Position: Unlocking the Full Potential of Data Science Requires Tabular Foundation Models, Agents, and Humans

Tianji Cong[1*] **Julian Martin Eisenschlos**[2*] **Daniel Gomm**[3,4*] **Leo Grinsztajn**[16*]
Andreas C. Müller[6*] **Anupam Sanghi**[7*] **Jan-Micha Bodensohn**[20,7] **Vadim Borisov**[18]
Michael Cochez[9] **Katharina Eggensperger**[8] **Floris Geerts**[10] **Myung Jun Kim**[5]
Andreas Kipf[11] **Xue Li**[3] **Olga Ovcharenko**[12] **Paolo Papotti**[13] **Lennart Purucker**[14]
Sebastian Schelter[12] **Immanuel Trummer**[15] **Gaël Varoquaux**[5,17] **Liane Vogel**[7]
Carsten Binnig[7,20†] **Madelon Hulsebos**[3†] **Frank Hutter**[16,19,14†]
[1]University of Michigan [2]University of Cordoba [3]Centrum Wiskunde & Informatica
[4]University of Amsterdam [5]SODA Team, Inria Saclay [6]Gray Systems Lab, Microsoft
[7]Technical University of Darmstadt [8]University of Tübingen [9]Vrije Universiteit Amsterdam
[10]University of Antwerp [11]University of Technology Nuremberg [12]BIFOLD & TU Berlin
[13]EURECOM [14]University of Freiburg [15]Cornell University [16]Prior Labs
[17]Probabl.ai [18]tabularis.ai [19]ELLIS Institute Tübingen [20]DFKI

Abstract

Despite its vast potential, data science remains constrained by manual workflows and fragmented tools. Meanwhile, foundation models have transformed natural language and computer vision - and are beginning to bring similar breakthroughs to structured data, particularly the ubiquitous tabular data central to data science. At the same time, there are strong claims that fully autonomous agentic data science systems will emerge. We argue that, rather than replacing data scientists, the future of data science lies in a new paradigm that amplifies their impact: collaborative systems that tightly integrate agents and tabular foundation models (TFMs) with human experts. In this paper, we discuss the potential and challenges of navigating the interplay between these three and present a research agenda to guide this disruption toward a more accessible, robust, and human-centered data science.

1 The Disruption of Data Science is Inevitable

Over the past two decades, data science has accelerated and improved decision-making. Fueled by increased data availability, advances in machine learning, easy-to-use software packages, and scalable compute infrastructure, data science has transformed applications across various domains. Yet, data science has not reached its full potential, especially for tabular data. Workflows remain fragmented across a diversity of tools and labor-intensive [119, 106], and the gap between technical experts and domain stakeholders continues to limit mutual understanding [117].

In recent years, foundation models have revolutionized text and images, and are starting to do the same in structured domains, such as tabular data. Tabular Foundation Models (TFMs), such as TabPFN [65] and TableGPT2 [129], demonstrate that models pretrained across heterogeneous tables can generalize well to new tables for predictive tasks, data wrangling, and beyond. Simultaneously, the rise of autonomous agentic systems capable of reasoning, tool use, and coding offers a perfect match to the many repetitive and procedural tasks that dominate tabular data science workflows. Together, foundation models and agents clearly herald a major shift for tabular data science as they have the potential to automate data science end-to-end, which is very different from prior, more siloed attempts in AutoML [69]. However, instead of replacing human experts by automation, we argue that the future of data science lies in a new paradigm that amplifies the impact of human experts:

[*]co-first authorship
[†]co-lead PI

Preprint.

Our Position

Owing to rapid developments in agentic data science and tabular foundation models, the role of data scientists is set to undergo massive disruption. Instead of chasing the "AI replacing data scientists" scenario, we argue that tabular foundation models, agents, and human experts, together, hold the key to unlocking the full potential of data science, if the risks and opportunities inherent to their interplay are navigated effectively.

While it is clear that agents and TFMs will continue to advance rapidly, this position paper argues that human involvement remains indispensable. Crucially, humans are often the sole source of institutional knowledge—knowledge that is deeply contextual, environment-specific, and typically undocumented and thus not accessible to either TFMs or agents. Furthermore, humans are essential for specifying success criteria, disambiguating edge cases, resolving blind spots in model reasoning, and serving as the final arbiters of correctness. They provide critical judgment in determining whether the outputs of automated systems are valid and trustworthy.

While the concept of keeping humans in the loop is not new, we go further, arguing for an approach in which agents, TFMs, and humans are combined as tightly coupled collaborators. When designed thoughtfully, this triad can complement and amplify each other's strengths, driving a new era of accelerated, robust data science. The central challenge we pose in this paper is how to orchestrate the collaboration, since navigating the interaction between agents, TFMs, and humans is far from trivial—especially as the capabilities of each component continue to evolve.

2 Tabular Data Science has Not Reached its Full Potential

Data science has matured substantially over the last 50 years, driving significant advancements across diverse sectors, including use cases such as enhanced diagnostic precision in healthcare [9, 81, 4], recommendation systems in retail [88, 41, 94, 15], and improved investment performance in finance [30, 59, 32] amongst many others. Despite its significance and widespread use, data science has by far not yet reached its full potential, with a substantial gap persisting between current practices and the ultimate potential it could have for organizations and society.

2.1 Challenges to Data Science

The Sobering Reality: High Failure Rates of Data Science projects. Recent analyses consistently indicate that a significant portion of data science and AI initiatives encounter substantial hurdles, often falling short of their intended business objectives or failing to achieve full production deployment [119, 137, 74, 140]. This considerable failure rate highlights systemic challenges prevalent throughout the data science lifecycle. These difficulties primarily originate from two critical areas: data-related issues and model-related issues [117, 141].

Under-utilized & Fragmented Data. A significant volume of data, often called *dark data* [17], remains under-utilized because of limited resources to navigate data, inadequate tools, or a lack of awareness regarding its inherent value. This widespread under-utilization of data directly contributes to suboptimal decisions, as countless business choices continue to be made based on incomplete information, rather than leveraging data-driven insights that could yield superior outcomes, cost savings, or new revenue streams. Furthermore, persistent integration challenges arise from data accessibility issues and data silos, with data locked in disparate systems [82, 124, 38, 3, 118, 93] as well as the lack of standardized definitions [126, 139].

Data Preparation as Impediment. Data preparation (collecting, integrating, cleaning, and transforming data) consumes a significant portion, approximately 80%, of a data scientist's time, leaving only 20% for actual analysis and model building [25, 124, 117, 119]. Despite long-standing efforts to address data quality and cleaning [110, 56, 37, 124, 70, 46], these remain substantial hurdles in practice. Overall, this leads to a diversion of highly skilled data scientists to often tedious and repetitive data work, which impedes high-impact tasks and responsibilities like model development and interpretation, leading to project delays and missed business opportunities.

Model Failures. Data scientists make use of models to capture the patterns in data, which can go wrong in many ways. Inadequate feature engineering–the process of selecting, transforming, and creating model inputs–can severely limit model performance. Other common pitfalls like over-

Table 1: Complementary strengths of human experts, alongside TFMs and LLM Agents. Only their combination covers the full spectrum of capabilities. ♠ indicates partial ability.

