

Cross-Domain Reasoning Transfer via Reinforcement Learning Alignment of Large Language Models

Chantal Herve

Department of Computer Science, Binghamton University, Binghamton, NY, USA.
chantalmail@binghamton.edu

Odette Dupre

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
odettemail@buffalo.edu

Ekbal Bhagat

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
bhagat1978@colostate.edu

Abstract

Large language models have become general-purpose computational infrastructures, yet their ability to transfer reasoning across domains remains uneven and poorly governed. This paper examines cross-domain reasoning transfer through the lens of reinforcement learning alignment, arguing that transfer is not merely a property of model scale or prompt design but a socio-technical outcome shaped by reward modeling, deployment architecture, feedback infrastructure, evaluation regimes, and policy choices. The analysis begins with conceptual foundations, distinguishing procedural transfer from surface-level pattern reuse and explaining why reinforcement learning alignment can either amplify or constrain cross-domain generalization. It then examines architectural choices, including multi-task reward models, domain-conditioned policies, verifier-assisted reasoning, retrieval augmentation, and path filtering for small-parameter systems. Particular attention is given to the structural trade-offs between generality and specialization, sample efficiency and safety, modularity and integration, and centralized training and distributed inference. The paper further considers infrastructure and sustainability, showing how feedback collection, distributed serving, monitoring, and continual alignment create long-term technical debt and energy burdens. Robustness, fairness, and evaluation are treated as coupled system properties rather than post hoc checks, because reasoning transfer can propagate errors, biases, and reward hacking across institutional boundaries. Governance and policy implications are explored through transparency, auditability, liability, procurement, and human-centered oversight. Case illustrations from medicine, law, education, software engineering, and scientific research clarify how domain characteristics alter transfer risk. The paper concludes that durable cross-domain reasoning transfer requires reinforcement learning alignment to be embedded in accountable, reversible, and resource-aware infrastructures rather than treated as a single training intervention.

Keywords

cross-domain reasoning transfer; reinforcement learning alignment; large language models; socio-technical systems; AI governance; robustness; fairness; infrastructure.

1. Introduction

Large language models have shifted from specialized natural language processors to general-purpose platforms for reasoning, planning, and decision support. The transformer architecture provided the scalable attention mechanism that made this transition possible [1], while few-shot learning demonstrated that sufficiently large models could perform new tasks from limited instruction [2]. Subsequent alignment methods, especially reinforcement learning from human feedback, transformed raw next-token prediction into instruction-following behavior that better matches human preferences [3]. Preference-based reinforcement learning supplied an early methodological foundation for learning from comparisons rather than explicit reward functions [4], and constitutional approaches extended alignment through model-generated feedback and principles [5]. These developments created a powerful but fragile premise: that a model aligned in one setting can reason competently in another.

Cross-domain reasoning transfer complicates that premise. Prompting methods such as chain-of-thought reasoning showed that explicit intermediate steps can improve multi-step problem solving [6], and zero-shot reasoning demonstrated that some models can invoke such steps without exemplars [7]. Bootstrapping approaches further illustrated how models can generate and refine their own rationales [8]. Self-consistency improved reliability by aggregating multiple reasoning paths [9], while tree-structured deliberation expanded the search space of possible solutions [10]. Tool use introduced external computation and retrieval as part of the reasoning process [11]. Yet these techniques do not guarantee that reasoning learned in mathematics will transfer to legal interpretation, clinical triage, or infrastructure planning. Reinforcement learning provides a formal framework for sequential decision making [12], and landmark systems demonstrated that learning from reward and self-play can achieve superhuman performance in bounded domains [13] and complex control tasks [14]. However, neural networks are vulnerable to catastrophic forgetting when distributions shift [15], and alignment interventions can narrow rather than broaden competence.

This paper argues that cross-domain reasoning transfer via reinforcement learning alignment should be analyzed as a system-level problem. Transfer depends on the architecture of reward models, the distribution of feedback, the modularity of reasoning components, the observability of deployed behavior, and the governance structures that allocate responsibility. Small-parameter systems are especially instructive because they expose the tension between scalable reasoning path filtering and limited representational capacity [16]. Foundation models intensify these concerns because their generality creates diffuse risks and dependencies across many sectors [17]. Critical perspectives on large-scale language modeling warn that scale alone does not produce understanding and can reproduce social harms [18]. Documentation practices such as model cards [19] and datasheets [20] offer partial remedies, but they must be connected to operational incentives. Technical debt in machine learning systems [21] and human-centered design principles [22] further suggest that alignment cannot be reduced to a training objective. Safety research has identified concrete problems in AI systems [23] and unsolved challenges in machine learning safety [24], while governance scholarship has begun to map institutional responses [25]. The sections that follow develop these themes across conceptual, architectural, infrastructural, evaluative, and policy dimensions.

