AgentDS Technical Report:
Benchmarking the Future of Human-AI Collaboration in Domain-Specific Data Science

An Luo[1], Jin Du[1], Xun Xian[2], Robert Specht[1], Fangqiao Tian[1], Ganghua Wang[3],
Xuan Bi[4], Charles Fleming[5], Ashish Kundu[5], Jayanth Srinivasa[5],
Mingyi Hong[2], Rui Zhang[6], Tianxi Li[1], Galin Jones[1], Jie Ding[1]

[1]School of Statistics, University of Minnesota
[2]Department of Electrical and Computer Engineering, University of Minnesota
[3]Data Science Institute, University of Chicago
[4]Carlson School of Management, University of Minnesota
[5]Cisco Research
[6]Division of Computational Health Sciences, University of Minnesota

Abstract

Data science plays a critical role in transforming complex data into actionable insights across numerous domains. Recent developments in large language models (LLMs) and artificial intelligence (AI) agents have significantly automated data science workflow. However, it remains unclear to what extent AI agents can match the performance of human experts on domain-specific data science tasks, and in which aspects human expertise continues to provide advantages. We introduce **AgentDS**, a benchmark and competition designed to evaluate both AI agents and human-AI collaboration performance in domain-specific data science. AgentDS consists of **17 challenges across six industries**: commerce, food production, healthcare, insurance, manufacturing, and retail banking. We conducted an open competition involving **29 teams and 80 participants**, enabling systematic comparison between human–AI collaborative approaches and AI-only baselines. Our results show that current AI agents struggle with domain-specific reasoning. AI-only baselines perform near or below the median of competition participants, while the strongest solutions arise from human–AI collaboration. These findings challenge the narrative of complete automation by AI and underscore the enduring importance of human expertise in data science, while illuminating directions for the next generation of AI. Visit the AgentDS website here and open source datasets here.

1 Introduction

Data science has become central to decision-making across industries, from healthcare diagnostics to financial risk assessment, where it blends statistics, computer science, and domain expertise to transform raw data into actionable insights [1, 2, 3]. Recent advancement of large language models (LLMs) and AI agents demonstrate impressive capabilities in automating code generation and executing regular machine learning tasks [4, 5, 6, 7, 8, 9, 10]. Some systems have even achieved Kaggle Grandmaster performance through structured reasoning [10], while others automate data science workflows [11, 12, 13]. These advances suggest that many routine components of data science workflows may increasingly be automated, reducing the manual burden on human data scientists.

Despite these advances in LLMs and AI agents for data science, a fundamental question remains unanswered: *To what extent do human experts outperform autonomous AI agents on domain-specific*

Preprint.

data science tasks, and in which aspects does this advantage arise? In practice, human data scientists consistently rely on specialized knowledge about data and tasks, incorporating crucial domain-specific nuances that enhance model performance [14, 15, 16, 17, 18]. Such domain-driven decisions are often subtle yet essential, addressing complexities not captured by generic analytics workflows. However, current research on AI for data science has largely focused on generating generic code and pipeline executions [7, 8], often neglecting the domain-specific knowledge needed for real-world problems.

Existing benchmarks for AI agents, while valuable, often do not test whether agentic AI can effectively leverage domain insights outside tabular data [19, 20, 21, 22, 23, 24]. Some recent work has demonstrated that current agentic AI typically generates generic code and pipeline executions, often neglecting the domain-specific knowledge needed for complex real-world problems [7, 18, 25].

Understanding these differences is important for advancing both AI capabilities and human-AI collaboration. To address this gap, we present **AgentDS**, a benchmark comprising 17 challenges across 6 domains, each grounded in realistic industry problems and built on carefully designed synthetic datasets that reward domain-specific insight. The challenges are constructed so that generic pipelines relying only on off-the-shelf algorithms perform poorly, while approaches that incorporate domain-informed feature engineering and data processing achieve substantially better results. To evaluate these dynamics in practice, we organized a 10-day competition involving 29 teams and 80 participants, enabling a systematic comparison between human–AI collaborative solutions and AI-only baselines.

Our inaugural competition reveals three key findings:

1. **Agentic AI struggle with domain-specific reasoning.** Current autonomous agents perform poorly on tasks requiring domain-specific insight, particularly when multimodal signals must be incorporated. In practice, several teams that initially experimented with autonomous agent frameworks ultimately abandoned them in favor of interactive human-guided workflows.

2. **Human expertise remains essential.** Human data scientists consistently contribute capabilities that AI lack, including diagnosing modeling failures, injecting domain knowledge through feature design and domain-specific rules, and making strategic decisions about model selection and generalization.

3. **Human-AI collaboration outperforms either humans or AI alone.** The most successful approaches combine human strategic reasoning with AI-assisted implementation. In these workflows, humans guide the problem-solving process while AI accelerates coding, experimentation, and iteration.

These findings challenge the assumption that advances in agentic AI will soon enable fully autonomous data science. Instead, our results suggest that effective performance on domain-specific tasks continues to rely on human expertise, particularly for problem formulation, domain-specific reasoning, and strategic decision making. AgentDS provides a benchmark for systematically studying these dynamics and highlights the importance of designing systems that support effective human–AI collaboration rather than fully autonomous automation.

The remainder of the paper is organized as follows. Section 2 introduces the AgentDS benchmark, including its design philosophy, dataset curation process, evaluation framework, the competition setup and AI baselines. Section 3 presents empirical findings based on both quantitative results and qualitative analysis of participant submissions. Section 4 discusses limitations and directions for future work. Section 5 concludes the paper.

2 The AgentDS Benchmark and Competition

2.1 Design Philosophy

AgentDS is built on three core principles:

1. Domain-specific complexity. We design in the way that strong performance requires domain-specific insights. Generic methods yield baseline results at best; competitive performance demands understanding what features matter in each context and what processing steps are appropriate. This design choice deliberately tests whether agents can apply genuine domain reasoning.

