2023. (Link).

Metaculus. Before 2030, will an AI complete the Turing Test in the Kurzweil/Kapor Longbet? Metaculus, 2024. Accessed: 2024-10-28. (Link).

METR.Guidelinesforcapabilityelicitation. *METR's Autonomy Evaluation Resources*,2024.(Link).

J. Michael, S. Mahdi, D. Rein, J. Petty, J. Dirani, V. Padmakumar, and S. R. Bowman. Debate helps supervise unreliable experts. *arXiv*, 2023. (Link).

Microsoft Azure. Abuse monitoring. Microsoft Azure AI Services Content Safety Article, 2024a. (Link).

Microsoft Azure. Prompt shields. Microsoft Azure AI Services Content Safety Article, 2024b. (Link).

Microsoft Threat Intelligence. Staying ahead of threat actors in the age of AI. Microsoft Security Blog, 2024. (Link).

T. Mikolov. Efficient estimation of word representations in vector space. *arXiv*, 2013. (Link).

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Neural information processing systems*, 26, 2013. (Link).

M. Milani Fard, K. Canini, A. Cotter, J. Pfeifer, and M. Gupta. Fast and flexible monotonic functions with ensembles of lattices. *Neural Information Processing Systems*, 29, 2016. (Link).

J. Miller, B. Chughtai, and W. Saunders. Transformer circuit faithfulness metrics are not robust. *arXiv*, 2024. (Link).

J. W. Milnor. Games against nature. *Decision Processes*, 1954. (Link).

T. M. Mitchell. The need for biases in learning generalizations. Technical report, Rutgers University, 1980. (Link).

T. M. Mitchell. How can AI accelerate science, and how can our government help?, July 2024. (Link).

G. E. Moore. Cramming more components onto integrated circuits. *IEEE*, 1965. (Link).

G. E. Moore et al. Progress in digital integrated electronics. In *Electron devices meeting*, volume 21, pages 11–13. IEEE, 1975. (Link).

F. Moorhouse and W. MacAskill. Preparing for the intelligence explosion. *Forethought*, 2025. Accessed: 2025-03-16. (Link).

J. Moos, K. Hansel, H. Abdulsamad, S. Stark, D. Clever, and J. Peters. Robust reinforcement learning: Areviewoffoundationsandrecentadvances. *Machine Learning & Knowledge Extraction* ,4(1): 276–315, 2022. (Link).

H. Moravec. When will computer hardware match the human brain? *Journal of evolution and technology*, 1(1):10, 1998. (Link).

M. R. Morris, J. Sohl-Dickstein, N. Fiedel, T. Warkentin, A. Dafoe, A. Faust, C. Farabet, and S. Legg. Levels of AGI: Operationalizing progress on the path to AGI. In *International Conference on Machine Learning*, 2023. (Link).

S.MorrisonandC.Winston. *The economic effects of airline deregulation*BrookingsInstitutionPress, 2010.

L. N. Moses and I. Savage. Aviation deregulation and safety: Theory and evidence*Journal of Transport Economics and Policy*, pages 171–188, 1990. (Link).

S. R. Motwani, M. Baranchuk, M. Strohmeier, V. Bolina, P. H. Torr, L. Hammond, and C. S. de Witt. Secret collusion among generative AI agents. *arXiv*, 2024. (Link).

C. A. Mouton, C. Lucas, and E. Guest. The operational risks of AI in large-scale biological attacks, 2024. (Link).

J. Mu and J. Andreas. Compositional explanations of neurons. *Neural Information Processing Systems*, 33:17153–17163, 2020. (Link).

A. Mueller, J. Brinkmann, M. Li, S. Marks, K. Pal, N. Prakash, C. Rager, A. Sankaranarayanan, A. S. Sharma, J. Sun, et al. The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability. *arXiv*, 2024. (Link).

J. Murphy, C. Hofacker, and R. Mizerski. Primacy and recency effects on clicking behavior. *Journal of computer-mediated communication*, 11(2):522–535, 2006. (Link).

S. Na, Y. J. Choe, D.-H. Lee, and G. Kim. Discovery of natural language concepts in individual units of cnns. *arXiv*, 2019. (Link).

D. S. Nagin. Deterrence: A review of the evidence by a criminologist for economists. *Annual Reviews in Economics*, 5(1):83–105, 2013.

R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv*, 2021. (Link).

N. Nanda. Attribution patching: Activation patching at industrial scale, 2023. (Link).

N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv*, 2023a. (Link).

