ICLR 2026 3rd Workshop on Navigating and Addressing Data Problems For Foundation Models (DATA-FM)

Abstract

The past year has witnessed remarkable advances in foundation models (FMs): new post-training paradigms such as reinforcement learning with verifiable rewards (RLVR) that strengthen reasoning, increasingly multimodal and agentic systems, and renewed attention to benchmark design and evaluation. Each of these advances depends on distinct data innovations: verifiable reward signals and reasoning traces for RLVR; aligned cross-modal corpora and interaction logs for multimodality and agency; and leak-resistant, representative test sets for evaluation. Together, these dependencies underscore the *continued centrality* of data as a design variable at the forefront of FM research. Meanwhile, longstanding challenges in data collection, curation, and synthesis remain unresolved, while concerns surrounding copyright, privacy, and fairness have only intensified. Building on the success of the first two DATA-FM workshops at ICLR 2024 and 2025, the third edition will revisit these persistent issues while highlighting emerging ones at the frontiers of post-training, multimodality, and evaluation. By convening researchers and practitioners across diverse research communities, DATA-FM seeks to advance understanding of data's evolving role in FMs and foster innovative solutions shaping the next generation of models and applications.

1 Workshop Summary

Foundation models (FMs) [Bommasani et al., 2021] have become a cornerstone of modern machine learning, demonstrating broad capabilities across domains through large-scale pretraining on diverse datasets. Data has been the key enabler of this progress, driving both the scale and adaptability that characterize FMs. At the same time, data is also the source of many of the most pressing challenges: how to sustain scalability in pre-training, improve efficiency in post-training, and ensure reliability and alignment across increasingly multimodal and heterogeneous data sources. These challenges are amplified by the unprecedented scale and limited transparency of FM training data, making traditional approaches inadequate. Addressing them requires new methods and sustained collaboration across technical, legal, and societal domains.

Persistent Issues. The first two editions of DATA-FM highlighted a set of fundamental data challenges that remain unresolved and continue to shape the development of foundation models. **Reliable and principled data collection and curation** [Liu et al., 2024, Maini et al., 2025b] at scale remain difficult: ensuring coverage and diversity, balancing domains, and mitigating undesirable artifacts such as bias and toxicity are still open problems. **Data attribution** [Wang et al., 2025, Deng et al., 2025] also remains pressing, as existing approaches face significant scalability challenges when applied to models and datasets of the size typical for FMs. **Privacy protection and copyright compliance** [Nasr et al., 2025, He et al., 2025, Liu et al., 2025, Henderson et al., 2023] have become even more urgent with ongoing legal cases and regulatory debates underscoring the tension between large-scale data use and individual or institutional rights. **Reproducibility** [Hardt, 2025, He and Lab, 2025] in data-centric FM research remains challenging, as outcomes are often sensitive to subtle preprocessing decisions and constrained by disparities in access to computational resources. Addressing these persistent challenges is critical for advancing both scientific understanding and responsible deployment of FMs.

Emergent Challenges. Over the past year, the research frontier of FMs has expanded, bringing new data challenges to the forefront. On the pre-training side, the concern on exhaustion of available web-scale corpora is driving the community to explore new paradigms [Maini et al., 2025a, Gao et al., 2025, Yang et al., 2025]. Advances in post-training, such as reinforcement learning with verifiable rewards (RLVR), are pushing the boundaries of model reasoning but also introduce significant data efficiency hurdles under the high inference costs of algorithms like GRPO [Guo et al., 2025]. Simultaneously, breakthroughs in multimodal learning, exemplified by Nano-Banana [Google Gemini Team, 2025], underscore an urgent need for principled strategies to construct and curate datasets that span diverse data modalities. Unlike text-only settings, integrating and balancing heterogeneous data sources (e.g., text, image, audio,

and interaction logs) introduces new challenges in alignment, scaling, and representation. As these capabilities grow, the science of evaluation itself has become a critical data problem; large-scale benchmarks like Humanity's Last Exam [Phan et al., 2025] reveal the limitations of current design, raising crucial questions about transferability. Finally, emerging work on pluralistic alignment [Sorensen et al., 2024] shows how data directly encodes competing objectives and values, shaping how FMs balance reliability and robustness. Together, these developments motivate the further expanded scope of the third DATA-FM workshop.