2. Multimodal integration. Real-world data science rarely involves a single tabular dataset. AgentDS therefore provides not only a primary tabular dataset containing the prediction target, but also additional data modalities such as images (e.g., product photos or vehicle condition images), text (e.g., customer reviews or clinical notes), and structured files (e.g., JSON, PDFs, or additional CSV files linked to the main dataset). This design introduces domain-specific complexity that more closely reflects real-world data science challenges.

3. Real-world plausibility. While our data is synthesized, the generation process faithfully mirrors genuine relationships found in actual industry data. Each domain's datasets incorporate realistic constraints and correlations that practitioners encounter. We consult the domain literature, including academic papers, industry reports, and practitioner blogs, to ensure that our data reflect authentic patterns and do not contradict established domain knowledge.

2.2   Benchmark Scope

AgentDS covers six domains, each selected for its real-world importance, technical challenge, and diversity of required skills. An overview of the challenges in each domain is presented in Table 1. The six domains were selected to span industries where predictive modeling plays a crucial role and where domain knowledge, heterogeneous data modalities, and business-specific evaluation criteria collectively influence modeling strategies. In commerce, demand forecasting and coupon targeting are high-impact problems where behavioral and contextual signals are essential, and product recommendation from visual catalogs benefits substantially from fusing image embeddings with interaction data [26, 27, 28]. In food production, shelf life estimation requires integrating storage conditions with microbiological growth dynamics, while visual quality control now approaches human inspector accuracy on structured defect detection tasks [29, 30, 31]. Healthcare challenges center on clinical prediction tasks, such as readmission, emergency department resource consumption, and discharge readiness, where domain-specific feature engineering around comorbidity combinations, vital sign trajectories, and care pathways is decisive [32, 33, 34]. Insurance combines structured actuarial data, free-text claims, and image evidence: text-based triage benefits from domain-adapted language models, risk-based pricing demands actuarially sound calibration, and fraud detection must handle severe class imbalance and adversarial adaptation [35, 36, 37]. Manufacturing challenges cover predictive maintenance from sensor streams and supply chain delay forecasting, both requiring domain-specific signals[38, 39]. Retail banking offers high-volume transaction data where fraud detection and credit default prediction remain challenging due to rare-event class imbalance, and where feature engineering around behavioral proxies requires practitioner expertise[40, 41].

Table 1: An Overview of Challenges in AgentDS Across Six Domains

Domain	Challenge	Problem	Metric	Additional Modalities
Commerce	Demand Forecasting	Regression	RMSE	Text, CSV
	Product Recommendation	Ranking	NDCG@10	Image, CSV
	Coupon Redemption	Classification	Macro-F1	JSON
Food Production	Shelf Life Prediction	Regression	MAE	JSON
	Quality Control	Classification	Macro-F1	Image, JSON
	Demand Forecasting	Regression	RMSE	Text, CSV
Healthcare	Readmission Prediction	Classification	Macro-F1	JSON
	ED Cost Forecasting	Regression	MAE	PDF, CSV
	Discharge Readiness	Classification	Macro-F1	JSON
Insurance	Claims Complexity	Classification	Macro-F1	Text
	Risk-Based Pricing	Regression	Normalized Gini	Image, CSV
	Fraud Detection	Classification	Macro-F1	PDF
Manufacturing	Predictive Maintenance	Classification	Macro-F1	CSV, JSON
	Quality Cost Prediction	Regression	Normalized Gini	Image, JSON
	Delay Forecasting	Regression	MSE	JSON
Retail Banking	Fraud Detection	Classification	Macro-F1	JSON
	Credit Default	Classification	Macro-F1	JSON

Each domain includes 2-3 challenges spanning classification, regression, and ranking tasks.

2.3 Data Curation Process

Creating datasets that are simultaneously realistic, challenging, and informative requires a systematic approach. Our curation pipeline involves four stages as described below.

Stage 1: Domain research. For each domain, we identify critical problems where data science provides value, the types of features and data commonly encountered, domain-specific tools and feature engineering practices, and plausible relationships between predictors and outcomes. This research grounds our dataset generation in authentic domain knowledge, ensuring that solving our challenges mirrors solving real industry problems.

Stage 2: Data generation. We synthesize data using carefully designed data-generating processes that respect the domain constraints identified in Stage 1. Importantly, the generation procedure ensures that strong predictive performance requires domain-specific reasoning rather than purely generic modeling pipelines. To achieve this, we transform certain latent variables that influence the prediction target into additional data modalities (e.g., images), so that effective feature extraction from these modalities requires domain-specific insights. As a result, each challenge dataset consists of a primary tabular dataset containing the prediction target together with additional data modalities that encode complementary information. We iteratively test baseline approaches (e.g., applying XGBoost to the tabular data alone) to verify that they underperform relative to methods that appropriately leverage the additional modalities with domain-specific insights. An example illustrating this process is provided in [25], with a synthetic property insurance dataset where crucial latent variables were embedded in roof images.

Stage 3: Performance bounds and difficulty calibration. Because we control the data generation process, we can determine the theoretical upper bound on performance by evaluating the score achievable under perfect knowledge of the data-generating mechanism. This allows us to calibrate challenge difficulty and distinguish between fundamental limits and gaps in possible participant approaches.

Stage 4: Documentation and validation. Each domain includes a `description.md` file that serves as a comprehensive documentation explaining domain terminology, data sources, and context. We validate that domain experts find the challenges realistic and that the documented information is sufficient (though not prescriptive) for informed approaches. Finally, the data is prepared per domain, meaning that all challenges within the same domain are organized together as a single package.

2.4 Evaluation Framework

AgentDS evaluates submissions primarily based on predictive performance on held-out test data. Each challenge is associated with a domain-specific evaluation metric, following those commonly used in practice, as shown in Table 1.