Capability	Humans		TFMs	LLM Agents
	Data Scientists	Domain Experts		
Domain knowledge understanding	✗	✓	✗	♠
Structural & statistical understanding	✓	✗	✓	✗
Contextual data science	✓	✗	✗	✗
Planning & ML tool execution	✓	✗	✗	✓
Goal alignment	✗	✓	✗	✗
Scalable automation	✗	✗	✓	✓

or underfitting can compromise model generalizability [55, 148, 132]. Beyond mere prediction performance, the validity of conclusions may be compromised by unaccounted-for confounders [1]. In addition to these, the misapplication of techniques often results in the selection of inappropriate or overly complex algorithms, undermining project success [34, 11]. But even if that succeeds, models used in dynamic environments degrade over time due to concept and data drift [42]. In all these cases, unnecessary model complexity creates technical debt [119].

2.2 Envisioning the Full Potential of Data Science

Many of the aforementioned challenges stem from the limited bandwidth of data scientists and the highly manual, repetitive work required to navigate the vast space of possible solutions, like selecting appropriate datasets and determining how to combine models with data. Recently, numerous efforts have thus aimed to address these issues and to automate various data-related tasks, including data exploration, transformation, and cleaning. Similarly, in model construction, AutoML systems [40, 35] have sought to automate the design and tuning of machine learning models. But mere automation, while reducing overhead, does not unlock the full potential of data science [27]. It risks reinforcing existing patterns while only marginally expanding the solution space. To truly innovate, we need systems that can creatively and effectively explore novel solutions. We believe that a tightly integrated combination of agents, TFMs, and human experts can enable the kind of exploration and creativity needed to catalyze data science. Moreover, such a system should empower a broader spectrum of users, from domain experts to data scientists, who engage with data systems in different ways to extract insights. The foundational building blocks are now in place to begin making this a reality.

3 TFMs and LLM Agents Fall Short in Isolation, and without Humans

Recent attempts to supercharge data science have focused primarily on automating individual steps of the data science workflow, with two promising directions emerging. LLM-based agents and their table-tuned variants [84, 128] can perform a wide range of tasks directly on tables but often lack a deep, structural understanding of tabular semantics and statistical reasoning. Conversely, the current generation of tabular foundation models [65, 108, 29, 63] better capture table structure but fall short in flexibility and task coverage. However, while automation helps to scale data science, we also show why human input remains indispensable and cannot be fully replaced. Table 1 summarizes the capabilities and limitations of LLM agents, TFMs, and human experts, serving as a reference throughout the discussion.

3.1 TFMs: Understand Tabular Structures, but Limited Capabilities

TFMs are purpose-built architectures designed to tackle tasks on tabular data out of the box or with minimal overhead. Their structure enables them to better understand table-specific patterns and relationships. Presently, TFMs typically fall into two categories:

- **Predictive TFMs** like TabPFN [64, 65], TabICL [108], CARTE [77], and other variants focus on predictive machine learning tasks like classification and regression.
- **Representation TFMs** like TaBERT [145], TURL [29], TaPas [63], and TableGPT2 [129] produce task-agnostic embeddings for downstream models or fine-tuning for specific tasks such as column type prediction, table question answering, and entity matching [6].

TFMs are Efficient but Limited. Importantly, both classes of TFMs are tabular-native; they employ both row-wise and column-wise attention within transformer-based frameworks, and they often natively process mixed numerical and categorical data without costly byte-pair tokenization. Synthetic dataset pre-training further imparts invariance to row and column order, as well as robustness to missing-value patterns, at scale. CARTE [77] and TARTE [78] additionally incorporate column names to bootstrap from knowledge graphs, and KumoRFM [111] also handles relational data. Moreover, TFMs are sample efficient in predictions and their architectures allow them to remain compact—typically within the 10–50M parameter range, which is two to four orders of magnitude smaller than frontier LLMs. As a result, they benefit from reduced inference latency, lower energy consumption, and a smaller carbon footprint [127, 71]. Nonetheless, TFMs lack many of the broader capabilities expected from true foundation models which are available in modern LLMs.

Why Today's TFMs Are Insufficient. The success of foundation models in other modalities stems from their broad task coverage, general-purpose representations, and ability to be adapted with minimal additional supervision [13, 2, 109]. In contrast, today's TFMs fall short of this ideal. Predictive TFMs are narrowly scoped to row-level classification or regression, addressing only a small fraction of the data science workflow, leaving many crucial tasks (e.g., data cleaning, which may require awareness of multiple rows or entire tables) unaddressed. Representation TFMs offer reusable embeddings across broader tasks, but they currently still require separate downstream models or training for each specific task [6]. This breaks the promise of end-to-end adaptability and adds complexity to deployment. As a result, today's TFMs do not yet match the versatility and integrative power that define true foundation models.

3.2 LLM Agents: Generalist Abilities, but Lacking Rigor

In contrast to TFMs, LLMs are generalists, excelling across a wide range of tasks. Their ability to follow natural language instructions and generalize from minimal examples based on their background knowledge makes them appealing as universal assistants throughout data science workflows. LLM *agents* further extend their capabilities to multi-step reasoning and tool use, enabling them to invoke external functions, APIs, or code—capabilities highly relevant for complex data science tasks. However, LLMs alone remain insufficient for tasks that require statistical reasoning, such as complex table understanding or prediction tasks.

LLMs Lack Rigorous Table Understanding. Although LLMs have achieved strong results on certain data wrangling [98] and exploration tasks (e.g., Text-to-SQL [123]), substantial evidence shows that they still lack a rigorous and reliable understanding of structured table data [23, 142]. While LLMs excel at predictive tasks with a handful of data points (due to their strong background knowledge) [60, 45], they are not capable of statistical reasoning for more than a few dozen data points. The limitations are especially pronounced in enterprise scenarios, where data is highly complex and domain-specific, and LLMs cause high computational costs when applied to large tables [12, 100]. Furthermore, LLMs' internal reasoning processes are opaque, making it difficult to interpret model outputs or systematically improve performance [54, 142].

Why Reasoning and Tools Are Not Enough. Enhancements such as retrieval-augmented generation and integrated tool use have been developed to address LLM challenges like hallucination and data unawareness. Yet, these methods remain limited. LLMs can hallucinate logic, struggle with complex multi-step reasoning, and often lack awareness of what data is needed to achieve specific analysis goals [20] or simply don't have relevant context information due to the undocumented nature of tacit knowledge. As a result, their execution capabilities are restricted, particularly when facing open-ended or loosely defined objectives [44].

3.3 Human Experts: Limited Bandwidth, but Indispensable Partners

Human experts face clear challenges, including limited time, cognitive bandwidth, and difficulty in systematically exploring large and complex solution spaces. A study among AI practitioners [117], for example, highlighted that data preparation work is "time-consuming, invisible to track, and often done under pressures to move fast due to margins — investment, constraints, and deadlines often came in the way of focusing on improving data quality." Compromises on problem formulation, data quality, and critical reflection on modeling validity are inevitable, leading to suboptimal data science outcomes. Nevertheless, humans bring expertise that remains critical for solving data science

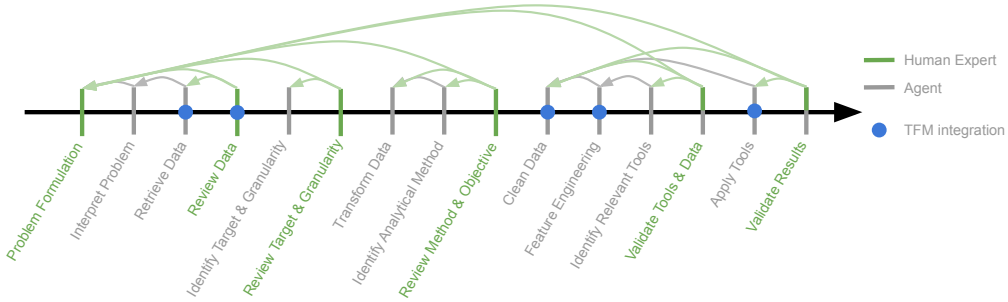

Figure 1: An example of what a data science workflow could look like. Different operations are executed by the human expert and agent, some with the help of tabular foundation models.

problems. This expertise broadly falls into two categories: domain expertise and data science expertise.