N. Nanda, A. Lee, and M. Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, Singapore, Dec 2023b. Association for Computational Linguistics. (Link).

N. Nanda, S. Rajamanoharan, J. Kramár, and R. Shah. Fact finding: Attempting to reverse-engineer factual recall on the neuron level. Alignment Forum, Dec 2023c. (Link).

National Cyber Security Centre (NCSC). Risk management: Threat modelling, 2023. Accessed: 2025-01-07. (Link).

S. Nevo, D. Lahav, A. Karpur, Y. Bar-On, H.-A. Bradley, and J. Alstott. *Securing AI model weights: Preventing theft and misuse of frontier models*. Rand Corporation, 2024. (Link).

R. Ngo, L. Chan, and S. Mindermann. The alignment problem from a deep learning perspective. In *International Conference on Learning Representations*, 2024. (Link).

NIH Office of Intramural Research. Dual-use research, 2023. (Link).

W. D. Nordhaus. Are we approaching an economic singularity? information technology and the future of economic growth. *American Economic Journal: Macroeconomics*, 13(1):299–332, 2021. (Link).

nostalgebraist. Interpreting GPT: the logit lens, 2020. (Link).

nostalgebraist. The case for CoT unfaithfulness is overstated, 2024. (Link).

B. Nour, M. Pourzandi, and M. Debbabi. A survey on threat hunting in enterprise networks. *IEEE Communications Surveys & Tutorials*, 25(4):2299–2324, 2023. (Link).

C. Olah. Interpretability dreams, 2023. (Link).

C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. Zoom in: An introduction to circuits. *Distill*, 2020. (Link).

S. M. Omohundro. The basic AI drives. In *Artificial General Intelligence*, 2008. (Link).

OpenAI. Openai preparedness framework (beta), 2023. (Link).

OpenAI. Disrupting deceptive uses of AI by covert influence operations, may 2024. Accessed: 2025-01-10. (Link).

OpenAI. Response to NIST Executive Order on AI, Feb 2024. (Link).

OpenAI. Disrupting malicious uses of AI by state-affiliated threat actors, 2024. (Link).

OpenAI. Learning to reason with LLMs, 2024. Accessed: 2024-10-23. (Link).

OpenAI. Building an early warning system for LLM-aided biological threat creation, Jan 2024a. Accessed: 2025-01-07. (Link).

OpenAI. o1 System Card. Technical report, OpenAI, Dec 2024b. (Link).

OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, et al. GPT-4 technical report. *arXiv*, 2023. (Link).

T. Ord. *The precipice: Existential risk and the future of humanity*. Hachette Books, 2020.

Our World In Data. Self-reported life satisfaction vs. GDP per capita, 2023. Our World in Data, 2023a. (Link).

Our World In Data. Literacy rate vs. GDP per capita, 2023. Our World in Data, 2023b. (Link).

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *arXiv*, 2022. (Link).

D. Owen. How predictable is language model benchmark performance? *arXiv*, 2024. (Link).

L. Pacchiardi, A. J. Chan, S. Mindermann, I. Moscovitz, A. Y. Pan, Y. Gal, O. Evans, and J. Brauner. How to catch an AI liar: Lie detection in black-box LLMs by asking unrelated questions. *arXiv*, Sep 2023. (Link).

D. Paleka, A. P. Sudhir, A. Alvarez, V. Bhat, A. Shen, E. Wang, and F. Tramèr. Consistency checks for language model forecasters. *arXiv*, 2024. (Link).

Q. Pang, J. Zhu, H. Möllering, W. Zheng, and T. Schneider. BOLT: Privacy-preserving, accurate and efficient inference for transformers. In *2024 IEEE Symposium on Security and Privacy*, 2024. (Link).

R. Y. Pang, A. Parrish, N. Joshi, N. Nangia, J. Phang, A. Chen, V. Padmakumar, J. Ma, J. Thompson, H. He, et al. QuALITY: Question answering with long input texts, yes! *arXiv*, 2021. (Link).

K. Park, Y. J. Choe, and V. Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv*, 2023a. (Link).

K. Park, Y. J. Choe, Y. Jiang, and V. Veitch. The geometry of categorical and hierarchical concepts in large language models. *arXiv*, 2024. (Link).

P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks. AI deception: A survey of examples, risks, and potential solutions. *arXiv*, 2023b. (Link).