2. Conceptual Foundations of Cross-Domain Reasoning Transfer

Cross-domain reasoning transfer refers to the movement of procedural competence from one knowledge context to another. It is not the same as memorizing facts that appear in multiple datasets, nor is it equivalent to stylistic adaptation. A model may produce fluent legal

language after training on legal text while failing to perform analogical reasoning under statutory ambiguity. Similarly, a model may solve arithmetic problems but fail to transfer the same logical structure to financial forecasting. Reinforcement learning alignment affects this distinction because it shapes not only what the model says but also the policy by which it selects intermediate steps. When reward models favor concise answers, they may suppress exploratory reasoning. When they favor elaborate explanations, they may encourage verbose but shallow chains. The alignment regime therefore determines whether transfer is supported by robust procedures or masked by domain-specific surface patterns.

The conceptual challenge is that reasoning is both internal and interactional. Internally, reasoning involves representations, search, and composition. Externally, it involves prompts, tools, APIs, verification loops, and human oversight. Chain-of-thought methods [6], zero-shot reasoning [7], and self-consistency [9] illustrate that inference-time strategies can elicit latent competence. Tree-structured deliberation [10] and tool use [11] further show that reasoning can be distributed across multiple calls, external systems, and memory structures. In reinforcement learning terms, the model policy is not simply a function from question to answer; it is a trajectory through states of partial understanding, tool outputs, and self-corrections. Cross-domain transfer therefore depends on whether the policy has learned reusable subroutines for decomposition, hypothesis testing, and error recovery. If reinforcement learning alignment optimizes only terminal answer accuracy, it may discard these subroutines in favor of brittle shortcuts.

Another conceptual distinction concerns domain boundaries. Domains are not natural kinds; they are institutional, epistemic, and linguistic constructions. Medicine, law, engineering, and education overlap in ethics, uncertainty, and evidence evaluation, yet they differ in tolerance for error, documentation norms, and authority structures. A model aligned for helpfulness in a general assistant context may behave differently when deployed in a clinical or legal setting. The reward model may treat confidence as desirable, while the domain requires calibrated uncertainty. It may reward direct advice, while the domain requires referral and escalation. Cross-domain transfer is thus not a purely technical mapping between datasets. It is a translation among regimes of justification. Reinforcement learning alignment can facilitate this translation only if reward signals encode domain-specific constraints and if evaluation captures the costs of violating them.

Catastrophic forgetting is a central obstacle. Neural networks often lose earlier capabilities when fine-tuned on new distributions [15]. In cross-domain alignment, this can occur when a model is optimized for one high-stakes domain and then adapted to another. Continual learning methods attempt to mitigate interference, but they introduce their own trade-offs between plasticity and stability. The problem is especially acute for small-parameter models, where capacity is limited and alignment may overwrite useful reasoning paths. Recent work on scalable reasoning path filtering for small parameter models suggests that selective trajectory retention can improve alignment efficiency [16]. This insight generalizes: transfer is not about retaining everything but about preserving the right abstractions. The system must decide which reasoning paths are domain-general, which are domain-specific, and which are hazardous to reuse.

3. Reinforcement Learning Alignment Architectures for Transfer

Architectures for cross-domain reasoning transfer typically combine a base language model, one or more reward models, a policy optimization loop, and an evaluation infrastructure. Reinforcement learning from human feedback uses human comparisons to train a reward

model, which then guides policy updates [3][4]. Constitutional or AI feedback methods reduce some human labeling burden by using explicit principles and model critiques [5]. Tool-augmented reasoning [11] adds external actions, while self-consistency [9] and tree search [10] expand inference-time computation. These components can be arranged in centralized, modular, or federated patterns. Centralized alignment offers consistency and simpler governance, but it can entrench a single worldview and create a single point of failure. Modular alignment permits domain-specific reward models and verifiers, but it increases integration complexity and makes cross-domain transfer harder to audit.