Domain Expertise. The knowledge to solve data-related problems is often only available in the institutional knowledge of teams and humans and is usually not documented (and thus not machine accessible) at all. Domain experts are thus critically required to help resolve ambiguity, uncover implicit assumptions, and adapt to evolving goals, tasks that rely on human insight, especially when important context is missing from the data. For example, only a local expert might know that in some radiology departments, but not in others, reports on follow-up scans omit previously-known findings [135]—a fact not encoded in structured data or documentation. Domain experts are also essential for interpreting outcomes, aligning analyses with undocumented clinical goals, or catching when something simply does not "feel right." These nuanced judgments rely on intuitions and tacit domain knowledge that current systems cannot capture [120].

Data Science Expertise. Contextual intelligence is crucial for effective and responsible data science. Data scientists are adept at framing the right problem given the organizational and societal context, selecting tools and methods that are both valid for the data at hand and appropriate for the task's objectives. Unlike automated systems, they can assess whether results make sense, especially in performance-critical or ethically sensitive settings [26]. For instance, a model predicting hospital readmissions may appear statistically accurate, but a data scientist might flag it as untrustworthy if it relies on biased care patterns, such as historically underserved patients receiving less follow-up, not because they were healthier, but due to systemic inequities. Moreover, they understand real-world deployment constraints, from latency requirements to interpretability, and can steer solutions accordingly. These capabilities do and should augment semi-automated systems [138].

4 Realizing the Potential of Data Science with Agents, TFMs, and Humans

LLM agents, TFMs, and human experts, when combined in tight collaboration, form a highly promising "architecture" for collaborative data science systems, covering a full spectrum of capabilities, shown in Table 1. Humans contribute intent, domain knowledge, and critical judgment for goal alignment. TFMs offer structure-aware understanding and statistical reasoning. Agents serve as orchestrators, able to plan and call tools and converse with humans. Any two without the other would leave critical gaps, be it a lack of alignment, understanding, or scalable automation of multi-step workflows. In the following, we discuss why and how to bring them together.

4.1 Synergies of Bringing Agents, TFMs, and Humans Together

Collaboration between human experts and TFM-equipped agents has the potential to elevate data science well beyond the capabilities of individual components. Collaborating with TFM-equipped agents can accelerate many time-consuming steps in data science. For instance, laborious data work can be augmented by agents leveraging specialized TFMs for data wrangling [43, 86, 98, 134] or data analysis [103]. Predictive TFMs like TabPFN [64, 65] alleviate the need for extensive hyperparameter tuning and modeling work for predictive tasks. Thus, agents applying TFMs can elevate human roles, enabling data scientists to create more value and empowering domain experts to contribute directly.

Unlocking Data Scientists for Critical Work. With LLM agents and TFMs handling more repetitive work, data scientists can dedicate their efforts to higher-value activities. Their expertise remains crucial in framing the problem by identifying the right questions and requisite data [11], especially since agents can misinterpret task details, leading to incorrect results [67]. Oversight is also vital for steering workflows, critically evaluating intermediate results, and ensuring alignment with project goals and real-world complexities. This is particularly important given agents' documented struggles with precise instruction adherence, task memory, and strategic planning in dynamic settings [67, 147]. The dynamic and uncertain nature of real-world analysis necessitates effective human-agent interaction, where data scientists navigate ambiguous user intents and adjust strategies based on intermediate findings [87]. Thus, the collaboration allows and requires data scientists to shift their focus towards scrutinizing the data science workflow through deeper critical thinking, robust and continuous validation, considering ethical consequences, and maintaining overall project integrity, which are all factors observed as major determinants of why data science projects succeed or fail to deliver long-term value [117, 141, 11]. Consequently, data scientists can not only increase their output but also significantly improve the quality and real-world impact of their work.

Empowering Domain Experts. TFM-equipped agents offer domain experts an accessible interface, potentially through natural language interactions or guided workflows, to directly engage in data analysis themselves [98]. Domain experts can directly contribute deep contextual knowledge and ensure practical alignment. To this end, they have to clearly articulate the problem, validate the agent's understanding & approach against their domain-specific knowledge, and critically assess whether the results align with real-world objectives and constraints. Elevating domain experts to direct collaborators minimizes divergence from project requirements and physical realities that are common in traditional settings where data scientists serve as intermediaries [141, 68]. Additionally, collaboration with TFM-equipped agents democratizes data analysis, enabling domain experts to perform limited data analysis work themselves, even in organizations without dedicated data scientists.

4.2 Operationalizing the Data Science workflow with TFM-equipped agents

Orchestration along abstract semantic operations. We envision that workflows are *dynamically* orchestrated across abstract semantic operations in data science processes, from problem formulation and data retrieval, to result validation. This allows for flexibility and adaptability across diverse tasks and domains, while enabling the optimization of individual operations and the deployment of context-dependent guardrails for limiting risks. Informed by discussions with data science experts, we have distilled a potential set of data science operations for these systems as illustrated in Figure 1, in line with preliminary systems [67, 54, 103] but with more explicit integration of human experts.

Adapting autonomy to operation and use case. The level of autonomy of the agent may vary based on the specific operation and use case. For example, the retrieval of relevant data (tables, documents, etc.) requires knowledge of the domain context and exploration of heterogeneous data sources [85], necessitating agents with a high level of autonomy, hence requiring a less constrained action space. In contrast, operations and use cases with higher risks require strong guidelines and involvement of human expertise. An example of a high-risk operation is the selection of the analysis tool or TFM for a task and dataset at hand. Benchmarks have shown the tendency of LLMs to take shortcuts and skip human input where needed in tabular data analysis tasks [144, 121]. For instance in [89], an LLM sneakily guesses a causal relationship from parametric knowledge instead of using the provided data.

Tool mapping and human oversight. To ensure responsible operation on high-risk tasks, agents should be constrained with a specified tool mapping and be verifiable and augmentable by humans. Dedicated primitives should explicitly handle operations like causal inference or bias checking during data preparation stages, thus enforcing methodological rigor and consistency. Similarly, humans can be integrated by mapping their interactions to certain primitives that describe how and when human experts should be involved (e.g., to disambiguate the meaning of data). Besides constraining the agent's action-space and thus enable better and more constrained reasoning, human data scientists as well as domain experts are key in evaluating the validity and appropriateness of the tool and model selection for the insight needs at hand. Yet, despite many calls for human oversight, the lack of a systematic design pattern specifying *when* or *how* human experts should be integrated into the workflow highlights the need for further research and practical solutions.

5 Risks to Navigate

The integration of LLM agents into end-to-end data science workflows raises significant technical, ethical, and operational concerns. As these agents increasingly automate tasks such as data processing, analysis, and even code execution, their limitations and failure modes must be critically examined.

Expertise and Automation Bias. LLM agents risk marginalizing human experts through automation bias, the tendency to over-trust model outputs [49]. Recent evidence shows AI assistants can be "detrimental to human skill learning and expertise" [92], as automated systems reduce opportunities for practice and skill development. This is already happening: predictions of radiology automation led to decreased training investments despite the continued need for human expertise [105]. Additionally, the human-AI gap creates risks in both directions: domain experts without statistical training may misuse AI tools, while data scientists may miss critical domain context.

