Understanding the Dataset Practitioners Behind Large Language Model Development

Crystal Qian*
cjqian@google.com
Google Research
New York City, NY, USA

Emily Reif*
ereif@google.com
Google Research
Seattle, WA, USA

Minsuk Kahng
kahng@google.com
Google Research
Atlanta, GA, USA

ABSTRACT

As large language models (LLMs) become more advanced and impactful, it is increasingly important to scrutinize the data that they rely upon and produce. What is it to be a dataset practitioner doing this work? We approach this in two parts: first, we define the role of "dataset practitioners" by performing a retrospective analysis on the responsibilities of teams contributing to LLM development at a technology company, Google. Then, we conduct semi-structured interviews with a cross-section of these practitioners (N=10). We find that although data quality is a top priority, there is little consensus around what data quality is and how to evaluate it. Consequently, practitioners either rely on their own intuition or write custom code to evaluate their data. We discuss potential reasons for this phenomenon and opportunities for alignment.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; *HCI theory, concepts and models.*

1 INTRODUCTION

As the state-of-the-art for large language models (LLMs) advances [40, 47], the field of relevant data analysis is rapidly evolving. Because the data used and produced by LLMs is largely unstructured, traditional statistical analyses are insufficient for rigorous evaluation [11, 45, 49]. Furthermore, as applications of these LLMs become more widely adopted and impactful [1, 7], there is a deeper need to qualitatively understand these datasets; for instance, to mitigate sociological biases, ensure safe outputs, and minimize harm.

We aim to identify the needs and challenges of those who want to understand unstructured, text-based datasets for LLM development: a group that we define as **dataset practitioners**. To develop this definition, we perform a retrospective analysis within Google, a

---

*Both authors contributed equally to this research.

technology company that is developing LLMs. We then conduct semi-structured interviews with a cross-section of practitioners (N=10) to better understand their workflows, tools, and challenges.

We find that practitioners increasingly prioritize data quality; however, there is no consensus on what constitutes "high quality" data. Despite the HCI and visualization researchers' active efforts to deliver relevant sensemaking methods and tools, data practitioners in aggregate do not appear to be adopting these solutions, instead relying either on cursory visual inspection of spreadsheets or custom analyses logic in notebooks to understand their data. There is demand for frameworks, consensus, and tooling in this space that is not being met. We discuss hypotheses for this observed phenomenon, and conclude with opportunities for further research and alignment.

2 RELATED WORK

2.1 Analyzing Analyzers

As data science has grown as a discipline, so have the amount of analyses [15, 23], surveys [53], and interviews [51] performed to capture the role of those who do this work.

Some notable highlights include Kandel et al. [27], which classifies the emerging role of the *data analysts* across different industrial sectors, such as healthcare and retail. Muller et al. [38] interviewed *data scientists* at IBM to capture different approaches to their work, and Crisan et al. [15] creates a taxonomy of job roles across *data workers*, such as *moonlighters, generalists* or *evangelists.*

Across these studies, the definitions of *data analysts* or *data workers* satisfy the breadth of work that we aim to capture in this inquiry. *Data scientist* is too narrow for our population. It does not encompass the specific challenges introduced by the new LLM-centered data regime, such as a rising need for qualitative evaluation methods or the broader range of job responsibilities within this role. These broader responsibilities might include, for example, creating new architecture to interpret data, or developing adjudication methods for human-labelled data.

2.2 Techniques and Tools

There have also been existing inquiries into the techniques and tools that practitioners use. Many data science workers interact with data in tabular formats, using tools such as Google Sheets or Microsoft Excel [9]. They may also writing code to perform custom analyses, commonly by using Python scripts or notebooks such as Google Colab or Jupyter [13, 29, 30, 46].

As large language models have become more salient, the space of applicable techniques and tools has increased. The field of explainable AI (XAI) [16, 17] has yielded new explainability [31, 44] and visualization techniques for natural language processing. These techniques can be packaged into frameworks and tools [2, 6, 27], such as Language Interpretability Tool [48], What-If Tool [52], and AI Fairness 360 [8] among many others [3, 5, 25, 32, 37]. However, these LLM-focused tools are relatively recent, and there is a lack of existing research assessing the extent of their adoption across industry and academia.

2.3 Curation Trends

Datasets relevant to LLM development have become increasingly composed of smaller, curated subsets that target address specific concerns, such as safety and fairness [35, 50]. The focus is increasingly on data quality [45] rather than quantity [19], though quantifying the criteria for data quality is an open problem [18, 24].