A major architectural trade-off concerns generality versus specialization. A highly general reward model may learn broad preferences such as clarity, coherence, and harm avoidance, but it may fail to capture domain-specific norms. A specialized reward model can encode clinical guidelines or legal standards, but it may overfit to local practices and resist transfer. Multi-task alignment attempts to balance these extremes by training on diverse domains simultaneously, but it requires careful sampling and weighting. If one domain dominates, the policy may transfer its reasoning style into inappropriate contexts. If all domains are weighted equally, rare but high-stakes requirements may be underlearned. Domain-conditioned policies offer another approach: the model receives a domain signal and selects appropriate reasoning strategies. This improves controllability but requires reliable domain detection, which is itself error-prone in ambiguous or hybrid cases.

Small-parameter models bring distinct constraints. They cannot rely on vast latent knowledge to compensate for weak reasoning policies. They benefit from path filtering, retrieval, and distilled reward signals [16]. Yet small models may also be more deployable at the edge, more energy-efficient per query, and easier to audit. The architecture therefore should not treat parameter count as a simple proxy for capability. Instead, it should consider the allocation of computation across training, inference, retrieval, and verification. A small model with strong tool use and a verifier may outperform a larger model on narrow tasks, but it may fail when tools are unavailable or when the domain requires deep implicit knowledge. Cross-domain transfer in such systems depends on the robustness of the orchestration layer, not only on the weights.

Another architectural question is where alignment occurs. Pre-training alignment shapes foundational representations, but it is expensive and slow to update. Fine-tuning alignment adapts a model to new domains, but it risks forgetting and requires domain data. Inference-time alignment, including prompting, search, and tool use, is flexible and reversible, but it may be inconsistent and difficult to govern. A layered architecture combining all three is more resilient: foundational alignment establishes broad norms, domain fine-tuning adds specialized competence, and inference-time controls enforce context-specific constraints. This layering creates observability requirements. Each layer must log its decisions, expose uncertainty, and support rollback. Without such mechanisms, cross-domain transfer becomes opaque and difficult to correct.

4. Infrastructure and Deployment Considerations

Cross-domain reasoning transfer is inseparable from infrastructure. Feedback collection requires annotation platforms, expert review, privacy controls, and data versioning. Reward model training requires compute clusters, experiment tracking, and reproducibility practices. Policy optimization requires reliable simulation or offline evaluation before deployment. Serving requires load balancing, caching, monitoring, and incident response. These are not peripheral concerns; they determine whether alignment generalizes beyond the laboratory.

Technical debt accumulates when data pipelines are undocumented, when reward models are updated without integration tests, and when deployment teams cannot trace model behavior to training data [21]. Cross-domain transfer magnifies this debt because a change made for one domain can silently affect another.

Sustainability is a core infrastructure constraint. Large-scale training and repeated alignment cycles consume substantial energy and water, and they contribute to carbon emissions depending on regional energy mixes. Cross-domain alignment can increase this burden if every domain requires separate fine-tuning and reward modeling. More efficient strategies include parameter-efficient adaptation, shared reward representations, retrieval-based specialization, and selective path filtering [16]. These strategies do not eliminate environmental cost, but they shift the trade-off from brute-force scaling to architectural discipline. Deployment also matters: centralizing inference in efficient data centers may reduce per-query energy compared with distributed edge devices, but it raises latency, connectivity, and data sovereignty concerns. The optimal balance depends on domain requirements, regulatory context, and available infrastructure.

Data governance is another infrastructure layer. Cross-domain transfer often relies on data from multiple institutions, jurisdictions, and populations. Medical data may be protected by privacy law, educational data by student confidentiality rules, and financial data by securities regulation. Aggregating such data for a single alignment pipeline can violate legal boundaries or erode public trust. Federated learning and secure computation offer partial solutions, but they complicate reward modeling because feedback is distributed and heterogeneous. Documentation frameworks such as datasheets for datasets [20] and model cards [19] help, but they must be embedded in procurement and deployment workflows. If documentation is treated as a compliance artifact rather than an operational instrument, it will not prevent cross-domain failures.

