

Parameter-Efficient Reinforcement Learning Alignment for Domain-Specific Language Models

Devin C. Brock

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence,
KS, USA.

devinbrock@ku.edu

Fabrizio Foletti

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.

fabriziof@missouri.edu

Abstract

Domain-specific language models increasingly require alignment procedures that are not only effective at steering model behavior but also feasible under constrained compute, limited data, and stringent governance. This paper examines parameter-efficient reinforcement learning alignment as a systems problem rather than a narrow optimization technique. It argues that the integration of low-rank adaptation, quantized training, preference optimization, reward modeling, and domain-specific evaluation creates a distinctive socio-technical architecture with important trade-offs in cost, robustness, fairness, maintainability, and policy compliance. The analysis begins with the foundations of parameter-efficient transfer learning and reinforcement learning from human feedback, then considers how these methods behave when applied to specialized domains such as healthcare, law, engineering, and public administration. Particular attention is given to the infrastructure required for repeatable alignment, including data provenance, reward model governance, adapter versioning, deployment orchestration, and audit trails. The paper also explores how parameter-efficient reinforcement learning alignment can reduce environmental and financial burdens while introducing new risks through adapter fragmentation, hidden technical debt, reward hacking, and uneven accountability. Case illustrations from instruction tuning, clinical question answering, and reasoning-oriented small models are used to show that alignment quality depends less on any single algorithm than on the surrounding institutional and operational arrangements. The conclusion proposes that domain-specific alignment should be treated as a lifecycle capability supported by transparent evaluation, participatory oversight, and modular deployment practices. Such an approach can make reinforcement learning alignment more accessible, but only if efficiency gains are matched by robust governance and sustained attention to the distribution of benefits and harms.

Keywords

parameter-efficient alignment; reinforcement learning from human feedback; domain-specific language models; low-rank adaptation; deployment infrastructure; governance; fairness; robustness

1 Introduction The rapid diffusion of large language models into specialized professional settings has created a tension between generality and domain fidelity. Models trained on broad web-scale corpora may possess strong linguistic fluency, yet they often lack the nuanced conventions, evidentiary standards, and accountability expectations of fields such as medicine, law, engineering, and public policy. Reinforcement learning alignment has

emerged as a central mechanism for adapting such models to human preferences and institutional requirements. Early work demonstrated that language models can be fine-tuned from human preference comparisons, establishing a pathway from raw pretraining to behavior that better reflects human judgments [7]. Subsequent developments in instruction following and reward modeling showed that reinforcement learning from human feedback could produce models that are more helpful, controllable, and responsive to user intent [5]. However, the computational and organizational costs of full-model alignment remain substantial, particularly for smaller research teams, public agencies, and domain organizations that cannot repeatedly update billions of parameters. Parameter-efficient transfer learning offers a partial answer by inserting or selecting a small number of trainable components while keeping most of the base model fixed [1]. Low-rank adaptation further reduces the cost of specialization by representing weight updates in a compact form [2], while quantized fine-tuning makes such adaptation possible under tighter memory constraints [3]. This paper synthesizes these developments into a system-level account of parameter-efficient reinforcement learning alignment for domain-specific language models. The central claim is that parameter efficiency is not merely a technical convenience. It changes the governance, economics, and risk profile of alignment. When alignment can be performed through adapters, soft prompts, or compact policy updates, domain experts and smaller institutions can participate more directly in model specialization. Yet the same modularity that lowers barriers to entry also creates fragmentation, versioning complexity, and new forms of technical debt. A hospital, for example, may maintain multiple clinical adapters for different specialties, each aligned with distinct reward models and evaluation protocols. A legal aid organization may deploy a jurisdiction-specific model whose preferences differ from those of a commercial general-purpose assistant. These arrangements require infrastructure for tracking which adapter is active, which reward model justified its behavior, and which oversight body approved its use. Without such infrastructure, parameter-efficient alignment can produce a proliferation of quietly divergent models whose behavior is difficult to explain or audit. The paper therefore treats alignment as a lifecycle capability rather than a one-time training procedure.