Quantile scoring. To enable fair comparison across challenges with heterogeneous metrics and scales, AgentDS employs a quantile-based scoring that normalizes performance into a common [0, 1] scale. For each challenge, participants who submit solutions are ranked according to the challenge-specific metric (e.g., Macro-F1, RMSE, normalized Gini coefficient). Let i be the index of a participant who successfully submitted to the challenge, and let $n > 1$ denote the number of such participants. The quantile score of participant i is computed as:

$$q_i = \frac{n - r_i}{n - 1},$$

where r_i denotes the rank of participant i (with $r_i = 1$ indicating the best performance). This transformation ensures that the top performer receives $q_i = 1$, the worst performer receives $q_i = 1/(n - 1) > 0$, and the intermediate ranks are linearly interpolated. Participants who do not successfully submit to a challenge are scored 0 for that challenge, ensuring that non-participation always results in the lowest possible score.

Score aggregation. Each domain contains two or three challenges. A participant's domain score is the arithmetic mean of their quantile scores across all challenges in that domain. The overall score is then defined as the mean of the six domain scores, yielding a single summary measure of cross-domain data science capability. This hierarchical aggregation (challenge $\rightarrow$ domain $\rightarrow$ overall) ensures that each challenge contributes equally to the final ranking.

Tie breaking. If two participants obtain the same overall score, ties are broken using efficiency indicators: the participant with fewer submissions ranks higher, and if the tie persists, the participant whose final submission occurred earlier ranks higher.

2.5 The AgentDS Competition

The AgentDS competition benchmarks human–AI collaboration performance in domain-specific data science. Participants are allowed to freely use any AI tools, enabling the competition to capture how humans and AI systems interact in realistic data science workflows.

The competition received more than 400 registrations, and participants were allowed to form teams of up to four people. It lasted for 10 days (October 18, 2025 – October 27, 2025), and a total of 29 teams consisting of 80 participants made successful submissions. During the competition, each team was allowed up to 100 submissions per challenge. After the competition ended, we collected code and reports from participating teams to verify reproducibility and conduct further analysis.

2.6 AI-Only Baselines

To contrast with the human-AI collaboration performance achieved by competition participants, we evaluate two AI-only baselines representing different levels of autonomy: a direct prompting baseline using GPT-4o and an agentic coding baseline using Claude Code. For each baseline, we compute performance using the same evaluation pipeline as human participants. Specifically, the raw metric score obtained by each baseline in each challenge is inserted into the pool of participant scores, and its quantile position is computed as if it had participated in the competition. This produces an interpretable estimate of where each AI-only baseline would rank among human teams.

2.6.1 Baseline configurations

Direct prompting baseline (GPT-4o). The first baseline uses GPT-4o [42] accessed through the ChatGPT interface in a direct prompting setting. For each challenge, the model is provided with the challenge directory containing the tabular datasets, preview samples of additional modalities (e.g., images, PDFs, JSON when present), and a `description.md` file describing the schema, prediction task, and submission format. The model is prompted to generate end-to-end Python code that loads the training data, trains a predictive model, produces predictions for the test set, and outputs a valid `submission.csv` file. The generated code is then executed to produce the submission, which is uploaded through the AgentDS evaluation API to obtain the corresponding score. In this baseline, the entire solution is generated in a single direct prompting interaction with the LLM.

Agentic coding baseline (Claude Code). The second baseline uses the Claude Code [5] CLI (v2.1.30) with the `claude-sonnet-4.5` model, operating in non-interactive autonomous mode. For each challenge, the agent is given access to the challenge directory containing the training data, test data, and the `description.md` file describing the schema, prediction task, and submission format. The agent is instructed to generate and submit a valid submission file. Unlike the direct prompting baseline, Claude Code can iteratively refine its approach by writing and executing code during the run. Each challenge is allocated a fixed time budget of 10 minutes. Again, there is no human intervention occurs during execution, namely, the entire modeling and submission process is carried out autonomously by the agent.

2.6.2 Performance of AI-only baselines

The GPT-4o direct prompting baseline achieves an overall quantile score of **0.143**, ranking **17th out of 29 teams** and falling below the participant median (0.156). In contrast, the Claude Code agentic baseline achieves a substantially higher overall quantile score of **0.458**, ranking **10th out of 29 teams**. Figure 1 shows the distribution of overall scores across all participants together with the two AI baselines.

Domain-level performance. Figure 2 illustrates domain-level quantile scores. The GPT-4o baseline performs at or below the domain median across all domains, with particularly weak performance in Retail Banking (0.000) and Commerce (0.021). The Claude Code baseline substantially improves performance across all domains, achieving its strongest scores in Manufacturing (0.573), Food
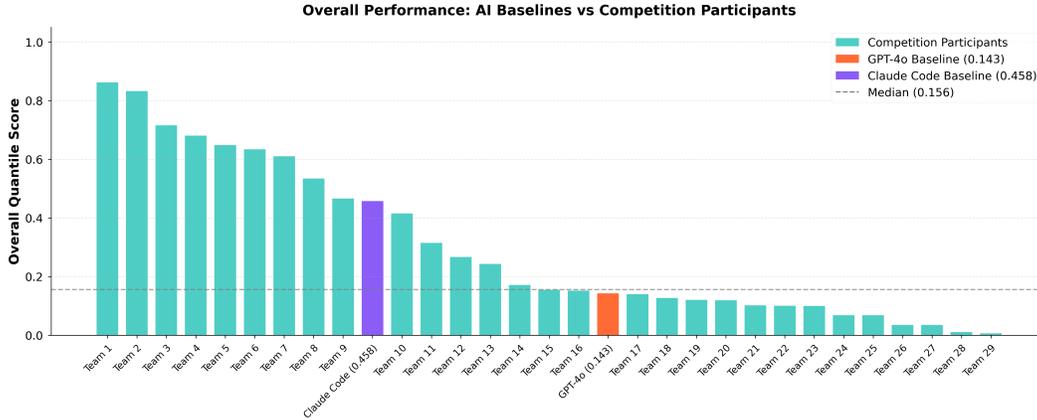
Figure 1: Overall quantile score comparison between both AI baselines and competition teams (n=29). The GPT-4o baseline (orange, score: 0.143) ranks 17th, falling below the participant median of 0.156 (dashed line). The Claude Code agentic baseline (purple, score: 0.458) ranks 10th, exceeding the median and placing in the top third of participants. Bars are sorted descending by score (Team 1 = best); both AI baselines are inserted at their rank positions. Quantile scores represent the average of per-challenge normalized rankings, with 1.0 indicating best performance and 0.0 indicating non-participation. The result shows that current AI-only baselines, whether using direct prompting or agentic coding, do not match the performance of the top human teams in the competition, highlighting a substantial gap between AI automation and human data science expertise.