3 RETROSPECTIVE ANALYSIS

To define the role of *data practitioner*, we conducted a retrospective analysis of teams working on developing LLMs at Google. This company's organizational structure is uniquely positioned to support a broad survey of the landscape as the technology stack is vertically-integrated [22]; that is, the relevant tooling, infrastructure, modeling, evaluation, and research are primary developed in-house. For example, Google has infrastructure teams that build custom software to deploy ML experiments on computational resources, tooling teams that create applications for interpreting model outputs, data teams that source and clean human data, modeling teams that improve LLM models across different modalities, and safety teams that focus on enforcing policies and model quality.

Using company-internal organizational charts and employee directories, we identified projects associated with the development of the company's core LLMs. We also conducted a meta-review of company-internal user studies around evaluating tools for data exploration. Applying a grounded theory methodology [14], we inductively applied a relational content analysis and synthesized common themes to develop a framework around data practitioning.

3.1 Defining the Dataset Practitioner

The *dataset practitioner* interacts with unstructured, text-based data for the purpose of developing large language models. The practitioner's day-to-day work can cover a broad range of tasks traditionally defined in roles such as software developer, machine learning engineer, data scientist, research scientist, product manager, or product counsel. The practitioner may prioritize these responsibilities concurrently, or switch gears along the model development lifecycle. They may do any of the following representative tasks:

- Curating a new dataset from scratch
- Creating a new benchmark dataset
- Cleaning a dataset by removing or fixing bad examples
- Analyzing a dataset (feedback, comments, etc) to find trends
- Understanding what bias issues might exist in the dataset
- Making a go/no go decision on whether to use a dataset to train a model
- Debugging a specific model error by finding relevant data

- Finding ways to improve models, try different datasets, and compare model results
- Identifying key metrics to define "quality" for a use case

Next, we give examples of datasets that they may explore. The term "dataset" traditionally implies static and well-curated data; we expand this notion to include *any* set of text examples, which may come from a variety of provenances (e.g. scraped, synthetically generated, curated by experts). We categorize these broadly:

(1) **Training datasets**
- **Pre-training data**: LLMs are pre-trained on huge amounts of data from webscrapes, books, and other giant corpora. The curation of these datasets is hugely impactful on the model's performance [34].
- **SFT and RLHF data**: Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) datasets are used to refine pre-trained LLMs [40, 47]. They are significantly smaller and more specialized than pre-training data, and can be used from an open-ended generation model to a specific use case—most notably, the chatbot interface that many productionized LLMs employ. LLMs can be fine-tuned for other specific products and use cases as well.

(2) **Datasets involved in model evaluation**
- **Benchmark evaluation data**: Benchmark datasets are created to test specific functionalities or behaviors of the model. One notable category of these are **safety benchmarks**, which test the model's ability to adhere to company policies and safety standards on concepts such as toxicity, hallucination, etc.
- **Model outputs**: Model outputs can be evaluated outside of the context of a specific benchmark. Side-by-side analysis of model outputs may be conducted against golden sets or outputs from a baseline model [26].
- **Outputs of in-context learning:** These are a specific subset of model outputs. In-context learning has allowed users to create new models with no golden data at all. These may then be evaluated by analyzing the outputs from multiple runs of a prompt.
- **Conversational data**: User interactions with LLM-based chatbots can be used to evaluate LLMs in the wild.

4 QUALITATIVE STUDY

4.1 Participants

Using our updated definition, we recruited ten dataset practitioners (N=10) within Google for our study.[1] We selected these participants with the criteria that their current work involves interacting with datasets for the purposes of developing large language models, and prioritized sampling participants from a variety of concentrations and backgrounds. These participants and their primary focus areas (tooling, modeling, or evaluation) are listed in *Table 1*. We validated our observation from *Section 3.1* that the domains of their work are fluid; participants who identified in one domain during our recruiting cycle demonstrated experience in many adjacent areas within the interview. For example, a practitioner formerly focused

---

[1]Note that Reif et al. [43] uses the same participant sample.