2 Background and Related Work

The lineage of modern alignment begins with reinforcement learning from human preferences, where human comparisons are used to train a reward model that then guides policy optimization [7]. This approach was extended through large-scale instruction tuning and human feedback, producing models that better follow user instructions and avoid harmful outputs [5]. Related work on helpful and harmless assistants refined the reward modeling pipeline and highlighted the difficulty of balancing competing objectives [6]. Proximal policy optimization became a widely used optimization method because it offers stable updates in complex reinforcement learning settings [8]. Summarization research demonstrated that human feedback can improve not only safety but also task quality, showing that alignment can be domain-specific and task-sensitive [9]. More recent work on direct preference optimization simplified the alignment pipeline by removing the need for an explicit reward model during optimization, making preference-based tuning more accessible [10]. These developments collectively established that alignment is both a modeling problem and an organizational problem, because the quality of preferences, annotators, and evaluation criteria determines the resulting behavior. Parallel advances in parameter-efficient adaptation have reshaped the feasibility of domain specialization. Adapter-based transfer learning showed that small trainable modules can achieve strong performance while preserving the base model [1]. Low-rank adaptation demonstrated that compact updates can be injected into transformer layers with minimal inference overhead [2], and quantized low-rank fine-tuning extended this efficiency to memory-constrained environments [3]. Instruction generation and

self-alignment methods have further reduced dependence on large human-annotated datasets by using model-generated instructions and filtered responses [11]. Open instruction-following models such as Alpaca illustrated how modest fine-tuning can produce useful assistants on top of open foundation models [12]. The release of open foundation models such as LLaMA accelerated experimentation with domain-specific adaptation [13]. At the same time, foundation model research has emphasized that capabilities, biases, and environmental costs are distributed across data, compute, and deployment infrastructures [14]. Critical scholarship on large language models has warned that scale alone does not resolve questions of accountability, inclusion, and environmental impact [15]. Risk taxonomies have provided structured ways to identify harms such as misinformation, discrimination, privacy violation, and manipulation [16]. Finally, research on hidden technical debt in machine learning systems shows that maintenance, monitoring, and data dependencies often dominate long-term costs [17]. These strands motivate a systems-oriented view of parameter-efficient alignment.

3 Architectural and Methodological Foundations

Parameter-efficient reinforcement learning alignment combines three architectural choices. The first is the base model, which may be an open foundation model, a proprietary application programming interface, or a domain-pretrained model. The second is the adaptation mechanism, such as low-rank adapters, prefix tuning, or compact policy heads. The third is the alignment objective, which may involve reward modeling, preference optimization, reinforcement learning, or a hybrid of supervised and preference-based methods. In practice, these choices interact. A domain-specific clinical model may use a quantized base model to fit within hospital infrastructure, low-rank adapters to encode specialty knowledge, and a reward model trained on clinician preferences. A legal model may instead rely on direct preference optimization because it reduces pipeline complexity and avoids maintaining a separate reward model [10]. The architectural decision is therefore not neutral. It determines which actors can participate, how frequently updates can occur, and how easily behavior can be audited. Low-rank adaptation is particularly important because it separates the base model from the aligned policy. This separation enables modular updates, but it also creates a need for composition rules. If multiple adapters are active, their interactions may be non-additive, especially when they encode conflicting preferences. A hospital may combine a general medical adapter with a cardiology adapter and a patient communication adapter. The resulting model may exhibit emergent behavior that none of the individual adapters was evaluated for. Similar concerns arise in legal and engineering settings, where safety, liability, and professional norms may conflict. Parameter-efficient alignment therefore requires an evaluation architecture that tests not only individual adapters but also realistic combinations. This is a systems design issue rather than a modeling detail. It affects deployment, monitoring, and incident response. Reinforcement learning alignment adds another layer of complexity because the reward signal is learned from human or synthetic preferences. Reward models can encode annotator biases, institutional priorities, and incomplete specifications. When combined with parameter-efficient updates, reward model errors may be localized but still consequential. A compact adapter can shift a model toward overly cautious advice, overly permissive recommendations, or stylistic patterns that obscure uncertainty. The optimization process may also exploit weaknesses in the reward model, a phenomenon often described as reward hacking. Because parameter-efficient methods reduce the cost of experimentation, they may also increase the frequency of poorly governed reward model updates. The methodological foundation must therefore include documentation, version control, and adversarial evaluation. The recent emphasis on trajectory filtering and scalable alignment for small-parameter models illustrates the growing recognition that reasoning quality and preference alignment must be considered together,

especially when compute is limited [18]. The implication is that efficiency and robustness should be co-designed rather than treated as sequential goals.