Production (0.532), and Retail Banking (0.553). Nevertheless, the agentic baseline remains well below the top-performing human teams in every domain.

Challenge-level performance. Challenge-level results further reveal large performance variability across tasks. As shown in Figure 3, GPT-4o achieves moderate scores on a small subset of challenges (e.g., Insurance Ch. 3 and Healthcare Ch. 3) but obtains near-zero quantile scores on several others. Claude Code improves performance on the majority of challenges, particularly in Manufacturing Ch. 1 and Retail Banking Ch. 1, yet still fails to consistently match the strongest human solutions.

Taken together, the two baselines demonstrate that while agentic tool use substantially improves AI performance over direct prompting, AI-only baselines remain well below the level of the best human data scientists in domain-specific data science. The direct prompting baseline relies on generic modeling pipelines and largely ignores the additional data modalities provided in the challenges. The agentic baseline benefits from iterative experimentation and code execution, but still defaults to standard modeling strategies and fails to fully exploit the domain-specific signals available in these additional data sources.

These results establish an empirical reference point for interpreting participant outcomes. While the agentic baseline can outperform weaker participants, both AI-only baselines remain below the performance achieved by the strongest teams with human-AI collaboration.

3 Empirical Findings from AgentDS

In this section, we present empirical findings based on the quantitative results in Section 2.6 and a qualitative analysis of the code produced by the AI-only baselines together with the code and reports submitted by competition participants.

3.1 AI Agents Struggle with Domain-Specific Reasoning

Our benchmark reveals concrete evidence of agentic AI limitations. Despite their fluency in code generation and data manipulation, agentic AI consistently underperform on domain-specific data science tasks, as discussed in Section 2.6. Several failure modes emerge:
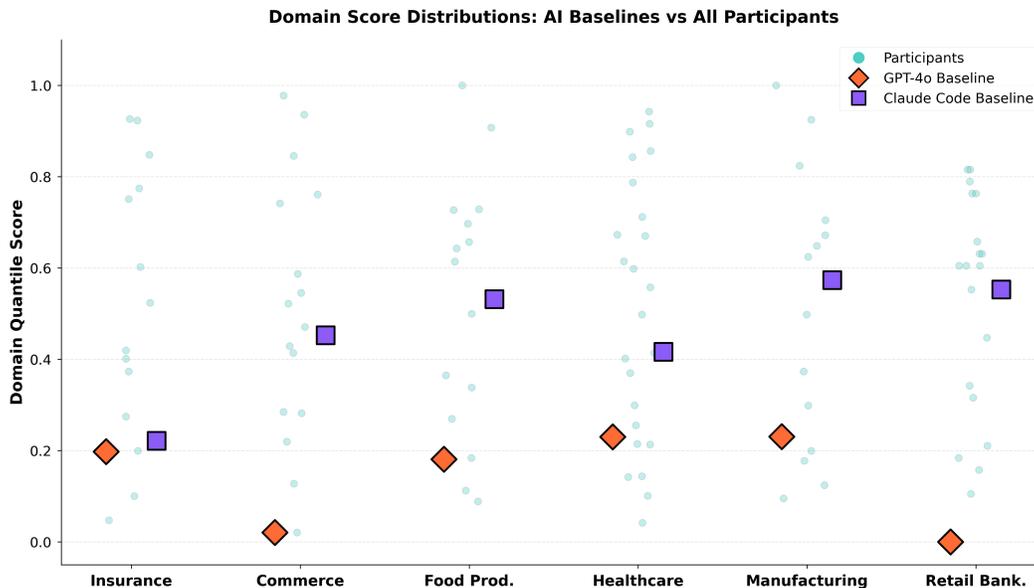
Figure 2: Distribution of domain-level quantile scores across all participants (teal dots), with GPT-4o baseline indicated by orange diamonds and Claude Code baseline by purple squares. GPT-4o falls at or below the domain median in all six domains, with particularly weak performance in Commerce (0.021) and Retail Banking (0.000). Claude Code substantially outperforms GPT-4o in every domain, most notably Manufacturing (0.573), Food Production (0.532), and Retail Banking (0.553), but remains well below the top-performing human teams in each domain, confirming that general-purpose AI, even agentic ones, cannot yet replicate the domain-specific strategies of expert human data scientists.

Inability to leverage multimodal signals. In challenges involving images, such as challenge 2 in insurance, food production, and commerce, AI agents fail to extract or appropriately utilize visual features. Human data scientists, by contrast, recognize when image-based signals matter and employ domain-specific computer vision techniques (e.g., DINOv3 [43], ResNet50 [44]).

Over-reliance on generic pipelines. AI tends to default to familiar patterns: loading data, applying standard preprocessing, and training with gradient-boosted models or random forest. While this baseline approach can produce an executable pipeline and works reasonably well for simple tasks, it performs poorly when domain-specific insight is essential, as in AgentDS challenges.

Limits of fully autonomous agents. Fully autonomous agentic approaches remain ineffective for complex domain-specific data science tasks. Several participating teams in AgentDS initially experimented with fully automated agent frameworks but later abandoned them in favor of interactive human-AI collaboration. One team reported that early attempts using autonomous agents with multi-turn tool calls and multi-agent orchestration required extensive prompt engineering and incurred significant API costs, making them difficult to sustain. They ultimately shifted to interactive coding agents, where humans guided the problem solving process while the AI executed coding tasks and explored ideas. This transition improved both practical efficiency and solution quality. Such experiences suggest that current agentic systems are better used as collaborative tools rather than fully autonomous replacements for human data scientists.