- **New paradigms of pre-training.** With web-scale corpora approaching saturation, sustaining progress in foundation model development requires rethinking how data is utilized and sourced. Two central questions arise: (1) **how to maximize information gain from existing data**, ensuring that each token contributes meaningfully and supports efficient scaling [Chang et al., 2024]; and (2) **where to obtain and how to integrate the next generation of data.** Addressing the latter calls for expanding beyond web-crawled text toward licensed and domain-specific corpora (e.g., scientific, legal, multilingual, or low-resource), consented interaction and telemetry logs, and instrumented environments with clear provenance and governance [Penedo et al., 2024, Cargnelutti et al., 2025]. Another frontier lies in **synthetic data for pre-training** [Maini et al., 2025a, Kang et al., 2025a]: understanding when and how to generate, select, and schedule synthetic examples; maintaining quality and diversity; mitigating model collapse; and effectively combining synthetic and real data. Tackling these challenges is essential for developing robust and data-efficient pre-training pipelines beyond the era of web-scale corpora.
- **Data efficiency in post-training.** As emerging post-training paradigms such as RLVR push foundation models toward more complex reasoning, their high computational cost makes data efficiency a critical bottleneck [Yu et al., 2025]. Three central questions follow: (1) how to select and organize supervised fine-tuning (SFT) data to optimize end-to-end post-training performance, given the complex interaction between SFT and RL stages [Kang et al., 2025b]; (2) how to prepare the most informative prompt mixtures offline; and (3) how to design effective selection and filtering mechanisms during online RL training to prioritize valuable interactions and avoid wasted computation [Hu et al., 2025]. Addressing these challenges is essential to making RL-based post-training both practical and scalable.
- **Multimodal data design and curation.** Recent breakthroughs in multimodal models (e.g., Genie-3 [Parker-Holder and Fruchter, 2025], Nano-Banana [Google Gemini Team, 2025]) highlight that progress depends not only on scaling data, but on developing a science of multimodal curation. Compared to purely textual domains such as math or coding, multimodal data is generally more difficult to source and validate. Key challenges include identifying appropriate visual representations to capture task-relevant information, balancing quality against web-scale quantity, correcting modality imbalances through mixture design, handling the inherent multiplicity of many-to-many cross-modal correspondences [Chun, 2025], and building benchmarks that probe deep, cross-modal reasoning.
- **Evaluation and Small-to-large Transferability.** When developing data curation pipelines, practitioners routinely rely on small-scale model experiments to make decisions. Yet the community has a limited understanding of whether conclusions from these small experiments reliably transfer to large-scale production training. This transferability problem is compounded by the subtle yet important issue in our definition of "high-quality dataset". In most data-centric studies, evaluation is performed by training separate models on datasets curated by different algorithms and then comparing their performance. This approach implicitly assumes that data quality can be assessed under a fixed training configuration. In practice, however, data, optimizers, and architectures interact in complex ways, making evaluations based on any single configuration unreliable. These challenges raise two fundamental questions for industry data practice: How can we reliably define and assess "data quality" through model training runs? And under what conditions do conclusions from small-scale experiments carry over to larger-scale training?
- **Emergent alignment questions.** Active research areas like Pluralistic alignment [Sorensen et al., 2024], model personalization [Ryan et al., 2025], and deliberative alignment [Guan et al., 2024] all depend fundamentally on data. Effective training requires responses or action trajectories that adapt to varied user profiles and latent preferences. Diverse datasets that capture different goals, styles, and priorities are essential for teaching models to account for user-level differences while also learning conflict-resolution strategies across instruction levels (e.g., prioritizing system messages over user messages [Wallace et al., 2025]).

Goals and Impact. The third DATA-FM workshop will convene researchers and practitioners spanning a wide range of communities, from CV, NLP, and robotics to specialists in pre- and post-training, as well as experts in law, governance, and applied domains. Through invited talks, contributed papers, spotlight talks, posters, and panels, the workshop will:

- Foster a comprehensive understanding of both persistent and emerging data challenges in FMs.
- Create a platform for interdisciplinary exchange and collaboration across technical and societal domains.
- Catalyze new research directions that address the evolving data problems shaping the next generation of foundation models.

Topics of Interest. We invite contributions on a broad range of topics, including but not limited to:

- Data curation: collection, cleaning, deduplication, selection, and mixture optimization. We especially encourage submissions targeting post-training and multimodal applications
- Data attribution, provenance, and valuation
- Data marketplaces and emerging economic models for data exchange
- Data scarcity, discovery, and sourcing strategies
- Synthetic data generation: quality assessment, diversity, and mitigation of model collapse
- Principled methodology for model evaluation and benchmark design
- Leveraging small-scale experiments to guide large-scale training, including scaling laws and scale-invariant techniques such as μP
- Data-centric approaches to alignment and AI safety
- Responsible data practices: privacy, security, copyright, and fairness
- Legal, regulatory, and governance frameworks for data in foundation models

2 Tentative Schedule and Plans

The workshop will prioritize in-person participation with online access for remote attendees. We have planned for six 30-min invited talks (22-min talk + 8-min Q&A), four 10-min spotlight talks selected from regular/position papers track submissions (7-min talk + 3-min Q&A), two 5-min spotlight talks selected from Tiny papers track submissions, two 75-min poster sessions, and a 30-min panel discussion.

2.1 Tentative Schedule

We provide a tentative schedule as follows, **with around 3.5 hours allocated for open discussions or networking** (poster sessions, coffee/lunch breaks, and panel discussions). As a core objective of the workshop is to provide a platform for exchanging insights, sharing ideas, and fostering collaboration, our schedule switches between talks and open discussions (poster sessions, panel discussion, coffee breaks) several times throughout the event. This arrangement allows the audience to easily connect with speakers following their talks, encouraging open discussion. It also facilitates cross-participation between different workshops and helps connect with a broader audience.

