Understanding and Supporting the ML Supply Chain through ML Bill of Materials

Trevor Stalnaker
William & Mary
Williamsburg, Virginia, USA
twstalnaker@wm.edu

***Abstract*—Within the last decade, the Machine Learning (ML) supply chain has emerged with increasing complexity. This dissertation focuses on identifying and resolving the challenges faced by various stakeholders in the ML supply chain, including those relating to provenance and compliance tasks. These challenges will be identified through a combination of surveys, interviews, mining studies, and literature reviews. They will be addressed by employing Machine Learning Bills of Material (MLBOM) accompanied with appropriate automated tooling solutions. Our anticipated contributions include developing a rich understanding of practitioner needs, undertaking a comprehensive evaluation of the current ML supply chain, and implementing novel tooling solutions to assist ML supply chain stakeholders.**

I. Problem and Research Statement

The use of machine learning (ML) models in software applications has increased dramatically in recent years, but training large ML models is often cost and resource prohibitive [1], leading developers to instead fine-tune pre-existing open-weight models [2]. This has resulted in a complex ML supply chain with various stages (*e.g.*, data curation, training, fine-tuning, *etc.*) and components (*e.g.*, training data, base models, prompts, *etc.*), each with a measurable effect on model output. Lee *et al.* have identified eight stages that make up this ML supply chain [3], which do not always occur in a linear sequence (*e.g.*, model output can be used as part of future training sets and multiple rounds of fine-tuning can be employed).

Problems in the traditional software supply chain, such as dependency tracking and license compliance, are also found in the ML supply chain [4], [5], but due to the immaturity of the ecosystem, some of these challenges may be exacerbated. For example, compliance is difficult when datasets are often poorly or incorrectly licensed (*e.g.*, contain examples with conflicting licenses) [6] and models are poorly or incompletely documented [7], [8], [9]. Yet, failure to comply with licensing terms of models, datasets, and even traditional software components can lead to reputational [10], [11], financial [12], [13], and/or legal damage [14], [15]. However, developers often struggle with compliance tasks and software licenses [16], [17], [18], [19], a problem made worse by the wide variety of licenses routinely used in the ML ecosystem. Models on Hugging Face (HF) [20] can be licensed under creative-commons, open-source, and ML-specific licenses [21]. ML-specific licenses are similar to open-source, but impose additional restrictions on model usage, typically revolving around ethical issues (similar to the BSD-3-Clause No Nuclear License [22]). It can be challenging to apply license terms to these new technologies and to understand how various licenses relate to each other. For example, it remains unclear what is meant by derivative work in this domain (*e.g.*, should a model be considered a derivative work of its training data?). Additionally, these documentation shortcomings (and the inherent black box nature of ML systems) impact component traceability and can also complicate the detection and resolution of security/privacy issues.

MLBOM (ML Bill of Materials) has been proposed to address some of these challenges [5], [4], [9]. An MLBOM is a specialized SBOM (Software Bill of Materials) that can be thought of as a nutrition fact block for an ML model, acting as a manifest that lists relevant training and licensing information, as well as a variety of other meta data, such as architecture details, usage requirements, and known short-comings. MLBOM is distinct from HF model cards [23], where information is not always available in a standardized, machine-readable format [7], [8], [9]. Both major SBOM standards, CycloneDX [24] and SPDX [25], offer support for MLBOM.

While MLBOMs have the potential to help various stakeholders address the challenges of an increasingly complex ML supply chain, it remains unknown how MLBOMs are produced and used in reality, what tooling and best practices are being employed, and what unmet needs still remain. While much effort has been spent on the development of MLBOM standards, significant gaps impeding adoption and effective usage likely remain in developer practices, tool support, and even the existing standards.

This dissertation hypothesizes that the various challenges in the ML supply chain can be resolved through the effective usage of MLBOM. We aim to support developers in the ML supply chain by understanding their current workflows, exploring the challenges they face, developing novel tooling solutions for the generation and consumption of detailed, useful MLBOM, and facilitating compliance tasks. The next section provides an overview of the research plan and various research directions.

