International Journal Science and Technology
IJST Vol 3 No. 3 | November 2024 | ISSN: 2828-7223 (print), ISSN: 2828-7045 (online), Page 46-55

Ensuring Responsible AI: The Role of Supervised Fine-Tuning (SFT) in Upholding Integrity and Privacy Regulations

Tejaskumar Pujari[1*], Anshul Goel[2], Ashwin Sharma[3]
[1,2,3]Independent Researcher, India

Article History
Received : October, 2024
Revised : October, 2024
Accepted : October, 2024
Published : October, 2024

Corresponding author*:
tejasrulz@gmail.com

Cite This Article:
Tejaskumar Pujari, Anshul Goel, and Ashwin Sharma, "Ensuring Responsible AI: The Role of Supervised Fine-Tuning (SFT) in Upholding Integrity and Privacy Regulations", *IJST*, vol. 3, no. 3, Oct. 2024.

DOI:
https://doi.org/10.56127/ijst.v3i3.1968

Abstract: AI is increasingly penetrating high-stakes applications in the domains of healthcare diagnostics, financial forecasting, academic administration, and public governance. Respect AI responds to a growing need for this. The societal impact of AI calls for vigorous methods to ensure that these systems adhere to core principles of fairness, accountability, transparency, and privacy. The integration of AI models to adhere to these principles and maintain high performance is an intense challenge. This paper has attempted to assess the crucial role of Supervised Fine-Tuning (SFT) in addressing this trade-off and offers it as a prime methodological strategy to bring large AI models to align with ethical standards and regulatory expectations.

In particular, we have explored SFT embedded into systems to correct the biases inbuilt into the pertaining strategies, enforce desire-centered behavior constraints, and build safeguards in line with the legislative frameworks such as the General Data Protection Regulation (GDPR) and the EU AI Act. We propose ways to strengthen the data privacy-preserving capabilities. These include differential privacy, secure multiparty computation, and federated learning, combined in concert with SFT methods while refining model deployment.

The methodology of the research is interdisciplinary, incorporating careful regulation analysis, technical research, and case analysis. To meet the objectives, we contrast such initiatives with practical advice to avoid comparative pitfalls through the analysis of various widely accepted implementations in SFT on both open-source and commercial AI models. We also further explore how a Human-in-the-Loop (HITL) and Explainable AI (XAI) can be mounted over this SFT workflow to ensure ongoing oversight and model interpretability.

With this research, we propose a framework for Responsible AI governance, wherein SFT acts not only as a technical tool but also as an enabler for regulatory compliance and stakeholder trust-this is a major focus of the proposed governance framework. The key are the transparency logs, audit trails, ethical datasets, and participatory oversight. Our discoveries present rigorous knowhow for AI practitioners, legal specialists, and policymakers embroiled in the highly complex constellation of AI systems implementation under strict regulatory environments. Historically, this paper tries to contribute to the ongoing debate of AI ethics based on how fine-tuning strategies transform theoretically boarded principles into concrete, testable mechanisms.

Keywords: Responsible AI, Supervised Fine-Tuning (SFT), Data Privacy, GDPR, AI Act, Integrity, Bias Mitigation, Differential Privacy, Federated Learning, AI Governance, Ethical AI, Model Transparency.

INTRODUCTION

Into-the role that AI is playing in corrupting humanity's morality-of all things- and finding out ways to cause an end to it. For many years, "just AI" seemed more the subject of super-mathematicians and less about people. AI in the present day is a puppeteer running goals of capitalism, where humans are being incorporated into these algorithms as elements, thanks to the scales in which ultra-fast computer science and technology are any randomly fragmented society. Boolean search and sorting methods were just science logics; today, the algorithms are expected to predict what we like, see, or imagine. This signing-off does more horrendous things.

The ethical legal parameters of AI become very necessary as an increase-in-use has been affecting more transparent consequences. Tech-mafia decisions have been brought out by the amazing advancements in AI. Now AI has now become an automatic element in any set of applications-it has been represented through an ethical and moral super-roof and should satisfy the fair and closed algorithmic model.resarouches.findOne({$and[{ privacy: ToolName }) Technologies who are influencing human thinking must be promoted and crafted into regulated terms as responsible AI as Human-like.