Domain	Participant ID	Focus Area
Tooling	T1	Tools for data annotation
	T2	Tools for data curation
	T3	Tools for data understanding
	T4	Pipeline infrastructure
Modeling	M1	Data curation.
	M2	Model architecture
	M3	Model refinement
Evaluation	R1	Robustness and abuse
	R2	Unsafe and sensitive content
	R3	Annotator ethnography

Table 1: Study participants and their current focus areas, grouped by domain.

on modeling shifted priorities to safety and fairness evaluation as their models became more scrutinized and regulated, and two tool-builders reported being driven to build tooling to address their own unmet needs in modeling.

4.2 Interview Protocol

Following recruitment and an informed consent process, we conducted semi-structured, one-on-one interviews with participants over video conferencing. Each 30-minute interview covered the following topics:

(1) *Understanding the use case:* Background, use case, product impact, research questions
(2) *Tools and techniques*: Awareness and usage of existing tools and pipelines, decision making, advantages and limitations, statistical and visual interpretability methods
(3) *User challenges*: Bottlenecks, unaddressed concerns

We curated the interview topics from prior contextual inquiries and protocols from similar research studies in defining data work [28, 33, 51]. By following a similar interview protocol, we hope to isolate the specific challenges faced in LLM-development.

We synthesized our findings through a thematic analysis [10]. Each interview was de-identified, transcribed, broken into excerpts, and coded. Thematic elements, behaviors, and representative quotes in this paper are saturated [4, 21], with a code repeated in at least three distinct transcriptions.

4.3 Findings

4.3.1 **Participants prioritize data quality.** Corroborating the prior work described in Section 2.3, we find that data quality—defining, finding, and identifying high-quality data—was unanimously the biggest user challenge and priority across all use cases (*Table 2, Challenges*).

> *Data, historically, has been around volume rather than quality.. we've had this big paradigm shift.* —T2

> "Quality is the big obstacle… [You need] a lot of high-quality data... there's no shortcut." —E1

Although data quality has always remained an important priority for data scientists, these concerns were addressable through tasks such as data cleaning [38] or feature engineering [15]. In the context of generative modeling, the evaluation metrics and consensus frameworks are less straightforward.

4.3.2 **However, practitioners rely largely on their own intuition to validate this data quality.** All participants reported that they would evaluate their data by scanning it visually in spreadsheet form; that is, they would look at a handful of examples.

> "I'll read the first 10 examples, and then maybe some in the middle." —E1

> "I eyeball data.. It's all my own intuition and kind of individually spot checking examples." —M2

Participants cited efficiency, customization, a short learning curve, and ease-of-sharing as reasons for their reliance on spreadsheets (*Table 2, Challenges*).[2] While these factors align with prior research on spreadsheet usage [9], the *ease-of-sharing* factor may particularly encourage practitioners to use spreadsheets for LLM development. Unlike the data analysts in Kandel et al. [27], who collaborated with "hacker"-types with scripting and coding proficiency, our participants reported needing to share data with a larger and more diverse set of stakeholders, such as directors and legal teams, to review high-stakes safety fine-tuning datasets.

4.3.3 **Or, practitioners will run custom analyses.** Seven of the nine participants mentioned also writing custom code in Python notebooks to explore their data, and in one instance even to train production models. Participants liked the customization of these notebooks [29], yet cited reliability, setup, efficiency, code management as pain points (*Table 2, Challenges*), validating results from other studies on Python notebook usage [13, 29, 30, 46].

The efficiency concerns around long-running computations in Python notebooks [13] may be further exacerbated as LLMs require more computational power; participants mentioned that "getting model servers up and running takes forever" (R1), "my queries [to LLM APIs] take a while" (E1), and they wished they had "infinite QPS (Queries Per Second) [for their LLM API]" (R2).

4.3.4 **Practitioners recognize the confirmation biases in their exploration practices.** The majority—if not all—of the data exploration is being done between visual inspection in spreadsheets and custom logic in Python notebooks, allowing the practitioner to look at whatever they would like. This degree of freedom exacerbates cognitive bias [12, 20, 24, 42]; for example, Miller et al. [36] mentions that "explainable AI uses only the researchers' intuition of what constitutes a 'good' explanation." Indeed, our participants admit to this confirmation bias in their practices:

> ""I eyeball that things make sense [in the data]." —M2

In fact, model developers reported that they did not look at training data unless their model outputs were surprising.

> "When the data is passed to the modeling side, we assume that the data team has fixed everything. Unless

---

[2]Interestingly and consistent with similar user studies, our participants emphasized that their reliance on visual inspection of spreadsheets were their *own* behaviors and not best practices. They suggested that other practitioners likely used more sophisticated tooling [41].