Model Misuse and Data Bias. Data processing workflows are susceptible to subtle but critical errors such as data leakage [76, 75, 50]. Although there is no direct evidence that LLM agents are more prone to these errors, a lack of knowledge of the provenance of the data and the compartmentalization of agentic workflows might increase the associated risks. The ethical and statistical hazards of using AI systems trained on biased or poorly understood datasets raise concerns about social discrimination and data misuse [36], particularly without domain-informed oversight.

Security and Execution Risks. Code-generating agents that execute SQL or Python pose significant safety risks. LLMs have been shown to generate vulnerable or harmful code even under seemingly benign prompts [102], which could be made much worse through new techniques such as training data poisoning [7]. The use of sandboxed execution environments and access management systems, adversarial testing [16], and specific delimitations of sensitive operations 4 are essential.

Privacy, Memorization. LLMs are known to memorize rare or sensitive data from their training corpora [57], which can be extracted with targeted prompting [19]. This raises serious concerns about compliance with privacy regulations such as GDPR and CCPA.[3] An agent who decides to train a predictive model on private data might then cause unexpected private data leakage.

Behavioral Pathologies. Reward hacking and sycophantic behavior has been regularly observed in reinforcement learning based agents [122, 104, 80, 8]. In data science, such behavior could lead to hard-to-detect test set leakage, multiple comparison gaming to generate fake insights, or suggesting complex and wasteful models to unsuspecting users.

Sustainability and Resource Cost. Finally, agentic workflows typically iterate trials and errors, requiring extensive inference cycles. This markedly increases computation costs and energy consumption. For context, a frontier model such as o3 or DeepSeek-R1 [28] is estimated to consume over 33 Wh per long prompt [71], and agentic workflows multiply this footprint through repeated queries. As usage scales, the environmental impact of these systems becomes increasingly untenable. Required infrastructure lead to concentrated resources worrisome from a political economy standpoint [136].

6 Research Agenda and Call to Action

The convergence of agentic systems, tabular foundation models, and human expertise offers an unprecedented opportunity to fundamentally reshape data science. Yet, safely realizing this vision demands a targeted and interdisciplinary research agenda—one that acknowledges the unique needs of data-centric workflows, prioritizes meaningful human-agent collaboration, and reorients the field toward long-term value creation rather than superficial benchmark gains. We outline below the concrete research directions we believe are necessary to enable this transition.

6.1 Evolving the Core Pillars of Modern Data Science

Agenda for Agents. A lot of the issues currently hindering data-science agents stem from more general issues plaguing agents, including long horizon planning, very high reliability and error correction to stay on rails, better tool use integration etc. These issues are the object of intense focus in academia and AI companies, and we believe there will be rapid progress on them. As described in section 4, we also believe that designing new high-level primitives for data science would make

[3] `https://gdpr-info.eu` and `https://oag.ca.gov/privacy/ccpa`

these agentic workflows much more efficient and safe. To this end, research on building agents optimized for interacting with humans, and deferring to humans or asking for clarification in critical moments is key. This involves accurately modeling user goals and expert knowledge [125], ideally in transparent and configurable ways [21], and developing efficient strategies to elicit such information while minimizing the user's cognitive and time burden [99, 79].

Agenda for Tabular Foundation Models. Previous work [133] already highlighted important research directions for predictive TFMs. While such models have improved quickly since then, we believe progress in two main areas is still essential: scalability and contextualization. Predictive TFMs such as TabPFN [65] and representation-focused models like TaPAs [63] are currently strictly constrained on the number of input data points and features, and incur high inference costs. To address these issues, there are several promising techniques that warrant deeper study, including architectural innovations like linear attention [146], retrieval-augmented methods and fine-tuning [131], prompt tuning approaches [39], or the use of hypernetworks [96]. Another severe constraint is that SOTA predictive TFMs like TabPFN [65] are purely statistical reasoners, without any contextual knowledge about the task, dataset or the semantics of columns and values. To close this gap, TFMs need to be able to ingest much more diverse inputs, in particular text descriptions, annotations of data and tasks, contextual metadata, and expert knowledge, for instance in the form of elicited probabilistic or causal priors. Moving in this direction requires bridging the gap between predictive TFMs, often trained on synthetic data with strong numerical signals, and representation TFMs, trained on contextually rich data, potentially through training on more diverse tables [91] or knowledge graphs [77]. Both model types have complementary strengths to offer: predictive TFMs would benefit from improved contextual understanding, while representation TFMs would benefit from a better understanding of numerical values and patterns [66], as well as missing tabular specific biases [23].

6.2 Building and Evaluating End-to-End Systems

Building an end-to-end system for assisted data science poses many challenges beyond agents and TFMs. Scoping, goal setting, evaluation and verifiability are aspects of the full system that are not addressed by research on agents or TFM themselves.

Beyond Prediction. Data science is more than data preparation and predictive modeling. Data science can be split in three groups of tasks [62]: (a) exploration, description, and hypothesis generation, i.e., getting to know your data and formulating what the questions of interest are, (b) answering causal questions about effects in the data, and (c) predictive tasks, usually as part of an automated decision-making process. A critical part of each of these is formulating the problem, identifying which tools are appropriate, and mapping the concepts required to apply the tool to the data. This crucial step is often assumed solved in machine learning benchmarks, which usually come with clearly defined tasks and metrics. The machine learning community so far has focused on selecting prediction methods when given a clear task, mapping the data schema to the task, and building ample benchmarks for this task [14, 52, 47]. While there are definitions for insight generation [31, 103], there is no accepted definition or metric for Exploratory Data Analysis tasks, though empirical benchmarks exist [58]. Benchmarks for causal inference usually assume well-phrased and grounded questions and are typically limited in scope. Mapping to the correct task and validating the choice of method has received little to no attention in all areas. With this broadened scope, it is usually impossible to simply rank algorithms based on their performance given the data, as the "correct" answer might be highly context-dependent and potentially unknowable.

More Realistic Evaluations. Current tabular benchmarks have significant discrepancies with real-life settings. They focus on well-framed problems, and often neglect important aspects like distribution shift across time [115], data-specific data preprocessing [115, 132], or finding the relevant data [18]. To evaluate agentic systems, benchmarks containing tables with varying levels of human preprocessing are important to understand the performance of the entire pipeline. In practice, data is usually part of multiple tables in a larger database schema [33], and selecting and aggregating the correct data is an everyday task for most data scientists. However, only a few benchmarks reflect this reality today [114, 147, 73]. This relational multi-table setup could be addressed within TFMs themselves or by the integration of TFMs with other agent tools like Text-to-SQL [83]. Furthermore, benchmarks should measure performance on realistic context-rich tables on which a mix of agents and TFMs can shine by retrieving useful information from parametric knowledge, databases, or knowledge graphs [147, 73, 72]. Adding human interactions into the workflow further complicates evaluation. New evaluation metrics and evaluation protocols, including varying degrees of human interaction

are needed. Such human-in-the-loop evaluations have been instrumental in biosecurity evaluating risks [101], and have been the gold standard of visualization and interpretability research [112, 143]. For tabular data analysis, the CollaborativeGym benchmark [121] provides a compelling starting point for such benchmarks, but work on much broader tasks is necessary.