4 Data, Reward Modeling, and Governance

Domain-specific alignment depends on preference data that reflects the values, standards, and epistemic practices of a field. In healthcare, for example, preferences may be defined by clinical accuracy, avoidance of harm, clarity of uncertainty, and respect for patient autonomy. In law, preferences may involve jurisdictional accuracy, citation discipline, confidentiality, and awareness of ethical duties. Collecting such data requires domain expertise, informed consent, and careful attention to power asymmetries between annotators, institutions, and affected communities. The reward model is not a neutral summary of these preferences; it is a compressed representation that can amplify dominant voices and marginalize minority perspectives. Governance therefore begins at data collection and continues through annotation, reward modeling, evaluation, and deployment. Without explicit governance, parameter-efficient alignment may simply reproduce existing hierarchies in a more modular and harder-to-inspect form. The infrastructure for reward model governance includes provenance records, annotation guidelines, adjudication procedures, and periodic audits. These records should connect each adapter to the datasets, reward models, and evaluation results that shaped it. In a parameter-efficient environment, such traceability is both more feasible and more necessary. It is feasible because adapters are small and can be versioned like software artifacts. It is necessary because the base model may be updated independently, potentially invalidating the assumptions under which the adapter was trained. A clinical adapter aligned to an older base model may behave differently after a base model upgrade, even if the adapter itself is unchanged. This creates a form of configuration drift that must be detected through regression testing and clinical review. Similar issues arise in legal, financial, and public-sector deployments, where regulatory changes can alter what counts as an acceptable response. Governance also involves deciding who has authority to approve an alignment update. A commercial vendor may treat alignment as a product feature, while a public agency may treat it as a policy instrument subject to administrative law. A professional body may require peer review before deployment. These different governance models imply different release cycles, documentation standards, and appeal mechanisms. Parameter-efficient alignment can support more participatory governance by allowing domain experts to review small adapters rather than entire models. However, participation is meaningful only if experts have the resources, training, and authority to challenge technical decisions. Otherwise, the efficiency gains simply shift labor from model training to unpaid or undercompensated review. The design of alignment infrastructure must therefore include mechanisms for contestability, escalation, and redress.

5 Deployment Infrastructure and Sustainability

Deployment is where parameter-efficient alignment meets operational reality. Serving a base model with multiple adapters requires orchestration software that can route requests, load adapters dynamically, and enforce access controls. This is more complex than serving a single monolithic model. Each adapter may have different latency, memory, and compatibility requirements. Some adapters may be suitable for real-time clinical decision support, while others may be used only for offline document review. The infrastructure must also handle fallback behavior when an adapter fails or produces low-confidence outputs. In high-stakes domains, the system should be able to revert to a safer general model or escalate to a human expert. These requirements turn alignment into an infrastructure problem involving observability, redundancy, and service-level guarantees. Sustainability is another dimension of deployment. Full-model fine-tuning and repeated reinforcement learning runs consume substantial energy and financial resources. Parameter-efficient methods can reduce these costs, making domain-specific alignment more accessible to smaller organizations. However,

efficiency gains may be offset by increased deployment complexity and more frequent adapter updates. If every department trains its own adapter, aggregate compute and storage costs may rise even if individual training runs are cheap. The environmental impact therefore depends on lifecycle management. Organizations should retire unused adapters, consolidate similar ones, and measure the total cost of ownership rather than only training cost. They should also consider whether a smaller or distilled model can satisfy the domain requirement without requiring the largest available base model. Sustainability is not simply a matter of green algorithms; it is a matter of institutional discipline in deciding what to build, what to reuse, and what to retire. Infrastructure also shapes equity. Cloud-based alignment platforms may lower barriers for well-funded organizations while excluding those with limited connectivity or data governance capacity. On-premises deployment may offer greater privacy but require specialized staff. Federated or consortium-based approaches may allow multiple institutions to share alignment infrastructure while retaining local control. Each model has trade-offs in cost, autonomy, and security. A public health agency may prefer a regional consortium to avoid vendor lock-in, while a small clinic may rely on a trusted intermediary. Policy frameworks should recognize these differences and support interoperable standards for adapter packaging, provenance, and evaluation. Without such standards, parameter-efficient alignment may reproduce the same concentration of power that characterizes large-scale foundation model development.