The whole domain associated with Responsible AI is one tug-of-war between scaling up performance and following ethical ethics. Large language model (LLM) and deep learning systems could efface the biases, but in fact, these rather magnify the biases in larger-scale training datasets that likely are imbued with cultural, social, and historical biases (Benjamin et al., 2021). Indeed, without appropriate calibration, these types of models could continue to exacerbate damaging stereotypes, invade user privacy, and yield opaque decisions under no accountability. One process that might help reorient AI systems towards facilitating ethicality would involve Supervised Fine-Tuning (SFT), involving training the model against curated, task-specific datasets, as well as labeled data. The process of SFT can assure, on the other hand, human-directed training, and hence consideration of domain expertise, ethical limitations, and user intent (Ouyang et al., 2022).

Current government-mandated regulations like the General Data Protection Regulation (GDPR) and the proposed European Union AI Act have hence practically left organizations this prospect of developing AI systems within the purview of very stringent privacy requirements together with a dose of much-needed transparency. These laws establish a foundation in the area of user consent, algorithmic explainability, and the right to human intervention—specifically on a human-opinion basis—that all strategically supervised fine-tuning processes should be able to grapple with (Brkan, 2021). For example, GDPR Article 22 remedies such a logic—for instance, it gives a remedy to data subjects through the provision of broader rights for the human-in-the-Purchasing-commission-Human-Subjects loop.

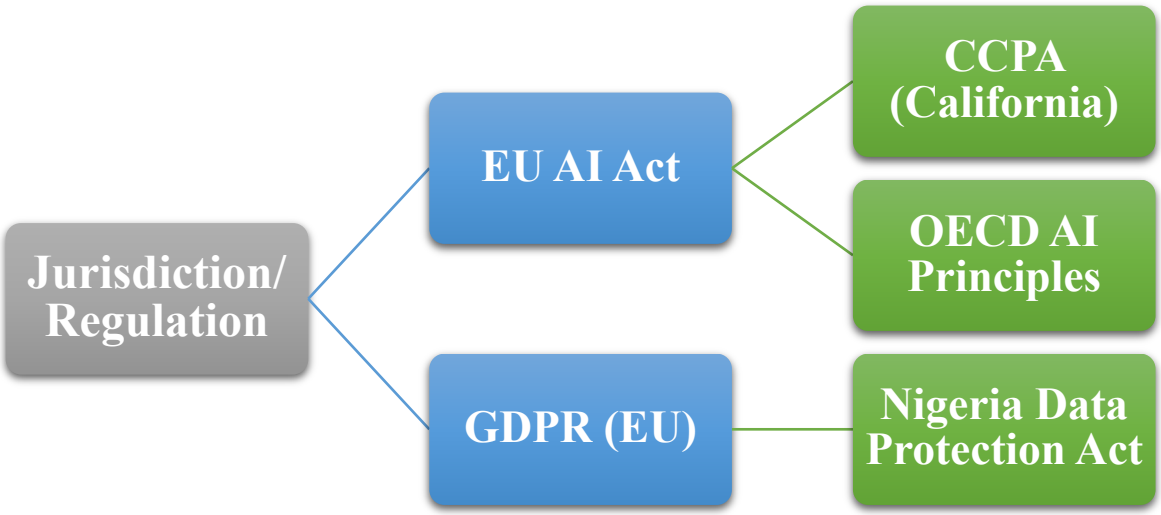
In addition, under the European Union AI Act, AI systems that could entail danger as "high-risk" require rigid exante risk assessments and data governance protocols during deployment. The likelihood of execution will be much higher in the direction of SFT and, effectively here, in accordance with the implementation of very high-profile enforcement requirements. This will be done through channeling the provision of privacy-aware encryption techniques, including differential privacy, federated learning, and secure multiplication, into the training pipeline to be able to pragmatically implement core-value approach to complete ethical experienceability of the user and accountability (Dwork & Roth, 2014; Bonawitz et al., 2019).