S. M. Park, K. Georgiev, A. Ilyas, G. Leclerc, and A. Mądry. Trak: attributing model behavior at scale. In *International Conference on Machine Learning*. JMLR, 2023c. (Link).

A. Parrish, H. Trivedi, N. Nangia, V. Padmakumar, J. Phang, A. S. Saimbhi, and S. R. Bowman. Two-turn debate doesn't help humans answer hard reading comprehension questions. *arXiv*, 2022a. (Link).

A. Parrish, H. Trivedi, E. Perez, A. Chen, N. Nangia, J. Phang, and S. R. Bowman. Single-turn debate does not help humans answer hard reading-comprehension questions. *arXiv*, 2022b. (Link).

O. Patel and R. Wang. Preventing memorized completions via white-box filtering. *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024. (Link).

S. G. Patil, T. Zhang, V. Fang, R. Huang, A. Hao, M. Casado, J. E. Gonzalez, R. A. Popa, I. Stoica, et al. GoEX: perspectives and designs towards a runtime for autonomous LLM applications. *arXiv*, 2024. (Link).

G. Paulo, A. Mallen, C. Juang, and N. Belrose. Automatically interpreting millions of features in large language models. *arXiv*, 2024. (Link).

A. Peppin, A. Reuel, S. Casper, E. Jones, A. Strait, U. Anwar, A. Agrawal, S. Kapoor, S. Koyejo, M. Pellat, et al. The reality of AI and biorisk. *arXiv*, 2024. (Link).

E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. Red teaming language models with language models. *arXiv*, 2022a. (Link).

E. Perez, S. Ringer, K. Lukošiū ̄tė ̇, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv*, 2022b. (Link).

J. Pfau, W. Merrill, and S. R. Bowman. Let's think dot by dot: Hidden computation in transformer language models. *arXiv*, 2024. (Link).

M. Phuong, M. Aitchison, E. Catt, S. Cogan, A. Kaskasoli, V. Krakovna, D. Lindner, M. Rahtz, Y. Assael, S. Hodkinson, et al. Evaluating frontier models for dangerous capabilities. *arXiv*, 2024. (Link).

M. Phute, A. Helbling, M. Hull, S. Peng, S. Szyller, C. Cornelius, and D. H. P. Chau. LLM self defense: By self-examination, LLMs know they are being tricked. *arXiv*, 2023. (Link).

S. Pinker. *The better angels of our nature: The decline of violence in history and its causes*. PenguinUK, 2011.

O. Popeti. Reducing misuse to an access problem. *Noema Research Blog*, Jun 2024. (Link).

G. Pruthi, F. Liu, S. Kale, and M. Sundararajan. Estimating training data influence by tracing gradient descent. *Neural Information Processing Systems*, 33:19920–19930, 2020. (Link).

X. Qi, Y. Zeng, T. Xie, P.-Y. Chen, R. Jia, P. Mittal, and P. Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv*, 2023. (Link).

X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang, and P. Mittal. Visual adversarial examples jailbreak aligned large language models. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 21527–21536, 2024a. (Link).

X. Qi, A. Panda, K. Lyu, X. Ma, S. Roy, A. Beirami, P. Mittal, and P. Henderson. Safety alignment should be made more than just a few tokens deep. *arXiv*, 2024b. (Link).

Qwen Team. QwQ: Reflect deeply on the boundaries of the unknown, 2024. (Link).

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI*, 1(8):9, 2019. (Link).

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *arXiv*, 2021. (Link).

A. Radhakrishnan. Anthropic Fall 2023 debate progress update, 2023. (Link).

A. Radhakrishnan, K. Nguyen, A. Chen, C. Chen, C. Denison, D. Hernandez, E. Durmus, E. Hubinger, J. Kernion, K. Lukošiūtė, et al. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv*, 2023a. (Link).

A. Radhakrishnan, B. Shlegeris, R. Greenblatt, and F. Roger. Scalable oversight and weak-to-strong generalization: Compatible approaches to the same problem, 2023b. (Link).

J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv*, 2021. (Link).

R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv*, 2023. (Link).

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

S. Rajamanoharan, A. Conmy, L. Smith, T. Lieberum, V. Varma, J. Kramár, R. Shah, and N. Nanda. Improving sparse decomposition of language model activations with gated sparse autoencoders. In *Neural Information Processing Systems*, 2024. (Link).

V. V. Ramasesh, A. Lewkowycz, and E. Dyer. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*, 2022. (Link).

D. Raposo, S. Ritter, B. Richards, T. Lillicrap, P. C. Humphreys, and A. Santoro. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv*, 2024. (Link).

