LLM Adoption in Data Curation Workflows: Industry Practices and Insights

Crystal Qian
Google DeepMind
New York, NY, USA
cjqian@google.com

Michael Xieyang Liu
Google DeepMind
Pittsburgh, PA, USA
lxieyang@google.com

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

Grady Simon*
Google DeepMind
San Francisco, CA, USA
grady.hsimon@gmail.com

Nada Hussein
Google DeepMind
Cambridge, MA, USA
nadahussein@google.com

Nathan Clement
Google DeepMind
London, United Kingdom
nclement@google.com

James Wexler
Google DeepMind
Cambridge, MA, USA
jwexler@google.com

Carrie J Cai
Google DeepMind
Mountain View, CA, USA
cjcai@google.com

Michael Terry
Google DeepMind
Cambridge, MA, USA
michaelterry@google.com

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

Abstract

As large language models (LLMs) grow more proficient at processing unstructured text data, they offer new opportunities to enhance data curation workflows. This paper presents findings from a user study involving 12 industry practitioners from various roles and organizations across a large technology company (N=12). The study examines their data curation workflows before and after LLM adoption, using two custom design probes that integrate LLMs into existing tools. Our study reveals a shift from heuristics-driven, bottom-up curation to insights-driven, top-down workflows supported by LLMs. To navigate increasingly complex data landscapes, practitioners supplement traditional subject-expert-created "golden datasets" with LLM-generated "silver" datasets and rigorously validated "super golden" datasets curated by diverse experts. This research highlights the transformative potential of LLMs in large-scale analysis of unstructured data and highlights opportunities for further tool development.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

ACM Reference Format:
Crystal Qian, Michael Xieyang Liu, Emily Reif, Grady Simon, Nada Hussein, Nathan Clement, James Wexler, Carrie J Cai, Michael Terry, and Minsuk Kahng. 2025. LLM Adoption in Data Curation Workflows: Industry Practices and Insights. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3706599.3719677

*Work done at Google.
†Corresponding author, now at Yonsei University.

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3719677

1 Introduction

As large language models (LLMs) continue to advance, their improved reasoning capabilities, enhanced summarization techniques, and growing context windows enable them to process and generate insights from complex and voluminous data more effectively than ever before [7, 25, 37–39, 46, 49]. These advancements present significant opportunities to improve data curation and analysis workflows, particularly for those working with unstructured, text-based datasets. At the same time, the complexity of such data has grown. Modern foundation models increasingly rely on text data throughout their pipelines, including pre-training, fine-tuning, human feedback, and evaluation [12, 38–40]. With data coming from increasingly diverse sources, such as LLM-generated content, *curating* it—ensuring its quality, coherence, and relevance through iterative refinement and evaluation—becomes even more critical and challenging, as reported by recent work [10, 14, 21, 23, 30, 48].

Motivated by these increasing capabilities and complexities, we contribute design probes and a user study (N=12) to explore the opportunities and challenges of adopting LLM-based workflows.[1]. Building on expert interviews that examined data practitioners' existing workflows [34], we incorporated feedback into developing two LLM-based design probes to address reported data curation challenges. Our study investigates how practitioners use these probes for data curation tasks – that is, actions performed on general, unstructured text-based datasets, such as user feedback, conversation logs, or customer survey responses, to generate insights. Key findings from our study include:

- **Emergence of multi-tiered dataset hierarchies**: The adoption of LLMs in data labeling tasks has introduced new dataset hierarchies. Practitioners supplement traditional *"golden datasets"*—high-quality datasets for model training and evaluation—with *"silver datasets,"* which primarily consist of LLM-generate labels. This aligns with findings that recent models rely on generated and synthetic training data [5, 22]. However, benchmarking

[1]This study was conducted in Q4 of 2024

LLMs against human performance demands rigorous evaluation datasets, motivating the construction of *"super-golden datasets"*—exceptionally high-quality datasets curated by teams of experts.
- **Shift from bottom-up to top-down data understanding**: Prior to LLMs, practitioners aggregated insights from granular, heuristics-based analyses, performing manual and repetitive data labelling and aggregation. Using LLMs, practitioners can focus on more strategic, high-level data analysis – generating high-level summaries upfront, and diving into details only when needed.

Additionally, we observed a growing trend of LLM reliance across multiple stages of the curation and analysis workflow, with a perceived increase in efficiency. However, there are also challenges and concerns hindering LLMs' widespread adoption, such as cost and output reliability. This research contributes to understanding the emerging role of LLMs technical workflows and offers considerations for further LLM-based tool development and evaluation.

2 Related Work

There is a growing trend of integrating LLMs into tools for data curation tasks [15, 36, 49]. Numerous prompting interfaces and bespoke LLM-based tools have emerged [8, 23, 24, 26, 28, 29, 31, 32, 42], addressing tasks such summarization and categorization. LLMs can be used to interactively cluster datasets [41], explain and label these clusters [44], qualitatively code and analyze data [4, 6, 11], and even expand existing datasets by generating synthetic examples [22, 45, 47]. Furthermore, researchers have proposed leveraging LLMs to evaluate the performance of models, a practice known as "LLM-as-a-judge" [49], along with tools that visualize results [16, 17]. This approach can be applied not only to assessing general-purpose LLMs but also to evaluating specific aspects such as safety, factuality, coherence, fluency, or other evaluation criteria [15, 19].

3 User Study Design

Despite the recent emergence of LLM-focused tools, there is limited research examining their adoption in industry data practices, likely due to the nascent nature of the technology [48]. To address this gap in understanding LLM integration in technical workflows, we conducted a user study with industry professionals who perform diverse data curation tasks, with the following research questions:

- **RQ1:** In what ways are LLMs being used by practitioners for data curation tasks?
- **RQ2:** What aspects of their workflows do practitioners feel are made more productive by LLMs?
- **RQ3:** Which parts of workflows do LLM-based tools support, and which tools do they replace or complement?