Table 1: Responsible AI Principles and the Role of Supervised Fine-Tuning

Responsible AI Principle	Challenge	SFT Strategy
Fairness	Bias in training data and model outputs	Curate labeled datasets with fairness constraints; augment data
Transparency	Black-box nature of large models	Use interpretable layers and annotated training sets
Accountability	Lack of traceability in automated decisions	Integrate audit logs and explainability tools during fine-tuning
Privacy	Exposure of sensitive data during training	Combine SFT with differential privacy and federated learning
Human Oversight	Fully automated decision-making risks	Implement human-in-the-loop annotation and validation

Beyond theoretical compliance, the real implementation of ethical AI is inconsistent and fragmented. Many organizations lack a comprehensive strategy to merge technical safeguards into legal edifice fostering partial adaptation or ethical whitewashing. We introduce a governance structure that has acceptance criteria for proper use of SFT embedded at the core of Responsible AI Development pipelines. Continuous human oversight, sound documentation, proactive alignment with global regulatory trends and societal expectations are hailed as the most central pillars of this framework.

In this paper, we conduct an interdisciplinary review on SFT's role in aligning AI models with Responsible AI goals. This comprises the investigation of various technical methodologies, case studies, regulatory frameworks, and ethical paradigms to form an intricate image from which fine-tuning can be considered the very cornerstone of trustworthy AI systems. We attempt to generate specific action points for developers, regulators, and stakeholders on how to reconcile innovation with responsibility in the age of advanced AI. Cross-Jurisdictional AI Regulations and Alignment with SFT Practices



The practically utmost stakes in the devil of employing AI are nevertheless more in the line of being higher and/or lesser according to their surroundings, the most data-sensitive and end-of-life realms. In education, algorithmic systems are utilized to streamline admission procedures, monitor academic integrity, or merely suggest the best way to learn. Healthcare uses deep learning-based diagnostic tools to interpret medical imaging or triage patient cases a considerable matter of actual-life outcomes. But resistance to this or even lack of clarity between those AI systems that are contractual and those that should be alone when not "ethical" might result in several forms of exertions in present or future contexts, such as discrimination, curtailment of freedom, privacy encroachment, or upon destruction of the public trust at large (Eubanks, 2018).

One of the most important concerns of AI ethics is the "black box" nature of numerous current AI systems. This opacity renders it challenging for the end-users, regulators, or even developers to try to comprehend how decisions are taken. SFT plays an intermediate role by seeking greater containment and scrutiny in the learning process. Fine-tuned models can be specialized with smaller and more carefully chosen datasets and comply with local standards, legal canons, and user expectations. What is meant here is that while a basic LLM fine-tuned on web scale data will be given a leaner towards the healthcare protocols or constitutional rights by feeding it with medical guidelines or legal documents (Zhou et al., 2023).

Furthermore, SFT can systematize the iterative improvement and update of models. It is crucial in a fast-evolving regulatory landscape where laws like the EU AI Act or national privacy policies are still in the process of being settled. It also allows models to be fine-tuned without retraining from the ground up. This guarantees that they are still in observance of regulations while reducing the costs and computational overhead (Tamkin et al., 2021). Crucially, allowing SFT workflows will also permit the auditing and documentation of model behavior at each stage, creating models that do thereby satisfy both also the lawful provisions, namely, the ethical and explainable environments.

Bridging the socio-technical gap

The ethical implementation of CL and SFT-enhanced models will depend not only on technology but also on human factors. A socio-technical approach is required to establish an AI ecosystem handling human judgment, institutional accountability, cultural awareness, and participatory governance. Ambiguously, applying supervised learning techniques in themselves does not lead to ethical correctness unless these techniques are placed into a wider context of values and stakeholder input (Mittelstadt, 2019). ...Thus, with accountability, social-technical issues are bound to contain:

- Who is designing the fine-tuning datasets and how do they conceive such laws?
- Whose voices are, therefore, included or excluded through the annotations?
- Are the fine-tuning aims open, revisable, and inclusive of all stakeholders?