		Tooling				Modeling			Eval.		
		T1	T2	T3	T4	M1	M2	M3	E1	E2	E3
Processes	Data discovery/creation: generating, annotating, synthesizing, scraping				x	x		x	x	x	x
	Data curation/evaluation: resampling, identifying outliers, safety	x	x	x		x		x	x	x	x
	Model development: architecture, algorithms, tuning						x	x			
	Model evaluation: metrics, benchmarking, ranking, compliance			x			x	x		x	
	Infrastructure/tooling: data pipelines, libraries	x	x	x	x	x		x			
Tools	Spreadsheets: table-views of data, Google Sheets	x	x	x	x	x	x	x	x	x	x
	Python notebooks: web-based IDEs, Jupyter notebooks, Google Colab				x	x	x	x	x	x	x
	Scripting tools: Custom Python scripts, command-line binaries		x	x				x	x		
	UI-based tools: data exploration, model interpretability					x				x	
Challenges	Assessing data quality: summarization, visualization, insights	x	x	x	x	x	x	x	x	x	x
	Version control: snapshots, reproducibility	x	x	x				x	x	x	
	Efficiency: start-up time, speed, latency, QPS	x		x	x			x	x		
	Customization: configurability, flexibility				x	x	x				x
	Reliability: service up-time/flakiness, serving story	x		x				x	x		
	Integration across tools: import/export to common datatypes/services		x		x					x	x
	Integration across people: within-team, cross-team, permissions		x		x	x					
	Learning curve: intuitive UI, easy-to-understand functionality								x	x	x

Table 2: This matrix categorizes our findings (inspired by Kandel et al. [27]). An 'x' in the cell indicates that a participant mentioned this specific topic in their interview. Topics are grouped by *Processes, Tools,* and *Challenges*, and participant are grouped by their domain from *Table 1*. All participants mentioned interacting with spreadsheets and cited data quality as a challenge in their work.

we train and it doesn't look right, then we'll [look at the data] and give the data team that feedback."   —M3

4.3.5 ***Participants have not converged upon other tools.*** Apart from Google Sheets and Python notebooks like Colab, no other tools garnered consensus among practitioners. Some practitioners employed additional methods, such as running a binary for calculating safety and toxicity thresholds, kicking off a pipeline to automatically classify their data, and using an user interface to visualize embeddings. However, these practices were not prevalent in our sample.

"Everyone is using a different thing, and getting everyone on the same page is really difficult."   —M1

The lack of alignment in tooling presents an organization challenge. As training datasets are increasingly composed of smaller datasets to leverage the expertise of specific subteams, greater collaboration across groups is necessary. This can lead to increased friction in adopting new tools and exploration patterns [27], as stakeholders and collaborators must transition to new tooling simultaneously, or migrate in a manner that preserves data sharing capabilities.

"With the new generative data— Many people are contributing with many different lenses. In practice, these [subsets] get built by random teams, they get added and nobody really reviews it because you can't."   —T4

5 DISCUSSION

The reason why practitioners have not aligned on alternative tooling is not obvious. Practitioners across all domains recognize that there is a gap in the workflow:

"Not having an easy-to-use-tool is a major bottleneck… Every time [that I make changes to data], I have to write a custom colab to ingest the new fields."   —M2

"There are no helpful tools from a qualitative researcher's perspective. I jump between spreadsheets, a CSV file and a colab… The long story short is that we haven't really found a very useful tool for this."   —E3

"Right now, if you want to curate high-quality data, you go through [each point] manually as an expert, which is not scalable [for] thousands of examples."   —T2

Practitioners are aware of and have tried the existing tools in this space. They are aligned on the properties that they want out of this tool (*Table 2, Challenges*), and these requests are being communicated to tooling teams:

"The kinds of requests we tend to get nowadays are about larger-scale dataset management, like mixture building. When you have a big selection, reviewing 10,000 rows is not what you want to do … That is much more amenable to summary review."   —T1

In response, tooling teams are evaluating and building tools to address these requests [3, 48, 52]. So, why is there a lack of alignment? We discuss hypotheses posed by two different domains of practitioners.