3.1 Design Probes

Our research questions focus on understanding participants' existing and anticipated workflows with LLMs. However, recent prior work on the data curation workflows of industry practitioners suggests a lack of alignment regarding LLM-based curation tools, with practitioners continuing to rely on spreadsheets (e.g., Google Sheets) and Python notebooks (e.g., Colab[2]) [35].

To foster ideation using tools that practitioners already trust and use, we created two design probes integrating LLMs - one within Google Sheets, and another within Colab. These probes harness LLM capabilities to address common curation challenges identified in prior work, such as flexibility, ease-of-use, and integration [2, 35].

3.1.1 Spreadsheet Design Probe. We developed an Apps Script application[3] that enables LLM prompting within spreadsheet cells (Figure 1). This application introduces a "`RUN_PROMPT`" function that sends a text prompt to an LLM model, with customizable API parameters in a separate sheet.

3.1.2 Computational Notebook Design Probe. We developed a Colab notebook with built-in libraries for LLM prompting, with configurable parameters through form fields. Figure 2 shows the example notebook. The library includes a "`run_classifier`" function that accepts a Pandas dataframe (i.e., `df`) and an instruction. The function calls the LLM and returns the dataframe with an additional column containing the LLM's outputs. Since Python notebooks offer greater flexibility than spreadsheets. We provide two additional features:

- **Summative analysis**: Users can query the LLM with an entire dataset (Figure 3).
- **Controlled generation**: This feature allows structured outputs [20, 23] (e.g. `yes` or `no`) for tabular queries.[4] In the spreadsheets probe, controlled generation can only be approximated with the inclusion of instructions in the prompt such as *Please output only "yes" or "no".*

3.2 User Study Participants

We recruited 12 participants (N=12; 5 female, 7 male) who work with text-based datasets within a large technology company. Our focus was on recruiting participants who handle general text-based data– such as customer feedback or conversation logs– rather than specialized domains like medical data, which may require bespoke and less generalizable curation practices. Our study sample was carefully curated to include industry experts across different job roles and six distinct product areas within the company. We categorize these participants into three roles:

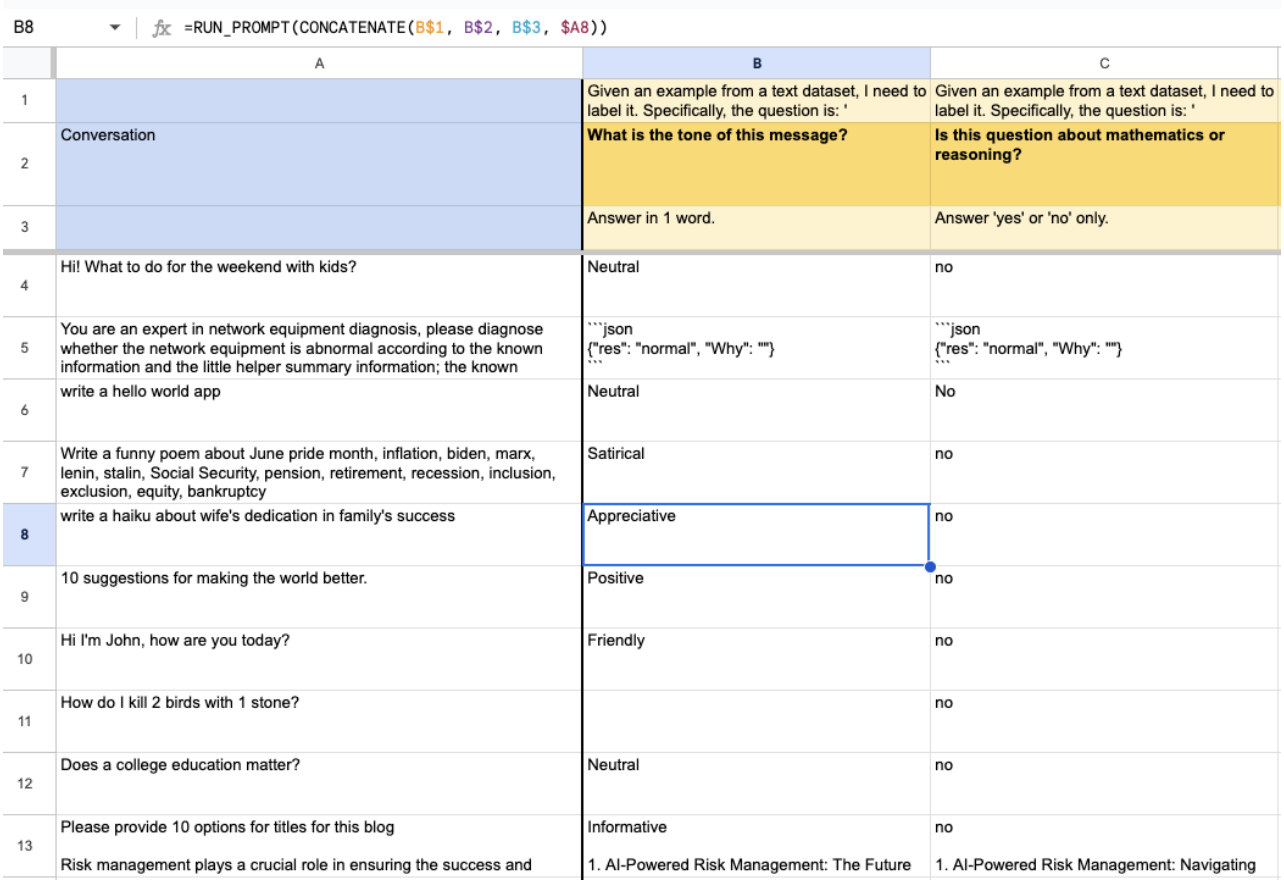
- **Technical roles** (T1–T4): Engineers and model developers who create and evaluate models for products.
- **Analytical and operational roles** (A1–A5): Domain experts, ethics researchers, and project leads who develop policies around products, primarily focused on safety.
- **Client-facing roles** (C1–C3): User experience researchers and survey experts who assess product usability.