Morning sessions:

- 9:00–9:05 Opening Remarks
- 9:05–9:35 Invited Talk 1 (Maria De-Arteaga)
- 9:35–10:05 Invited Talk 2 (Kelvin Guu)
- 10:05–11:20 Poster Session 1
- 11:20–11:30 Coffee Break

- 11:30–12:00 Invited Talk 3 (Hanna Hajishirzi)
- 12:00–12:10 Spotlight Presentation 1
- 12:10–12:20 Spotlight Presentation 2
- 12:20–12:30 Spotlight Presentation (Tiny Papers)
- 12:30–1:30 Lunch Break

Afternoon sessions:

- 1:30–2:00 Invited Talk 4 (Junyang Lin)
- 2:00–2:30 Invited Talk 5 (Baharan Mirzasoleiman)
- 2:30–2:40 Spotlight Presentation 3
- 2:40–2:50 Spotlight Presentation 4
- 2:50–4:05 Poster Session 2
- 4:05–4:35 Invited Talk 6 (Sébastien Bubeck)
- 4:35–5:05 Panel Discussion (Moderator: Matthew Jagielski)
- 5:05–5:10 Closing Remarks

Our workshop will conclude around **5:10 PM** to maintain a manageable schedule for attendees. Prior to the workshop, we will put talk and poster titles/abstracts on the workshop website to allow for choice of attendance based on content.

2.2 Highlights of Our Plans

Invited Talks & Spotlight Presentations. **(1) Remote participant accommodation:** To accommodate online participants, all invited talks and spotlight presentations will be livestreamed, with questions from online participants facilitated through moderated discussions. **(2) Extended Q&A session:** Following last year's practice, we will reserve 8-minute Q&A sessions after each invited talk to encourage deeper interaction between speakers and the audience. **(3) Selection of spotlight presentations:** We aim to provide presentation opportunities for high-quality submissions to showcase the latest advancements in the field. Spotlight presentations will be selected based on reviewer nominations and discussions within the organizing committee, with conflicts of interest being carefully managed. **(4) Spotlight presentations from the Tiny Papers track:** Our workshop will continue the Tiny Papers Track (see Section 2.3), designed to increase the visibility of research by underrepresented, under-resourced, and early-career researchers. We will select two high-quality submissions from this track for a 5-min spotlight presentation during the workshop, presented alongside the spotlight presentations from the regular track.

Poster Sessions. **(1) In-person and accessibility considerations:** The poster sessions will be held in person to maximize engagement and interaction among participants. To ensure accessibility for those unable to attend on-site, we will invite authors to upload a PDF version of their poster and an optional short video presentation, which will be featured on the workshop website (see Section 2.3). **(2) Length:** Following last year's workshop, we will maintain 75-minute poster sessions to foster deeper discussions and stronger networking opportunities.

2.3 Paper Submission and Review

Regular/Position Papers Track. Our workshop welcomes research and position papers addressing data challenges for foundation models, with options for both long (10-page) and short (4-page) submissions in ICLR template. Discouraging submissions of published work: We discourage submissions of previously published work at major ML venues (e.g., ICLR, ICML, NeurIPS); the papers accepted to the ICLR main conference cannot present at our workshop.

Tiny Papers Track. Building on last year's practice, our workshop will again feature a Tiny Papers Track to support underrepresented, under-resourced, and early-career researchers who may not yet have the means to submit full papers. This track is intended for work at the early stages of a project: for example, a concise but self-contained theoretical result, a novel observation from preliminary experiments, or a fresh perspective on an existing problem. The goal is to foster early-stage ideas and provide a platform for researchers to receive constructive feedback and guidance as they develop their work further.

Optional Videos. We encourage the authors of accepted papers from both tracks to upload a pre-recorded video about their works, enabling remote attendees to access the content flexibly.

Paper Review & Conflict of Interests. The paper review process will follow a double-blind format, ensuring anonymity for both authors and reviewers. We will reach out to diverse reviewers to provide a broad range of expertise. Each submission will be reviewed by at least 3 reviewers, and decisions will be made transparently by the organizing committee. We leverage OpenReview profiles to prevent conflicts of interest. The final list of accepted papers will be published on the workshop website.

Tentative Schedule of Paper Submission. We will follow the suggested dates by ICLR.

- Workshop paper submission deadline: January 30, 2026
- Workshop paper notification date: February 27, 2026
- Camera-ready and (optional) posters and video recordings upload: April 3, 2026

3 Invited Speakers and Panelists

We have invited 6 speakers, listed below with their tentative talk themes, covering existing key data challenges for foundation models and offering diverse perspectives to enrich the discussion. **All invited speakers have expressed strong interest in the workshop, with five confirming their participation and one pending.**

✓ Maria De-Arteaga (Associate Professor, ESADE Business School): Data-centric approaches for trustworthy ML

✓ Kelvin Guu (Research Scientist Director, Google DeepMind): Data attribution for Gemini

✓ Hanna Hajishirzi (Associate Professor, University of Washington): Open data curation and benchmark development

✓ Junyang Lin (Principal Scientist, Alibaba Qwen): Data efforts for Qwen

✓ Baharan Mirzasoleiman (Assistant Professor, UCLA): Theoretical foundation of data-efficient learning

⌛ Sébastien Bubeck (Member of Technical Staff, OpenAI): Synthetic data

(✓: confirmed; ⌛: pending confirmation)

Our invited speakers represent a balanced mix of leading voices from industry and academia, reflecting the workshop's goal of bridging practical challenges and theoretical foundations in data-centric foundation model research. They include senior researchers from major industry labs (OpenAI, Google DeepMind, Alibaba), as well as faculty members across multiple career stages (Assistant to Associate Professors) from top universities (ESADE, University of Washington, and UCLA). A detailed discussion of speaker and topic diversity is presented in Section 7.