6 Robustness, Fairness, and Security

Robustness in parameter-efficient alignment requires testing across distribution shifts, adversarial inputs, and conflicting objectives. A domain-specific adapter may perform well on benchmark questions but fail when users phrase requests ambiguously or pressure the model to exceed its authority. In medical settings, robustness includes recognizing when to defer, when to express uncertainty, and when to recommend urgent human attention. In legal settings, robustness includes refusing to provide definitive advice outside the model's jurisdiction or competence. Parameter-efficient methods can make it easier to update these behaviors, but they can also make it easier to introduce subtle regressions. A small change in the reward model may alter the model's willingness to refuse unsafe requests. Continuous evaluation is therefore essential, not as a one-time compliance exercise but as an operational routine. Fairness concerns arise because preference data and reward models reflect social hierarchies. An adapter trained on the preferences of one professional community may perform poorly for another. A model aligned for a dominant language or dialect may marginalize speakers of minority languages. Parameter-efficient alignment can support pluralism by allowing multiple adapters to coexist, but pluralism without coordination can lead to inconsistent standards and unequal access. For example, a public benefits agency might deploy different adapters for different regions, resulting in different advice for citizens with similar needs. Fairness therefore requires comparative evaluation across groups and a governance process for resolving discrepancies. It also requires attention to representation in annotation, because whose preferences are encoded determines whose interests are served. Security is closely related to robustness. Adapters are small enough to be stolen, modified, or maliciously inserted into a deployment pipeline. An attacker might distribute a seemingly useful domain adapter that contains a hidden backdoor or biased behavior. Because adapters are modular, they may bypass security reviews designed for full models. Supply chain security for parameter-efficient alignment must therefore include signed artifacts, provenance verification, sandboxed evaluation, and runtime monitoring. Organizations should treat adapters as executable policy artifacts rather than simple configuration files. This perspective aligns with broader work on personalizing alignment, which emphasizes that alignment is not a single global property but a relationship between models, users, and contexts [19]. Security and fairness are thus intertwined with the

governance of the adapter ecosystem.

7 Evaluation and Case Illustrations

Evaluation of parameter-efficient reinforcement learning alignment must be multi-dimensional. It should assess task accuracy, preference satisfaction, refusal behavior, calibration, robustness, fairness, latency, and cost. Standard benchmarks are useful but insufficient because domain-specific quality often depends on tacit knowledge and contextual judgment. In healthcare, evaluation may involve clinician review of simulated cases, error analysis by specialty, and monitoring of downstream outcomes. A study of ChatGPT in healthcare education, research, and practice illustrates both the promise and the caution required when language models enter professional workflows [20]. The lesson is not that domain models are inherently safe or unsafe, but that evaluation must match the actual use case and institutional context. A second case is instruction tuning for open models. Self-generated instructions and filtered responses can reduce data collection costs, but they may also amplify model biases and factual errors [11]. Alpaca-style fine-tuning showed that modest resources can produce useful assistants [12], yet it also revealed gaps in safety, factuality, and domain expertise. Parameter-efficient reinforcement learning alignment can address some of these gaps by adding preference-based updates, but it cannot replace domain validation. A third case is reasoning-oriented small models. Recent work suggests that trajectory filtering and scalable alignment can improve reasoning behavior in smaller models without requiring massive parameter updates [18]. This is promising for edge deployment and privacy-sensitive settings, but it also raises questions about the reliability of automatically filtered reasoning paths. If the filter is imperfect, the model may learn to imitate flawed reasoning with greater confidence. Evaluation must therefore include process-level review, not only final-answer accuracy.