5.0.1 ***The toolmakers' hypothesis: the world is new.*** When tool developers (T1-T4) described exploration workflows, they explained that there was a lack of alignment because the field is new:

> *"The pace is very frenetic right now.. tools are fast-changing…"* —T1

> *"There's been a big step function in the NLP world.. it just takes a while to figure out what tools people need and what all use cases."* —T2

Two observations from our interviews may support this claim. First, practitioners are using spreadsheets. Perhaps in the absence of a ground truth for unstructured data, practitioners prefer to rely on their own intuition. Similarly, without a definitive framework for qualitative data exploration, practitioners are sticking to the tools they know. Adopting new practices takes effort (see *Table 2, Challenges > Learning curve*), and spreadsheets have been tried-and-true from the previous state-of-the-art when visually spot-checking data and conducting statistical analyses were sufficient.

Second, our participants described a landscape where there was a lack of alignment [18, 20] across multiple topics such as objectives, metrics, and benchmarks, suggesting that the field and its principles are still emerging. The following are representative quotes from participants:

- **Data quality**:
  - T1, on LLM prompts: *"There's so many competing definitions of prompt quality…it's a research north star that happens to be a major product priority. How can we improve this extremely important data set?'*
  - M1, on training data: *"The quality of data is subjective; a lot of people disagree…one person thinks it's really high-quality data, but there's no objective."*
  - T3, on evaluation data: *"There's not a framework for evaluating [data].. in a perfect world, there is well-articulated behavior (tone, subject matter, objective results).."*
- **Metrics**:
  - M1: *"[Consider] search rankings… what makes for a good benchmark, how do we come to an agreement?"*
  - E1: *"If you're doing simple classification, it's easy to measure accuracy or precision or recall. But with generative models, evaluation is very subjective. Even the output of the model is subjective, so then, what's going into the model- it's really hard to say, is this better or worse?"*
- **Safety**:
  - T2: *"Think about safety data curation…people can't agree on criteria, let alone apply that criteria at scale."*
- **Communication**:
  - T3: *"What [data practitioners are] actually doing and what they communicate that they need are two very different things. What are they actually trying to do?"*

This lack of alignment is amplified as teams collaborate more closely [39, 53]. Even if one team in the development pipeline identifies their quality evaluation parameters, there needs to be further agreement at the inter-team level.

5.0.2 ***The model developers' hypothesis: there's no tool that works for my use case.*** Modeling and evaluation practitioners speculated that alignment was unlikely due to custom needs and requirements (*Section 3.1*).

> *"I think why [a spreadsheet is] so universal is that it's so basic.. you can customize it to give this affordance that other tools may not give you.. it's simple."* —E1

> *"We have tried so many [tools]. These tools are limiting is because they offer you exploration on only one aspect of [the data]…For me, they're too specific."* —M2

Interestingly, when asked about the custom requirements for their use cases, practitioners listed similar requirements, which suggest that there may be opportunities for shared methods and evaluation frameworks. Some of these requirements include:

- Summarizing salient features of a dataset and identifying the corresponding data slices (6 participants)
- Ensuring safety of outputs/respecting toxicity thresholds (4 participants)
- Evaluating numeric distributions on text/token length (3 participants)

It is likely the case that both the toolmakers' and model developers' hypotheses are true to some extent. There may be select opportunities for alignment as the field matures, and there are likely other problems that will require custom solutions. For example, there are specific tools being developed to address challenges that persist across datasets, such as safety and toxicity classification [8].

6 CONCLUSIONS AND FUTURE WORK

In this study, we aimed to identify the needs of those who are exploring unstructured, text-based datasets for the purpose of developing LLMs. To define this population of **dataset practitioners**, we conducted a retrospective analysis on teams working on LLM development. We then interviewed a broad cross-section of these practitioners to better understand their use cases and challenges.

Through our retrospective analysis, we found that the dataset practitioner takes on a fluid role that is not well-defined in current literature on data workers. We hope that our contribution of defining this population and their use cases will enable the HCI community to better assess and support their needs.

In our interviews, we found that data quality is unanimously the top priority, but quality is subjective. Further research should explore what data quality means in different contexts, and how the same data can be high-quality or low-quality depending on the situation and perspective. Clarifying subjectivity across conceptual frameworks, evaluations, and workflows in this domain remains a top priority, potentially achieved through standardizing metrics (e.g. toxicity, distributions of relevant safety features, data diversity) and evaluation criteria.

