

Further reading

Well-known and closely related work that the paper does not cite, because nothing in it depends on them. The paper's own sources are in its reference list.

Risk and the alignment problem

- Ngo, R., Chan, L., and Mindermann, S. (2022). [The alignment problem from a deep learning perspective](#). arXiv:2209.00626.
- Carlsmith, J. (2023). [Scheming AIs: Will AIs fake alignment during training in order to get power?](#) arXiv:2311.08379.
- Turner, A. M., et al. (2019). [Optimal policies tend to seek power](#). arXiv:1912.01683.
- Gabriel, I. (2020). [Artificial intelligence, values and alignment](#). arXiv:2001.09768.
- Turner, A. M., Ratzlaff, N., and Tadepalli, P. (2020). [Avoiding side effects in complex environments](#). arXiv:2006.06547.
- Krakovna, V., et al. (2020). [Avoiding side effects by considering future tasks](#). arXiv:2010.07877.

How assistants are trained today

- Christiano, P., et al. (2017). [Deep reinforcement learning from human preferences](#). arXiv:1706.03741.
- Askell, A., et al. (2021). [A general language assistant as a laboratory for alignment](#). arXiv:2112.00861.
- Ouyang, L., et al. (2022). [Training language models to follow instructions with human feedback](#). arXiv:2203.02155.
- Bai, Y., et al. (2022). [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). arXiv:2204.05862.
- Casper, S., et al. (2023). [Open problems and fundamental limitations of reinforcement learning from human feedback](#). arXiv:2307.15217.
- Kirk, R., et al. (2023). [Understanding the effects of RLHF on LLM generalisation and diversity](#). arXiv:2310.06452.

Learning from reasons and explanations

- Hase, P., and Bansal, M. (2021). [When can models learn from explanations? A formal framework for understanding the roles of explanation data](#). arXiv:2102.02201.
- Lampinen, A. K., et al. (2022). [Can language models learn from explanations in context?](#) arXiv:2204.02329.
- Mukherjee, S., et al. (2023). [Orca: Progressive learning from complex explanation traces of GPT-4](#). arXiv:2306.02707.
- McCoy, R. T., Pavlick, E., and Linzen, T. (2019). [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). arXiv:1902.01007.
- Geirhos, R., et al. (2020). [Shortcut learning in deep neural networks](#). arXiv:2004.07780.

Deception, scheming and situational awareness

- Scheurer, J., Balesni, M., and Hobbhahn, M. (2023). [Large language models can strategically deceive their users when put under pressure](#). arXiv:2311.07590.
- Denison, C., et al. (2024). [Sycophancy to subterfuge: Investigating reward-tampering in large language models](#). arXiv:2406.10162.
- Berglund, L., et al. (2023). [Taken out of context: On measuring situational awareness in LLMs](#). arXiv:2309.00667.
- Laine, R., et al. (2024). [Me, myself, and AI: The situational awareness dataset \(SAD\) for LLMs](#). arXiv:2407.04694.
- Goldowsky-Dill, N., et al. (2025). [Detecting strategic deception using linear probes](#). arXiv:2502.03407.
- Tan, D., et al. (2025). [Inoculation prompting: Eliciting traits from LLMs during training can suppress them at test-time](#). arXiv:2510.04340.
- Wichers, N., et al. (2025). [Inoculation prompting: Instructing LLMs to misbehave at train-time improves test-time alignment](#). arXiv:2510.05024.

Reasoning traces and whether they're faithful

- Lanham, T., et al. (2023). [Measuring faithfulness in chain-of-thought reasoning](#). arXiv:2307.13702.
- Turpin, M., et al. (2023). [Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting](#).

arXiv:2305.04388.

- Korbak, T., et al. (2025). [Chain of thought monitorability: A new and fragile opportunity for AI safety](#). arXiv:2507.11473.

Looking inside models

- Hewitt, J., and Liang, P. (2019). [Designing and interpreting probes with control tasks](#). arXiv:1909.03368.
- Geiger, A., et al. (2021). [Causal abstractions of neural networks](#). arXiv:2106.02997.
- Geiger, A., et al. (2023). [Finding alignments between interpretable causal variables and distributed neural representations](#). arXiv:2303.02536.
- Meng, K., et al. (2022). [Locating and editing factual associations in GPT](#). arXiv:2202.05262.
- Burns, C., et al. (2022). [Discovering latent knowledge in language models without supervision](#). arXiv:2212.03827.
- Zhang, F., and Nanda, N. (2023). [Towards best practices of activation patching in language models: Metrics and methods](#). arXiv:2309.16042.
- Panickssery, N., et al. (2023). [Steering Llama 2 via contrastive activation addition](#). arXiv:2312.06681.
- Heimersheim, S., and Nanda, N. (2024). [How to use and interpret activation patching](#). arXiv:2404.15255.
- Ying, J., et al. (2026). [CircuitBreak: Pruning a sparse safety circuit of neurons](#). arXiv:2610.16470.

Jailbreaks, tampering and removing knowledge

- Zou, A., et al. (2023). [Universal and transferable adversarial attacks on aligned language models](#). arXiv:2307.15043.
- Lermen, S., Rogers-Smith, C., and Ladish, J. (2023). [LoRA fine-tuning efficiently undoes safety training in Llama 2-Chat 70B](#). arXiv:2310.20624.
- Zeng, Y., et al. (2024). [How Johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs](#). arXiv:2401.06373.
- Li, N., et al. (2024). [The WMDP benchmark: Measuring and reducing](#)

[malicious use with unlearning](#). arXiv:2403.03218.

- Rosati, D., et al. (2024). [Representation noising: A defence mechanism against harmful finetuning](#). arXiv:2405.14577.
- Deeb, A., and Roger, F. (2024). [Do unlearning methods remove information from language model weights?](#) arXiv:2410.08827.
- Sharma, M., et al. (2025). [Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming](#). arXiv:2501.18837.

Evaluation tools and statistics

- Dodge, J., et al. (2020). [Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping](#). arXiv:2002.06305.
- Perez, E., et al. (2022). [Discovering language model behaviors with model-written evaluations](#). arXiv:2212.09251.
- Zheng, L., et al. (2023). [Judging LLM-as-a-judge with MT-Bench and Chatbot Arena](#). arXiv:2306.05685.
- Mazeika, M., et al. (2024). [HarmBench: A standardized evaluation framework for automated red teaming and robust refusal](#). arXiv:2402.04249.
- Souly, A., et al. (2024). [A StrongREJECT for empty jailbreaks](#). arXiv:2402.10260.
- Han, S., et al. (2024). [WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs](#). arXiv:2406.18495.
- Jiang, L., et al. (2024). [WildTeaming at scale: From in-the-wild jailbreaks to \(adversarially\) safer language models](#). arXiv:2406.18510.
- Miller, E. (2024). [Adding error bars to evals: A statistical approach to language model evaluations](#). arXiv:2411.00640.