Unpacking Trust Dynamics in the LLM Supply Chain: An Empirical Exploration to Foster Trustworthy LLM Production & Use [Supplementary Material]

Agathe Balayn
ServiceNow, Trento University, Delft University of Technology
The Netherlands
a.m.a.balayn@tudelft.nl

Mireia Yurrita
m.yurritasemperena@tudelft.nl
Delft University of Technology
The Netherlands

Fanny Rancourt
fanny.rancourt@servicenow.com
ServiceNow
Canada

Fabio Casati
ServiceNow, University of Trento
Switzerland, Italy
fabio.casati@servicenow.com

Ujwal Gadiraju
Delft University of Technology
The Netherlands
u.k.gadiraju@tudelft.nl

1 ADDITIONAL INFORMATION ABOUT THE LLM SUPPLY CHAIN

1.1 Many, diverse, entities populated the LLM-supply chains

- *Technical artifacts*: LLM-based services were developed along multiple phases and relied on several technical components. In use, they were composed of a fine-tuned model for the application at hand, but also user interfaces for their end-users, workflows for post-filtering the outputs of the LLMs (e.g., for toxicity), logging data and monitoring potential issues, infrastructures to support the computations and data needs, etc. Such services might even be composed of several fine-tuned models chained together [7]. Building each of these in-use components required various build-up components, such as pre-training datasets, the foundation model on which the model was fine-tuned, a fine-tuning dataset, training scripts, data processing scripts, more infrastructures to handle data and computations, etc. The granularity of these components is not set: sub-components make up these in-use-components and build-up-components, e.g., a training dataset is made of data samples and annotations, a filtering workflow might be made of several filters for gender bias, race bias, toxicity, etc. *P33 "[Deployer organization], as a trusted platform, is easier to get started with and to deploy use cases on than building a new system from scratch, and having to work through your own security and firewalls and load balancers, and all the complexity behind standing up a new web application."*
- *Provider*: An entity might participate in producing any technical component used downstream to produce and maintain the final LLM. Often, it was not the same entity that worked on the different components of the LLM. Policy documents [1, 6] make the useful distinction between the "developer" (or "provider") and the "deployer" of the LLM, where the developer creates a fine-tuned LLM and the deployer makes this LLM ready for use, e.g., integrating it within a software stack, adding filters in the output of the LLM to circumvent potential offensiveness. We add the "assessor" category that deals with evaluating the LLM or its components during or after development, according to various requirements, e.g., to make sure of the legal compliance of the LLM and assess whether the system is ready for use. Assessors were involved with the assessment of different aspects of the AI system, be it the quality of the outputs of the AI system (e.g., accuracy, fairness), the system performance (e.g., speed), or the legality and ethics of its process (e.g., model training on non-copyrighted data). Note that at a finer granularity level, providers adopted specific roles vis-a-vis the components they were in charge of (e.g., developers, researchers, designers).
- *Consumer*: This could either be an independent user or an organization that provided access to an LLM-based service to its employees or external users. The consumer employed the LLM-based service to conduct a certain task for their personal or professional work. Integrating the LLM into existing workflows might require additional efforts, e.g., to merge the application with existing software. Here we defined consumers and producers with regard to the final LLM. Yet, producers could themselves be considered consumers of other technical artifacts created by other producers (e.g., an LLM fine-tuner, while being the producer of the fine-tuned LLM, might be the consumer of training data samples and their annotations). In that sense, the nature of an organization did not depend on whether it was an LLM consumer or producer, it could be both for different systems and different technical artifacts.
- *Indirect stakeholders*: Other entities which did not explicitly participate in the LLM supply chain, could be impacted by the LLM due to negative externalities [3] or could simply have an opinion about the LLM and its impact on society [10]. LLMs impacted indirect stakeholders in several ways. Indirect impact dealt with the environmental impact of the training and deployment of the system on various communities [11], the potential privacy infringement of the data used to train and evaluate the AI systems [2], etc. Direct impacts dealt with issues related to the inputs and

Please use nonacm option or ACM Engage class to enable CC licenses
This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '25, April 26-May 1, 2025, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713787

outputs of the LLM. One might unintentionally be exposed to the potentially problematic (e.g., offensive) outputs of the LLM. Besides, one might be the content-subject in an LLM query or in the training data, and become at risk. Such risk could revolve around misrepresentation of the content-subject (e.g., due to hallucination), infringement of their data rights (the LLM prompt might contain information about them, that might be logged in by the system potentially without their consent [9], e.g., to further fine-tune the AI model), or other privacy infringement (the LLM user might get access to their private information adversarially or not, because of information leakage problems and hallucinations that LLMs suffer from [12]).