These practitioners' roles involve extracting insights from text data, including tasks such as labeling text to identify trends, summarizing findings, developing training and evaluation datasets for foundational model development, or analyzing user notes for abuse detection. A detailed description of their job responsibilities can be found in Appendix A.

[2]https://colab.research.google.com

[3]https://developers.google.com/apps-script

[4]https://ai.google.dev/gemini-api/docs/structured-output?lang=python, https://spec.openapis.org/oas/v3.0.3#schema



B8 =RUN_PROMPT(CONCATENATE(B$1, B$2, B$3, $A8))

	A	B	C
1		Given an example from a text dataset, I need to label it. Specifically, the question is: '	Given an example from a text dataset, I need to label it. Specifically, the question is: '
2	Conversation	**What is the tone of this message?**	**Is this question about mathematics or reasoning?**
3		Answer in 1 word.	Answer 'yes' or 'no' only.
4	Hi! What to do for the weekend with kids?	Neutral	no
5	You are an expert in network equipment diagnosis, please diagnose whether the network equipment is abnormal according to the known information and the little helper summary information; the known	```json {"res": "normal", "Why": ""} ```	```json {"res": "normal", "Why": ""} ```
6	write a hello world app	Neutral	No
7	Write a funny poem about June pride month, inflation, biden, marx, lenin, stalin, Social Security, pension, retirement, recession, inclusion, exclusion, equity, bankruptcy	Satirical	no
8	write a haiku about wife's dedication in family's success	Appreciative	no
9	10 suggestions for making the world better.	Positive	no
10	Hi I'm John, how are you today?	Friendly	no
11	How do I kill 2 birds with 1 stone?		no
12	Does a college education matter?	Neutral	no
13	Please provide 10 options for titles for this blog Risk management plays a crucial role in ensuring the success and	Informative 1. AI-Powered Risk Management: The Future	no 1. AI-Powered Risk Management: Navigating

Figure 1: The tabular LLM prompting interface within the spreadsheet design probe. The cells in column A include conversations (i.e., questions to AI agents asked by crowd users) from the Chatbot Arena Conversation Dataset [27]. The header of the second column (B1-B3) contains an instruction that users of the probe can specify. The cells in the column are automatically populated with LLM outputs, generated by running an LLM query that combines the specified instruction from the header with the corresponding data in column A (e.g., `=RUN_PROMPT(CONCATENATE(B1, B2, B3, A8))`). Column C shows another prompt.

3.3 User Study Protocol

We conducted individual, hour-long sessions with the participants via video conferencing. At the beginning of the session, each participant received a dedicated copy of both the spreadsheet and notebook design probes (Figures 1, 2, and 3), which contained an excerpt of 100 entries from the Chatbot Arena Conversation Dataset.[5]

Each session began with a brief interview to understand the participant's use case and background, followed by an introduction and tutorial on the design probes. Participants then shared their screens for real-time observation. They explored and explained their current approaches to curation tasks such as summative analysis, categorization, and numerical analysis. Discussions focused on existing workflows, the current and potential role of LLMs, and how interfaces like the design probes might fit into their workflow.

Participants were encouraged to think aloud as they interacted with the spreadsheet and notebook probes. There was no fixed time allocation for each probe, and participants were free to move back and forth between them as needed. To analyze the transcripts, three researchers then developed a coding scheme for thematic analysis [3], and refined themes based on these codes [1, 13]. The study was approved by our institution's IRB.

[5]https://huggingface.co/datasets/lmsys/chatbot_arena_conversations

4 Results

4.1 RQ1: Data curation tasks using LLMs

Participants valued the flexibility of LLMs and reported utilizing LLMs across a variety of generalized curation tasks, such as:

- **Classification**: Prior to adopting LLMs, participants reported relying upon wordlists, manual searches, and regular expressions for classification tasks. These methods were prone to errors caused by missing typos, acronyms, translations, or synonyms. Furthermore, these methods were limited to only the wordlists that were either manually curated by experts or generated by existing tools like safety classifiers. LLMs offer a valuable alternative for classification tasks where pre-existing classifiers are not available.