Two primary data exploration patterns emerge: visually inspecting data in spreadsheets, which lacks scalability, and crafting custom analyses in Python notebooks, which is high-effort. Both practices are susceptible to confirmation bias. However, the community has yet to reach a consensus on alternative best-practices to for data exploration, possibly due to the nascent nature of the field or the custom needs of the practitioners. There are opportunities to determine the specific areas where prioritizing either flexibility or specificity is most beneficial; these opportunities can be addressed by formalizing evaluation frameworks in the evolving landscape, and developing flexible tooling for custom analysis.

> *"There's a fundamental chicken and egg problem...there's no tooling so people don't use tooling so tooling doesn't develop."* —T2

REFERENCES

[1] Malak Abdullah, Alia Madain, and Yaser Jararweh. 2022. ChatGPT: Fundamentals, Applications and Social Impacts. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. 1–8. https://doi.org/10.1109/SNAMS58071.2022.10062688

[2] Namita Agarwal and Saikat Das. 2020. Interpretable Machine Learning Tools: A Survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, Canberra, Australia, 1528–1534. https://doi.org/10.1109/SSCI47803.2020.9308260

[3] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, 337–346. https://doi.org/10.1145/2702123.2702509

[4] Hikari Ando, Rosanna Cousins, and Carolyn Young. 2014. Achieving Saturation in Thematic Analysis: Development and Refinement of a Codebook,. *Comprehensive Psychology* 3 (2014), 03.CP.3.4. https://doi.org/10.2466/03.CP.3.4 arXiv:https://doi.org/10.2466/03.CP.3.4

[5] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena Glassman. 2023. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. arXiv:2309.09128 [cs.HC]

[6] Narges Ashtari, Ryan Mullins, Crystal Qian, James Wexler, Ian Tenney, and Mahima Pushkarna. 2023. From Discovery to Adoption: Understanding the ML Practitioners' Interpretability Journey. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (<conf-loc>, <city>Pittsburgh</city>, <state>PA</state>, <country>USA</country>, </conf-loc>) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2304–2325. https://doi.org/10.1145/3563657.3596046

[7] Maria Teresa Baldassarre, Danilo Caivano, Berenice Fernandez Nieto, Domenico Gigante, and Azzurra Ragone. 2023. The Social Impact of Generative AI: An Analysis on ChatGPT. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good* (Lisbon, Portugal) *(GoodIT '23)*. ACM, 363–373. https://doi.org/10.1145/3582515.3609555

[8] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv preprint arXiv:1810.01943* (2018). http://arxiv.org/abs/1810.01943

[9] David Birch, David Lyford-Smith, and Yike Guo. 2018. The Future of Spreadsheets in the Big Data Era. arXiv:1801.10231 [cs.CY]

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[12] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[13] Souti Chattopadhyay, Ishita Prasad, Austin Z. Henley, Anita Sarma, and Titus Barik. 2020. What's Wrong with Computational Notebooks? Pain Points, Needs, and Design Opportunities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, 1–12. https://doi.org/10.1145/3313831.3376729

[14] Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21.

[15] Anamaria Crisan, Brittany Fiore-Gartland, and Melanie Tory. 2021. Passing the Data Baton : A Retrospective Analysis on Data Science Work and Workers. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1860–1870. https://doi.org/10.1109/TVCG.2020.3030340

[16] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A Survey of the State of Explainable AI for Natural Language Processing. *CoRR* abs/2010.00711 (2020). arXiv:2010.00711 https://arxiv.org/abs/2010.00711

[17] Arun Das and Paul Rad. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *CoRR* abs/2006.11371 (2020). arXiv:2006.11371 https://arxiv.org/abs/2006.11371

[18] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [stat.ML]

[19] Hugh Durrant-Whyte. 2015. Data, Knowledge and Discovery: Machine Learning meets Natural Science. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. ACM, 7. https://doi.org/10.1145/2783258.2785467

[20] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, IEEE, Turin, Italy, 80–89.

[21] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods* 18, 1 (2006), 59–82. https://doi.org/10.1177/1525822X05279903 arXiv:https://doi.org/10.1177/1525822X05279903

[22] Kathryn Rudie Harrigan. 1985. Vertical integration and corporate strategy. *Academy of Management journal* 28, 2 (1985), 397–425.

[23] Harlan Harris, Sean Murphy, and Marck Vaisman. 2013. *Analyzing the analyzers: An introspective survey of data scientists and their work*. O'Reilly Media, Inc.