1.2 The entities connected at several junctions of the supply chain

- *Junctions for the production of technical artifacts.* Producers had to make the first decision to pour efforts and invest in producing a technical artifact. Later on, many decisions were made towards the production of the artifact. These could be decisions internal to the producer (e.g., which base model to use), or decisions involving multiple producers (e.g., deciding to rely on one data annotation organization or another).
- *Junctions at the intersection between production and consumption.* Later on, once a technical artifact had been developed, the producer needed to decide that this artifact was satisfying enough to release it. Then, the producer also needed to make various access-related decisions, such as deciding on the ways consumers will access this artifact, which consumers will be granted access, etc.
- *Junctions in the consumption of the technical artifacts.* Finally, consumers decided for or against adopting a technical artifact (e.g., based on whether it functioned, or whether it reduced cost), and use was granted or pushed toward individual end-users. Indirect stakeholders and the individual end-users, in this context, could decide to contest the production and/or consumption of the technical artifact.

These junctions were not only relevant to the final LLM, but also to the technical artifacts required to develop this LLM-based service. The actors could greatly vary across junctions and supply chains, be they organizational or individual (e.g., an organization decided to adopt the LLM of another organization, or a product manager in this first organization actually took the decision).

1.3 The entities and junctions involved in the supply chain were extremely complex

No two supply chains were identical. They differed in terms of entities and artifacts involved, and of the relations between them.

- *Organizational complexity*: A single organization could play multiple roles vis-a-vis the same technical artifact, as developer and deployer, or even consumer. An organization could also play several roles vis-a-vis different artifacts (e.g., a company could specialize in deploying AI systems for various applications, and hence maintained several LLMs at a time). Several organizations could bear the same types of role within the same supply chain, e.g., of developer for the foundation or the fine-tuned model.
- *Individual complexity*: An individual entity could adopt multiple roles vis-a-vis one or multiple organizations or artifacts. For instance, someone employed by a consumer organization and hence used an LLM professionally, could also use the same or a different LLM outside the context of the company, e.g., using the free version of ChatGPT to prepare their homework. Besides, they could be users of an LLM and producers of a different one. For instance, AI developers relied on LLM-code-generators to produce the code for the LLM their organization developed, and marketing teams used LLMs to produce content to advertise the LLM their organization developed.
- *Technical complexity*: One technical component of an LLM could be used to develop more than one application (e.g., public datasets).

2 ADDITIONAL FINDINGS

Because of similarities with prior literature, we did not expand on the findings below in the main text.

2.1 The trustee's factors related to the technical artifacts that impact trust

The trust factors that trustors typically paid attention to are aligned with those that prior works have investigated and that communicate the trustworthiness of an AI system [5], and directly relate to the trust expectations and vulnerabilities above. The **ability** of the LLM-based service is for instance illustrated by precise considerations around the accuracy of the service's outputs. An AI engineer (P12) discussed the model's accuracy as a priority to develop their generative AI offerings: *"It would be really important to have accuracy: it builds trust. If it fails miserably, trust will go down quickly."*; while a product manager (P21) in a provider organization insisted on their need to generally know about the LLM's capabilities: *"Transparency means knowing the capabilities of the model, where it performs well or not. With that, I will be able to better trust it and utilize it."* Interestingly, despite trustor's expectations sometimes being at the organizational level, only two consumer organizations went beyond the properties of the LLM-based systems and attempted at evaluating the ability of the systems to fulfill organizational benefits and avoid organizational risks. They attempted at measuring increased productivity by testing the system with real users over several weeks.

As for the **process integrity**, while prior works discussed integrity for non-LLM systems and focused on explainability mechanisms [5], our participants discussed the source of the information outputted by the LLM-based service and the advantages of retrieval-augmented LLM generation [4], and the suitability of the system's individual components. For instance, a UX researcher (P6) discussed how they prompted their provider organization to design meaningful experiences for the end-users, instead of focusing on pleasing the product managers of the consumer organizations *"We are trying to push for including transparency and even onboarding in the service. Otherwise, people are just not gonna adopt the product because they won't trust it since they can't understand where it comes from and how the results are generated."* The **intention benevolence** is reflected by considerations around the ethicality of the system and its production. For instance, another UX researcher (P5) pointed out the necessity to account for both the

ability and benevolence of the service: *P5 "If we think in terms of product adoption and trust building experience, we want to make sure that our LLM is performing very well, but also obviously from the responsibile AI side, it should be safe before we roll it out."*

2.2 The factors related to the trustor that impact trust

Differences in the extent of trust are due to the trustors' personal perceptions and expectations of the trustee.