T. Räuker, A. Ho, S. Casper, and D. Hadfield-Menell. Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE conference on secure and trustworthy machine learning (SATML)*, pages 464–483. IEEE, 2023. (Link).

J. Reason. The contribution of latent human failures to the breakdown of complex systems. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 327(1241):475–484, 1990. (Link).

A. Rechkemmer and M. Yin. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–14, 2022. (Link).

L. Reid. AI overviews: About last week, May 2024. (Link).

D. Rein. NYU code debates update/postmortem, 2024. (Link).

D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv*, 2023. (Link).

F. Ren, A. Aliper, J. Chen, H. Zhao, S. Rao, C. Kuppe, I. V. Ozerov, M. Zhang, K. Witte, C. Kruse, et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models. *Nature Biotechnology*, pages 1–13, 2024. (Link).

A. Rey, K. Le Goff, M. Abadie, and P. Courrieu. The primacy order effect in complex decision making. *Psychological Research*, 84(6):1739–1748, 2020.

F. Roger. Coup probes: Catching catastrophes with probes trained off-policy, 2023. (Link).

F. Roger and R. Greenblatt. Preventing language models from hiding their reasoning. *arXiv*, 2023. (Link).

A. Rogers, O. Kovaleva, and A. Rumshisky. A primer in BERTology: what we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2021. (Link).

B. Romera-Paredes, M. Barekatain, A. Novikov, M. Balog, M. P. Kumar, E. Dupont, F. J. Ruiz, J. S. Ellenberg, P. Wang, O. Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. (Link).

L. Ross, D. Greene, and P. House. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, 13(3):279–301, 1977. (Link).

Y. Ruan, C. J. Maddison, and T. Hashimoto. Observational scaling laws and the predictability of language model performance. *arXiv*, 2024. (Link).

L. Ruis, M. Mozes, J. Bae, S. R. Kamalakara, D. Talupuru, A. Locatelli, R. Kirk, T. Rocktäschel, E. Grefenstette, and M. Bartolo. Procedural knowledge in pretraining drives reasoning in large language models. *arXiv*, 2024. (Link).

C. Rushing and N. Nanda. Explorations of self-repair in language models. *arXiv*, 2024. (Link).

S. Ruder. An overview of gradient descent optimization algorithms. *arXiv*, 2016. (Link).

S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017. (Link).

M. Samvelyan, S. C. Raparthy, A. Lupu, E. Hambro, A. H. Markosyan, M. Bhatt, Y. Mao, M. Jiang, J. Parker-Holder, J. Foerster, T. Rocktäschel, and R. Raileanu. Rainbow teaming: Open-ended generation of diverse adversarial prompts, 2024. (Link).

A. Sandberg. An overview of models of technological singularity. *The transhumanist reader: Classical and contemporary essays on the science, technology, and philosophy of the human future*, pages 376–394, 2013. (Link).

S. Sanyal, S. Addepalli, and R. V. Babu. Towards data-free model stealing in a hard label setting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15284–15293, 2022. (Link).

N. Saphra and A. Lopez. Understanding learning dynamics of language models with svcca. *arXiv*, 2018. (Link).

W. Saunders, C. Yeh, J. Wu, S. Bills, L. Ouyang, J. Ward, and J. Leike. Self-critiquing models for assisting human evaluators. *arXiv*, 2022. (Link).

R. Schaeffer, B. Miranda, and S. Koyejo. Are emergent abilities of large language models a mirage? *arXiv*, 2023. (Link).

R. Schaeffer, H. Schoelkopf, B. Miranda, G. Mukobi, V. Madan, A. Ibrahim, H. Bradley, S. Biderman, and S. Koyejo. Why has predicting downstream capabilities of frontier AI models with scale remained elusive? *arXiv*, 2024. (Link).

T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Prioritized experience replay. *arXiv*, 2015. (Link).

P. Schoenegger, P. S. Park, E. Karger, S. Trott, and P. E. Tetlock. AI-augmented predictions: LLM assistants improve human forecasting accuracy. *ACM Transactions on Interactive Intelligent Systems*, 2024. (Link).

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv*, 2017. (Link).

R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Why did you say that? *arXiv*, 2016. (Link).

A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020. (Link).