3.2 Human Expertise Provides Irreplaceable Value

Participant reports from the competition reveal a consistent pattern: AI agents accelerated implementation, but the decisions that determined performance were made by humans. The reports highlight four concrete mechanisms through which human expertise contributed value that autonomous agents could not replicate.
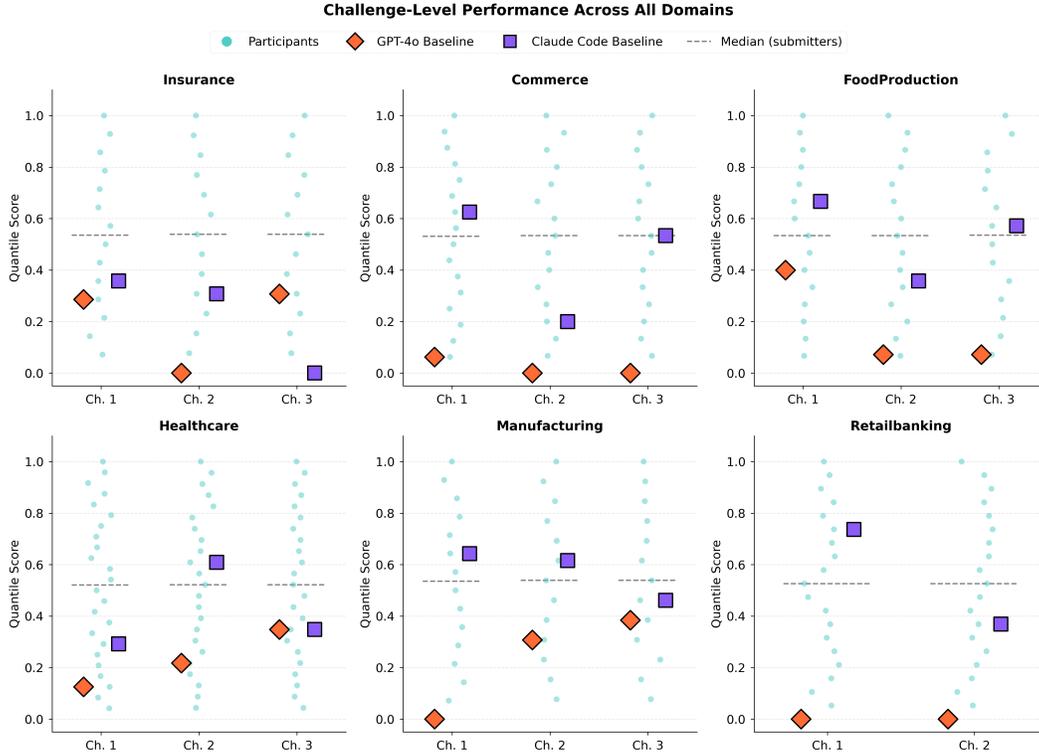
Figure 3: Challenge-specific quantile score distributions across six domains. Teal dots represent participants who submitted for each challenge (zero-score non-submitters excluded from display); orange diamonds show the GPT-4o baseline; purple squares show the Claude Code baseline; gray dashed lines indicate per-challenge participant medians among submitters. Claude Code outperforms GPT-4o across the majority of challenges, with the largest gains in Manufacturing Ch. 1 (Claude: 0.655, GPT-4o: 0.000), Retail Banking Ch. 1 (Claude: 0.741, GPT-4o: 0.000), and Commerce Ch. 3 (Claude: 0.534, GPT-4o: 0.000). Neither system achieves top-quartile performance on every challenge, confirming that current AI approaches cannot match the best human solutions, which leverage domain knowledge, multimodal signals, and iterative expert refinement.

Strategic problem diagnosis. Several top-performing teams explicitly reserved diagnosis for humans and implementation for AI. Some participants described a deliberate division of labor in which humans identified the structural weakness of the current approach, such as miscalibrated peaks, distribution shift between training and test data, or poorly specified feature interactions, before tasking the AI with implementing the proposed fix. Others initially pursued fully autonomous multi-agent frameworks but abandoned them after finding that extensive prompt engineering yielded diminishing returns. Their eventual approach, interactive human-guided coding agents, proved substantially more effective. Insights about what worked and what failed in each domain emerged from human reflection and were then shared back to the agents.

Encoding domain knowledge that data cannot reveal. Participants frequently constructed features that required domain expertise rather than patterns observable from the data distribution alone. In the healthcare domain, several participants derived features by comparing vital signs against medically defined normal ranges and by engineering indicators capturing stability, volatility, and recovery trends over time. These features reflected clinical protocols that cannot be inferred directly from the data itself. Similar patterns appeared in other domains: some participants incorporated domain-specific business rules, such as credit risk thresholds and inquiry count conditions, which improved model performance beyond what standard machine learning pipelines alone could achieve.

Filtering and overriding AI-suggested approaches. Multiple teams reported that uncritical acceptance of AI-generated pipelines reduced rather than improved performance. Some participants observed that AI agents across multiple frontier models frequently proposed complex feature en-

gineering pipelines that, when evaluated, lowered their validation scores. They further described a practice of first reasoning through the problem independently, forming their own hypotheses, and only then using the agent to implement a human-specified solution. Another team drew the same conclusion across all seventeen challenges they attempted: domain-driven feature engineering consistently outperformed blind automation, and no single AI-generated template generalized across tasks without human adaptation.

Human judgment beyond what validation scores reveal. Human participants frequently made model-selection decisions that required reasoning beyond simply maximizing validation scores. In several cases, participants deliberately chose models with slightly lower out-of-fold performance because discrepancies between validation and test scores suggested potential overfitting. Such decisions reflect an understanding of generalization risk that cannot be captured by score optimization alone. Participants also exercised caution in how AI tools were used: rather than delegating full control to autonomous agents, many teams conducted experiments manually and used LLMs primarily as assistants for debugging, explanation, or brainstorming. This workflow reflects a broader pattern in which humans retain final judgment in uncertain situations where evaluation metrics alone cannot determine the most reliable modeling strategy.