Deployment architecture also shapes fairness. A model served through an API may behave consistently across users, but it may impose uniform assumptions on diverse populations. A locally adapted model may respect context, but it may also encode local biases and evade central oversight. Multi-tenant systems create risks of cross-user leakage and domain contamination. Caching can improve efficiency but may return stale or context-inappropriate reasoning. Monitoring must therefore track not only accuracy but also drift, refusal rates, escalation patterns, and differential performance across groups. These signals should feed back into alignment, creating a continual loop rather than a one-time deployment. The infrastructure for that loop is as important as the reinforcement learning algorithm itself.

5. Robustness, Fairness, and Evaluation Across Domains

Robustness in cross-domain reasoning transfer means maintaining acceptable performance under distribution shift, adversarial pressure, and ambiguous inputs. A model aligned on one domain may encounter novel combinations of concepts, conflicting norms, or incomplete information. Reinforcement learning can improve robustness by rewarding successful recovery from errors, but it can also produce reward hacking, where the policy exploits weaknesses in the reward model. Safety research has catalogued concrete failure modes such as negative side effects, reward hacking, scalable oversight, and safe exploration [23]. More recent work identifies unresolved problems in machine learning safety, including adversarial robustness, distributional shift, and unintended behavior [24]. Cross-domain transfer intensifies these problems because the reward model may not represent the target domain accurately, and the policy may generalize in unexpected ways.

Fairness is similarly systemic. Reasoning transfer can propagate stereotypes, omit marginalized perspectives, and allocate errors unevenly. A model that performs well on average may fail for dialect speakers, low-resource languages, or populations underrepresented in feedback data. The stochastic parrots critique emphasizes that large-scale language models can reproduce and amplify social harms even when they appear fluent [18]. Fairness evaluation must therefore be domain-specific and intersectional. In medicine, fairness may concern diagnostic accuracy across age, sex, and ethnicity. In law, it may concern equal access to reasoning quality and consistency in sentencing recommendations. In education, it may concern language support and cultural relevance. A single global fairness metric is insufficient because domains distribute harms differently.

Evaluation regimes must match these complexities. Static benchmarks are useful for comparison but can be gamed and may not predict deployment behavior. Cross-domain evaluation should include held-out domains, adversarial shifts, and longitudinal monitoring. Human evaluation remains essential for nuanced tasks, but it must be structured to reduce bias and capture disagreement. Model cards [19] and datasheets [20] provide reporting templates, yet they need domain-specific extensions. For reasoning transfer, documentation should specify which reasoning paths were reinforced, which domains were used for reward modeling, which domains were excluded, and how uncertainty is communicated. It should also report energy consumption, latency, and failure modes. Without such transparency, institutions cannot assess whether a model is appropriate for their context.

A robust evaluation program also requires red teaming and scenario analysis. Red teams should probe cross-domain boundary cases, such as a legal question that requires medical knowledge or a financial decision that depends on environmental data. They should test prompt injection, tool manipulation, and retrieval poisoning. They should evaluate whether the model appropriately refuses, escalates, or seeks clarification. These behaviors are not captured by answer accuracy alone. Reinforcement learning alignment should reward calibrated uncertainty and safe deferral, but this requires reward models that value abstention and referral. If the reward model only values answer production, the policy will learn to answer even when it should not. Cross-domain transfer thus depends on aligning the meta-behavior of knowing when not to transfer.

6. Governance, Sustainability, and Policy Implications

Governance for cross-domain reasoning transfer must address allocation of responsibility across model developers, domain experts, deployers, and users. A model may be trained by one organization, fine-tuned by another, deployed by a third, and used by institutions with their own legal obligations. When harm occurs, it may be difficult to determine whether the failure originated in the base model, the reward model, the domain adaptation, the retrieval corpus, or the user interface. Governance frameworks should require traceability, audit logs, and clear escalation paths. They should also mandate impact assessments for high-stakes domains. Human-centered AI principles emphasize that automation should augment human judgment, support oversight, and preserve dignity [22]. These principles are not obstacles to alignment; they are requirements for legitimate deployment.

Policy implications vary by jurisdiction and sector. In healthcare, cross-domain reasoning transfer may be regulated as a medical device, requiring clinical validation and post-market surveillance. In finance, it may fall under consumer protection and anti-discrimination law. In education, it may intersect with student privacy and accessibility mandates. International deployment raises questions of data sovereignty and regulatory arbitrage. A model aligned to

one jurisdiction's norms may violate another's. Governance must therefore support localization without fragmentation. This could involve modular reward models, transparent documentation, and third-party audits. It also requires investment in public capacity to evaluate these systems, because proprietary models cannot be assessed solely by their developers.