Panel Discussion. The workshop will feature a panel discussion focused on the current challenges of data-centric approaches for foundation models. We plan to invite Matthew Jagielski (Anthropic) as the panel moderator. All confirmed invited speakers are welcome to join as panelists, with the final list to be determined closer to the workshop date. To ensure a wide range of perspectives from diverse research communities, we may also invite additional experts working at the intersection of computer science and other disciplines, such as Catherine Brobston (Harvard) who leads the Institutional Data Initiative at Harvard Law School Library to broaden data access. We may also invite experts from specific application domains, such as Jun-Yan Zhu (CMU) and Saining Xie (NYU), who work on computer vision, and Dorsa Sadigh (Stanford), who works on robotics.

3.1 Biographies of Speakers and Panel Moderator

Maria De-Arteaga is an Associate Professor in the Department of Data, Analytics, Technology, and Artificial Intelligence at ESADE Business School. Previously, she was an Assistant Professor in the Information, Risk, and Operations Management Department at the University of Texas at Austin, where she was also a core faculty member of the Machine Learning Laboratory. She received a joint PhD in Machine Learning and Public Policy from Carnegie Mellon University. Her research examines the risks and opportunities of deploying machine learning for decision support in high-stakes settings. Her work has been recognized with the Best Thematic Paper Award at NAACL'19 and the Innovation Award on Data Science at Data for Policy'16, and featured by UN Women and Global Pulse in their report Gender Equality and Big Data: Making Gender Data Visible. She is also a recipient of a 2020 Google Award for Inclusion Research, a 2018 Microsoft Research Dissertation Grant, and was named an EECS Rising Star in 2019.

Kelvin Guu is a Research Scientist Director at Google DeepMind, where he leads research on natural language processing and machine learning. He has contributed to influential advances in language models, including retrieval-augmented models (REALM) and instruction-following models (FLAN). His recent work focuses on data attribution. Kelvin received his Ph.D. in Statistics from Stanford University, advised by Percy Liang, and a B.S. in Mathematics from Duke University.

Hanna Hajishirzi is the Torode Family Associate Professor in the Allen School at the University of Washington, where she leads the H2Lab, and a Senior Director of NLP at AI2. She received her Ph.D in Computer Science from University of Illinois at Urbana-Champaign, and spent a year as Postdoctoral associate at Disney Research and CMU. Hajishirzi's current research delves into various areas within NLP and AI, with a particular focus on understanding and pushing the boundaries of large language models. She has published more than 140 scientific articles in top-tier journals and conferences in ML, AI, NLP, and Computer Vision. She is a recipient of 2020 Alfred Sloan Fellowship, 2021 NSF CAREER award, 2019 Intel rising star award, 2018 Allen Distinguished Investigator award, 2023 Academic Achievement UIUC Alumni award, 2024 innovator of the year award finalist by GeekWire, and several research faculty awards from industry. The work from her lab has been nominated or received best paper awards at conferences and have been featured in a variety of magazines and newspapers including New York Times, Forbes, NPR, MIT Technology Review, Geekwire, Wired Magazine, and more.

Junyang Lin is a Principal Scientist at Alibaba Group and leads the Qwen Team. His research focuses on natural language processing and multimodal representation learning, with an emphasis on large-scale pretraining. His team has recently released and open-sourced the Qwen series, including the large language model Qwen, the vision-language model Qwen-VL, and the audio-language model Qwen-Audio. He has also contributed to the development of widely used open-source models such as OFA and Chinese-CLIP. His current goal is to build a multimodal AI system advancing toward a generalist agent.

Baharan Mirzasoleiman is an Assistant Professor in the Computer Science Department at UCLA, where she leads the BigML research group. She is also a Visiting Faculty Researcher at Google Research. Her research aims to address sustainability, reliability, and efficiency of machine learning. She is mainly working on improving the big data quality, by developing theoretically rigorous methods to select the most beneficial data for efficient and robust learning. Besides, she is also interested in improving the models and learning algorithms. The resulting methods are broadly applicable for learning from massive datasets across a wide range of applications, such as medical diagnosis and environment sensing. Baharan received the ETH medal for Outstanding Doctoral Thesis, were recognized as a Rising Star in EECS by MIT, and were awarded the NSF Career Award.

Sébastien Bubeck is a Member of Technical Staff at OpenAI. Before joining OpenAI, he served as Vice President of AI and a Distinguished Scientist at Microsoft. He was previously an Assistant Professor in the Department of Operations Research and Financial Engineering at Princeton University. His research has been recognized with numerous awards, including Best Paper Awards at NeurIPS 2018 and 2021, Best Paper Award at STOC 2023, and Best Student Paper Awards at ALT 2018 and 2023. He is also a recipient of the Sloan Research Fellowship in Computer Science and the Jacques Neveu Prize for the best French Ph.D. thesis in Probability and Statistics.