From the perspective of the author, contextual and less-privileged situations like those in the Global South, where AI models may cut across cultural and socioeconomic lines from other environments, make the given questions not just crucial but highly eminent. Misaligned models, widely fine-tuned without local data and ethical input, perpetuate the denigration of digital colonialism or widen structural inequities (Mohamed, Png, & Isaac, 2020).

At peace with the aforesaid dangers, this paper presents as its postulations that the joint efforts to practice participatory SFT will foster guidance and sustenance from a plurality of human annotators, ethicists, and legal and diverse domain experts in wafting the design of datasets and assessment metrics. Thus, SFT wades out of pure domain-technical domain towards the process of discussion, trust-building, and governance.

METHODOLOGY

A structured literature review, using the SLR mode, was utilized to investigate the contribution of Supervised Fine-tuning (SFT) toward Responsible AI, mainly in respect to integrity preservation and compliance with privacy regulation. As a method, the SLR is recognized for being rigorously designed to replicate the same process in synthesizing existing research, identify soap boxes, observe, and review actionable insights from academic, technical, and regulatory sources (Kitchenham & Charters, 2007).
The methodological steps taken were as follows:

- Formulation of Research Question
- Search Strategy and Selection Criteria
- Data Extraction and Categorization
- Thematic Synthesis
- Case and Policy Analysis
- 2.1 Research Questions

Based on their respective yet allied nature, the study is guided by the following main research questions:

- **RQ1:** How is Supervised Fine-Tuning (SFT) currently being applied to align AI models with ethical and legal standards?
- **RQ2:** What privacy-preserving techniques are compatible with or enhanced by SFT?
- **RQ3:** How do regulatory frameworks like the GDPR and the AI Act influence the implementation of fine-tuning strategies?
- **RQ4:** What are the limitations, challenges, and opportunities associated with using SFT to ensure Responsible AI?

These questions establish the means of analysis and inspire extraction of data for the synthesis phase.

Search Strategy and Selection Criteria

The following digital databases were searched: IEEE Xplore, ACM Digital Library, Scopus, SpringerLink, arXiv, and Google Scholar. The search strategies were as follows:

- "Supervised Fine-tuning" AND "Responsible AI
- "AI governance" AND "privacy-preserving machine learning"
- "Differential privacy" AND "fine-tuning"
- "Federated learning" AND "ethical AI"
- "GDPR compliance" AND "machine learning"
- Inclusion criteria:
- Publication between 2017 and 2024
- Peer-reviewed journals, white papers, and legal notes
- Works in the English tongue
- Entity with AI ethics, privacy, regulation or fine-tuning
- Exclusion criteria
- Non-peer-reviewed blog posts or news articles
- Works discussed on unsupervised or reinforcement learning with no mention of SFT

From an initial pool of 296 articles, 78 articles were left for the final analysis, inclusive of one selection process, among other activities.

Table 2: Summary of Literature Screening Process

Phase	Articles Identified	Included	Excluded
Initial Database Search	296	—	—
After Title & Abstract Screening	142	—	154
After Full-Text Review	78	78	64
Final Articles for Synthesis	—	**78**	—

