

Multi-Objective Alignment of Small Language Models for Accuracy, Safety, and Efficiency

Karl-Heinrich Scheuermann

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.

hellokarlheinrich@unr.edu

Rozalia Wabik

Department of Computer Science, University of North Texas, Denton, TX, USA.

rozalia1989@unt.edu

Indy van Stralen

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.

indy.vanstralen@buffalo.edu

Abstract

Small language models are increasingly positioned as practical alternatives to very large foundation models in settings where latency, cost, privacy, and environmental footprint matter. Yet their alignment cannot be reduced to a single objective such as benchmark accuracy or refusal behavior. This paper examines multi-objective alignment for small language models through a system-level lens that integrates accuracy, safety, and efficiency with governance, infrastructure, robustness, fairness, and sustainability. It argues that the central challenge is not selecting one optimum but designing architectures, evaluation regimes, and deployment policies that make trade-offs visible and manageable across the model lifecycle. The analysis considers compact model families, distillation, parameter-efficient adaptation, quantization, preference optimization, and reasoning-path filtering as interlocking mechanisms rather than isolated techniques. It also discusses how safety alignment can degrade under compression, how efficiency gains can shift costs from computation to maintenance and oversight, and how fairness concerns become more acute when small models are deployed as gatekeepers in resource-constrained environments. The paper proposes that multi-objective alignment should be treated as a socio-technical governance problem, with continuous monitoring, transparent reporting, staged deployment, and context-sensitive evaluation. The conclusion emphasizes that small models can support more distributed and sustainable artificial intelligence infrastructures only if alignment frameworks explicitly account for the structural tensions among accuracy, safety, efficiency, and public accountability.

Keywords

small language models, multi-objective alignment, safety, efficiency, governance, sustainability, robustness, fairness.

1. Introduction

The rapid expansion of language model capabilities has produced a strong narrative of scale, in which larger parameter counts, larger training corpora, and larger compute budgets are presumed to yield better performance across a wide range of tasks. Foundational analyses of foundation models have documented both the promise and the concentration of technical,

economic, and political power associated with this trajectory [1]. At the same time, critical scholarship has warned that indiscriminate scaling can reproduce and amplify social harms while obscuring the environmental and labor costs embedded in data and computation [2]. These concerns have created renewed interest in small language models, which are not merely smaller versions of large systems but different socio-technical objects with distinct deployment profiles, governance requirements, and failure modes.

Alignment research has traditionally focused on making models helpful, harmless, and honest through instruction tuning, human feedback, and preference optimization [3]. Constitutional and artificial feedback approaches have further attempted to reduce reliance on extensive human annotation by encoding normative principles into the training loop [4]. Reinforcement learning from human preferences established an influential paradigm for optimizing model behavior against learned reward signals [5], while reasoning-oriented prompting demonstrated that model behavior can change substantially when intermediate reasoning is elicited [6]. However, these methods were largely developed for large models and are often evaluated through benchmarks that privilege accuracy, fluency, or aggregate safety rates. Small models introduce a different constraint structure: their limited capacity makes alignment losses more likely, their deployment contexts are often edge-based or privacy-sensitive, and their efficiency advantages can be undermined if safety requires expensive external oversight.

This paper argues that multi-objective alignment for small language models must be understood as a system-level design problem. Accuracy, safety, and efficiency are not independent dials. Increasing safety through refusal behavior can reduce helpfulness and accuracy on legitimate tasks. Increasing efficiency through quantization or distillation can degrade calibration, increase brittleness, and weaken safety filters. Increasing accuracy through specialized fine-tuning can narrow generalization and create new vulnerabilities. These trade-offs are further shaped by infrastructure, governance, and policy choices that determine who can audit a model, where it runs, how often it is updated, and what recourse exists when it causes harm. The paper therefore develops an interdisciplinary account that connects model architecture to deployment sustainability, robustness, fairness, and public accountability.

2. Conceptual Foundations and Structural Trade-offs

Small language models occupy a broad technical space. They may be pretrained compact models, distilled derivatives of larger systems, pruned and quantized versions of existing models, or domain-specific architectures trained from scratch. Open model families such as LLaMA demonstrated that relatively accessible model weights can stimulate distributed innovation [7], while later small-model efforts such as TinyLlama pursued practical inference under constrained resources [8]. These developments matter because they shift the locus of alignment from centralized laboratories to a wider array of institutions, including universities, public agencies, startups, and community organizations. Once small models circulate as reusable artifacts, alignment becomes not only a training problem but also a maintenance and governance problem.