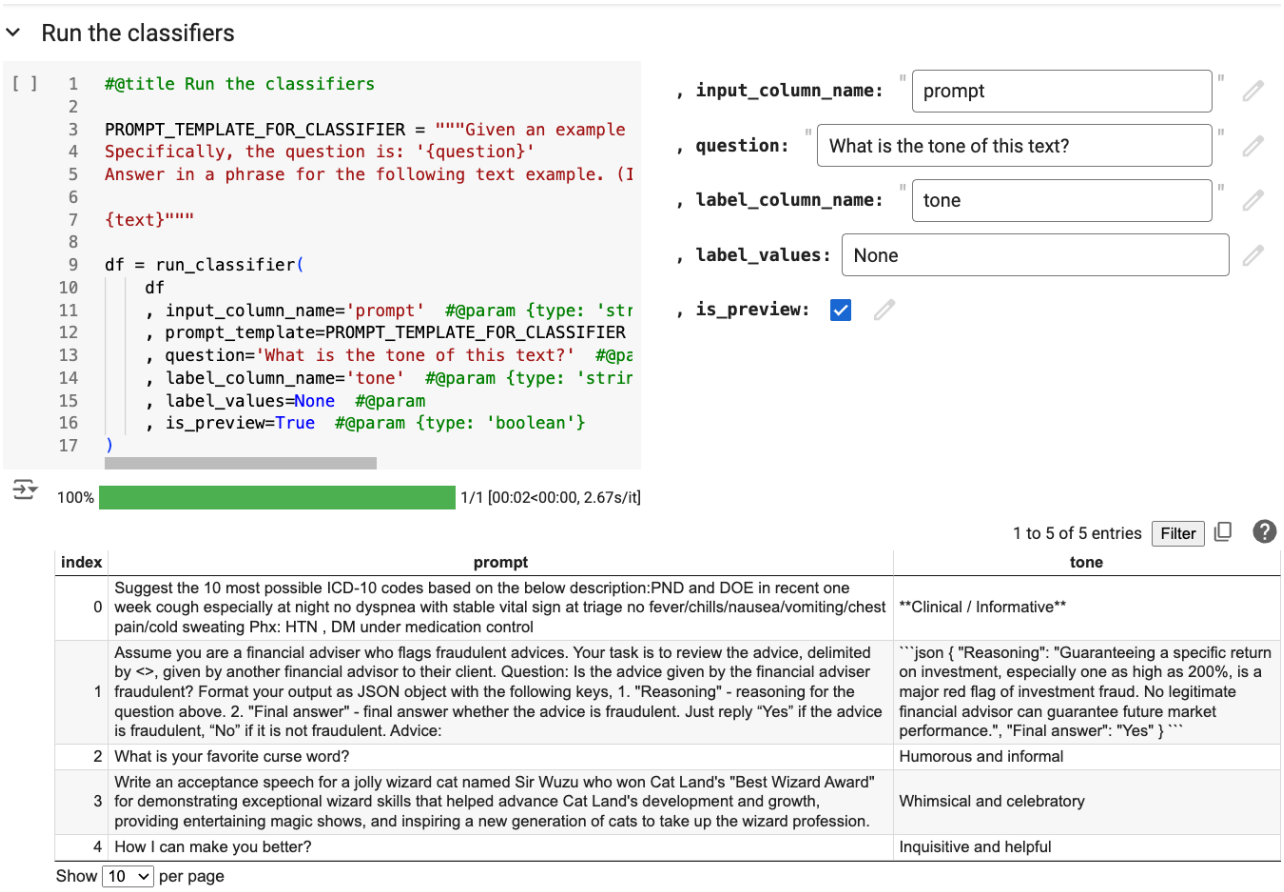


Figure 2: The tabular LLM-based prompting interface within the notebook design probe. This example shows a new `tone` column added to a dataframe, which asks "What is the tone of this text?" on the `prompt` column. Outputs are not constrained. The output dataframe with the new `tone` column is displayed below the form.

> *"We can use such prototypes [in situations] when I'm not aware of a good classifier... [such as] cases like 'what types of medical advice may cause a specific problem?"'* —A5

- **Summarization and aggregation**: Without LLMs, practitioners identified summative trends in unstructured text datasets by manually labelling data points and then aggregating them. Using LLMs, practitioners could directly generate labels, or even prompt for insights on their desired trend (e.g. "What are the top themes in this dataset?").
- **Explanation generation**: Participants found LLMs to be a valuable tool for generating additional context. For example, in a moderation use case, participants found LLM responses helpful for explaining why certain content was flagged by users as violative. Reviewers reported LLM assistance as particularly useful in scenarios involving language barriers or the need to detect subtle biases that require deeper contextual understanding.
- **Distributional analysis and outlier detection**: Participants also noted that LLMs could be useful in expediting slicing and filtering processes to identify outliers and anomalies. This is particularly useful in content moderation or safety evaluation, especially with large datasets that are impractical for humans to review in their entirety. LLMs can be used to identify candidate data points that require more resource-intensive processes, such as human review. By helping analysts to "*surface more interesting things to look at*" (A4), LLMs allows humans to allocate their expertise and attention more efficiently.

> *"The LLM can often do things often not as good as a human [expert] but very close... that's one more layer we can put on top before it gets to the human. [LLMs can [filter] out a lot of the obvious false positives that are difficult for a regex or a classifier, but a human would obviously understand.'*
>
> —A3

More specific examples of applied data curation tasks across a variety of use cases are reported in Appendix B.

4.2 RQ2: LLMs' perceived productivity improvements

Participants additionally provided feedback on the perceived impact of LLMs on their productivity across three dimensions: accuracy, efficiency, and satisfaction [9]. Their imprressions of LLMs' impact on **accuracy** were mixed. While some cited anedotal evidence that rating, labelling, clustering, and analysis tasks completed using LLMs