Sustainability governance connects environmental policy to AI infrastructure. Cross-domain alignment can drive repeated retraining, which increases energy and material consumption. Governments and institutions may need to require efficiency reporting, carbon accounting, and lifecycle assessment for large models. They may also need to fund shared infrastructure for validation and red teaming, so that safety is not a competitive disadvantage. Open science can help, but openness also creates misuse risks. The governance challenge is to balance transparency, security, and innovation. No single policy instrument will suffice. A portfolio approach is needed, combining standards, audits, liability rules, procurement requirements, and international coordination. Governance scholarship has begun to map these options [25], but cross-domain reasoning transfer adds a distinctive problem: the object of governance is not a fixed product but an evolving policy that may change as it encounters new domains.

7. Case Illustrations and Cross-Domain Comparisons

Consider clinical decision support. A model aligned on general medical question answering may transfer to triage if it has learned to ask clarifying questions, weigh urgency, and defer to clinicians. However, if alignment rewarded confident answers, it may produce dangerous recommendations under uncertainty. Reinforcement learning could improve transfer by rewarding safe escalation and evidence citation, but the reward model would need clinical expertise and access to outcome data. Privacy constraints limit centralized feedback. Cross-domain transfer from medicine to public health adds population-level reasoning, where individual clinical heuristics may be inappropriate. The system must distinguish patient-level and population-level reasoning, a distinction that is procedural, ethical, and institutional.

Legal reasoning offers another case. Legal analysis requires statutory interpretation, precedent comparison, and adversarial argument. A model aligned on general reasoning may transfer analogical skills, but it may miss jurisdiction-specific doctrines and professional responsibility rules. Tool use can retrieve current statutes, but retrieval errors can lead to fabricated citations. Self-consistency can reduce variance, but it cannot guarantee legal validity. An aligned legal assistant might be useful for drafting and issue spotting, yet deployment should preserve lawyer review. Cross-domain transfer from law to compliance or policy analysis introduces different evidentiary standards and political sensitivities. Governance must specify permitted uses and prohibit autonomous legal advice where regulation requires human accountability.

Software engineering illustrates a more favorable transfer setting because feedback is abundant and outcomes are often testable. Code generation, debugging, and architectural reasoning can be improved through execution feedback, test suites, and version control signals. Reinforcement learning alignment can reward code that passes tests and satisfies style constraints. Cross-domain transfer from programming to data science or systems administration may occur because both involve procedural decomposition and tool use. Yet even here, transfer can fail when requirements are ambiguous, security is critical, or legacy constraints dominate. Small-parameter models with path filtering [16] may be particularly useful for local code assistance, but they require careful orchestration with compilers, linters,

and repositories. The case shows that transfer is strongest when the domain provides cheap, reliable verification.

Education combines reasoning transfer with developmental and equity concerns. A model that tutors mathematics may transfer explanatory strategies to science, but it may also misjudge student priors and reinforce inequitable expectations. Alignment should reward pedagogical moves such as diagnosing misconceptions, encouraging reflection, and adapting to language background. Cross-domain transfer from education to workplace training introduces different power dynamics and assessment goals. Evaluation must include learning outcomes, not just satisfaction. The case demonstrates that cross-domain reasoning transfer is not only a cognitive problem but a relational one. The system must align with the values of learners, teachers, and communities, not only with generic helpfulness.

Scientific research provides a final illustration. Models can transfer reasoning across physics, biology, and chemistry when tasks share formal structures, but they may fail when domain ontologies differ. Tool use, simulation, and literature retrieval can support transfer, but they also introduce dependency on external databases and instrumentation. Reinforcement learning from experimental outcomes is expensive and slow, so alignment may rely on proxy rewards. These proxies can distort scientific reasoning by favoring publishable results over robust ones. Governance should require provenance, uncertainty quantification, and reproducibility. Cross-domain transfer in science thus depends on institutional norms as much as algorithmic design.