Embracing a New Flexibility of Goals. TFMs based on the principles of PFNs [97] enable addressing previously separate tasks in a unified framework. Access to the causal mechanism during pretraining allows for new training objectives that are more aligned with practical requirements and can be easily adjusted to suit new scenarios. An example of exploring this newfound flexibility are the drift-resilient TabPFN [61] which is robust to distribution shifts by using the known causal model during pretraining, and FairPFN [113] which addresses counterfactual fairness. AVICI [90] goes even further and infers the causal graph in-context, allowing for complex inferences. These are just the starting points for training models to answer more general inference questions, including causal ones, even outside of the usual i.i.d. framework. TFMs are also able to produce models of a given architecture using in-context learning [96] focusing on interpretable model families, such as Generalized Additive Models [95]. These constrained architectures enforce white-box prediction functions and could extend to domain-specific families meaningful to experts. While TFMs can widen the range of scenarios we can address computationally, agents can help in selecting appropriate methods, validate assumptions, and guard against common mistakes, such as target leakage or non-i.i.d. data. Without the semantic understanding provided by LLM-based agentic systems, numerical tools alone can not detect these failure cases [62].

End-to-end Verifiability. To ensure trustworthy and aligned outcomes any computation and reasoning done through TFM or agent needs to be fully transparent to the human user. Only when human experts can review the data, context, and reasoning used for a conclusion can they fully trust and build upon the results. While there is a long line of work on data provenance [51, 48, 22, 107, 50] and LLM reasoning [5, 24, 10], as well as a growing interest in interpreting TFMs [116], combining these to produce explanations of end-to-end systems that are faithful and verifiable is an open challenge.

Deployment. Many challenges in data science projects only arise during deployment when transitioning from the well-defined development environment to the brittle physical world [117, 141]. Therefore, the collaborative framework should extend to deployment, accounting for monitoring deployed models, detecting drifts, and model updates. Novel capabilities, such as the ability of agentic systems to re-use pre-trained TFMs via in-context learning in different applications may open up opportunities to simplify deployment infrastructure.

7 Alternative Views

Alternative view 1: *LLM-powered agentic systems will be enough. There are so many people working on LLMs, LLMs will learn to reason better than TFMs soon, rendering TFMs unnecessary.*
Rebuttal: LLM architectures designed for sequences are inherently misaligned with the concept of columns in a table; thus, standard LLMs cannot be as efficient for tables as the specially adapted architectures in TFMs. Concepts from LLMs will be extremely important, but adapted for TFMs.

Alternative view 2: *We should entirely automate data science. This is already a reality [53, 130].*
Rebuttal: These works demonstrate that it is feasible to build fully automated *predictive* data science agents; yet, we argue that this risks jeopardizing the integrity and quality of results in critical tasks and may thus outweigh their benefits. Hence, we advocate to supercharge humans who can bypass the tedious steps and focus on reviewing and validating the steps taken. Also, data science is much more than prediction, and the remaining tasks are much harder to automate fully.

8 Conclusion

The future of data science lies in a collaborative paradigm that integrates Tabular Foundation Models, agentic systems, and human expertise. TFMs provide structure-aware statistical reasoning, agents enable orchestration and accessibility, and humans ensure contextual understanding and critical oversight. This synergy can overcome current workflow fragmentation and unlock greater efficiency and quality, but must be guided carefully to avoid risks like automation bias and model misuse. We call for research into scalable, context-aware TFMs, robust human-agent collaboration, realistic benchmarks, and transparent, verifiable systems. Rather than replacing data scientists, these tools should amplify their impact.

Acknowledgements

This paper builds on discussions held during the Dagstuhl Seminar 25182 "Challenges and Opportunities of Table Representation Learning". MH, DG, and XL acknowledge support from a grant from NWO (grant NGF.1607.22.045) and a gift from SAP. LP acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under SFB 1597 (SmallData), grant number 499552394. FH acknowledges the financial support of the Hector Foundation. MC is partially funded by the Elsevier Discovery Lab, partially funded by the Graph-Massivizer project, funded by the Horizon Europe programme of the European Union (grant 101093202), and supported by a gift from Accenture LLP. His work on this publication is in part based upon work from COST Action CA23147 GOBLIN - Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu).

References

[1] Judith Abécassis, Élise Dumas, Julie Alberge, and Gaël Varoquaux. From prediction to prescription: Machine learning and causal inference for the heterogeneous treatment effect. *Annual Review of Biomedical Data Science*, 8, 2025.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Anastasia Ailamaki, Samuel Madden, Daniel Abadi, Gustavo Alonso, Sihem Amer-Yahia, Magdalena Balazinska, Philip A Bernstein, Peter Boncz, Michael Cafarella, Surajit Chaudhuri, et al. The cambridge report on database research. *arXiv preprint arXiv:2504.11259*, 2025.

[4] Ahmed Al Kuwaiti, Khalid Nazer, Abdullah Al-Reedy, Shaher Al-Shehri, Afnan Al-Muhanna, Arun Vijay Subbarayalu, Dhoha Al Muhanna, and Fahad A Al-Muhanna. A review of the role of artificial intelligence in healthcare. *Journal of personalized medicine*, 13(6):951, 2023.

[5] Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, et al. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025.

[6] Gilbert Badaro, Mohammed Saeed, and Paolo Papotti. Transformers for tabular data representation: A survey of models and applications. *Transactions of the Association for Computational Linguistics*, 11:227–249, 2023.

[7] Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1505–1521. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/bagdasaryan.

[8] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation, 2025. URL https://arxiv.org/abs/2503.11926.

[9] Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, 77(2):81–97, 2008.

[10] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for AI safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

[11] Lucas Bernardi, Themistoklis Mavridis, and Pablo Estevez. 150 successful machine learning models: 6 lessons learned at booking.com. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1743–1751, 2019.

[12] Jan-Micha Bodensohn, Ulf Brackmann, Liane Vogel, Anupam Sanghi, and Carsten Binnig. Unveiling challenges for LLMs in enterprise data engineering. *arXiv preprint arXiv:2504.10950*, 2025.

[13] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[14] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE transactions on neural networks and learning systems*, 2022.

[15] Eric T Bradlow, Manish Gangwar, Praveen Kopalle, and Sudhir Voleti. The role of big data and predictive analytics in retailing. *Journal of retailing*, 93(1):79–95, 2017.

[16] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. Toward trustworthy ai development: Mechanisms for supporting verifiable claims. Workingpaper, arXiv, United States, April 2020. URL https://arxiv.org/abs/2004.07213. Version 1 uploaded 15th April 2020, Version 2 uploaded 20th April 2020.

[17] Michael Cafarella, Ihab F Ilyas, Marcel Kornacker, Tim Kraska, and Christopher Ré. Dark data: Are we solving the right problems? In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 1444–1445. IEEE, 2016.

[18] Riccardo Cappuzzo, Aimee Coelho, Felix Lefebvre, Paolo Papotti, and Gael Varoquaux. Retrieve, merge, predict: Augmenting tables with data lakes, May 2024.

[19] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

[20] Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, et al. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*, 2025.

[21] Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, Martin Wattenberg, and Fernanda Viégas. Designing a dashboard for transparency and control of conversational AI, October 2024.

[22] Zaheer Chothia, John Liagouris, Frank McSherry, and Timothy Roscoe. Explaining outputs in modern data analytics. Technical report, ETH Zurich, 2016.

[23] Tianji Cong, Madelon Hulsebos, Zhenjie Sun, Paul Groth, and H. V. Jagadish. Observatory: Characterizing embeddings of relational tables. *Proceedings of the VLDB Endowment*, 17(4): 849–862, December 2023. ISSN 2150-8097. doi: 10.14778/3636218.3636237.