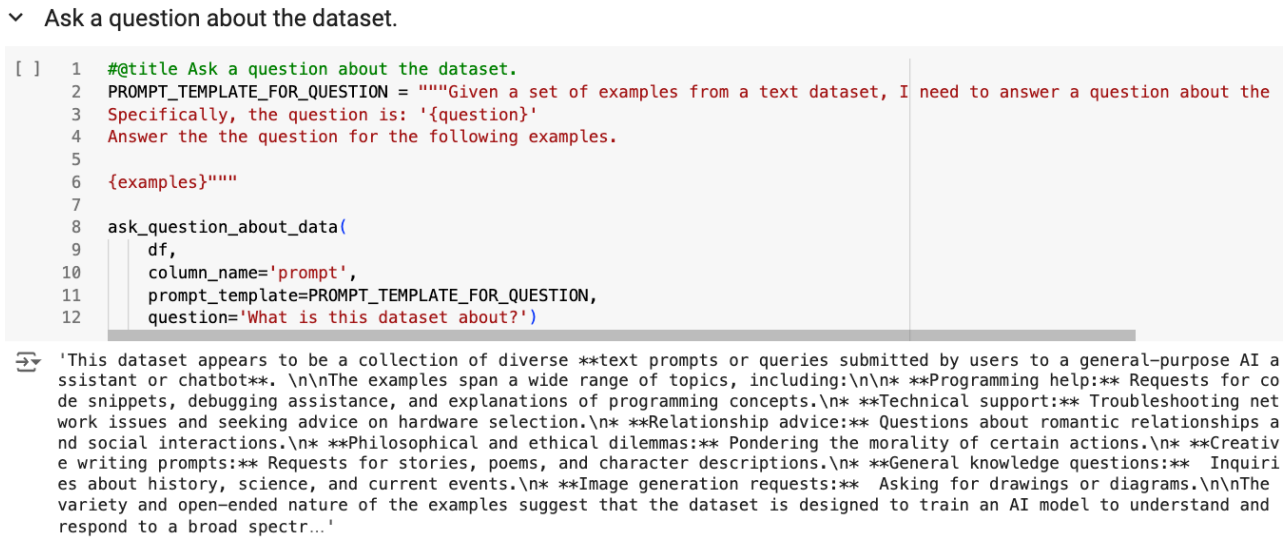


Figure 3: The summative LLM-based prompting interface within the notebook design probe. The example illustrates querying "What is this dataset about?" for the `prompt` column of a dataframe.

performed comparably to humans – and in some cases, showed higher inter-rater agreement – this might be partially attributed to LLMs' self-consistency [43].

> *"[We ran a] manual inter-rater reliability exercise… we slightly agreed more with human codes (compared to LLMs), but the agreement metrics were only 60%, 70%. This tells us that [accuracy wasn't high to begin with] …so I wanted to get out of the business of coding."*
> —C2, on using LLMs for survey coding

Participants agreed that LLMs provided significant **efficiency gains**, particularly for repetitive coding tasks. Tasks that previously took hours could be completed in minutes, freeing up time for higher-level tasks like refining taxonomies and reviewing policies.

> *"What's important to me is… what can I do to speed up the workflow? I'm trying to make it more efficient and faster for someone to create a prompt that allows you to go from 80% precision/80% recall to 90/90 [on my classification task]…My goal is to go from zero to essentially a fully functional classifier in hours."* —A1

While **satisfaction** with the Python notebook LLM design probe was limited to participants with technical backgrounds, all participants expressed **satisfaction** with the spreadsheet probe, noting its low learning curve and ease of collaboration.

> *"I can train other people up on it very easily, whereas there's a [learning curve] for Colab. I'm working with other analysts who aren't as technical… so I'm trying to use tools that are easier for other people."* —A2

LLMs' lower barrier to entry can improve collaboration: Participants highlighted the importance of embedding LLMs in familiar platforms for seamless collaboration. While many teams had developed LLM tools within Python notebooks, these solutions were often developer-centric and limited in the context of cross-functional teams. The spreadsheet probe was praised for improving accessibility and reducing the need to convert outputs for non-technical collaborators, fostering a more collaborative workflow.

> *"My product manager doesn't use LLMs for things that they could… we have to run things in Colab and share them with the [PMs] and go back and forth."* —T2

4.3 RQ3: LLMs' role in workflows – benefits and limitations

Participants reported that they would use LLMs within the spreadsheet probe for analysis on smaller datasets. For handling larger amounts of data or more complex tasks, such as concatenating outputs from multiple columns or building automated workflows, participants showed a clear preference for the notebook probe. Participants cited scalability and latency as limitations of LLM-based methods, although these concerns may be reduced as the technology advances.

Barriers to adopting LLM-based tooling more broadly include the fast-evolving nature of the field, stability, and reliability concerns. A few participants cited that they had not considered using LLMs for tabular data analysis before because large-scale analysis was only recently supported. For example, S1 explained that *"It didn't occur to us to [use LLMs]… the long context [capabilities] are new.'* For tasks involving sensitive content, safety-tuned responses by pre-trained models may be insufficient [33]. Finally, participants hesitated to use LLMs for tasks requiring precise, deterministic outputs, particularly where hallucinations or biases would be unacceptable:

> *"I would never use quotes spit out by the LLM as examples…I would go pull it myself."* —A1

Despite these limitations, participants reported that LLMs were broadly and rapidly utilized within their teams, re-defining both data curation workflows and the hierarchies of unstructured datasets. We discuss these trends in the following section.

5 Discussion

5.1 Emerging Workflow Trends

5.1.1 From bottom-up aggregation to top-down extraction. Traditionally, practitioners have performed bottom-up data analyses. They label and categorize individual data points, and then aggregate them to identify trends [35]. LLMs now upend this process, allowing practitioners to gain high-level insights from the start, extracting individual data points only when granular analysis is needed. However, as users grow more accustomed to incorporating LLMs for this purpose, there is a risk that this adoption may lead to a decrease in the rigor of validation processes.

> *"If [you were] a new team going straight to LLMs, there's a risk that you don't know when things are off. When I saw strange words [in an LLM summary], I did a data pull to verify that this was wrong. I deeply [knew that the summary was wrong] already because I've read through so much of [the data] before."* —C3

5.1.2 Expanded scope for data practitioners. LLM usage is also reshaping the role of data practitioners. While some roles involving data gathering and manual coding may be increasingly automated, experts such as policy analysts and safety analysts reported using LLMs to expand the scope of their work. Previously, collecting data labels and annotations took weeks, requiring experts to define labeling rules and wait for human annotators to execute them. Now, LLMs enable a much faster understanding of th edata, streamlining the process.