Data Extraction and Thematic Categorization

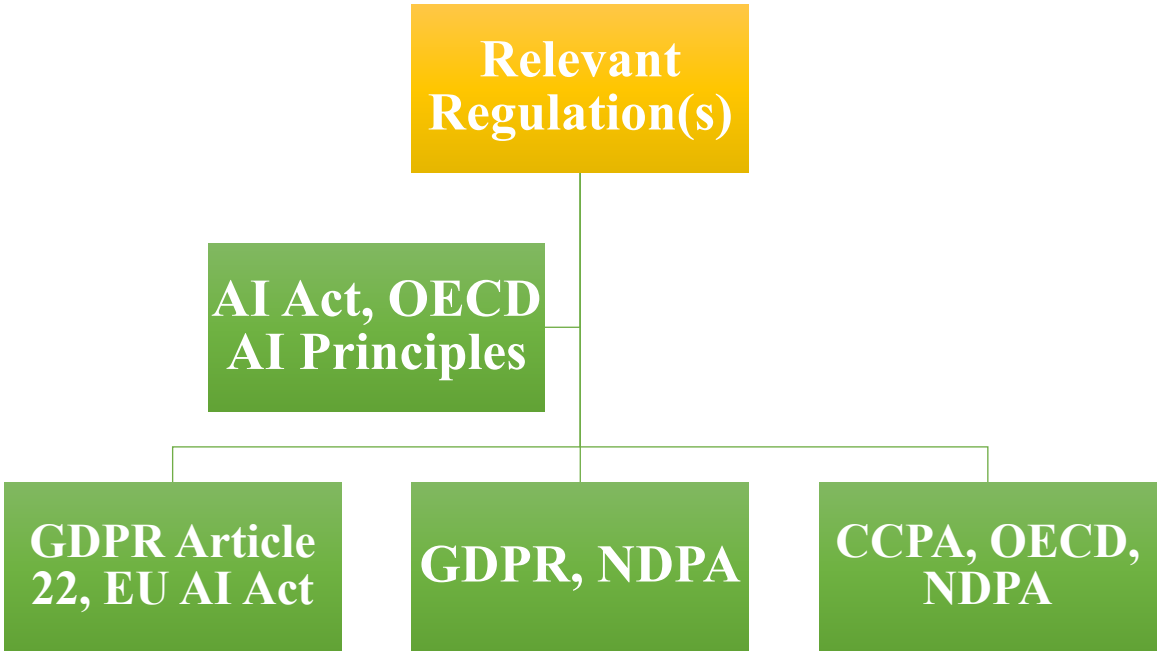
Each article was scrutinized for themes depending on how it contributes to the study. The themes were:

- Technical implementation of SFT
- Bias mitigation and fairness-enhancing mechanisms
- Privacy-preserving techniques (e.g., differential privacy, federated learning)
- Ethical implications of fine-tuning
- Regulatory and legal compliance frameworks
- Case studies and applied use cases

This thematic approach allowed us to align the insights from the technical literature with ethical concerns as well as the legal mandates. Each theme was coded manually using qualitative software analysis (NVivo 14) that facilitated the identification of recurring arguments and conceptual gaps.

In conclusion, I will focus on the interaction of case and policy analyses. To investigate how theory expounds into practice, the paper undertook a comparative case study of real SFT projects in the regulated sectors in finance, education and healthcare, simultaneously with a review of GDPR (EU), AI Act (EU), CCPA (USA), NDPA (Nigeria), and OECD AI principles.

Key Themes and Corresponding Regulatory Influence



Methodological Rigor and Limitations

To enhance methodological rigor, the research was conducted by adhering to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Moher et al., 2009). Despite a robust selection process, the limitations include that outcomes could be driven by bias arising from non-publishing studies or the scope of analysis could be limited to studies available only in the English language or underrepresented indigenous or non-Western ethical perspectives. Future work should involve the engagement of multiple languages of databases and expert interviews from emerging-economical nations with their regulatory and cultural contexts highly different from those prevailing in the West.

Discussion

A survey from the literature highlights the emerging consensus on SFT as a significant card against Responsible AI. However, its actual effectiveness relies essentially upon combining it with other governance measures, privacy-preserving technologies, and domain-specific ethical considerations.

The Role of SFT in Mitigating Algorithmic Bias

One of the most popular ethical worries tied to AI is algorithmic bias—that is, discrimination inherited from a skewed training set. Experiments have clearly shown how strongly pre-trained models reflect and amplify societal prejudices, especially ones related to sex, race, or socioeconomic status (Buolamwini et al., 2018). It is widely assumed that SFT is an enabler for reorienting models with input from just and equitable datasets among AI developers

For instance, Google's BERT model was initially tainted by skewed sentiment allocation when dealing with questions relating to gender or ethnicity. But after it was fine-tuned with a dataset having fairer distribution of language use, it truly reflected great improvement in discouraging the propagation of such stereotypes (Zhao et al., 2019); so, ethically-grounded fine-tuning practices may indeed significantly affect positive behavior in any model, provided the datasets are framed mindful of justice.