[24] Bernease Herman. 2019. The Promise and Peril of Human Evaluation for Model Interpretability. arXiv:1711.07414 [cs.AI]

[25] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics* 25, 8 (2018), 2674–2693. https://doi.org/10.1109/TVCG.2018.2843369

[26] Minsuk Kahng, Ian Tenney, Mahima Pushkarna, Michael Xieyang Liu, James Wexler, Emily Reif, Krystal Kallarackal, Minsuk Chang, Michael Terry, and Lucas Dixon. 2024. LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models. *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*.

[27] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Enterprise Data Analysis and Visualization: An Interview Study. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2917–2926. https://doi.org/10.1109/TVCG.2012.219

[28] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '20)*. ACM, 1–14. https://doi.org/10.1145/3313831.3376219

[29] Mary Beth Kery, Bonnie E. John, Patrick O'Flaherty, Amber Horvath, and Brad A. Myers. 2019. Towards Effective Foraging by Data Scientists to Find Past Analysis Choices. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, 1–13. https://doi.org/10.1145/3290605.3300322

[30] Mary Beth Kery, Marissa Radensky, Mahima Arya, Bonnie E. John, and Brad A. Myers. 2018. The Story in the Notebook: Exploratory Data Science using a Literate Programming Tool. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Montreal QC</city>, <country>Canada</country>, </conf-loc>) *(CHI '18)*. ACM, 1–11. https://doi.org/10.1145/3173574.3173748

[31] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. PMLR, 2668–2677. https://proceedings.mlr.press/v80/kim18d.html

[32] Biagio La Rosa, Graziano Blasilli, Romain Bourqui, David Auber, Giuseppe Santucci, Roberto Capobianco, Enrico Bertini, Romain Giot, and Marco Angelini. 2023. State of the art of visual analytics for explainable deep learning. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 319–355.

[33] Catherine Li, Talie Massachi, Jordan Eschler, and Jeff Huang. 2023. Understanding the Needs of Enterprise Users in Collaborative Python Notebooks: This paper examines enterprise user needs in collaborative Python notebooks through a dyadic interview study. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI EA '23)*. ACM, Article 402, 7 pages. https://doi.org/10.1145/3544549.3573843

[34] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2023. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. arXiv:2305.13169 [cs.CL]

[35] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6, Article 115 (2021), 35 pages. https://doi.org/10.1145/3457607

[36] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[37] Yao Ming, Huamin Qu, and Enrico Bertini. 2019. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 342–352. https://doi.org/10.1109/TVCG.2018.2864812

[38] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, 1–15. https://doi.org/10.1145/3290605.3300356

[39] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. 2022. Collaboration challenges in building ML-enabled systems: communication, documentation, engineering, and process. In *Proceedings of the 44th International Conference on Software Engineering* (Pittsburgh, Pennsylvania) *(ICSE '22)*. ACM, 413–425. https://doi.org/10.1145/3510003.3510209

[40] OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[41] James W. Pennebaker. 2011. The secret life of pronouns. *New Scientist* 211, 2828 (2011), 42–45. https://doi.org/10.1016/S0262-4079(11)62167-2

[42] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *Proceedings of international conference on intelligence analysis* 5 (2005), 2–4.

[43] Emily Reif, Crystal Qian, James Wexler, and Minsuk Kahng. 2024. Automatic Histograms: Leveraging Language Models for Text Dataset Exploration. *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*.

[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, 1135–1144. https://doi.org/10.1145/2939672.2939778

[45] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) *(CHI '21)*. ACM, Article 39, 15 pages. https://doi.org/10.1145/3411764.3445518

[46] Aurélien Tabard, Wendy E. Mackay, and Evelyn Eastmond. 2008. From individual to collaborative: the evolution of prism, a hybrid laboratory notebook. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (San Diego, CA, USA) *(CSCW '08)*. ACM, 569–578. https://doi.org/10.1145/1460563.1460653

[47] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[48] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. arXiv:2008.05122 [cs.CL]

[49] Krzysztof Wach, Cong Doanh Duong, Joanna Ejdys, Rūta Kazlauskaitė, Pawel Korzynski, Grzegorz Mazurek, Joanna Paliszkiewicz, and Ewa Ziemba. 2023. The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review* 11, 2 (2023), 7–30.

[50] Kiri Wagstaff. 2012. Machine Learning that Matters. arXiv:1206.4656 [cs.LG]

[51] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 211 (2019), 24 pages. https://doi.org/10.1145/3359313

[52] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 56–65. https://doi.org/10.1109/TVCG.2019.2934619

[53] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.