> *"Prior to the advent of using LLMs, I was more of a consumer of data provided by others, as opposed to having the ability to create and identify the data that I was using.'* —A2

5.2 Emerging Dataset Hierarchies

Traditionally, "golden datasets," meticulously labeled by human experts, have been a standard of data for model training and evaluation. However, participants reported that the capabilities of LLMs have enabled more sophisticated tiers of small, high-quality datasets:

(1) **Silver datasets**: While human-labeled "golden" datasets remain crucial, there is a growing trend to complement them with "silver" datasets generated by LLMs, particularly for high-stakes labeling tasks.

> *"We would never use LLMs to classify the entire [data] corpus of hundreds of millions of instances…However, we're currently trying zero-shot/few-shot prompting to complement our classifications on important [data instances]. We'd still have golden output by human raters, but complemented with a silver output by LLMs for the high-traffic data, and a cheap and flexible classifier for the remaining data.'* —A1

(2) **Super-golden datasets**: Comparing LLMs to human performance necessitates even more rigorous ground-truth. "Super golden data" are created by diverse teams of experts including product managers, policy makers, and engineers. They are critical for fine-tuning and evaluating LLM performance; However, developing these super-golden datasets is both time-consuming and resource-intensive, often taking on the order of weeks.

> *"It's very expensive to compare an LLM with humans because where is the ground truth coming from? You need a higher authority of human rater, like super golden labels. It's a mix of product managers, policy makers, and [engineers] from our team. It takes a long time to label even 500 examples.'* —T4

5.3 Future Work and Directions

Our findings on emerging trends in workflows and dataset formation highlight several promising directions for future research. Below, we outline critical areas for further exploration and actionable guidance based on our study.

Safety and dataset evaluation. The growing use of silver datasets – datasets curated by practitioners using LLM-generated labels – raises concerns about quality and bias. Validation processes must evolve to ensure reliability, similar to how safety and responsibility efforts seek to validate LLM outputs. Future research should explore methods for verifying classification results, such as conducting error analysis, auditing for potential biases and stereotypes, and maintaining diversity within these datasets.

As the outputs of LLMs may be opaquely generated, integrating human-in-the-loop workflows will ensure better validation, reliability, and quality control. HCI will be essential in designing effective oversight mechanisms to enhance interpretability and trust in LLM-generated data.

Evolving paradigms. We anticipate that the current emphasis on creating small, high-quality datasets will remain a focus for the foreseeable future. The current approach to refining the existing "golden" dataset paradigm has resulted in a complex landscape that includes variations such as silver and super-golden datasets, and this space of data will likely grow.

With increasing oversight, privacy concerns, and fairness priorities, dataset development trends will likely continue to shift from bottom-up to top-down approaches, addressing predefined policies and target distributions.

5.4 Study Limitations

This study was conducted within the context of a single company, utilizing specific internal infrastructures and particular cultural and operational practices. While our study utilized a diverse population across many company organizations, and the findings aligned with prior research [18], further work is needed to validate their generalizability.

Furthermore, the scope of this work was constrained to individuals primarily involved in generalized, largely analytical data curation tasks. This may not capture the full range of experiences, including those of specialized data practitioners working with domain-specific datasets (such as medical or biological data) or data of different modalities, and data workers and crowd workers, whose work also involves text-based datasets.

6 Conclusions

We undertook this study to understand the opportunities of incorporating LLM-based methods into data curation workflows. However, it quickly became clear that the question was not *if* practitioners

were open to using LLMs, but rather, *how*. We observed a rapidly growing reliance on LLMs for a wide variety of tasks, such as classification, summarization, explanation, and outlier detection, especially in cases where efficiency is prioritized. LLMs are enabling practitioners to move away from heuristics-based, bottom-up data aggregation and toward insights-first, top-down analyses, marking a fundamental transformation in how practitioners engage with their data. This shift underscores the need for robust definitions and frameworks for evaluating data quality.

As the landscape evolves, we anticipate a shift toward more systematic, policy-driven dataset creation, with human-in-the-loop workflows ensuring transparency and quality. This progression could lead to "super-golden" datasets—small yet highly refined collections that set new standards for data curation in the era of foundation models. At the same time, the use of of LLM-curated "silver" datasets introduces concerns about quality, bias, and validation. Silver datasets require robust error analysis and auditing to ensure diversity and fairness. As LLMs become integral to data workflows, human oversight will remain essential for ensuring transparency, accountability, and reliability in data-driven insights.

Acknowledgments

We thank our pilot and study participants for their time, and the People + AI Research (PAIR) team at Google DeepMind, especially Lucas Dixon, Andy Coenen, and Alex Fiannaca.

References

[1] 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. doi:10.2466/03.CP.3.4 _eprint: https://doi.org/10.2466/03.CP.3.4.

[2] 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 (DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2304–2325. doi:10.1145/3563657.3596046

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

[4] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. doi:10.48550/arXiv.2306.14924 arXiv:2306.14924 [cs, stat].

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[6] Stefano De Paoli. 2024. Performing an Inductive Thematic Analysis of Semi-Structured Interviews With a Large Language Model: An Exploration and Provocation on the Limits of the Approach. *Social Science Computer Review* 42, 4 (Aug. 2024), 997–1019. doi:10.1177/08944393231220483 Publisher: SAGE Publications Inc.

[7] Zackary Okun Dunivin. 2024. Scalable Qualitative Coding with LLMs: Chain-of-Thought Reasoning Matches Human Performance in Some Hermeneutic Tasks. doi:10.48550/arXiv.2401.15170 arXiv:2401.15170 [cs].

[8] Raymond Fok, Joseph Chee Chang, Tal August, Amy X. Zhang, and Daniel S. Weld. 2024. Qlarify: Recursively Expandable Abstracts for Directed Information Retrieval over Scientific Papers. https://arxiv.org/abs/2310.07581 _eprint: 2310.07581.