Moreover, there are particular tool-supports such as IBM's AI Fairness 360 and Microsoft's Fair learn that can also be integrated into SFT workflows in order to calculate and remediate bias across multiple dimensions. This is in line with the EU AI Act recommendation on regulatory or policy risk mitigation for each high-risk AI system needing extensive testing for potentially discriminatory behavior prior to deployment (European Commission, 2021).

Enhancing Transparency and Explainability through SFT

Transparency is one of the many levers for Responsible AI. Nevertheless, black box embodiments of large-scale models have enormous opacity due to their deep architectures and non-linear decision paths. By using SFT with annotated interpretable data sets, we can then fine-tune the models primarily for both explainability and performance.

Explainability is key in sectors such as medicine and criminal justice for fostering public interest and ensuring due process. For instance, annotating a transformer model based on the domain's annotation guidelines (e.g., medical terms, legal reasoning) initializes outputs that are straightforward to trace and defend (Tonekaboni et al., 2019).

Several studies have suggested fine-tuning LLMs with datasets containing rationales or justifications for classification decisions. Such a technique, cry individuating as fine-tuning, not only increases the model's transparency but also allows for counterfactual analysis, an important element in algorithmic audits.

Combining Privacy-Preserving Techniques

Increased legislative scrutiny over user data warrants the necessity of integrating privacy-ameliorating techniques directly into the training process for developers. SFT can significantly put its weight behind this endeavor because more often than not, it involves much smaller and better-controlled datasets compared to the pertaining phase.

The privacy-preservation methods apt for SFT would include the following:

Differential Privacy (DP): Involves noise being added to the fine-tuning process statistically with an aim to prevent individual data leakage while keeping overall trends intact (Dwork & Roth, 2014).

Federated Learning (FL): Allows models to be trained fine with decentralized user data and, thereby, no raw user data be carried by the centralized server (Kairouz et al., 2021). Synthetic Data Augmentation: Involved the application of GANs or Probabilistic Models to generate training data replicating distributions in real life without exposure of sensitive information.

Adding these methods into SFT will help developers produce systems that are compliant and secure. For example, Apple's on-device AI models are fine-tuned using the federated learning model, ensuring that any user behavior data that is confidential never leaves the device and falls into the realms of CCPA and GDPR compliance.

Table 3: Privacy Techniques and Their Integration with SFT

Privacy Technique	SFT Integration	Regulatory Alignment
Differential Privacy	Adds noise during gradient updates in fine-tuning	GDPR Article 5 (Data Minimization)
Federated Learning	Enables distributed fine-tuning on edge devices	CCPA (Right to Data Control)

Privacy Technique	SFT Integration	Regulatory Alignment
Synthetic Data Generation	Uses generated data for supervised tasks without risk	AI Act (Data Governance Obligations)
Data Anonymization	Removes personal identifiers from fine-tuning datasets	NDPA Nigeria (Data Security)

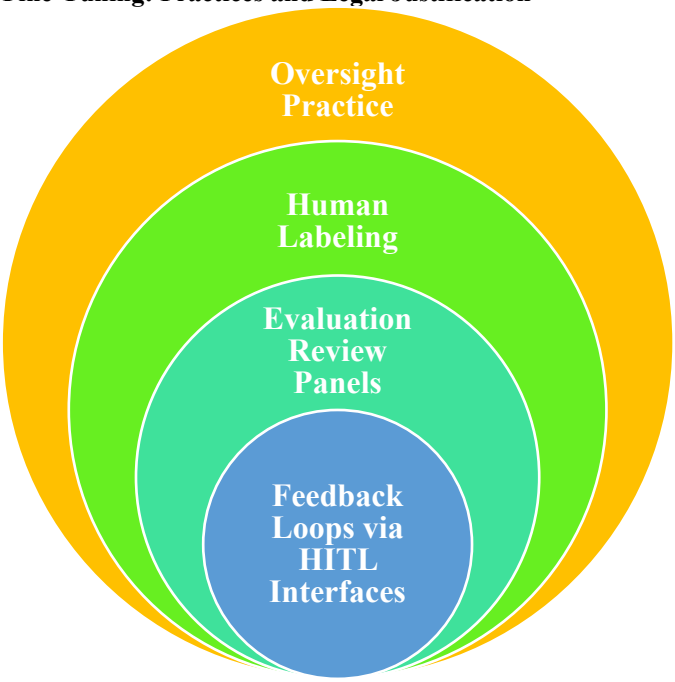
Human-in-the-Loop Oversight and Continuous Auditing