Distillation is a central mechanism in this ecosystem. Knowledge distillation transfers behavior from a larger teacher model to a smaller student model, potentially preserving much of the teacher’s functional competence at lower inference cost [9]. DistilBERT and TinyBERT illustrated how transformer models can be compressed for faster and cheaper deployment [10], [11]. Yet distillation is not a neutral compression of capability. The student model may inherit teacher biases without inheriting the teacher’s broader context, safety

training, or uncertainty calibration. A small model may reproduce confident but inappropriate answers because it has learned surface patterns rather than robust reasoning. The structural trade-off is therefore not simply size versus performance but fidelity versus interpretability, compression versus adaptability, and local efficiency versus systemic risk.

Multi-objective alignment also involves temporal trade-offs. A model aligned for immediate safety may perform poorly when deployed in a new domain with different norms and informational needs. A model optimized for low latency may be unable to support the deliberative reasoning required for high-stakes decisions. A model fine-tuned for cost efficiency may require more frequent human review, thereby shifting costs from computation to oversight. These dynamics resemble the hidden technical debt observed in machine learning systems, where quick engineering choices create long-term maintenance burdens [17]. For small models, such debt can be especially consequential because deployers may lack the resources to monitor drift, update safety filters, or respond to emerging harms.

3. Architecture and Multi-Objective Alignment Mechanisms

Architectural choices shape which alignment objectives are feasible. Parameter-efficient adaptation methods such as low-rank adaptation allow a small model to be specialized without full retraining [13], while quantized fine-tuning methods reduce memory requirements and enable adaptation on modest hardware [12]. These techniques are often framed as efficiency wins, but they also affect safety and accuracy. Low-rank updates may concentrate learning in a narrow subspace, making the model highly capable in one domain while brittle outside it. Quantization can introduce numerical noise that disproportionately affects rare tokens, safety-critical instructions, or marginalized language varieties. The result is that efficiency mechanisms are also alignment mechanisms, even when they are not described as such.

Preference-based alignment adds another layer. Human feedback and constitutional methods can shape model behavior toward desired norms [3], [4], [5]. In small models, however, the capacity to represent nuanced normative distinctions is limited. A small model may learn a blunt refusal policy that blocks legitimate requests or, conversely, may fail to generalize safety training to paraphrased harmful prompts. Alignment therefore depends on data curation, reward modeling, and evaluation design as much as on model size. If reward models are trained predominantly on the preferences of a narrow group, the resulting small model may encode that group’s assumptions as universal defaults. This is a fairness problem, but it is also an architectural problem because the model has fewer parameters with which to represent competing values or context-dependent exceptions.

Reasoning-path filtering and scalable alignment for small parameter models represent an emerging response to these constraints. Recent work on trajectory filtering and alignment for small parameter models suggests that selective reasoning paths can improve alignment without requiring the full computational overhead of larger systems [18]. The system-level implication is that small models may benefit from hybrid architectures that combine compact parametric knowledge with external tools, retrieval, or verifiers. Such hybrids can preserve efficiency while offloading high-stakes reasoning to auditable components. However, hybrid designs introduce new dependencies: retrieval corpora can be poisoned, external tools can fail, and orchestration logic can become an opaque point of control. Multi-objective alignment must therefore account for the entire inference pipeline, not only the model weights.

4. Safety, Robustness, and Governance

Safety alignment for small language models cannot be evaluated solely through static refusal benchmarks. Robustness requires testing under distribution shift, adversarial prompting, multilingual variation, domain-specific jargon, and resource constraints. A model that refuses harmful requests in a controlled evaluation may comply when prompted through indirect speech, role-play, or code-switching. Small models may be more susceptible to such failures because they have less capacity to separate surface form from intent. The risk taxonomy for language models identifies harms ranging from information hazards and bias to misinformation and malicious use [16]. These risks are not eliminated by reducing model size; they are redistributed across deployment contexts and oversight arrangements.

Governance is essential because safety is not a property that can be fully certified at training time. A small model released as open weights can be fine-tuned, quantized, merged with other models, or embedded in applications beyond the original developer’s control. This diffusion creates accountability gaps. Who is responsible when a compact model deployed in a public benefits office produces discriminatory recommendations? Who audits a model that runs on a local device without centralized logging? Who ensures that safety patches reach downstream users? These questions point toward governance mechanisms such as model cards, licensing conditions, update channels, and third-party auditing. They also point toward infrastructure choices, because auditability depends on whether inference is observable and whether version histories are preserved.