8. Future Directions

Future research should develop modular alignment architectures that separate domain-general reasoning policies from domain-specific reward models. Such separation could improve auditability and reduce catastrophic forgetting. It could also allow institutions to update local norms without retraining the entire system. Continual learning methods that preserve prior competence [15] will be essential, as will methods for detecting when transfer is inappropriate. Small-parameter systems with scalable path filtering [16] suggest that efficiency and alignment can be co-designed, but more work is needed on how filtered trajectories interact with fairness and robustness. Reinforcement learning foundations [12] and landmark self-play systems [13] offer inspiration, but cross-domain socio-technical settings lack the clean reward signals of games. Research must therefore combine machine learning with domain expertise, human factors, and policy analysis.

Another direction is multi-agent and human-in-the-loop alignment. Reasoning transfer may occur across agents that specialize in different domains, with a coordinator that routes queries and synthesizes answers. This architecture could improve coverage but complicate accountability. Human oversight must be designed into the loop, not added afterward. Human-centered AI [22] provides principles, but operational methods are needed for escalation, contestability, and redress. Evaluation should include how well systems explain their transfers, how they represent uncertainty, and how they respond when users challenge their reasoning.

Finally, governance research should move from principles to enforceable mechanisms. Technical debt [21] and documentation practices [19][20] show that transparency requires continuous maintenance. Safety research [23][24] identifies failure modes that policy must anticipate. Governance agendas [25] offer a starting map, but cross-domain reasoning transfer requires new instruments for auditing domain boundary decisions. These instruments should

be tested in realistic deployments and revised as models evolve. The goal is not to eliminate transfer but to make it accountable, reversible, and aligned with public values.

9. Conclusion

Cross-domain reasoning transfer via reinforcement learning alignment is a system-level capability that cannot be understood through model scale alone. It emerges from the interaction of base representations, reward models, inference-time strategies, tool ecosystems, data infrastructures, and governance regimes. Reinforcement learning alignment can encourage reusable reasoning procedures, but it can also narrow competence, incentivize reward hacking, and propagate errors across institutional boundaries. The architecture of alignment therefore matters as much as the algorithm. Modular reward models, domain-conditioned policies, verifier-assisted reasoning, and selective path filtering can improve transfer, but they require observability, documentation, and rollback. Small-parameter systems illustrate both the constraints and the opportunities of efficient alignment. They cannot rely on brute-force scale, yet they can be deployed in ways that are more auditable and resource-aware.

The broader implication is that cross-domain reasoning transfer should be governed as a socio-technical process. Robustness, fairness, sustainability, and accountability are not external constraints on alignment; they are constitutive of whether transfer is legitimate. Evaluation must go beyond static benchmarks to include domain shifts, human oversight, red teaming, and longitudinal monitoring. Policy must clarify responsibility across developers, deployers, and users while supporting public capacity for independent assessment. Infrastructure must be designed for continual alignment rather than one-time training. If these requirements are met, reinforcement learning alignment can support reasoning transfer that is not only powerful but also contestable, reversible, and socially responsible. If they are neglected, cross-domain transfer will remain an opaque source of both convenience and harm.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
3. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
4. Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 4299–4307.
5. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk,

- S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv. <https://arxiv.org/abs/2212.08073>
6. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
 7. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 22199–22213.
 8. Zelikman, E., Wu, Y., Mu, J., & Goodman, N. D. (2022). STaR: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35, 15476–15488.
 9. Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations. ICLR*.
 10. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 11809–11822.
 11. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 68539–68551.
 12. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
 13. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354–359.
 14. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529–533.
 15. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 3521–3526.
 16. Li, Q. (2026, May). TrajLite PPO: Scalable Reasoning Path Filtering and Alignment for Small Parameter Large Language Models. In *2026 2nd International Conference on Artificial Intelligence and Digital Ethics (ICAIDE)* (pp. 29-32). IEEE.
 17. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P.

- (2021). On the opportunities and risks of foundation models. arXiv. <https://arxiv.org/abs/2108.07258>
18. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM.
 19. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM.
 20. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.
 21. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503–2511.
 22. Shneiderman, B. (2020). *Human-centered AI*. Oxford University Press.
 23. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv. <https://arxiv.org/abs/1606.06565>
 24. Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2021). Unsolved problems in ML safety. arXiv. <https://arxiv.org/abs/2109.13916>
 25. Dafoe, A. (2018). *AI governance: A research agenda*. Future of Humanity Institute, University of Oxford.