Revisions are not a one-off process but benefit from continuous feedback and real-time supervision. Human-in-the-loop (HITL) systems permit domain experts to undertake the supervisory role. They design data labeling, training, and evaluation tasks, so far as ethically acceptable. In so doing, the human-in-the-loop concept enhances performance and also maintains compliance with current legal requirements for human oversight in decision-making systems (GDPR Article 22).

Recent work by Ouyang et al. (2021) has shown that fine-tuning using RLHF offers a model that tends to display behavior more along human lines. Combining RLHF with SFT has shown clearly beneficial effects when dealing with tasks related to value-sensitive decisions —like educational assessment or ranking job candidates.

Audit trails must be maintained for all steps of fine-tuning in order to maintain traceability. This includes lists of annotated datasets, the agreements for the same, and both the training and learning of the fine-tuned models; retrospective studies on data origin will be extremely useful in this context as proof to demonstrate that questionable materials weren't used in the fine-tuning exercise.

Human Oversight in Fine-Tuning: Practices and Legal Justification



Challenges and Future Opportunities

Even though applying SFT in the realm of Responsible AI offers room for some praise, it does not come without challenges. Key problems include:

Data Access and Ownership

To do fine-tuning, one needs access to high-quality datasets, which are often proprietary or sensitive.

Annotation Quality

Foregrounding good ethical tuning is the judgment of annotators, who might instead bring an influence of their own.

Scalability

Bringing privacy-preserving SFT to scale remains an immense computational challenge.

Regulatory Ambiguity

New regulations and laws spurred by new case law might result in shifting requirements, necessitating constant legal-technical bipartisanship.

However, there are quite a few opportunities for bettering the situation. Open instruction-tuned models have started cropping up. Model Cards are also becoming a norm. We also have many technological paths to scalable, privacy-compliant fine-tuning with Low-Rank Adaptation (LoRA).

CONCLUSION

The urgency for informing their responsible design and use could not be underscored more as artificial intelligence continues to penetrate into sensitive realms of human life, like those of healthcare, education, finance, and public governance. This paper takes a close look at Supervised Fine Tuning (SFT) as a tool central for aligning AI models with ethics and evolving legal standards.

- From this extensive review of related literature and policy, it can be gathered that SFT holds copious promises for:
- Reduction of bias and reinforcing fairness through offering developers an opportunity to recalibrate a pertained model with datasets that are of great ethical value;
- Enabling transparency and interpretability through more transparent and controlled behavior;
- Respect for privacy and regulatory compliance by calling for implementations like differential privacy and federated learning;
- Continuous governance and accountability through feedback from human operators, traceable documentation, and post-deployment audits.

However, any advancement through SFT is highly dependent upon the environment where it is performed. Without high-quality, diverse, and ethically annotated data, the most sophisticated fine-tuning strategies shall only contribute to reinforcing the existing inequities. Moreover, technical interventions must complement legal frameworks, social engagement, and organizational ethics.

In conclusion, SFT is mainly understood not as a technical solution but as a governance strategy that embodies the principles of Responsible AI: promoting user trust and ensuring societal well-being through maintaining the integrity of models.

Recommendations

Based on the findings of the research, the following are the recommendations for developers, policymakers, and AI governance bodies:

A. For AI Developers and Engineers

Incorporating Ethical Design Right from the Start

Integrate ethical design into dataset collection and annotation during SFT, paying attention to fairness, inclusiveness, and domain relevance.

Default to Privacy-Preserving Techniques

For bias control on personal or sensitive data, preserve the privacy of the individual model through, in particular, differential privacy or federated learning while complying with the GDPR, NDPA, and CCPA.