J. Sevilla and E. Roldán. Training compute of frontier AI models grows by 4-5x per year, 2024. (Link).

J. Sevilla, T. Besiroglu, B. Cottier, J. You, E. Roldán, P. Villalobos, and E. Erdil. Can AI scaling continue through 2030?, 2024. Accessed: 2024-10-23. (Link).

R. Shah. AI alignment 2018-19 review. Alignment Forum, Jan 2020. (Link).

R. Shah, N. Gundotra, P. Abbeel, and A. Dragan. On the feasibility of learning, rather than assuming, human biases for reward inference. In *International Conference on Machine Learning*, volume 97, pages 5670–5679. PMLR, jun 2019. (Link).

R. Shah, P. Freire, N. Alex, R. Freedman, D. Krasheninnikov, L. Chan, M. D. Dennis, P. Abbeel, A. Dragan, and S. Russell. Benefits of assistance over reward learning, 2020. (Link).

R. Shah, V. Varma, R. Kumar, M. Phuong, V. Krakovna, J. Uesato, and Z. Kenton. Goal misgeneralization: Whycorrectspecificationsaren'tenoughforcorrectgoals. *arXiv*,2022.(Link).

R. Shah, S. Pour, A. Tagade, S. Casper, J. Rando, et al. Scalable and transferable black-box jailbreaks forlanguagemodelsviapersonamodulation. *arXiv*,2023.(Link).

T. R. Shaham, S. Schwettmann, F. Wang, A. Rajaram, E. Hernandez, J. Andreas, and A. Torralba. A multimodalautomatedinterpretabilityagent.In *International Conference on Machine Learning*, 2024. (Link).

M. Shanahan, K. McDonell, and L. Reynolds. Role play with large language models. *Nature*, 623 (7987):493–498, 2023. (Link).

A. S. Sharma, N. Sarkar, V. Chundawat, A. A. Mali, and M. Mandal. Unlearning or concealment? a critical analysis and evaluation metrics for unlearning in diffusion models. *arXiv*, 2024a. (Link).

M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, E. DURMUS, Z. Hatfield-Dodds, S. R. Johnston, S. M. Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024b. (Link).

M. Sharma, M. Tong, J. Mu, J. Wei, J. Kruthoff, S. Goodfriend, E. Ong, A. Peng, R. Agarwal, C. Anil, et al. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. *arXiv*, 2025. (Link).

X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang. "Do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv*, 2023. (Link).

T. Shevlane and A. Dafoe. The offense-defense balance of scientific knowledge: Does publishing AI research reduce misuse? *arXiv*, 2020. (Link).

T. Shevlane, S. Farquhar, B. Garfinkel, M. Phuong, J. Whittlestone, J. Leung, D. Kokotajlo, N. Marchal, M. Anderljung, N. Kolt, et al. Model evaluation for extreme risks. *arXiv*, 2023. (Link).

B. Shlegeris. How to prevent collusion when using untrusted models to monitor each other. Alignment Forum, 2024a. (Link).

B. Shlegeris. Win/continue/lose scenarios and execute/replace/audit protocols, 2024b. (Link).

B. Shneiderman. *Human-Centered AI*. OUP Oxford, 2022. ISBN 9780192660008. (Link).

O. Shorinwa, Z. Mei, J. Lidard, A. Z. Ren, and A. Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *arXiv*, 2024. (Link).

A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv*, 2016. (Link).

C. Si, N. Goyal, S. T. Wu, C. Zhao, S. Feng, H. Daumé III, and J. Boyd-Graber. Large language models help humans verify truthfulness–except when they are convincingly wrong. *arXiv*, 2023. (Link).

N. Y. Siegel, O.-M. Camburu, N. Heess, and M. Perez-Ortiz. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. *arXiv*, 2024. (Link).

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. (Link).

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017. (Link).

D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018. (Link).

H. A. Simon. Rational choice and the structure of the environment. *Psychological review*, 63(2):129, 1956. (Link).

H. A. Simon. *The new science of management decision*. Harper & Row, 1960.

K. Simonyan. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv*, 2013. (Link).

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014. (Link).

K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025. (Link).

M. Skjuve, A. Følstad, K. I. Fostervold, and P. B. Brandtzaeg. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies*, 168:102903, 2022. (Link).

D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv*, 2017. (Link).

L. Smith. The strong feature hypothesis could be wrong, 2024. (Link).

C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv*, 2024. (Link).

N. Soares, B. Fallenstein, S. Armstrong, and E. Yudkowsky. Corrigibility. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, 2015. (Link).