[9] Nicole Forsgren, Margaret-Anne Storey, Chandra Maddila, Thomas Zimmermann, Brian Houck, and Jenna Butler. 2021. The SPACE of Developer Productivity: There's more to it than you think. *Queue* 19, 1 (March 2021), Pages 10:20–Pages 10:48. doi:10.1145/3454122.3454124

[10] André Freitas and Edward Curry. 2016. Big data curation. *New horizons for a data-driven economy: A roadmap for usage and exploitation of big data in Europe* (2016), 87–118. Publisher: Springer International Publishing.

[11] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. CollabCoder: A Lower-barrier, Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–29. doi:10.1145/3613904.3642002

[12] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the Science of Language Models. https://arxiv.org/abs/2402.00838 _eprint: 2402.00838.

[13] 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. doi:10.1177/1525822X05279903 _eprint: https://doi.org/10.1177/1525822X05279903.

[14] Lei Han, Tianwa Chen, Gianluca Demartini, Marta Indulska, and Shazia Sadiq. 2023. A data-driven analysis of behaviors in data curation processes. *ACM Transactions on Information Systems* 41, 3 (2023), 1–35. Publisher: ACM New York, NY.

[15] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. doi:10.48550/arXiv.2312.06674 arXiv:2312.06674 [cs].

[16] 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. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3613905.3650755

[17] Minsuk Kahng, Ian Tenney, Mahima Pushkarna, Michael Xieyang Liu, James Wexler, Emily Reif, Krystal Kallarackal, Minsuk Chang, Michael Terry, and Lucas Dixon. 2025. LLM Comparator: Interactive Analysis of Side-by-Side Evaluation of Large Language Models. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 503–513. doi:10.1109/TVCG.2024.3456354

[18] 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. doi:10.1109/TVCG.2012.219

[19] Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. doi:10.48550/arXiv.2405.01535 arXiv:2405.01535 [cs].

[20] Terry Koo, Frederick Liu, and Luheng He. 2024. Automata-based constraints for language model decoding. doi:10.48550/arXiv.2407.08103 arXiv:2407.08103 [cs].

[21] Tzu-Sheng Kuo, Aaron Lee Halfaker, Zirui Cheng, Jiwoo Kim, Meng-Hsin Wu, Tongshuang Wu, Kenneth Holstein, and Haiyi Zhu. 2024. Wikibench: Community-driven data curation for ai evaluation on wikipedia. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. doi:10.48550/arXiv.2304.08485 arXiv:2304.08485.

[23] Michael Xieyang Liu, Frederick Liu, Alexander J. Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J. Cai. 2024. "We Need Structured Output": Towards User-centered Constraints on Large Language Model Output. doi:10.1145/3613905.3650756 arXiv:2404.07362 [cs].

[24] Michael Xieyang Liu, Savvas Petridis, Vivian Tsai, Alexander J. Fiannaca, Alex Olwal, Michael Terry, and Carrie J. Cai. 2025. Gensors: Authoring Personalized Visual Sensors with Multimodal Foundation Models and Reasoning. doi:10.1145/3708359.3712085 arXiv:2501.15727 [cs].

[25] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D. Gordon. 2023. "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–31. doi:10.1145/3544548.3580817

[26] Michael Xieyang Liu, Tongshuang Wu, Tianying Chen, Franklin Mingzhe Li, Aniket Kittur, and Brad A. Myers. 2024. Selenite: Scaffolding Online Sensemaking with Comprehensive Overviews Elicited from Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3613904.3642149

[27] LMSYS. 2024. Chatbot Arena Conversations Dataset. https://huggingface.co/datasets/lmsys/chatbot_arena_conversations

[28] Qianou Ma, Hua Shen, Kenneth Koedinger, and Tongshuang Wu. 2024. How to Teach Programming in the AI Era? Using LLMs as a Teachable Agent for Debugging. Vol. 14829. 265–279. doi:10.1007/978-3-031-64302-6_19 arXiv:2310.05292

[cs].

[29] Xiao Ma, Swaroop Mishra, Ariel Liu, Sophie Ying Su, Jilin Chen, Chinmay Kulkarni, Heng-Tze Cheng, Quoc Le, and Ed Chi. 2024. Beyond ChatBots: ExploreLLM for Structured Thoughts and Personalized Model Responses. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3613905.3651093

[30] 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 (CHI '19)*. ACM, 1–15. doi:10.1145/3290605.3300356 Place: Glasgow, Scotland Uk.

[31] Chris Parnin, Gustavo Soares, Rahul Pandita, Sumit Gulwani, Jessica Rich, and Austin Z. Henley. 2023. Building Your Own Product Copilot: Challenges, Opportunities, and Needs. doi:10.48550/arXiv.2312.14231 arXiv:2312.14231 [cs].

[32] Savvas Petridis, Michael Xieyang Liu, Alexander J. Fiannaca, Vivian Tsai, Michael Terry, and Carrie J. Cai. 2024. In Situ AI Prototyping: Infusing Multimodal Prompts into Mobile Settings with MobileMaker. In *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 121–133. doi:10.1109/VL/HCC60511.2024.00023 ISSN: 1943-6106.

[33] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! https://arxiv.org/abs/2310.03693 _eprint: 2310.03693.

[34] Crystal Qian, Michael Xieyang Liu, Emily Reif, Grady Simon, Nada Hussein, Nathan Clement, James Wexler, Carrie J. Cai, Michael Terry, and Minsuk Kahng. 2024. The Evolution of LLM Adoption in Industry Data Curation Practices. doi:10.48550/arXiv.2412.16089 arXiv:2412.16089 [cs].

[35] Crystal Qian, Emily Reif, and Minsuk Kahng. 2024. Understanding the Dataset Practitioners Behind Large Language Model Development. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM.

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

[37] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP models with CheckList. doi:10.48550/arXiv.2005.04118 arXiv:2005.04118 [cs].