Matthew Jagielski is a Member of Technical Staff at Anthropic. His research centers on the privacy, security, and memorization properties of large-scale machine learning systems. Before joining Anthropic, he was a Research Scientist at Google DeepMind. He received his Ph.D. in Computer Science from Northeastern University, where he was advised by Alina Oprea and Cristina Nita-Rotaru. His work has been recognized with an Outstanding Paper

Award at NeurIPS 2023, a Best Paper Award at USENIX Security 2023, and a Runner-Up for the Caspar Bowden Award at PETS 2023.

4 Organizers and Program Committee

Members of the Organizing Team.

- Dr. Zheng Xu (Staff Research Scientist, Meta) [🎓]
- Prof. Ruoxi Jia (Assistant Professor, Virginia Tech) [🎓, 🌐]
- Prof. Martin Jaggi (Associate Professor, EPFL) [🎓, 🌐]
- Dr. Monica Ribero (Research Scientist, Google) [🎓, 🌐]
- Pratyush Maini (Founding Member, DatologyAI) [🎓, 🌐]
- Jiachen T. Wang (Ph.D. Student, Princeton University) [🎓, 🌐]
- Luxi He (Ph.D. Student, Princeton University) [🎓, 🌐]
- Yuzheng Hu (Ph.D. Student, UIUC) [🎓, 🌐]

Organizing Experiences. The organizing team combines researchers with extensive experience in running academic events and first-time workshop organizers. Among them, Dr. Monica Ribero, Luxi He and Yuzheng Hu are serving as workshop organizers for the first time.

Dr. Zheng Xu and Prof. Ruoxi Jia bring substantial organizing experience. Dr. Xu has organized the ICLR Synthetic Data and Data Access Workshop (2025), the KDD Workshop on Federated Learning for Data Mining and Graph Analytics (2025, 2024), the ICLR Privacy Regulation and Protection in Machine Learning Workshop (2024), the ICML Workshop on Federated Learning and Analytics in Practice (2023), the TTIC Workshop on New Frontiers in Federated Learning (2023), and the Google Federated Learning and Analytics Workshop (2020). Prof. Jia has organized the ICLR Workshops on Navigating and Addressing Data Problems for Foundation Models (2025, 2024), the AsiaCCS Workshop on Secure and Trustworthy Deep Learning Systems (2022), the ICML Workshop on Economics of Privacy and Data Labor (2020), the Workshop on AI for Energy-Cyber-Physical Systems (2018), and the Workshop on Smart Buildings as Enablers for a Smarter Grid (2016). In addition, Prof. Martin Jaggi served as Tutorial Chair for ICML (2024, 2023), Pratyush Maini organized the Workshop on Large Language Model Memorization at ACL (2025), and Jiachen T. Wang organized the ICLR Workshop on Navigating and Addressing Data Problems for Foundation Models (2025).

Overall, six of the eight organizers are new to the DATA-FM series in an organizing capacity, bringing fresh perspectives while complementing the continuity offered by returning organizers. We provide a detailed discussion of organizer diversity in Section 7.

4.1 Biographies

Zheng Xu is a staff research scientist working on GenAI privacy at Meta. Previously, he was a staff research scientist working on federated learning and privacy at Google. He earned his Ph.D. in optimization and machine learning from University of Maryland, College Park, in 2019. Before that, he got his master's and bachelor's degree from the University of Science and Technology of China. He has published 40+ papers at top research conferences and journals with 18K+ citations, and received two best student paper awards. He is a co-author of Advances and Open Problems in Federated Learning, and a lead author of A Field Guide to Federated Optimization, both of which resulted from 20+ collaborators in workshop discussions. He is a co-organizer of the Google Federated Learning and Analytics Workshop 2020, TTIC workshop on New Frontiers in Federated Learning 2023, KDD workshop on Federated Learning for Data Mining and Graph Analytics 2024 and 2025, and the lead organizer of Federated Learning and Analytics in Practice Workshop at ICML 2023, Privacy Regulation and Protection in Machine Learning Workshop at ICLR 2024, and Synthetic Data and Data Access Workshop at ICLR 2025.

Ruoxi Jia is an assistant professor in the Bradley Department of Electrical and Computer Engineering at Virginia Tech. Her research interests span machine learning, security, privacy, and cyber-physical systems, with a recent focus on data-centric and trustworthy AI. Her work has earned her several prestigious awards and fellowships, including

the NSF CAREER Award and the Best Social Impact Paper Award at ACL. Her research has been featured in prominent media outlets such as The New York Times, IEEE Spectrum, and MIT Technology Review. Her work on data valuation and selection has been adopted by companies in the financial sector and tech industry.

Martin Jaggi is an Associate Professor at EPFL, heading the Machine Learning and Optimization Laboratory. Before joining EPFL, he was a post-doctoral researcher at ETH Zurich, at the Simons Institute in Berkeley, and at École Polytechnique in Paris. He has earned his PhD in Machine Learning and Optimization from ETH Zurich in 2011, and a MSc in Mathematics also from ETH Zurich. He is a co-founder of EPFL's Applied Machine Learning Days, and a Fellow of the European ELLIS network.