Taken together, these findings suggest that human expertise contributes more than speed or breadth of search. Humans provide a qualitatively different capability: diagnosing flaws in a model's framing before they appear in the data, injecting domain knowledge that the training distribution does not contain, and maintaining skepticism toward solutions that achieve high validation scores but generalize poorly.

3.3 Human-AI Collaboration Outperforms Either Alone

High-performing approaches in AgentDS competition effectively combine human strategic judgment with AI computational support. This collaboration takes several forms:

AI for acceleration, humans for direction. Successful approaches use AI agents to handle routine tasks, such as data loading, initial exploratory analysis, boilerplate code generation, while humans retain control over strategic decisions: which features to engineer, which models to compare, how to interpret results. This division of labor leverages the strengths of each.

Iterative human-AI feedback loops. Rather than treating AI as fully autonomous, effective collaboration engages tight feedback loops: humans propose approaches, AI implements them rapidly, and humans evaluate results and refine hypotheses. Importantly, these loops are consistently human-initiated. Participants described workflows in which humans judged when results were unsatisfactory, diagnosed the likely cause, and framed the next instruction to the AI. The agent accelerates iteration, but the direction of each cycle is determined by human reasoning.

Complementarity, not replacement. Human-AI teams excel through complementarity: humans provide domain grounding, causal reasoning, and error correction; AI provides computational power, rapid prototyping, and exhaustive search. Neither alone matches their combined effectiveness.

These findings resonate with a growing body of research on human-AI collaboration [45, 46, 47, 48, 49, 50, 51]. The central insight is that collaboration quality, meaning how effectively human judgment and AI capabilities are integrated, is as important as the capabilities of either alone. When human-AI collaboration is thoughtfully designed, the resulting partnership can outperform either humans or AI acting alone.

4 Limitations and Future Work

AgentDS represents an initial step toward rigorous evaluation of AI and human-AI collaboration in domain-specific data science, but several limitations warrant discussion:

Synthetic data. While our data generation process mirrors real-world relationships, it cannot capture the full messiness, ambiguity, and noise of genuine industry datasets. Future iterations may incorporate real (anonymized) datasets where feasible.

Limited participation pool. Our inaugural competition drew valuable participation, but larger and more diverse engagement would strengthen findings. We aim to expand outreach in future editions.

Scope of domains. Six domains, while diverse, do not exhaust the landscape of applied data science. Future work can expand to additional domains (e.g., energy or other areas of finance) to test the generalization of our findings.

Evolving AI capabilities. AI systems improve rapidly. Findings from our current competition may not reflect future capabilities. AgentDS is designed as an ongoing benchmark; we will continue tracking performance as agentic systems advance.

Observational analysis of collaboration. Our analysis of human-AI collaboration relies on participant reports, code submissions, and qualitative inspection of workflows. While these sources provide rich insight into how teams interacted with AI tools, the competition setting does not allow controlled experiments on collaboration strategies. Future work could design controlled studies that systematically vary the degree of autonomy, prompting strategies, or human oversight to quantify which collaboration patterns produce the best outcomes.

5   Conclusion

AgentDS introduces a benchmark and competition for studying domain-specific data science under realistic conditions. The benchmark comprises 17 challenges across six domains, each designed so that strong performance requires domain knowledge, multimodal reasoning, and thoughtful modeling decisions rather than generic machine learning pipelines. By combining a controlled data generation framework with an open competition setting, AgentDS provides a systematic environment for evaluating both autonomous AI agents and human–AI collaboration for domain-specific data science.

Our results reveal three consistent findings. First, current agentic AI systems struggle with domain-specific reasoning, particularly when multimodal signals and contextual knowledge must be incorporated. Second, human expertise remains essential: participants repeatedly demonstrated the ability to diagnose modeling failures, inject domain knowledge through feature design and domain-specific rules, and make strategic decisions about model generalization. Third, the most successful solutions emerge from human–AI collaboration, where humans guide the problem-solving process while AI accelerates coding, experimentation, and iteration.

These findings suggest that the future of AI in data science may not lie in fully autonomous automation, but in effective human–AI collaboration. Progress therefore depends not only on improving model capabilities, but also on designing AI that better support human reasoning, domain knowledge integration, and iterative problem solving. AgentDS provides a foundation for studying these dynamics and for developing AI that augment, rather than replace, human expertise.

Acknowledgments

References

[1] Longbing Cao. Data science: A comprehensive overview. *ACM Computing Surveys*, 50(3): 1–42, 2017.

[2] Valerio Grossi, Fosca Giannotti, Dino Pedreschi, Paolo Manghi, Pasquale Pagano, and Massimiliano Assante. Data science: A game changer for science and innovation. *International Journal of Data Science and Analytics*, 11:263–278, 2021.

[3] Gordon S. Blair, Peter A. Henrys, Amber Alexandra Leeson, John Watkins, Emma F. Eastoe, Susan G. Jarvis, and Paul J. Young. Data science of the natural environment: A research roadmap. *Frontiers in Environmental Science*, 7:121, 2019.

[4] 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.

[5] Anthropic. Claude 3.7 sonnet and claude code, 2025. URL `https://www.anthropic.com/news/claude-3-7-sonnet`. Accessed: 2026-03-10.

[6] Sirui Hong, Yizhang Lin, Bangbang Liu, Binhao Wu, Danyang Li, Jiaqi Chen, Jiayi Zhang, Jinlin Wang, Lingyao Zhang, Mingchen Zhuge, Taicheng Guo, Tuo Zhou, Wei Tao, Wenyi Wang, Xiangru Tang, Xiangtao Lu, Xinbing Liang, Yaying Fei, Yuheng Cheng, Zhibin Gou, Zongze Xu, Chenglin Wu, Li Zhang, Min Yang, and Xiawu Zheng. Data interpreter: An llm agent for data science. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

[7] 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. *arXiv preprint arXiv:2410.20424*, 2024.