II. Proposed Research

A. Understanding the State of the ML Supply Chain

We propose a combination of surveys, interviews, and mining studies to understand (1) the ML supply chain practices, needs, and challenges of various stakeholder groups (*e.g.*, developers, AI engineers, data scientists, compliance professionals, procurement personnel, regulatory bodies, *etc.*), (2) how MLBOMs

are being produced and consumed to support the supply chain, and (3) how MLBOMs and associated tooling can be used to provide better support.

1) Investigating Stakeholder Perspectives: Through a combination of surveys and interviews with various stakeholder groups we will determine how MLBOMs are currently used by practitioners, what obstacles are encountered during their creation/consumption, which metadata fields are considered most important, and what barriers exist to wide-spread adoption. Participants will be identified from our professional network, mailing lists, relevant social platforms, online communities, software repositories, and model hubs.

2) Mining Studies: Next, we will conduct supplementary mining studies consistent with, but improving upon prior work [7], [8]. Our mining studies will provide a current snapshot of the rapidly evolving ML supply chain, which, when combined with survey and interview insights, will help identify and prioritize key challenges facing the community. Beyond triangulating these challenges, we also aim to understand dependency relationships between models, assess license compliance, and evaluate the completeness of model documentation. Our prior surveys and interviews will also inform data filtering decisions aimed at preserving the integrity of the mined dataset, since we suspect that some model uploaders may just be using model hubs, like Hugging Face, as an intermediary during ad hoc processes involving Google Collab and other tools.

B. MLBOM for the Supply Chain

We aim to support stakeholders across the supply chain by developing practices and technology to produce/consume accurate, complete MLBOMs while also informing the future development of existing MLBOM standards.

1) Producing MLBOM: Based on the results of the initial surveys, interviews, and mining studies, we will propose tools, techniques, and best practices that can help developers to produce more useful MLBOMs, focusing primarily on those metadata fields identified as most important to consumers. In the traditional software supply chain, consumers are often left to produce SBOMs for their dependencies, which can be an error prone, time-consuming process [4]. This problem is exacerbated in the ML context given the inherent difficulties of determining what was contained in a model's training set, what training parameters were used, and what training techniques were employed. If this information is not supplied by the model trainer, it is difficult, if not impossible, to obtain post-hoc. This is exceptionally problematic since the information provided by model trainers is often incomplete. Further, traditional SBOM generation tools, that rely on build files and static analysis, will also fail. To this end, our solutions will try to automatically produce as much of the MLBOM as possible, reducing the burden placed on the model trainer and increasing the likelihood of sufficient documentation. To achieve this, we will focus on four questions: (1) Which fields should be prioritized for automatic completion? (2) What techniques and strategies can be used to extract information for automatically filling those fields? (3) How can we best facilitate user-tool interactions to gather additional information and validate automatically populated fields? (4) What institutional and development practices/processes, if any, must be in place to foster effective MLBOM creation regardless of tool capabilities? Our solutions will also aim to balance the need for completeness in MLBOM with the risk of introducing incorrect or misleading information.

2) Consuming MLBOM: Here we aim to specifically address the needs of MLBOM consumers, which might differ from the SBOM consumption needs of standard software developers. Different stakeholders will likely desire different information from an MLBOM. For example, a model user might care about known limitations and runtime requirements, whereas a model fine-tuner may care more about architecture information and training parameters, since when fine-tuning models, it is best practice to format the new dataset to match the original training set, which requires information on the formatting of that original training data. Combining an analysis of existing MLBOM tools with the results from our previous surveys and interviews, we will identify the short-comings of the state-of-the-art MLBOM consumption tools and then propose better alternatives that more closely align with stakeholder needs.