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

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. doi:10.48550/arXiv.2302.13971 arXiv:2302.13971 [cs].

[40] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. doi:10.48550/arXiv.2307.09288 arXiv:2307.09288 [cs].

[41] Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2023. Large Language Models Enable Few-Shot Clustering. doi:10.48550/arXiv.2307.00524 arXiv:2307.00524 [cs].

[42] Chenglong Wang, Bongshin Lee, Steven Drucker, Dan Marshall, and Jianfeng Gao. 2024. Data Formulator 2: Iteratively Creating Rich Visualizations with AI. doi:10.48550/arXiv.2408.16119 arXiv:2408.16119 [cs].

[43] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=1PL1NIMMrw

[44] Zihan Wang, Jingbo Shang, and Ruiqi Zhong. 2023. Goal-Driven Explainable Clustering via Language Descriptions. doi:10.48550/arXiv.2305.13749 arXiv:2305.13749 [cs].

[45] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S. Weld. 2021. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. doi:10.48550/arXiv.2101.00288 arXiv:2101.00288 [cs].

[46] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 75–78. doi:10.1145/3581754.3584136

[47] Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. 2022. SynthBio: A Case Study in Human-AI Collaborative Curation of Text Datasets. doi:10.48550/arXiv.2111.06467 arXiv:2111.06467 [cs].

[48] Dora Zhao, Morgan Klaus Scheuerman, Pooja Chitre, Jerone T. A. Andrews, Georgia Panagiotidou, Shawn Walker, Kathleen H. Pine, and Alice Xiang. 2024. A Taxonomy of Challenges to Curating Fair Datasets. doi:10.48550/arXiv.2406.06407 arXiv:2406.06407 [cs].

[49] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, 46595–46623.

A Study Participants

Table 1 describes the participants recruited for this user study, as well as their responses to a screening survey. Participants evaluated their familiarity with relevant tools (spreadsheets, Colab, Python) on a five-point scale to provide context for their usage patterns. Python and Colab usage were less common in client-facing roles but prevalent in technical and analytical roles. Notably, Colab was utilized by some participants without extensive Python experience.

Table 1: Descriptions of participants in the user study (N=12).

Participant	Product Area	Job Description	Tool Familiarity		
			Sheets	Colab	Python
[T] Technical roles					
T1	Foundation models	Evaluates prompt expansion text generation models	5	5	5
T2	Foundation models	Inspects text-datasets for LLM post-training	5	5	5
T3	Foundation models	Works on post-training a variety of LLM models	4	4	5
T4	Content platforms	Builds safety classifiers for content	5	5	5
[A] Analytical / operational roles					
A1	Trust & safety	Works on detecting abuse content at scale across products	4	4	1
A2	Trust & safety	Develops golden datasets for scaled abuse detection	4	4	2
A3	Content platforms	Analyzes user notes to detect violative content	4	4	3
A4	Responsible AI	Analyzes and creates safety datasets for text-to-image generation	5	5	4
A5	Responsible AI	Designs evaluation metrics of datasets	3	3	4
[C] Client-facing roles					
C1	User experience	Analyzes behavioral survey data for product users	5	1	1
C2	User experience	Evaluates custom feedback survey data for accounting teams	5	2	2
C3	User experience	Develops customer-facing feedback surveys	5	2	2

B Reported Use Cases

Table 2: Participants' current and anticipated LLM usage cases within their product areas.

Product Area, Participants	Description	LLM Usage and Examples
Foundation models T1, T2, T3	T1, T2, and T3 curate data for training, fine-tuning, and evaluating LLMs on a variety of use cases, such as safety evaluation and image generation.	**Summarization**: "Which topics are extremely prevalent in this dataset?" **Distributional analysis**: "How diverse are the responses generated by raters?" "Are these prompts duplicates or near-duplicates?" **Categorization**: "There are 10 categories: scientific, factuality, writing…which categories fit this prompt?" "Is this prompt about a person? Yes, no, or maybe?"
Trust & safety A1, A2	A1 and A2 are policy experts who create golden datasets of carefully curated violative content, such as hate speech or violent extremism, to detect abuse at-scale across products.	**Summarization**: "Here's a dataset of user comments. Please cluster them, give a description of what's in the cluster, and examples from the cluster itself, in the style of a business analyst." **Categorization**: "Was the third-party vendor who flagged this content as violating a policy correct?" **Probabilistic classification**: "What is the probability of this text violating the policy?" **Distributional analysis / Explanation**: "Identify things [in this text] that violate [these policies], explain why."
Content platforms A3, T4	A3 and T4 build safety policies and classifiers around violative content, using text-based data such as captions, content metadata, and user commentary.	**Classification**: "Does this content have violative content in it?" "Is this classification safe, risky, or unsafe?" "Is the report on this content actionable?" **Explanation**: "Why was this content considered harmful?"
Responsible AI A4, A5	A4 and A5 create and analyze safety evaluation datasets for downstream tasks such as model safety evaluation. Their work may include designing metrics or interacting with rater pools.	**Summarization**: "What are the top violative themes in this dataset? **Classification**: "Is this text about kids?" "Here are 5 policies: which might this violate?" "On a scale of 1-10, what is the complexity of this prompt?" **Text generation**: "What are some synonyms for this sensitive term?"
User experience C1, C2, C3	C1, C2, and C3 develop client-facing surveys to evaluate a broad range of products. They interact with large-scale survey responses and operational metrics, and report insights to leadership.	**Summarization**: "What are the top 5 issues that customers have mentioned?" **Classification**: "Which of the 100 products is this feedback addressing?" "What theme fits this open-ended survey response?" "Is this feedback positive or negative?" **Extraction**: "Pull quotes that add context to each theme." "Which of the data is about networking issues?"