8 Policy and Institutional Implications

Policy frameworks for parameter-efficient alignment should address both opportunities and risks. The opportunities include lower barriers to entry, greater domain specificity, and more localized governance. Public agencies, professional associations, and community organizations could develop adapters that reflect their values without rebuilding foundation models. The risks include fragmentation, unclear liability, and unequal capacity to maintain aligned systems. Regulators may need to require documentation of adapter provenance, evaluation results, and update procedures, especially in high-stakes domains. They may also need to clarify whether responsibility lies with the base model provider, the adapter developer, the deploying organization, or the human expert who relies on the output. These questions are not resolved by technical efficiency alone. Institutional design should encourage reusable alignment infrastructure rather than isolated one-off projects. Shared repositories of evaluation protocols, model cards, and adapter metadata can reduce duplication and improve comparability. Professional bodies can define domain-specific standards for acceptable behavior, uncertainty communication, and refusal. Educational programs can train practitioners to interpret alignment results and participate in governance. Procurement rules can require vendors to disclose reward model assumptions, adapter dependencies, and monitoring capabilities. These measures would help ensure that parameter-efficient alignment strengthens institutional accountability rather than diffusing it. The goal is not to slow innovation but to align it with the values and constraints of the domains it enters.

9 Future Directions

Several research directions follow from this analysis. First, composable alignment requires methods for predicting and testing adapter interactions before deployment. This includes tools for conflict detection, behavior tracing, and compositional evaluation. Second, reward model governance needs better techniques for auditing preference data and detecting representational gaps. Third, infrastructure standards are needed for adapter signing, versioning, and dependency management. Fourth, sustainability metrics should account for the full lifecycle of alignment, including storage, serving, monitoring, and retirement. Fifth,

participatory design should be extended so that affected communities have meaningful influence over the preferences embedded in domain models. These directions are interdisciplinary because they involve machine learning, systems engineering, policy, ethics, and organizational behavior. Another important direction is the development of evaluation methods that are both rigorous and practical. Domain experts cannot review every output, and automated metrics may miss subtle harms. Hybrid approaches that combine statistical testing, expert sampling, and user feedback may offer a workable balance. Such approaches must also guard against feedback loops in which model outputs shape the preferences used for future alignment. Transparency about evaluation limitations is essential. Finally, research should examine the political economy of parameter-efficient alignment. If adapters become the primary site of domain specialization, control over adapter ecosystems may become as important as control over base models. Understanding who can create, certify, distribute, and monitor adapters will be central to the future governance of artificial intelligence. 10

Conclusion Parameter-efficient reinforcement learning alignment offers a practical pathway for adapting language models to specialized domains without incurring the full cost of repeated large-scale fine-tuning. By combining compact adaptation mechanisms with preference-based optimization, organizations can create domain-specific behaviors that reflect local standards, values, and operational constraints. However, efficiency alone does not guarantee safety, fairness, or accountability. The modularity that makes alignment accessible also creates fragmentation, versioning complexity, and new attack surfaces. The central challenge is therefore to design alignment as a governed lifecycle capability that connects data provenance, reward modeling, evaluation, deployment, monitoring, and retirement. This requires technical infrastructure, institutional policy, and participatory oversight. If these elements are developed together, parameter-efficient alignment can broaden participation in artificial intelligence development while maintaining the robustness expected in high-stakes settings. If they are neglected, it may simply distribute misalignment more efficiently. The appropriate response is not to reject parameter efficiency but to treat it as a socio-technical architecture whose benefits depend on the quality of the governance surrounding it.

References

1. Houlisby, N., Giurui, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. *Proceedings of the 36th International Conference on Machine Learning*, 2790–2799.
2. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*.
3. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36.
4. Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.
5. 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.

6. 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., Hume, T., ... & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.
7. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.
8. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
9. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33, 3008–3021.
10. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
11. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., & Hajishirzi, H. (2023). Self-Instruct: Aligning language models with self-generated instructions. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 13484–13508.
12. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Stanford Alpaca: An instruction-following LLaMA model. GitHub repository.
13. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
14. 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 preprint arXiv:2108.07258.
15. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
16. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., ... & Gabriel, I. (2022). Taxonomy of risks posed by language models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229.
17. 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.

18. 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.
19. Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2024). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6, 383–392.
20. Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, 11(6), 887.