Monica Ribero is a Research Scientist at Google NYC specializing in data privacy. Her work centers on developing algorithms to bound sensitive data exposure, audit privacy-preserving systems, and create interpretable measures of risk. She earned her PhD in Electrical and Computer Engineering from the University of Texas at Austin, where her research with Dr. Haris Vikalo explored private machine learning through federated learning techniques and differential privacy. Before that, she received a BS in Mathematics from Universidad de Los Andes in Bogotá, Colombia.

Pratyush Maini is a founding member of DatologyAI and a Ph.D. candidate in the Machine Learning Department at Carnegie Mellon University. In his work, he has developed scalable and performant methods for improving the quality of data that we train machine learning models on. He has also developed methods that allow us to evaluate, locate, and mitigate the memorization of data points by neural networks. His works have been recognized through a best paper award nomination at NeurIPS, and multiple oral and spotlight talks at major ML conferences.

Jiachen T. Wang is a final-year Ph.D. student at Princeton University, where he is advised by Prof. Prateek Mittal and collaborates closely with Prof. Ruoxi Jia at Virginia Tech. His research focuses on developing theoretical foundations and practical tools for trustworthy machine learning from a data-centric perspective. His contributions have been recognized through 9 oral/spotlight presentations at top AI/ML venues. He received ICLR Outstanding Paper Honorable Mention in 2025, Apple PhD Fellowship in 2025, and Rising Star in Data Science in 2024.

Luxi He is a Ph.D. student at Princeton Language and Intelligence (PLI), co-advised by Prof. Danqi Chen and Prof. Peter Henderson. She is interested in studying the impact of data on the language model life cycle, with recent works including investigations into the impact of benign fine-tuning data on model safety and metadata conditioning for accelerating language model pretraining. She is also broadly interested in AI alignment. Her research has received spotlight and oral recognitions at major conferences. Luxi co-leads the Princeton Alignment and Safety Seminar (PASS). Before Princeton, she graduated from Harvard College Magna Cum Laude with highest honors in computer science and math, and holds a concurrent Master's in applied math.

Yuzheng Hu is a final-year Ph.D. student in the Siebel School of Computing and Data Science at the University of Illinois Urbana-Champaign. He is interested in understanding the role of data in generative AI and improving its use, with a focus on data attribution, differential privacy, and synthetic data. He received his bachelor's degree in math from Peking University and has spent time at Alibaba, Jane Street, Simons Institute, and Google Research.

5 Anticipated Audience Size

Based on the increasing participation and successful outcome through the first to the second edition of DATA-FM workshop at ICLR 2024 and 2025, and given the growing research interests in the intersection of foundation models and data-centric machine learning, we expect the following audience size:

1. **Number of attendees**: We anticipate around 75 attendees in the room at all times and more than 1000 audiences in total throughout the event day.
2. **Number of submissions**: We anticipate approximately 150 paper submissions based on the increasing trend observed in the last two editions of the workshop.

6 Plan to Get an Audience for the Workshop

To ensure the success of the workshop and maximize engagement with potential audiences, the organizing team plans to dedicate substantial effort to promoting this event. We consider the following approaches:

1. Create a comprehensive workshop webpage with all relevant information, including the schedule, speakers, submission guidelines, and updates.
2. Advertise the workshop in the upcoming machine learning conferences (e.g., NeurIPS 2025) to increase visibility to relevant audiences.
3. Send email announcements and calls for papers to relevant mailing lists in both academia and industry to ensure broad awareness.
4. Share information on social media platforms such as X and LinkedIn, as well as advertise on relevant Slack workspaces, to maximize reach.
5. Pay special attention to mailing lists related to minority and underrepresented groups in AI, ensuring inclusivity and broader representation. The organizers' home institutions can help to reach a diverse set of participants and encourage attendance from different disciplines.
6. Encourage accepted paper authors to promote their participation and attract their own networks to the workshop.
7. Prior to the workshop, we will put talk and poster titles up publicly to allow for choice of attendance based on content.

7 Diversity Commitment

We are committed to fostering diversity in both our organizing team and invited speakers, encompassing gender, affiliation, geographic location, career stage, expertise, and cultural background.

Diversity in the Organizing Team. Our eight-member organizing team reflects diversity across gender, geography, culture, sector, career stage, and research expertise. Three organizers are women (Luxi, Ruoxi, and Monica), contributing to gender balance and mentorship representation. Prof. Jaggi is based in Europe while Monica has a Latin American background, contributing to the team's geographic and cultural diversity. The organizers represent academia, large technology companies, and startups: four Ph.D. students (CMU, Princeton, and UIUC, including one also representing DatologyAI), two faculty members (EPFL and Virginia Tech), and two industry research scientists (Google and Meta). The team also balances early-career researchers with senior leaders in the field, fostering mentorship, creativity, and cross-sector collaboration that benefit both workshop participants and the broader community. Collectively, their expertise covers a broad spectrum of data-related research areas, such as AI safety and alignment, privacy, data attribution, data curation, and synthetic data.