A. Souly, Q. Lu, D. Bowen, T. Trinh, E. Hsieh, S. Pandey, P. Abbeel, J. Svegliato, S. Emmons, O. Watkins, et al. A strongREJECT for empty jailbreaks. In *Neural Information Processing Systems*, 2024. (Link).

J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv*, 2014. (Link).

A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv*, 2023. (Link).

D. Stander, Q. Yu, H. Fan, and S. Biderman. Grokking group multiplication with cosets. *arXiv*, 2023. (Link).

G.Stankevičiū ̄tė ̇.WhatisEnhancedDueDiligence(EDD)?[WithExamples],dec2023.(Link).

A. C. Stickland, A. Lyzhov, J. Pfau, S. Mahdi, and S. R. Bowman. Steering without side effects: Improvingpost-deploymentcontroloflanguagemodels. *arXiv*,2024.(Link).

N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano. Learning to summarize from human feedback. *arXiv*, 2022. (Link).

C. Summerfield, L. Argyle, M. Bakker, T. Collins, E. Durmus, T. Eloundou, I. Gabriel, D. Ganguli, K. Hackenburg, G. Hadfield, et al. How will advanced AI systems impact democracy? *arXiv*, 2024. (Link).

C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. (Link).

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. (Link).

J. Sušnik and P. van der Zaag. Correlation and causation between the un human development index and national and personal wealth and resource exploitation. *Economic research-Ekonomska istraživanja*, 30(1):1705–1723, 2017. (Link).

R. Sutton. The bitter lesson. Incomplete Ideas (blog), 2019. (Link).

M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv*, 2022. (Link).

K. Tadepalli. What we know about economic growth in LMICs. Effective Altruism Forum, Dec 2023. (Link).

J. Taylor. Quantilizers: A safer alternative to maximizers for limited optimization. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016. (Link).

J. Taylor, E. Yudkowsky, P. LaVictoire, and A. Critch. Alignment for advanced machine learning systems. *Ethics of artificial intelligence*, pages 342–382, 2016. (Link).

A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. (Link).

Thomson Reuters Legal. Regulatory compliance on adverse media screening, feb 2024. (Link).

R. Tian, L. Sun, A. Bajcsy, M. Tomizuka, and A. D. Dragan. Safety assurances for human-robot interaction via confidence-aware game-theoretic human models. *arXiv*, 2021. (Link).

C. Tigges, M. Hanna, Q. Yu, and S. Biderman. LLM circuit analyses are consistent across training and scale. *arXiv*, 2024. (Link).

P. Trammell and A. Korinek. Economic growth under transformative AI. Technical report, National Bureau of Economic Research, 2023. (Link).

J. Treutlein, D. Choi, J. Betley, S. Marks, C. Anil, R. B. Grosse, and O. Evans. Connecting the dots: LLMs can infer and verbalize latent structure from disparate training data. *Advances in Neural Information Processing Systems*,37:140667–140730, 2024. (Link).

J.-B. Truong, P. Maini, R. J. Walls, and N. Papernot. Data-free model extraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4771–4780, 2021. (Link).

Y.-L. Tsai, C.-Y. Hsu, C. Xie, C.-H. Lin, J.-Y. Chen, B. Li, P.-Y. Chen, C.-M. Yu, and C.-Y. Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv*, 2023. (Link).

A. M. Turner. Self-fulfilling misalignment data might be poisoning our AI models. The Pond (blog), Mar 2025. (Link).

A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid. Activation addition: Steering language models without optimization. *arXiv*, 2023. (Link).

N. L. Turner, A. Jermyn, and J. Batson. Measuring feature sensitivity using dataset filtering, 2024. (Link).

M. Turpin, J. Michael, E. Perez, and S. Bowman. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Neural Information Processing Systems*, 36, 2024. (Link).

A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

J. Uesato, R. Kumar, V. Krakovna, T. Everitt, R. Ngo, and S. Legg. Avoiding tampering incentives in deep RL via decoupled approval. *arXiv*, 2020. (Link).

UK Ministry of Defence. Safety management requirements for defence systems: Part 1: Requirements. Defence Standard 00-56, UK Ministry of Defence, Feb 2017. (Link).

F. Urbina, F. Lentzos, C. Invernizzi, and S. Ekins. Dual use of artificial intelligence-powered drug discovery. *Nature Machine Intelligence*, 4:189–191, 2022. (Link).