- *Two trustors could judge the value of one trustee's characteristics differently*. For instance, one trustor considered the ability of one LLM as acceptable and meeting their expectations because the accuracy reached a high percentage. Instead, another trustor doubted it because they did not think accuracy was measured appropriately. Similarly, some perceived the developer organization making some information transparent as a sign of integrity, while others remained more critical and saw it as a lack of integrity because not all information was transparent.
- *Different trustors presented different expectations towards a same trustee and activity*. They attributed different relative importance to the trustee's characteristics. For instance, many consumer organizations trusted deployer organizations based on signs of integrity and ability (both at the organizational- and LLM-levels), while indirect stakeholders and individual end-users attached more attention to the benevolence of the deployer and consumer organizations. The employees of the developer organization in turn emphasized benevolence concerning potential job displacement and integrity concerning the deployment of an LLM.
- These differences came from trustors' inherent subjectivity, their natural propensity for trust in organizations and in technology, the ways they perceived AI systems, their advantages and potential concerns. Trustors' knowledge of LLMs (e.g., their functioning and potential harm) also impacted the types of trustworthiness cues they were looking for. *P34 "Continuing to educate everybody on what's the best use of AI and what is trustworthy AI, it will allow each team to make better calls of what to use."*

2.3 The contextual factors that impact trust

- *The stakes of the trust activity directly impacted the relation.* While being aware of several LLM issues (e.g., unfairness and brittleness), consumer organizations were primarily concerned with accuracy issues and data leakages because of the risks such issues might cause to their business. In turn, the considerations of end-users and indirect stakeholders revolved around the risks of using LLMs for themselves (e.g., the offensiveness of the outputs and potential discrimination towards their community). They did not necessarily match those of the consumer organizations.
- *Additional contextual factors impacted trustors' expectations*. The expectations of the consumer organizations towards the LLM differed depending on the environment in which individual users would use this LLM. Particularly, when users are employees of the consumer organization, we found expectations in terms of accuracy, while in cases where the users are external to the organization, we also found expectations concerning the offensiveness of the LLM outputs –although mishaps can happen, employees could be trained, informed or notified so that harm is minimized in the first case.
- *Past interactions with other similar trustees impacted trust*. A deployer organization trusted the LLM for text translation from one producer, and consequently also trusted the LLM for text summarization of this producer.
- *Role multiplicity impacted trust.* For trustees with multiple roles, their ability, benevolence, integrity characteristics vis-a-vis their different roles impacted each other. For instance, if an organization was both developing and deploying the LLM, the end-user could expect more of this organization than if it had a single role. Besides, if a technical artifact was used to develop two different LLMs and one of them had a bad reputation, trust towards this artifact and the second LLM could decrease. As for trustors with multiple roles, having different trust relations with an LLM in different qualities was likely to impact their overall trust towards AI and individual LLMs.

2.4 Miscellaneous findings about miscalibrated trust, distrust, and trust for adversarial purposes

- *Miscalibrated trust due to lack of expertise and misinterpretations*: Certain participants believed that fine-tuned models are trustworthy as long as the underlying foundation model is, which is not always the case in practice [8]. Others, simply because of the presence of explanations for the outputs of the systems, had a stronger belief in the integrity of the organization and system, despite these explanations not necessarily being informative.
- *Impossibility of trust*: Tensions between technical feasibility and these meta-properties surfaced. For instance, while certain participants asked for extensive transparency about technical details of the LLM supply chain, this revealed to be impossible to provide technically due to the temporality of the supply chain.
- *Trust dynamics that do not foster trustworthy LLM supply chains*: Some activities in the trust relations contribute to maintaining the company's trade secrets about the performance of the LLMs (that is potentially low) or simply about the LLMs' technical details –i.e., a producer organization or a manager trusting its employees or end-users not to release to the public any information about the LLM.

REFERENCES

[1] BSA (The Software Alliance). [n. d.]. *AI Developers and Deployers: An Important Distinction.* https://www.bsa.org/files/policy-filings/03162023aidevdep.pdf

[2] Abeba Birhane. 2020. Algorithmic colonization of Africa. *SCRIPTed* 17 (2020), 389.

[3] Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020. POTs: protective optimization technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 177–188.

[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[5] Q Vera Liao and S Shyam Sundar. 2022. Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1257–1268.

[6] Tambiama Madiega. 2021. Artificial intelligence act. *European Parliament: European Parliamentary Research Service* (2021).
[7] Besmira Nushi, Ece Kamar, and Eric Horvitz. 2018. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 6. 126–135.
[8] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! arXiv:2310.03693 [cs.CL]
[9] Siladitya Ray. 2024. Samsung bans chatgpt among employees after sensitive code leak. https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/
[10] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III, Jesse Dodge, Ellie Evans, Sara Hooker, et al. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. *arXiv preprint arXiv:2306.05949* (2023).
[11] Aimee Van Wynsberghe. 2021. Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics* 1, 3 (2021), 213–218.
[12] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing* (2024), 100211.