Diversity in Invited Speakers and Discussion Topics. The six invited speakers reflect diversity across gender, geography, culture, sector, and research focus. Three speakers are women (Maria, Hanna, and Baharan), contributing to a gender-balanced and visible representation. Geographically, the lineup spans multiple continents, including Asia (Junyang Lin), Europe (Prof. De-Arteaga), and North America. Culturally, the speakers represent a wide range of backgrounds, including the Middle East (Prof. Mirzasoleiman), East Asia (Junyang Lin), Europe (Sébastien Bubeck), and Latin America (Prof. De-Arteaga), bringing regional perspectives that highlight the global nature of data-centric AI research. The speakers bridge industry and academia, featuring experts from frontier AI labs (OpenAI, Google DeepMind), the open-source community (Alibaba Qwen), and leading universities across North America and Europe (ESADE, UWashington, and UCLA). Together, they bring perspectives on key data problems in frontier model development (synthetic data, data attribution), open-source practices for models and data, benchmark development, theoretical foundations of data-efficient learning, and the societal and ethical dimensions of data in AI. This integration of technical and human-centered perspectives ensures inclusive and multifaceted discussions that advance a holistic understanding of data's role in shaping foundation models.

Representation from Latin America. Our organizing team and invited speakers include strong representation from Latin America, including Monica Ribero and Maria De-Arteaga, ensuring geographic diversity and highlighting perspectives from the region where ICLR 2026 will take place.

Registration Fee & Travel Grants for Junior Researchers, Local Participants, and Underrepresented Groups. To make our workshop more accessible, we aim to offer grants for registration fees and travel expenses to participants who may face financial barriers to attending. Priority will be given to junior researchers, local participants from Brazil and across Latin America, and historically underrepresented groups. We are seeking sponsorship from DatologyAI, Google, and Meta to support these initiatives and encourage broader participation.

Hybrid Meeting Format. Recognizing the challenges posed by varying time zones in a hybrid meeting format, we will incorporate a series of actions to ensure wide participation. If the accepted paper author cannot attend the workshop in-person, we will display posters on behalf of authors. We also encourage the authors to upload a pre-recorded video about their works and put them on the workshop website in advance, enabling other attendees to access the content flexibly. Live sessions (invited talks and panel discussion) will be lived streamed through Zoom.

8 Virtual Access to Workshop Materials and Outcomes

All relevant information will be regularly updated on a dedicated webpage.

Accepted Papers, Posters, and Optional Videos. Although the workshop is non-archival, we will use OpenReview for the double-blind review process and will host the accepted papers. The posters and videos will be uploaded to the workshop website (with authors' consent).

Slides and Recordings. With the consent of the speakers, we will provide workshop slides and recordings to all attendees after the event. We will also offer a summarization of relevant code repositories, datasets, and additional reading materials. We will host a YouTube channel to publish the workshop recordings as well as the optional videos submitted by paper authors (with their consent). This will help disseminate the content to a broader audience.

After the workshop, we will write a position paper summarizing the new insights and future directions discussed during the presentations and panel discussions. This paper will serve as a valuable reference for both attendees and the broader research community.

9 Related Workshops

This is the third edition of the DATA-FM workshop, following our first and second events at ICLR 2024 and 2025. In this new iteration, we focus on both persistent and emerging data challenges during the past year. Research on data for foundation models is still in its early stages, and DATA-FM distinguishes itself from existing workshops and series in several key aspects.

Workshops on Data-Centric ML. Recent years have witnessed the successful launch and growing popularity of workshops in data-centric ML such as the DMLR workshop series (ICML'23, ICLR'24, ICML'24), the DCAI workshop (NeurIPS'21), and DataPerf (ICML'22). While these workshops engage the broader field of data-centric ML, the DATA-FM workshop places a dedicated emphasis on the pressing and increasingly important challenges specific to data problems in *foundation models*. Given the unique difficulties and rising significance of foundation model research, a dedicated workshop is essential for reshaping the research landscape and fostering focused discussions.

Workshops on Data Attribution and Curation. The Attributing Model Behavior at Scale workshops (NeurIPS'23, NeurIPS'24) primarily explored efficient data attribution techniques for large-scale models and datasets. Complementary to this, the new DataWorld workshop on Data Curation at ICML 2025 focuses on frameworks for systematic data curation across domains. Attribution and curation are closely connected: attribution methods reveal which data points most strongly shape model behavior, and these insights can in turn guide curation strategies by identifying what data should be included, refined, or excluded. Together, they form a feedback loop that supports more principled dataset design and improvement. The DATA-FM workshop aims to bridge these topics with other critical issues such as copyright and governance, fostering a more holistic understanding of data challenges in foundation models.

Workshops on Foundation Models. The past two years have also seen a surge of workshops dedicated to foundation models, such as Long-Context Foundation Models (ICML'24, ICML'25), Reliable and Responsible Foundation Models (ICLR'24, ICML'25), and the Theoretical Foundations of Foundation Models (TF2M) (ICML'24). While these workshops emphasize different aspects of FM development, DATA-FM focuses specifically on the data layer, offering a platform to tackle data-centric challenges that cut across training, post-training, evaluation, and deployment.

Research focusing on data offers promising opportunities to address some of the key challenges in foundation models and their deployment. We are committed to supporting research on data-related challenges in the FM era and aim to continue this workshop series, as interest and usage in this area are expected to grow in the coming years.