[24] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

[25] Tamraparni Dasu and Theodore Johnson. *Exploratory data mining and data cleaning*. John Wiley & Sons, 2003.

[26] T De Bie, L De Raedt, J Hernández-Orallo, HH Hoos, P Smyth, and CKI Williams. Automating data science: Prospects and challenges. 65 (3), 76–87, 2021.

[27] Tijl De Bie, Luc De Raedt, José Hernández-Orallo, Holger H. Hoos, Padhraic Smyth, and Christopher K. I. Williams. Automating data science. *Commun. ACM*, 65(3):76–87, February 2022. ISSN 0001-0782. doi: 10.1145/3495256. URL https://doi.org/10.1145/3495256.

[28] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[29] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. TURL: table understanding through representation learning. *Proc. VLDB Endow.*, 14(3):307–319, 2020. doi: 10.5555/3430915.3442430. URL http://www.vldb.org/pvldb/vol14/p307-deng.pdf.

[30] Yue Deng, Feng Bao, Youyong Kong, Zhiquan Ren, and Qionghai Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE transactions on neural networks and learning systems*, 28(3):653–664, 2016.

[31] Rui Ding, Shi Han, Yong Xu, Haidong Zhang, and Dongmei Zhang. Quickinsights: Quick and automatic discovery of insights from multi-dimensional data. In *Proceedings of the 2019 international conference on management of data*, pages 317–332, 2019.

[32] Matthew F Dixon, Igor Halperin, and Paul Bilokon. *Machine learning in finance*, volume 1170. Springer, 2020.

[33] Till Döhmen, Radu Geacu, Madelon Hulsebos, and Sebastian Schelter. SchemaPile: A large collection of relational database schemas. *Proceedings of the ACM on Management of Data*, 2 (3):1–25, 2024.

[34] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[35] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data, 2020. URL https://arxiv.org/abs/2003.06505.

[36] Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, Inc., USA, 2018. ISBN 1250074312. URL https://dl.acm.org/doi/10.5555/3208509.

[37] Wenfei Fan and Floris Geerts. *Foundations of Data Quality Management*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2012.

[38] Raul Castro Fernandez, Pranav Subramaniam, and Michael J. Franklin. Data market platforms: Trading data assets to solve data problems. *Proc. VLDB Endow.*, 13(11):1933–1947, 2020.

[39] Benjamin Feuer, Robin Tibor Schirrmeister, Valeriia Cherepanova, Chinmay Hegde, Frank Hutter, Micah Goldblum, Niv Cohen, and Colin White. TuneTables: Context optimization for scalable prior-data fitted networks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FOfU3qhcIG.

[40] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/11d0e6287202fced83f79975ec59a3a6-Paper.pdf.

[41] Robert Fildes, Paul Goodwin, Michael Lawrence, and Konstantinos Nikolopoulos. Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International journal of forecasting*, 25(1):3–23, 2009.

[42] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

[43] Haotian Gao, Shaofeng Cai, Tien Tuan Anh Dinh, Zhiyong Huang, and Beng Chin Ooi. CtxPipe: Context-aware data preparation pipeline construction for machine learning. *Proceedings of the ACM on Management of Data*, 2(6):1–27, 2024.

[44] Yunfan Gao, Yun Xiong, Yijie Zhong, Yuxi Bi, Ming Xue, and Haofen Wang. Synergizing RAG and reasoning: A systematic review. *arXiv preprint arXiv:2504.15909*, 2025.

[45] Josh Gardner, Juan C. Perdomo, and Ludwig Schmidt. Large scale transfer learning for tabular data via language modeling, 2024. URL https://arxiv.org/abs/2406.12031.

[46] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019.

[47] Pieter Gijsbers, Marcos LP Bueno, Stefan Coors, Erin LeDell, Sébastien Poirier, Janek Thomas, Bernd Bischl, and Joaquin Vanschoren. AMLB: an AutoML benchmark. *Journal of Machine Learning Research*, 25(101):1–65, 2024.

[48] Boris Glavic and Gustavo Alonso. Perm: Processing provenance and data on the same data model through query rewriting. In *2009 IEEE 25th International Conference on Data Engineering*, pages 174–185. IEEE, 2009.

[49] Kate Goddard, Abdul Roudsari, and Jeremy Wyatt. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association : JAMIA*, 19:121–7, 06 2011. doi: 10.1136/amiajnl-2011-000089. URL https://pubmed.ncbi.nlm.nih.gov/21685142/.

[50] Stefan Grafberger, Paul Groth, Julia Stoyanovich, and Sebastian Schelter. Data distribution debugging in machine learning pipelines. *The VLDB Journal*, 31(5):1103–1126, 2022.

[51] Todd J Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 31–40, 2007.

[52] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

[53] Antoine Grosnit, Alexandre Maraval, James Doran, Giuseppe Paolo, Albert Thomas, Refinath Shahul Hameed Nabeezath Beevi, Jonas Gonzalez, Khyati Khandelwal, Ignacio Iacobacci, Abdelhakim Benechehab, Hamza Cherkaoui, Youssef Attia El-Hili, Kun Shao, Jianye Hao, Jun Yao, Balazs Kegl, Haitham Bou-Ammar, and Jun Wang. Large language models orchestrating structured reasoning achieve kaggle grandmaster level, 2024. URL https://arxiv.org/abs/2411.03562.

[54] Yang Gu, Hengyu You, Jian Cao, Muran Yu, Haoran Fan, and Shiyou Qian. Large language models for constructing and optimizing machine learning workflows: A survey. *arXiv preprint arXiv:2411.10478*, 2024.

[55] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[56] Alon Halevy, Anand Rajaraman, and Joann Ordille. Data integration: The teenage years. In *Proceedings of the 32nd international conference on Very large data bases*, pages 9–16, 2006.

[57] Abhimanyu Hans, John Kirchenbauer, Yuxin Wen, Neel Jain, Hamid Kazemi, Prajwal Singhania, Siddharth Singh, Gowthami Somepalli, Jonas Geiping, Abhinav Bhatele, and Tom Goldstein. Be like a goldfish, don't memorize! Mitigating memorization in generative LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=DylSyAfmWs.

[58] Xinyi He, Mengyu Zhou, Xinrun Xu, Xiaojun Ma, Rui Ding, Lun Du, Yan Gao, Ran Jia, Xu Chen, Shi Han, Zejian Yuan, and Dongmei Zhang. Text2Analysis: A benchmark of table question answering with advanced data analysis and unclear queries. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18206–18215. AAAI Press, 2024.

[59] James B Heaton, Nick G Polson, and Jan Hendrik Witte. Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1):3–12, 2017.

[60] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5549–5581. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/hegselmann23a.html.

[61] Kai Helli, David Schnurr, Noah Hollmann, Samuel Müller, and Frank Hutter. Drift-resilient tabPFN: In-context learning temporal distribution shifts on tabular data. *Advances in Neural Information Processing Systems*, 37:98742–98781, 2024.

[62] Miguel A Hernán, John Hsu, and Brian Healy. A second chance to get causal inference right: A classification of data science tasks. *Chance*, 32(1):42–49, 2019.

[63] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. TaPas: Weakly supervised table parsing via pre-training. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.398. URL https://aclanthology.org/2020.acl-main.398/.

[64] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023.

[65] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, January 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08328-6.