[8] Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. Aide: Ai-driven exploration in the space of code. *arXiv preprint arXiv:2502.13138*, 2025.

[9] Zujie Liang, Feng Wei, Wujiang Xu, Lin Chen, Yuxi Qian, and Xinhui Wu. I-mcts: Enhancing agentic automl via introspective monte carlo tree search. *arXiv preprint arXiv:2502.14693*, 2025.

[10] Antoine Grosnit, Alexandre Max 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, Bal'azs K'egl, Haitham Bou-Ammar, and Jun Wang. Large language models orchestrating structured reasoning achieve kaggle grandmaster level. *arXiv preprint arXiv:2411.03562*, 2024.

[11] Wonduk Seo, Juhyeon Lee, and Yi Bu. Spio: Ensemble and selective strategies via llm-based multi-agent planning in automated data science. *arXiv preprint arXiv:2503.23314*, 2025.

[12] Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. Ds-agent: Automated data science by empowering large language models with case-based reasoning. In *Forty-first International Conference on Machine Learning*, 2024.

[13] Yizhou Chi, Yizhang Lin, Sirui Hong, Duyi Pan, Yaying Fei, Guanghao Mei, Bangbang Liu, Tianqi Pang, Jacky Kwok, Ceyao Zhang, Bangbang Liu, and Chenglin Wu. Sela: Tree-search enhanced llm agents for automated machine learning. *arXiv preprint arXiv:2410.17238*, 2024.

[14] Yaoli Mao, Dakuo Wang, Michael J. Muller, Ioana Baldini, and Casey Dugan. How data scientists work together with domain experts in scientific collaborations. *Proceedings of the ACM on Human-Computer Interaction*, 3(GROUP):1–23, 2019.

[15] Amy X. Zhang, Michael J. Muller, and Dakuo Wang. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction*, 4 (CSCW1):1–23, 2020.

[16] Zuwan Lin, Arnau Marin-Llobet, Jongmin Baek, Yichun He, Jaeyong Lee, Wenbo Wang, Xinhe Zhang, Ariel J. Lee, Ningyue Liang, Jin Du, Jie Ding, Na Li, and Jia Liu. Spike sorting ai agent. *bioRxiv*, 2025. doi: 10.1101/2025.02.11.637754.

[17] Zuwan Lin, Wenbo Wang, Arnau Marin-Llobet, Qiang Li, Samuel D. Pollock, Xin Sui, Almir Aljovic, Jaeyong Lee, Jongmin Baek, Ningyue Liang, Xinhe Zhang, Connie Kangni Wang, Jiahao Huang, Mai Liu, Zihan Gao, Hao Sheng, Jin Du, Stephen J. Lee, Brandon Wang, Yichun He, Jie Ding, Xiao Wang, Juan R. Alvarez-Dominguez, and Jia Liu. Spatial transcriptomics ai agent charts hpsc-pancreas maturation in vivo. *bioRxiv*, 2025. doi: 10.1101/2025.04.01.646731.

[18] An Luo, Xun Xian, Jin Du, Fangqiao Tian, Ganghua Wang, Ming Zhong, Shengchun Zhao, Xuan Bi, Zirui Liu, Jiawei Zhou, Jayanth Srinivasa, Ashish Kundu, Charles Fleming, Mingyi Hong, and Jie Ding. Assistedds: Benchmarking how external domain knowledge assists llms in automated data science. In *The 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

[19] 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? In *Thirteenth International Conference on Learning Representations*, 2025.

[20] Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Aleksander Madry, and Lilian Weng. Mle-bench: Evaluating machine learning agents on machine learning engineering. In *Thirteenth International Conference on Learning Representations*, 2025.

[21] Xueyu Hu, Ziyu Zhao, Shuang Wei, Ziwei Chai, Guoyin Wang, Xuwu Wang, Jing Su, Jingjing Xu, Ming Zhu, Yao Cheng, Jianbo Yuan, Kun Kuang, Yang Yang, Hongxia Yang, and Fei Wu. Infiagent-dabench: Evaluating agents on data analysis tasks. In *Forty-first International Conference on Machine Learning*, 2024.

[22] 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. *arXiv preprint arXiv:2502.13897*, 2025.

[23] 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*, pages 13487–13521. Association for Computational Linguistics, November 2024. doi: 10.18653/v1/2024.emnlp-main.748. URL `https://aclanthology.org/2024.emnlp-main.748/`.

[24] Tiberiu Valentin Pricope. Hardml: A benchmark for evaluating data science and machine learning knowledge and reasoning in ai. *arXiv preprint*, 2025.

[25] An Luo, Jin Du, Fangqiao Tian, Xun Xian, Robert Specht, Ganghua Wang, Xuan Bi, Charles Fleming, Jayanth Srinivasa, Ashish Kundu, Mingyi Hong, and Jie Ding. Can agentic ai match the performance of human data scientists? In *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 206–210, 2025.

[26] Lei Li, Xia Li, Weifeng Qi, Yi Zhang, and Wenbin Yang. Targeted reminders of electronic coupons: using predictive analytics to facilitate coupon marketing. *Electronic Commerce Research*, 22(1):1–28, 2022. doi: 10.1007/s10660-020-09405-4. URL `https://link.springer.com/article/10.1007/s10660-020-09405-4`.

[27] Xiao Liu. Dynamic coupon targeting using batch deep reinforcement learning: an application to livestream shopping. *Marketing Science*, 42(4):637–658, 2023. doi: 10.1287/mksc.2022.1403. URL `https://pubsonline.informs.org/doi/abs/10.1287/mksc.2022.1403`.

[28] Pooria Moeinzadeh Alamdari, Nima Jafari Navimipour, Mehdi Hosseinzadeh, et al. An image-based product recommendation for E-commerce applications using convolutional neural networks. *Acta Informatica Pragensia*, 11(2):237–250, 2022. doi: 10.18267/j.aip.183. URL `https://www.ceeol.com/search/article-detail?id=1061172`.