III. Anticipated Contributions

This dissertation is intended to support the development and compliance activities of various stakeholders in the ML supply chain, including those working on both open-source and proprietary projects, by effectively employing MLBOM and appropriate tooling solutions. Our first contribution will include a thorough evaluation of the current ML supply chain through surveys, interviews, and mining studies, leading to a deeper understanding of practitioner needs. Further, this contribution will elucidate which fields different stakeholder groups require from MLBOM, how consumers intend to use MLBOMs, and what obstacles prevent the effective creation of accurate MLBOM. Our second contribution will focus on the development of novel tools and techniques for producing and consuming MLBOM that will assist stakeholders in their activities and address challenges they face.

All datasets, where possible, will be made available for verifiability. Survey and interview questions will be drafted and analyzed using best practices from the field [26], [27], [28], [29], [30], [31], [32]. We will collect and curate a dataset of ground truth MLBOM on which to evaluate tooling solutions. Tooling solutions will also be evaluated against existing state-of-the-art MLBOM generation/consumption tools. Tool effectiveness and usefulness will be evaluated through user studies. Finally, source code for any tooling solutions will also be made available so that the claims made can be independently verified. We will conduct a triangulation of findings across the different studies.

Our solutions will be actionable and practical, allowing the various stakeholders in the ML supply chain to make more informed decisions about the models and datasets they intend to use or iterate on. Our goal is to make the process of tracking dependencies, maintaining compliance, and resolving other ML supply chain issues more efficient, accurate, and reliable.

REFERENCES

[1] C. S. Smith, "What large models cost you – there is no free ai lunch," https://www.forbes.com/sites/craigsmith/2023/09/08/what-large-models-cost-you--there-is-no-free-ai-lunch/, 1 2024.

[2] S. in the Office of Technology, "What large models cost you – there is no free ai lunch," https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/open-weights-foundation-models, 7 2024.

[3] K. Lee, A. F. Cooper, and J. Grimmelmann, "Talkin' 'bout ai generation: Copyright and the generative-ai supply chain," *arXiv preprint arXiv:2309.08133*, 2023.

[4] T. Stalnaker, N. Wintersgill, O. Chaparro, M. Di Penta, D. M. German, and D. Poshyvanyk, "Boms away! inside the minds of stakeholders: A comprehensive study of bills of materials for software systems," in *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*, 2024, pp. 1–13.

[5] B. Xia, T. Bi, Z. Xing, Q. Lu, and L. Zhu, "An empirical study on software bill of materials: Where we stand and the road ahead," *arXiv preprint arXiv:2301.05362*, 2023.

[6] S. Longpre, R. Mahari, A. Chen, N. Obeng-Marnu, D. Sileo, W. Brannon, N. Muennighoff, N. Khazam, J. Kabbara, K. Perisetla *et al.*, "The data provenance initiative: A large scale audit of dataset licensing & attribution in ai," *arXiv preprint arXiv:2310.16787*, 2023.

[7] W. Jiang, N. Synovic, P. Jajal, T. R. Schorlemmer, A. Tewari, B. Pareek, G. K. Thiruvathukal, and J. C. Davis, "Ptmtorrent: A dataset for mining open-source pre-trained model packages. in 2023 ieee/acm 20th international conference on mining software repositories (msr). 57–61," 2023.

[8] W. Jiang, J. Yasmin, J. Jones, N. Synovic, J. Kuo, N. Bielanski, Y. Tian, G. K. Thiruvathukal, and J. C. Davis, "Peatmoss: A dataset and initial analysis of pre-trained models in open-source software," in *2024 IEEE/ACM 21st International Conference on Mining Software Repositories (MSR)*. IEEE, 2024, pp. 431–443.

[9] F. Pepe, V. Nardone, A. Mastropaolo, G. Bavota, G. Canfora, and M. Di Penta, "How do hugging face models document datasets, bias, and licenses? an empirical study," in *Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension*, 2024, pp. 370–381.

[10] T. Claburn, "John deere urged to surrender source code under gpl," https://www.theregister.com/2023/03/17/john_deere_sfc_gpl/, 3 2023, accessed: 2023-14-09.

[11] N. Gunningham, R. A. Kagan, and D. Thornton, "Social license and environmental protection: Why businesses go beyond compliance," *Law & Social Inquiry*, vol. 29, no. 2, pp. 307–341, 2004, https://doi.org/10.1111/j.1747-4469.2004.tb00338.x.