[66] Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. The Tabular Foundation Model TabPFN Outperforms Specialized Time Series Forecasting Models Based on Simple Features. https://arxiv.org/abs/2501.02945v2, January 2025.

[67] Yiming Huang, Jianwen Luo, Yan Yu, Yitong Zhang, Fangyu Lei, Yifan Wei, Shizhu He, Lifu Huang, Xiao Liu, Jun Zhao, and Kang Liu. DA-Code: Agent data science code generation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, October 2024. doi: 10.18653/v1/2024.emnlp-main.748.

[68] Madelon Hulsebos, Wenjing Lin, Shreya Shankar, and Aditya Parameswaran. It took longer than i was expecting: Why is dataset search still so hard? In *Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics*, pages 1–4, 2024.

[69] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, editors. *Automated Machine Learning - Methods, Systems, Challenges*. Springer, 2019.

[70] Ihab F. Ilyas and Xu Chu. *Data Cleaning*, volume 28 of *ACM Books*. ACM, 2019.

[71] Nidhal Jegham, Marwen Abdelatti, Lassad Elmoubarki, and Abdeltawab Hendawi. How hungry is AI? Benchmarking energy, water, and carbon footprint of LLM inference, 2025. URL https://arxiv.org/abs/2505.09598.

[72] Xingyu Ji, Parker Glenn, Aditya G. Parameswaran, and Madelon Hulsebos. TARGET: Benchmarking table retrieval for generative tasks, May 2025.

[73] Liqiang Jing, Zhehui Huang, Xiaoyang Wang, Wenlin Yao, Wenhao Yu, Kaixin Ma, Hongming Zhang, Xinya Du, and Dong Yu. DSBench: How far are data science agents from becoming data science experts?, April 2025.

[74] Mayur P Joshi, Ning Su, Robert D Austin, and Anand K Sundaram. Why so many data science projects fail to deliver. *MIT Sloan Management Review*, 62(3), 2021.

[75] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, August 2023. ISSN 2666-3899. doi: 10.1016/j.patter.2023.100804. URL https://www.cell.com/patterns/abstract/S2666-3899(23)00159-9. Publisher: Elsevier.

[76] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data*, 6(4), December 2012. ISSN 1556-4681. doi: 10.1145/2382577.2382579. URL https://doi.org/10.1145/2382577.2382579.

[77] Myung Jun Kim, Léo Grinsztajn, and Gaël Varoquaux. CARTE: Pretraining and transfer for tabular learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 23843–23866, Vienna, Austria, July 2024. JMLR.org.

[78] Myung Jun Kim, Félix Lefebvre, Gaëtan Brison, Alexandre Perez-Lebel, and Gaël Varoquaux. Table foundation models: on knowledge pre-training for tabular learning, 2025. URL https://arxiv.org/abs/2505.14415.

[79] Neville K. Kitson and Anthony C. Constantinou. Causal discovery using dynamically requested knowledge. *Knowledge-Based Systems*, 314:113185, 2025. doi: 10.1016/j.knosys.2025.113185.

[80] Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: The flip side of AI ingenuity. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/, 2020.

[81] Clemens Scott Kruse, Rishi Goswamy, Yesha Jayendrakumar Raval, and Sarah Marawi. Challenges and opportunities of big data in health care: A systematic review. *JMIR medical informatics*, 4(4):e5359, 2016.

[82] Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246, 2002.

[83] Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C. C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. Can LLM Already Serve as A Database Interface? A BIg Bench for Large-Scale Database Grounded Text-to-SQLs, November 2023.

[84] Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. Table-GPT: Table fine-tuned GPT for diverse table tasks. *Proc. ACM Manag. Data*, 2(3):176, 2024. doi: 10.1145/3654979. URL https://doi.org/10.1145/3654979.

[85] Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. Chain-of-Knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*, 2024.

[86] Xue Li and Till Döhmen. Towards efficient data wrangling with LLMs using code generation. In *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning*, pages 62–66, 2024.

[87] Ziming Li, Qianbo Zang, David Ma, Jiawei Guo, Tuney Zheng, Minghao Liu, Xinyao Niu, Yue Wang, Jian Yang, Jiaheng Liu, Wanjun Zhong, Wangchunshu Zhou, Wenhao Huang, and Ge Zhang. AutoKaggle: A multi-agent framework for autonomous data science competitions, November 2024.

[88] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

[89] Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. Are LLMs capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9215–9235, Bangkok, Thailand and virtual meeting, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.548.

[90] Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. *Advances in Neural Information Processing Systems*, 35:13104–13118, 2022.

[91] Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Hamidreza Kamkari, Alex Labach, Jesse C Cresswell, Keyvan Golestan, Guangwei Yu, Maksims Volkovs, and Anthony L Caterini. TabDPT: Scaling tabular foundation models. *arXiv preprint arXiv:2410.18164*, 2024.

[92] Brooke N. Macnamara, Ibrahim Berber, M. Cenk Çavuşoğlu, Elizabeth A. Krupinski, Naren Nallapareddy, Noelle E. Nelson, Philip J. Smith, Amy L. Wilson-Delfosse, and Soumya Ray. Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? *Cognitive Research: Principles and Implications*, 9:46, 2024. doi: 10.1186/s41235-024-00572-8. URL https://doi.org/10.1186/s41235-024-00572-8.

[93] Mohammad Mahdavi, Ziawasch Abedjan, Raul Castro Fernandez, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, and Nan Tang. Raha: A configuration-free error detection system. In *Proceedings of the 2019 International Conference on Management of Data*, pages 865–882, 2019.

[94] James Manyika. Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, 1, 2011.

[95] Andreas C Mueller, Julien Siems, Harsha Nori, David Salinas, Arber Zela, Rich Caruana, and Frank Hutter. GAMformer: Exploring in-context learning for generalized additive models. In *NeurIPS 2024 Third Table Representation Learning Workshop*, 2024.

[96] Andreas C Mueller, Carlo A Curino, and Raghu Ramakrishnan. MotherNet: Fast training and inference via hyper-network transformers. In *International Conference on Learning Representations*, 2025.

[97] Samuel Müller, Noah Hollmann, Sebastian Pineda-Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[98] Avanika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. Can foundation models wrangle your data? *Proceedings of the VLDB Endowment*, 16(4):738–746, December 2022. ISSN 2150-8097. doi: 10.14778/3574245.3574258.

[99] Anthony O'Hagan, Catharine E. Buck, Alireza Daneshkhah, Jim Eiser, Paul H. Garthwaite, David J. Jenkinson, Jeremy E. Oakley, and Tim Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons, Chichester, UK, 2006.

[100] Simone Papicchio, Paolo Papotti, and Luca Cagliero. QATCH: benchmarking SQL-centric tasks with table representation learning models on your data. In *Advances in Neural Information Processing Systems*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/62a24b69b820d30e9e5ad4f15ff7bf72-Abstract-Datasets_and_Benchmarks.html.

[101] Tejal Patwardhan. Building an early warning system for LLM-aided biological threat creation. https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/, February 2024.

[102] Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. Asleep at the keyboard? assessing the security of GitHub Copilot's code contributions. *Commun. ACM*, 68(2):96–105, January 2025. ISSN 0001-0782. doi: 10.1145/3610721. URL https://doi.org/10.1145/3610721.

[103] Alberto Sánchez Pérez, Alaa Boukhary, Paolo Papotti, Luis Castejón Lozano, and Adam Elwood. An LLM-based approach for insight generation in data analysis. In *Proc. of the Association for Computational Linguistics*, pages 562–582. ACL, April 2025. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.24/.

[104] Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847/.