M. Vaccaro, A. Almaatouq, and T. Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, pages 1–11, 2024. (Link).

T. van der Weij, F. Hofstätter, O. Jaffe, S. F. Brown, and F. R. Ward. AI sandbagging: Language models can strategically underperform on evaluations. In *International Conference on Learning Representations*, 2025. (Link).

A. Variengien and E. Winsor. Look before you leap: A universal emergent decomposition of retrieval tasks in language models. *arXiv*, 2023. (Link).

H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction 7, CSCW1 (2023)*, 2023. (Link).

A. Vaswani. Attention is all you need. *Neural Information Processing Systems*, 2017. (Link).

A. S. Vezhnevets, J. P. Agapiou, A. Aharon, R. Ziv, J. Matyas, E. A. Duéñez-Guzmán, W. A. Cunningham, S. Osindero, D. Karmon, and J. Z. Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *arXiv*, 2023. (Link).

J. Vig and Y. Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv*, 2019. (Link).

J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer, and S. Shieber. Investigating gender biasinlanguagemodelsusingcausalmediationanalysis. *Neural information processing systems*, 33: 12388–12401, 2020. (Link).

P. Villalobos and D. Atkinson. Trading off compute in training and inference, 2023. Accessed: 2024-10-21. (Link).

J. Vincent. Microsoft's bing is an emotionally manipulative liar, and people love it. The Verge, 2023. (Link).

K. Vonnegut. *Cat's Cradle*. Holt, Rinehart and Winston, 1963. (Link).

E. Wallace, A. Williams, R. Jia, and D. Kiela. Analyzing dynamic adversarial training data in the limit. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 202–217, 2022. (Link).

E. Wallace, K. Xiao, R. Leike, L. Weng, J. Heidecke, and A. Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions. *arXiv*, 2024. (Link).

S. Wan, C. Nikolaidis, D. Song, D. Molnar, J. Crnkovich, J. Grace, M. Bhatt, S. Chennabasappa, S. Whitman, S. Ding, et al. CYBERSECEVAL 3: Advancing the evaluation of cybersecurity risks and capabilities in large language models. *arXiv*, 2024. (Link).

F. Wang and C. Rudin. Falling rule lists. In *Artificial intelligence and statistics*, pages 1013–1022. PMLR, 2015. (Link).

G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv*, 2023a. (Link).

H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023b. (Link).

K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. *arXiv*, 2022. (Link).

P. Wang, D. Zhang, L. Li, C. Tan, X. Wang, K. Ren, B. Jiang, and X. Qiu. InferAligner: inference-time alignment for harmlessness through cross-model guidance. *arXiv*, 2024a. (Link).

X. Wang, H. Kim, S. Rahman, K. Mitra, and Z. Miao. Human-LLM collaborative annotation through effective verification of LLM labels. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024b. (Link).

Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv*, 2023c. (Link).

Z. Wang, A. Ku, J. Baldridge, T. L. Griffiths, and B. Kim. Gaussian process probes (GPP) for uncertainty-aware probing. *arXiv*, 2023d. (Link).

A. Wei, N. Haghtalab, and J. Steinhardt. Jailbroken: How does LLM safety training fail? *Neural information processing systems*, 36, 2024. (Link).

J. Wei and K. Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv*, 2019. (Link).

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D.Metzler,etal.Emergentabilitiesoflargelanguagemodels. *arXiv*,2022a.(Link).

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought promptingelicitsreasoninginlargelanguagemodels. *Neural information processing systems*, 35: 24824–24837, 2022b. (Link).

J. Wen, V. Hebbar, C. Larson, A. Bhatt, A. Radhakrishnan, M. Sharma, H. Sleight, S. Feng, H. He, E. Perez, et al. Adaptive deployment of untrusted llms reduces distributed threats. *arXiv*, 2024. (Link).

J. Wentworth. Oversight misses 100% of thoughts the AI does not think. Alignment Forum, 2022. (Link).

T. White. Sampling generative networks. *arXiv*, 2016. (Link).

White House. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. White House, Washington, D.C., October 2023. (Link).

White House. Readout of President Joe Biden's meeting with president Xi Jinping of the People's Republic of China. White House Website, Nov 2024. (Link).

M. Wiescher and K. Langanke. Nuclear astrophysicists at war. *Natural Sciences*, 4(1):e20230023, 2024. (Link).

H. Wijk, T. Lin, J. Becker, S. Jawhar, N. Parikh, T. Broadley, L. Chan, M. Chen, J. Clymer, J. Dhyani, et al. RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. *arXiv*, 2024. (Link).