[12] A. Vance, "The defenders of free software," https://www.nytimes.com/2010/09/26/business/26ping.html, 9 2010, accessed: 2023-14-09.

[13] G. Gross, "Open-source legal group strikes again on busybox, suing verizon," https://www.computerworld.com/article/2537947/open-source-legal-group-strikes-again-on-busybox--suing-verizon.html, 12 2007, accessed: 2023-14-09.

[14] T. Claburn, "Gpl legal battle: Vizio told by judge it will have to answer breach-of-contract claims," https://www.theregister.com/2022/05/16/vizio_gpl_contract/, 5 2022, accessed: 2023-14-09.

[15] J. Vincent, "The lawsuit that could rewrite the rules of ai copyright," https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data, 11 2022, accessed: 2023-14-09.

[16] D. M. German and A. E. Hassan, "License integration patterns: Addressing license mismatches in component-based development," in *Proceedings of the 2009 IEEE 31st International Conference on Software Engineering (ICSE'09)*. IEEE, 2009, pp. 188–198, https://doi.org/10.1109/ICSE.2009.5070520.

[17] G. Gangadharan, V. D'Andrea, S. De Paoli, and M. Weiss, "Managing license compliance in free and open source software development," *Information Systems Frontiers*, vol. 14, pp. 143–154, 2012, https://doi.org/10.1007/s10796-009-9180-1.

[18] D. A. Almeida, G. C. Murphy, G. Wilson, and M. Hoye, "Investigating whether and how software developers understand open source software licensing," *Empirical Software Engineering*, vol. 24, pp. 211–239, 2019, https://doi.org/10.1007/s10664-018-9614-9.

[19] C. Vendome, D. M. German, M. Di Penta, G. Bavota, M. Linares-Vásquez, and D. Poshyvanyk, "To distribute or not to distribute? why licensing bugs matter," in *Proceedings of the 40th International Conference on Software Engineering (ICSE'18)*, 2018, pp. 268–279, https://doi.org/10.1145/3180155.3180221.

[20] H. F. Inc., "Hugging face https://huggingface.co," 2024. [Online]. Available: https://huggingface.co

[21] H. Face, "Licenses," https://huggingface.co/docs/hub/en/repositories-licenses, 2024.

[22] "Bsd 3-clause no nuclear license," https://spdx.org/licenses/BSD-3-Clause-No-Nuclear-License.html, accessed: 2025-29-01.

[23] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[24] "Cyclonedx specifications." [Online]. Available: https://github.com/CycloneDX/specification

[25] "SPDX specifications," The Linux Foundation. [Online]. Available: https://spdx.dev/specifications/

[26] S. L. Pfleeger and B. A. Kitchenham, "Principles of survey research: part 1: turning lemons into lemonade," *ACM SIGSOFT Software Engineering Notes*, vol. 26, no. 6, pp. 16–18, 2001.

[27] B. A. Kitchenham and S. L. Pfleeger, "Principles of survey research part 2: designing a survey," *ACM SIGSOFT Software Engineering Notes*, vol. 27, no. 1, pp. 18–20, 2002.

[28] ——, "Principles of survey research: part 3: constructing a survey instrument," *ACM SIGSOFT Software Engineering Notes*, vol. 27, no. 2, pp. 20–24, 2002.

[29] ——, "Principles of survey research part 4: questionnaire evaluation," *ACM SIGSOFT Software Engineering Notes*, vol. 27, no. 3, pp. 20–23, 2002.

[30] ——, "Principles of survey research: part 5: populations and samples," *ACM SIGSOFT Software Engineering Notes*, vol. 27, no. 5, pp. 17–20, 2002.

[31] ——, "Principles of survey research part 6: data analysis," *ACM SIGSOFT Software Engineering Notes*, vol. 28, no. 2, pp. 24–27, 2003.

[32] D. Spencer, *Card sorting: Designing usable categories*. Rosenfeld Media, 2009.