Robustness and safety also interact with efficiency. Safety filters, retrieval checks, and human escalation pathways consume computation and attention. If these mechanisms are removed to meet latency or cost targets, the system may appear efficient while transferring risk to users and frontline workers. A more sustainable approach treats safety as part of the system’s operating cost rather than as an optional add-on. This may require tiered deployment, in which low-risk tasks use fully local small models and high-risk tasks invoke additional verification. Such tiering is not a technical fix alone; it is a policy decision about acceptable risk, transparency, and recourse. Governance must therefore specify which tasks are suitable for autonomous small-model inference and which require human accountability.

5. Efficiency, Infrastructure, and Sustainability

Efficiency is often the primary justification for small language models. Lower parameter counts reduce memory, energy, and latency, enabling deployment on edge devices, personal computers, and modest servers. These benefits are real, but they must be assessed across the full lifecycle. Training, fine-tuning, inference, storage, networking, and eventual decommissioning all contribute to environmental impact. Research on energy and policy considerations for deep learning has shown that computational costs are unevenly distributed and often underreported [14]. Green artificial intelligence scholarship has argued for efficiency metrics that go beyond accuracy and include energy, carbon, and hardware utilization [15]. For small models, the sustainability case is strongest when efficiency gains are measured against actual deployment patterns rather than assumed from parameter count alone.

Infrastructure shapes whether efficiency translates into sustainability. A small model running on a battery-powered device may reduce data center load but increase e-waste if devices are replaced frequently. A small model served from a centralized cloud may be efficient in hardware utilization but dependent on energy-intensive cooling and network transmission. A federated deployment may improve privacy but complicate monitoring and updates. These are not merely engineering details; they determine who bears the costs of alignment. If safety

updates require new hardware, then resource-constrained users may be excluded. If efficiency optimizations depend on proprietary accelerators, then public institutions may become locked into specific vendors. Multi-objective alignment must therefore include infrastructure justice as a dimension of sustainability.

Efficiency also has cognitive and institutional dimensions. A model that produces answers quickly may encourage overreliance, reducing opportunities for verification. A model that is cheap to run may be deployed at scale before its limitations are understood. A model that is easy to fine-tune may proliferate domain-specific variants with inconsistent safety properties. The apparent efficiency of small models can therefore create new forms of technical debt, as noted in broader machine learning systems research [17]. Sustainable alignment requires lifecycle governance: documentation of training data and adaptation methods, clear versioning, monitoring for drift, and processes for deprecation. Without these, the environmental savings of small models may be offset by repeated remediation, public distrust, and institutional churn.

6. Evaluation and Multi-Objective Trade-offs

Evaluating multi-objective alignment requires metrics that reflect accuracy, safety, efficiency, robustness, fairness, and sustainability without collapsing them into a single score. Aggregate accuracy can hide severe failures on minority dialects, safety-critical instructions, or low-resource languages. Average latency can hide long-tail delays that matter in emergency contexts. A safety refusal rate can hide over-refusal that burdens users seeking legitimate information. Evaluation must therefore be disaggregated by task, population, deployment setting, and risk level. It must also include uncertainty, because small models may be less calibrated and more prone to confident errors. Evaluation regimes should combine automated benchmarks, red-teaming, user studies, and longitudinal monitoring.

Multi-objective trade-offs are unavoidable, but they can be made explicit. One approach is to define acceptable thresholds for each objective and then search for configurations that satisfy them, rather than optimizing a single weighted score. Another is to use staged deployment, where a small model operates in a limited domain with human review and gradually expands as evidence accumulates. A third is to design ensembles or routing systems in which a compact model handles routine cases and escalates uncertain or high-risk cases to larger models or human experts. These approaches do not eliminate trade-offs, but they relocate them into governance decisions that can be contested and revised. The recent focus on scalable reasoning path filtering for small parameter models illustrates how alignment can be improved through selective computation [18], yet such methods still require evaluation that asks who benefits, who is excluded, and what happens when the filter fails.

Cross-domain comparisons are instructive. In healthcare, a small model may support triage but must not replace clinical judgment without accountability. In education, a small model may provide tutoring but must avoid reinforcing inequities or exposing student data. In public administration, a small model may draft documents but must comply with due process and non-discrimination requirements. Each domain has different accuracy thresholds, safety norms, and efficiency constraints. A single alignment recipe is therefore inadequate. What is needed is a portfolio of alignment strategies and a governance framework that matches model capability to context. This framework must also address fairness, because small models are often proposed for resource-constrained settings where errors can have outsized consequences.