References

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL https://crfm.stanford.edu/assets/report.pdf.

Matteo Cargnelutti, Catherine Brobston, John Hess, Jack Cushman, Kristi Mukk, Aristana Scourtas, Kyle Courtney, Greg Leppert, Amanda Watson, Martha Whitehead, et al. Institutional books 1.0: A 242b token dataset from harvard library's collections, refined for accuracy and usability. *arXiv preprint arXiv:2506.08300*, 2025.

Tyler A Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. Scalable influence and fact tracing for large language model pretraining. *arXiv preprint arXiv:2410.17413*, 2024.

Sanghyuk Chun. Multiplicity is an inevitable and inherent challenge in multimodal learning. *arXiv preprint arXiv:2505.19614*, 2025.

Junwei Deng, Yuzheng Hu, Pingbang Hu, Ting-Wei Li, Shixuan Liu, Jiachen T. Wang, Dan Ley, Qirun Dai, Benhao Huang, Jin Huang, Cathy Jiao, Hoang Anh Just, Yijun Pan, Jingyan Shen, Yiwen Tu, Weiyi Wang, Xinhe Wang, Shichang Zhang, Shiyuan Zhang, Ruoxi Jia, Himabindu Lakkaraju, Hao Peng, Weijing Tang, Chenyan Xiong, Jieyu Zhao, Hanghang Tong, Han Zhao, and Jiaqi W. Ma. A Survey of Data Attribution: Methods, Applications, and Evaluation in the Era of Generative AI. working paper or preprint, August 2025. URL https://hal.science/hal-05230469.

Tianyu Gao, Alexander Wettig, Luxi He, Yihe Dong, Sadhika Malladi, and Danqi Chen. Metadata conditioning accelerates language model pre-training. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DdMMzlI5YE.

Google Gemini Team. Nano banana: Image editing in gemini just got a major upgrade. https://gemini.google/overview/image-generation/, 2025. Accessed: October 11, 2025.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Moritz Hardt. The emerging science of machine learning benchmarks. *Manuscript. https://mlbenchmarks. org*, 2025.

Horace He and Thinking Machines Lab. Defeating nondeterminism in llm inference. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250910. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Luxi He, Yangsibo Huang, Weijia Shi, Tinghao Xie, Haotian Liu, Yue Wang, Luke Zettlemoyer, Chiyuan Zhang, Danqi Chen, and Peter Henderson. Fantastic copyrighted beasts and how (not) to generate them. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ftHNJmogT1.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79, 2023.

Yuzheng Hu, Fan Wu, Haotian Ye, David Forsyth, James Zou, Nan Jiang, Jiaqi W Ma, and Han Zhao. A snapshot of influence: A local data attribution framework for online reinforcement learning. *arXiv preprint arXiv:2505.19281*, 2025.

Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sengupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. Demystifying synthetic data in llm pre-training: A systematic study of scaling laws, benefits, and pitfalls. *arXiv preprint arXiv:2510.01631*, 2025a.

Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. Quagmires in sft-rl post-training: When high sft scores mislead and what to use instead. *arXiv preprint arXiv:2510.01624*, 2025b.

Ken Liu, Christopher A. Choquette-Choo, Matthew Jagielski, Peter Kairouz, Sanmi Koyejo, Percy Liang, and Nicolas Papernot. Language models may verbatim complete text they were not explicitly trained on. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=bLcXkIasck.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. Datasets for large language models: A comprehensive survey. *arXiv preprint arXiv:2402.18041*, 2024.

Pratyush Maini, Vineeth Dorna, Parth Doshi, Aldo Carranza, Fan Pan, Jack Urbanek, Paul Burstein, Alex Fang, Alvin Deng, Amro Abbas, et al. Beyondweb: Lessons from scaling synthetic data for trillion-scale pretraining. *arXiv preprint arXiv:2508.10975*, 2025a.

Pratyush Maini, Sachin Goyal, Dylan Sam, Alex Robey, Yash Savani, Yiding Jiang, Andy Zou, Zacharcy C Lipton, and J Zico Kolter. Safety pretraining: Toward the next generation of safe ai. *arXiv preprint arXiv:2504.16980*, 2025b.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vjel3nWP2a.

Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/, August 2025. Accessed: October 11, 2025.

Guilherme Penedo, Hynek Kydlícek, Loubna Ben Allal, and Thomas Wolf. Fineweb: decanting the web for the finest text data at scale. *HuggingFace. Accessed: Jul*, 12, 2024.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Michael J. Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. SynthesizeMe! inducing persona-guided prompts for personalized reward models in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina

Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8045–8078, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.397. URL https://aclanthology.org/2025.acl-long.397/.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=gQpBnRHwxM.

Eric Wallace, Kai Yuanqing Xiao, Reimar Heinrich Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions, 2025. URL https://openreview.net/forum?id=vf5M8YaGPY.

Jiachen T. Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data shapley in one training run. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HD6bWcj87Y.

Zitong Yang, Aonan Zhang, Hong Liu, Tatsunori Hashimoto, Emmanuel Candès, Chong Wang, and Ruoming Pang. Synthetic bootstrapped pretraining. *arXiv preprint arXiv:2509.15248*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.