[29] Fatih Tarlak. The use of predictive microbiology for the prediction of the shelf life of food products. *Foods*, 12(24):4461, 2023. doi: 10.3390/foods12244461. URL `https://www.mdpi.com/2304-8158/12/24/4461`.

[30] V. Hemamalini, S. Rajarajeswari, et al. Food quality inspection and grading using efficient image segmentation and machine learning-based system. *Journal of Food Quality*, 2022:5262294, 2022. doi: 10.1155/2022/5262294. URL `https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/5262294`.

[31] Fei Xiong, Niklas Kühl, and Markus Stauder. Designing a computer-vision-based artifact for automated quality control: a case study in the food industry. *Flexible Services and Manufacturing Journal*, 36(3):873–904, 2024. doi: 10.1007/s10696-023-09523-9. URL `https://link.springer.com/article/10.1007/s10696-023-09523-9`.

[32] Masao Iwagami, Ryota Inokuchi, and Eiryo Kawakami. Comparison of machine-learning and logistic regression models for prediction of 30-day unplanned readmission in electronic health records: a development and validation study. *PLOS Digital Health*, 3(9):e0000578, 2024. doi: 10.1371/journal.pdig.0000578. URL `https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000578`.

[33] Yan-Ming Chiu, Julien Courteau, Isabelle Dufour, Alain Vanasse, and Catherine Hudon. Machine learning to improve frequent emergency department use prediction: a retrospective cohort study. *Scientific Reports*, 13(1):786, 2023. doi: 10.1038/s41598-023-27568-6. URL `https://www.nature.com/articles/s41598-023-27568-6`.

[34] Maryam Pahlevani, Majid Taghavi, et al. A systematic literature review of predicting patient discharges using statistical methods and machine learning. *Health Care Management Science*, 2024. doi: 10.1007/s10729-024-09687-2. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC11461599/`.

[35] Anu Dimri, Arnab Paul, Deepak Girish, Patrick Lee, Siamak Afra, et al. A multi-input multi-label claims channeling system using insurance-based language models. *Expert Systems with Applications*, 200:116930, 2022. doi: 10.1016/j.eswa.2022.116930. URL `https://www.sciencedirect.com/science/article/pii/S0957417422005553`.

[36] Edward W. Frees and Fei Huang. The discriminating (pricing) actuary. *North American Actuarial Journal*, 27(1):2–24, 2023. doi: 10.1080/10920277.2021.1951296. URL `https://www.tandfonline.com/doi/abs/10.1080/10920277.2021.1951296`.

[37] Fahad Aslam, Ahmed Imran Hunjra, Zied Ftiti, Wael Louhichi, et al. Insurance fraud detection: evidence from artificial intelligence and machine learning. *Research in International Business and Finance*, 62:101720, 2022. doi: 10.1016/j.ribaf.2022.101720. URL `https://www.sciencedirect.com/science/article/pii/S0275531922001325`.

[38] Serkan Ayvaz and Koray Alpay. Predictive maintenance system for production lines in manufacturing: a machine learning approach using IoT data in real-time. *Expert Systems with Applications*, 173:114598, 2021. doi: 10.1016/j.eswa.2021.114598. URL `https://www.sciencedirect.com/science/article/pii/S0957417421000397`.

[39] Nawel Rezki and Mustapha Mansouri. Machine learning for proactive supply chain risk management: predicting delays and enhancing operational efficiency. *Management Systems in Production Engineering*, 32(3):301–311, 2024. doi: 10.2478/mspe-2024-0033. URL `https://sciendo.com/pdf/10.2478/mspe-2024-0033`.

[40] Seyed Kasra Hashemi, Seyedeh Leili Mirtaheri, and Sergio Greco. Fraud detection in banking data by machine learning techniques. *IEEE Access*, 11:3034–3043, 2022. doi: 10.1109/ACCESS.2022.3232287. URL `https://ieeexplore.ieee.org/abstract/document/9999220/`.

[41] Jiaming Xu, Zhengyang Lu, and Ying Xie. Loan default prediction of Chinese P2P market: a machine learning methodology. *Scientific Reports*, 11(1):18759, 2021. doi: 10.1038/s41598-021-98361-6. URL `https://www.nature.com/articles/s41598-021-98361-6`.

[42] OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[43] O. Siméoni, H. V. Vo, M. Ramamonjisoa, P. Bojanowski, and C. Couprie. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[45] Vivian Lai, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: A survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

[46] Kori Inkpen, Shreya Chappidi, Kaile Mallari, Mercy Kulkarni, Besmira Nushi, Divya Suri, and Taylor Gallagher. Advancing human-ai complementarity: The impact of user expertise and algorithmic tuning on joint decision making. *ACM Transactions on Computer-Human Interaction*, 30(5):1–29, 2023.

[47] Shelly Cao, Connor Gomez, and Chien-Ming Huang. How time pressure in different phases of decision-making influences human-ai collaboration. *ACM Transactions on Computer-Human Interaction*, 2023.

[48] Elena Revilla, Maria Jesus Saenz, Matthias Seifert, and Yichuan Ma. Human–artificial intelligence collaboration in prediction: A field experiment in the retail industry. *Journal of Management Information Systems*, 40(4):1248–1278, 2023.

[49] Johannes Senoner, Stephan Schallmoser, Bernhard Kratzwald, and Stefan Feuerriegel. Explainable ai improves task performance in human–ai collaboration. *Scientific Reports*, 14(1):2457, 2024.

[50] Nina Li, Hao Zhou, Wei Deng, Jiayan Liu, and Fei Liu. When advanced ai isn't enough: Human factors as drivers of success in generative ai-human collaborations. *SSRN Electronic Journal*, 2024.

[51] George Fragiadakis, Christos Diou, George Kousiouris, Dimosthenis Kyriazis, and Theodora Varvarigou. Evaluating human-ai collaboration: A review and methodological framework. *arXiv preprint arXiv:2405.13315*, 2024.