World Bank. The state of global learning poverty: 2022 update, Jun 2022. (Link).

T. Wu, L. Luo, Y.-F. Li, S. Pan, T.-T. Vu, and G. Haffari. Continual learning for large language models: A survey. *arXiv*, abs/2402.01364, 2024a. (Link).

W. Wu, L. Jaburi, J. Drori, and J. Gross. Unifying and verifying mechanistic interpretations: A case study with group operations. *arXiv*, 2024b. (Link).

Y. Wu, F. Roesner, T. Kohno, N. Zhang, and U. Iqbal. SecGPT: An execution isolation architecture for LLM-based systems. *arXiv*, 2024c. (Link).

K. Wynroe, D. Atkinson, and J. Sevilla. Literature review of transformative artificial intelligence timelines, 2023. (Link).

Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025. (Link).

M. Xia, S. Malladi, S. Gururangan, S. Arora, and D. Chen. Less: Selecting influential data for targeted instruction tuning. *arXiv*, 2024. (Link).

Z. Xu, F. Jiang, L. Niu, J. Jia, B. Y. Lin, and R. Poovendran. SafeDecoding: Defending against jailbreak attacks via safety-aware decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5587–5605. Association for Computational Linguistics, Aug 2024. (Link).

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

L. Yu, V. Do, K. Hambardzumyan, and N. Cancedda. Robust LLM safeguarding via refusal feature adversarialtraining. *arXiv*,2024.(Link).

Z. Yuan, H. Yuan, C. Li, G. Dong, K. Lu, C. Tan, C. Zhou, and J. Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv*, 2023. (Link).

E. Yudkowsky. Recursive self-improvement, 2008. (Link).

E.Yudkowsky.Intelligenceexplosionmicroeconomics.*MachineIntelligenceResearchInstitute*,2013. (Link).

Z. Yun, Y. Chen, B. A. Olshausen, and Y. LeCun. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. *arXiv*, 2021. (Link).

E. Zelikman, Y. Wu, J. Mu, and N. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Neural information processing systems*, volume 35, pages 15476–15488. Curran Associates, Inc., 2022. (Link).

W. Zeng, Y. Liu, R. Mullins, L. Peran, J. Fernandez, H. Harkous, K. Narasimhan, D. Proud, P. Kumar, B. Radharapu, et al. ShieldGemma: Generative AI content moderation based on gemma. *arXiv*, 2024. (Link).

Q. Zhan, R. Fang, R. Bindu, A. Gupta, T. Hashimoto, and D. Kang. Removing RLHF protections in gpt-4 via fine-tuning. *arXiv*, 2023. (Link).

F. Zhang and N. Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv*, 2023. (Link).

K. Zhang and A. B. Aslan. AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence*, 2:100025, 2021. (Link).

L. Zhang, A. Hosseini, H. Bansal, M. Kazemi, A. Kumar, and R. Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv*, 2024a. (Link).

Y. Zhang, Q. V. Liao, and R. K. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020. (Link).

Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021. (Link).

Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang. A survey of sparse representation: algorithms and applications. *IEEE access*, 3:490–530, 2015. (Link).

Z. Zhang, F. Wang, X. Li, Z. Wu, X. Tang, H. Liu, Q. He, W. Yin, and S. Wang. Does your LLM truly unlearn? an embarrassingly simple approach to recover unlearned knowledge. *arXiv*, 2024b. (Link).

Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021. (Link).

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Neural Information Processing Systems*, 36, 2024. (Link).

C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, et al. LIMA: less is more for alignment. *arXiv*,2023.(Link).

D. Ziegler, S. Nix, L. Chan, T. Bauman, P. Schmidt-Nielsen, T. Lin, A. Scherlis, N. Nabeshima, B. Weinstein-Raun, D. de Haas, et al. Adversarial training for high-stakes reliability. *Neural information processing systems*,35:9274–9286, 2022. (Link).

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *arXiv*, 2019. (Link).

A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv*, 2023a. (Link).

A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv*, 2023b. (Link).

A. Zou, L. Phan, J. Wang, D. Duenas, M. Lin, M. Andriushchenko, J. Z. Kolter, M. Fredrikson, and D. Hendrycks. Improving alignment and robustness with circuit breakers. In *Neural Information Processing Systems*, 2024. (Link).

R. Zwetsloot and A. Dafoe. Thinking about risks from AI: Accidents, misuse and structure. The Lawfare Institute, 2019. (Link).