Document the Model Cards and Audio Log in Detail

Document every step of SFT, starting from data sources and annotation logic and ending with metrics, to make transparent models for future investigation.

Use Human-in-the-Loop Architectures

Domain experts must be able to ensure a continuous assessment and improvement of model behavior both during and after the fine-tuning process.

B. For Policy Makers and Regulators

Mandating Transparent SFT Pipelines

Insert clear norms in forthcoming AI laws such as AI Act that point out that an extensive record of the fine-tuning pipeline and detailed audits are mandatory.

Supporting Datasets That Are Open and Well-Annotated

Putting special attention on creating open-access platforms for datasets having annotations that possess ethical value, this is especially true in the context of partially supported or otherwise underrepresented languages and regions.

Standardization of Impact Assessment
Here, the intention is to ensure pre-deployment Algorithmic Impact Assessments for models that have undergone fine-tuning for high-risk domains, while developing human oversight policies.
Creating an Ecosystem of Multistakeholder Collaborations
Encourage the development of panels with representatives from various sectors, as in ethicists, law experts, civil society, and technologists, for co-building norms around fine-tuning.

C. For Educational and Research Institutions
Promotion in Interdisciplinary Curricula
Make sure to train next-generation AI professionals to think across legal, ethical, and technological boundaries, primarily focusing on real-life cases of fine-tuning.

Research in Contextual Fine-Tuning
Invest in learning more about cultural, language, and local ethical implications on fine-tuned model outputs for varying populations.

Open-Source Contributions to Tooling
Develop and provide tools for ethical SFT like those supporting bias detection during fine-tuning and fairness-aware loss functions or privacy risk estimators."

Final Thoughts
As global societies shift toward frontline acceptance of AI, it becomes vital that tech progress not come at the expense of human dignity, agency, and legal rights. Supervised fine-tuning provides a window for injecting values into machine learning models but such function should come with an array of responsible oversight and inclusive governance.
With the embodiment of SFT within a broader ethical structure that celebrates accountability, transparency, and participation, we may edge closer to a world where AI acts (and not just operates) in alignment with the values we hold dear.

References
[1] Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-Muslim bias in large language models. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 389–400.
[2] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 149–159.
[3] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 1–15.
[4] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407.
[5] European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (AI Act).
[6] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1), 1–15.
[7] Ghallab, M. (2023). Ethical AI governance: From principles to practice. *AI & Society*, 38(4), 897–911.
[8] Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(3), e12432.
[9] IBM. (2022). AI Fairness 360 toolkit. Retrieved from https://aif360.mybluemix.net/
[10] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
[11] Kairouz, P., McMahan, H. B., & Avent, B. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1), 1–210.
[12] Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 633–705.
[13] Leike, J., Krueger, D., & Everitt, T. (2023). Aligning language models with human values. *Journal of Artificial Intelligence Research*, 77, 1–32.
[14] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.
[15] Nigeria Data Protection Commission (NDPC). (2023). Nigeria Data Protection Act.

[16] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
[17] Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.
[18] Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results. *AAAI/ACM Conference on AI, Ethics, and Society*, 429–435.
[19] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., & Barnes, P. (2020). Closing the AI accountability gap. *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, 33–44.
[20] Rocher, L., Hendrickx, J. M., & de Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1), 3069.
[21] Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4), 105–114.
[22] Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1310–1321.
[23] Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, 361(6404), 751–752.
[24] Tolan, S., Miron, M., Gomez, E., & Castillo, C. (2022). Measuring and mitigating unfairness in AI. *Journal of Artificial Intelligence Research*, 74, 1–43.
[25] Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Proceedings of the Machine Learning for Healthcare Conference*, 359–380.
[26] U.S. National Institute of Standards and Technology (NIST). (2023). AI Risk Management Framework.
[27] Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2), 2053951717743530.
[28] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop*, 353–355.
[29] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2019). Gender bias in contextualized word embeddings. *Proceedings of NAACL-HLT*, 629–634.
[30] Zhou, K., & Suresh, H. (2023). The ethics of adapting language models across domains. *Ethics and Information Technology*, 25(2), 141–158.