7. Fairness, Policy, and Deployment Practice

Fairness in small language models is both a data and an infrastructure problem. Compact models may be trained on datasets that underrepresent certain languages, dialects, disabilities, or cultural contexts. Distillation can amplify these gaps if the teacher model already performs poorly on those groups. Quantization and pruning can further degrade performance on rare or complex inputs. When such models are deployed in public services, they may systematically disadvantage the people who most need support. Fairness therefore cannot be reduced to equal aggregate performance; it requires attention to error distribution, access to recourse, and the power asymmetries between model deployers and affected communities.

Policy implications follow from these concerns. Regulators may require risk assessments, transparency reports, and human oversight for certain uses of small language models. Procurement rules can mandate energy reporting, data provenance, and independent auditing. Open-weight releases can support innovation and scrutiny, but they can also complicate safety enforcement. A balanced policy approach might distinguish between low-risk and high-risk applications, require stronger documentation for the latter, and fund public infrastructure for testing and monitoring. It might also support community participation in alignment, so that norms are not defined solely by technology vendors or narrow expert groups. Such policies must be adaptable because small-model capabilities and deployment patterns are changing rapidly.

Deployment practice should reflect these governance principles. Organizations should inventory where small models are used, classify use cases by risk, and establish escalation pathways for uncertain or harmful outputs. They should monitor for drift, document fine-tuning and quantization steps, and maintain the ability to roll back updates. They should train frontline staff to recognize model limitations and provide users with clear avenues for appeal. These measures may seem costly, but they are part of the true cost of alignment. A small model that is cheap to run but expensive to govern is not efficient in any meaningful sense. The objective is not to maximize one dimension but to sustain a trustworthy service over time.

8. Future Research Directions

Future research should develop evaluation suites that treat accuracy, safety, efficiency, robustness, fairness, and sustainability as jointly measured outcomes. Such suites should include longitudinal studies of deployed small models, not only one-time benchmark results. They should examine how alignment properties change after quantization, distillation, domain adaptation, and model merging. They should also investigate hybrid systems that combine small models with retrieval, symbolic tools, and human review. The design of these systems requires expertise from computer science, law, policy, human-computer interaction, and domain practice. Multi-objective alignment is therefore an interdisciplinary research program, not a narrow optimization problem.

Another direction is the development of governance mechanisms for distributed small-model ecosystems. This includes technical standards for model provenance, update notification, and audit logs, as well as legal frameworks for liability and redress. It also includes economic mechanisms that fund public-interest evaluation and maintenance. If small models are to support more democratic and sustainable artificial intelligence infrastructures, then the benefits of efficiency must be coupled with investments in accountability. Otherwise, the same concentration dynamics that characterize large-model development may reappear in fragmented form, with many small models but few meaningful channels for public oversight.

Conclusion

Multi-objective alignment of small language models requires a system-level perspective that treats accuracy, safety, and efficiency as interdependent and politically situated objectives. Small models offer important opportunities for privacy, latency, cost, and environmental sustainability, but these opportunities are not automatic. Compression can weaken safety, efficiency can shift costs to oversight, and distributed deployment can create accountability gaps. Alignment must therefore encompass architecture, evaluation, infrastructure, governance, and policy. The most promising path is not to search for a single optimal configuration but to build frameworks that make trade-offs explicit, support staged and context-sensitive deployment, and preserve human recourse. Small language models can contribute to a more plural and sustainable artificial intelligence ecosystem, but only if their alignment is understood as a long-term socio-technical commitment rather than a one-time technical adjustment.

References

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv. <https://arxiv.org/abs/2108.07258>
2. 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. <https://doi.org/10.1145/3442188.3445922>
3. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 27730–27744). Curran Associates.
4. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv. <https://arxiv.org/abs/2212.08073>
5. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4299–4307). Curran Associates.
6. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 24824–24837). Curran Associates.
7. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv. <https://arxiv.org/abs/2302.13971>
8. Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). TinyLlama: An open-source small language model. arXiv. <https://arxiv.org/abs/2401.02385>
9. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv. <https://arxiv.org/abs/1503.02531>
10. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv. <https://arxiv.org/abs/1910.01108>

11. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., ... Liu, Q. (2020). TinyBERT: Distilling BERT for natural language understanding. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 4163–4174). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.372>
12. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. In Advances in Neural Information Processing Systems (Vol. 36, pp. 10088–10115). Curran Associates.
13. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv. <https://arxiv.org/abs/2106.09685>
14. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1355>
15. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63. <https://doi.org/10.1145/3381831>
16. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., ... Gabriel, I. (2022). Taxonomy of risks posed by language models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 214–229). ACM. <https://doi.org/10.1145/3531146.3533088>
17. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... Young, M. (2015). Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems (Vol. 28, pp. 2503–2511). Curran Associates.
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.