[105] Filippo Pesapane, Mattia Codari, and Francesco Sardanelli. Artificial intelligence in medical imaging: threat or opportunity? radiologists again at the forefront of innovation in medicine. *European Radiology Experimental*, 2(1):1–10, 2018. doi: 10.1186/s41747-018-0061-6. URL https://eurradiolexp.springeropen.com/articles/10.1186/s41747-018-0061-6.

[106] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. Data management challenges in production machine learning. In *Proceedings of the 2017 ACM international conference on management of data*, pages 1723–1726, 2017.

[107] Fotis Psallidas and Eugene Wu. Smoke: Fine-grained lineage at interactive speed. *arXiv preprint arXiv:1801.07237*, 2018.

[108] Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data, February 2025.

[109] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[110] Erhard Rahm, Hong Hai Do, et al. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[111] Kumo.AI Research. KumoRFM: A foundation model for in-context learning on relational data. https://kumo.ai/research/kumo_relational_foundation_model.pdf, 2025. Accessed: 2025-05-21.

[112] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "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, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016.

[113] Jake Robertson, Noah Hollmann, Noor Awad, and Frank Hutter. FairPFN: A tabular foundation model for causal fairness. 2025.

[114] Joshua Robinson, Rishabh Ranjan, Weihua Hu, Kexin Huang, Jiaqi Han, Alejandro Dobles, Matthias Fey, Jan E. Lenssen, Yiwen Yuan, Zecheng Zhang, Xinwei He, and Jure Leskovec. RelBench: A benchmark for deep learning on relational databases, July 2024.

[115] Ivan Rubachev, Nikolay Kartashev, Yury Gorishniy, and Artem Babenko. TabReD: A benchmark of tabular machine learning in-the-wild, August 2024.

[116] David Rundel, Julius Kobialka, Constantin von Crailsheim, Matthias Feurer, Thomas Nagler, and David Rügamer. Interpretable machine learning for TabPFN. In *World Conference on Explainable Artificial Intelligence*, pages 465–476. Springer, 2024.

[117] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "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*, pages 1–15, Yokohama Japan, May 2021. ACM. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445518.

[118] Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. Automating large-scale data quality verification. *Proceedings of the VLDB Endowment*, 11(12):1781–1794, 2018.

[119] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28, 2015.

[120] Shreya Shankar, Rolando Garcia, Joseph M. Hellerstein, and Aditya G. Parameswaran. Operationalizing machine learning: An interview study, September 2022.

[121] Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. Collaborative Gym: A framework for enabling and evaluating human-agent collaboration, January 2025.

[122] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

[123] Liang Shi, Zhengju Tang, Nan Zhang, Xiaotong Zhang, and Zhi Yang. A survey on employing large language models for text-to-sql tasks. *arXiv preprint arXiv:2407.15186*, 2024.

[124] Michael Stonebraker, Ihab F Ilyas, et al. Data integration: The current status and the way forward. *IEEE Data Eng. Bull.*, 41(2):3–9, 2018.

[125] James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295, 2024. doi: 10.1038/s41562-024-01882-z.

[126] Diane M Strong, Yang W Lee, and Richard Y Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, 1997.

[127] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355/.

[128] Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, Haoxuan Lan, Jiaming Tian, Jing Yuan, Junbo Zhao, Junlin Zhou, Kaizhe Shou, Liangyu Zha, Lin Long, Liyao Li, Pengzuo Wu, Qi Zhang, Qingyi Huang, Saisai Yang, Tao Zhang, Wentao Ye, Wufang Zhu, Xiaomeng Hu, Xijun Gu, Xinjie Sun, Xiang Li, Yuhang Yang, and Zhiqing Xiao. TableGPT2: A large multimodal model with tabular data integration, November 2024.

[129] Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, et al. TableGPT2: A large multimodal model with tabular data integration. *arXiv preprint arXiv:2411.02059*, 2024.

[130] Pecan Team. URL https://www.pecan.ai/.

[131] Valentin Thomas, Junwei Ma, Rasa Hosseinzadeh, Keyvan Golestan, Guangwei Yu, Maksims Volkovs, and Anthony Caterini. Retrieval & fine-tuning for in-context tabular models, 2024. URL https://arxiv.org/abs/2406.05207.

[132] Andrej Tschalzev, Sascha Marton, Stefan Lüdtke, Christian Bartelt, and Heiner Stuckenschmidt. A data-centric perspective on evaluating machine learning models for tabular data, August 2024.

[133] Boris Van Breugel and Mihaela Van Der Schaar. Position: Why tabular foundation models should be a research priority. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 48976–48993, Vienna, Austria, July 2024. JMLR.org.

[134] Alexander van Renen, Mihail Stoian, and Andreas Kipf. DataLoom: Simplifying data loading with LLMs. *Proc. VLDB Endow.*, 17(12):4449–4452, 2024. doi: 10.14778/3685800.3685897. URL https://www.vldb.org/pvldb/vol17/p4449-renen.pdf.

[135] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):48, 2022.

[136] Gael Varoquaux, Alexandra Sasha Luccioni, and Meredith Whittaker. Hype, sustainability, and the price of the bigger-is-better paradigm in ai. *arXiv preprint arXiv:2409.14160*, 2024.

[137] Bill Waid. Solving the last mile problem for data science project success, 2019. URL https://www.forbes.com/councils/forbestechcouncil/2019/07/23/solving-the-last-mile-problem-for-data-science-project-success/. (23.07. 2019). Forbes.

[138] Dakuo Wang, Q Vera Liao, Yunfeng Zhang, Udayan Khurana, Horst Samulowitz, Soya Park, Michael Muller, and Lisa Amini. How much automation does a data scientist want? *arXiv preprint arXiv:2101.03970*, 2021.

[139] Kristin Weber, Boris Otto, and Hubert Österle. One size does not fit all — a contingency approach to data governance. *Journal of Data and Information Quality (JDIQ)*, 1(1):1–27, 2009.

[140] Joyce Weiner. *Why AI/data science projects fail: How to avoid project pitfalls*. Springer Nature, 2022.

[141] Jens Westenberger, Kajetan Schuler, and Dennis Schlegel. Failure of AI projects: Understanding the critical factors. *Procedia Computer Science*, 196:69–76, 2022. ISSN 18770509. doi: 10.1016/j.procs.2021.11.074.

[142] Cornelius Wolff and Madelon Hulsebos. How well do LLMs reason over tabular data, really?, 2025. URL https://arxiv.org/abs/2505.07453.

[143] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics*, 22(1):649–658, 2015.

[144] Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. TheAgentCompany: Benchmarking LLM agents on consequential real world tasks, December 2024.

[145] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. TaBERT: Pretraining for joint understanding of textual and tabular data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8413–8426. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.745. URL https://doi.org/10.18653/v1/2020.acl-main.745.

[146] Yuchen Zeng, Wonjun Kang, and Andreas C Mueller. TabFlex: Scaling tabular learning to millions with linear attention. In *NeurIPS 2024 Third Table Representation Learning Workshop*, 2024. URL https://openreview.net/forum?id=f8aganC0tN.

[147] Dan Zhang, Sining Zhoubian, Min Cai, Fengzu Li, Lekang Yang, Wei Wang, Tianjiao Dong, Ziniu Hu, Jie Tang, and Yisong Yue. DataSciBench: An LLM agent benchmark for data science, February 2025.

[148] Alice Zheng and Amanda Casari. *Feature engineering for machine learning: principles and techniques for data scientists*. O'Reilly Media, Inc., 2018.