ISSN:2790-1688 Volume-14-(2025)

From Assistants to Agents: The Evolution of Large Language Models in Data Science Workflows

Xiyuan Yin

School of mathematics, Renmin University of China

johnnyyin2000@foxmail.com

Abstract. This paper presents a comprehensive overview of the evolution of data science from a statistics-centric discipline to a machine learning–driven field, culminating in the current integration of large language models (LLMs). It identifies key limitations in traditional LLM applications—such as limited cross-domain adaptability, lack of interpretability, and workflow rigidity—and explores recent innovations addressing these challenges. Three representative frameworks—R&D-Agent, SPIO, and Agent Laboratory—illustrate LLMs' transition from assistive tools to autonomous agents capable of planning, executing, and optimizing entire data science workflows. These systems leverage dual-agent cooperation, modular architectures, and self-correcting capabilities to improve performance in end-to-end data analysis and scientific research. The paper concludes by outlining future priorities, including domain-specific customization, standardized agent evaluation, and improved interpretability, all of which are essential for the next generation of intelligent, autonomous data science systems.

Keywords: Large Language Models (LLMs), Data Science, Technical Evolution.

1. Introduction

Since its initial proposal by Peter Naur in 1974, the concept of data science has evolved over more than four decades into an emerging discipline grounded in statistics, machine learning, and data visualization, while maintaining close integration with diverse industries [1]. Due to the limitations of traditional statistics, model-driven statistical approaches are shifting toward data-driven machine learning. Researchers outside data science must master machine learning methodologies, or risk being replaced. As the latest achievement in deep learning and one of the technologies most impactful to productivity development, large language models (LLMs) are gradually becoming a novel methodology in data science research. LLMs refer to models that comprehend and process human language through massive parameters, built upon pretraining tasks such as masked language modeling and autoregressive prediction. Their core mechanism is modeling contextual semantics and probability distributions from large-scale textual data. Although LLMs have demonstrated remarkable adaptability across numerous domains, their application in data science still faces significant limitations, including insufficient domain knowledge when processing cross-disciplinary data and the persistent lack of interpretability since their inception. This paper outlines the developmental trajectory of data science, introduces applications and recent advances of LLMs within this field, and discusses prospects for future research directions.

2. Developmental Trajectory of Data Science

2.1 The Formative Period (1974–2009)

In 1974, Peter Naur formally defined "Data Science" for the first time, regarded as the origin of this field [1]. In 1993, J.M. Chambers critiqued the limitations of traditional statistics and advocated for a new paradigm of "learning from data."[2] In 2001, W.S. Cleveland explicitly proposed positioning data science as an expansion of statistics [3]. By this point, statistics and computer science had jointly established the dual theoretical foundations of data science. On the technical front, Brewer's CAP theorem, proposed in 2000, and Pritchett's BASE principle, introduced in 2008, catalyzed the transformation of distributed data management from theoretical perfectionism to

constraint-oriented pragmatism [4]. This paradigm shift emphasized practicality and engineering feasibility, becoming a defining characteristic that distinguishes data science from traditional data processing.

As the primary subject of data science research, the big data domain also saw its initial development during this period. In 2001, Gartner analyst Doug Laney proposed the "3V" characteristic model of big data—Volume, Velocity, and Variety—while Clive Humby coined the resource metaphor "data is the new oil" in 2006 [5, 6]. These contributions formed the early theoretical pillars of data science and underscored the significance of big data.

2.2 The Development and Maturation Period (2010–2019)

In 2010, data scientist Drew Conway proposed the Venn diagram model for data science, defining it for the first time as a tripartite interdisciplinary field integrating mathematical-statistical knowledge, domain-specific expertise, and hacking expertise [7]. This theory systematically elucidated the cross-disciplinary nature of data science, laying a crucial theoretical foundation for the field and continuously stimulating academic discourse. An empirical study by McAfee and colleagues in 2012 demonstrated that top-tier enterprises (the top third of their industries) adopting data-driven decision-making achieved 5% higher productivity and 6% greater profit margins than competitors, quantitatively validating the commercial value of data science for the first time [8]. In 2013, IBM expanded Laney's "3V" model by adding the dimension of Veracity, forming the "4V" big data characteristic framework, thereby elevating data quality to a core strategic enterprise concern [9].

In 2015, Michael I. Jordan's team asserted that machine learning constitutes the shared core of data science and artificial intelligence [10]. Its data-intensive methodologies drive evidence-driven decision-making transformations across healthcare, manufacturing, and finance. With the evolution of AI technologies, key capabilities such as machine perception, cognitive computing, and natural language processing are progressively integrating into the data science framework. Concurrently, data science methodologies accelerated their penetration into vertical domains including wind energy, agriculture, and materials science, giving rise to two research paradigms: Domain-Agnostic Data Science, focusing on universal theories and methodological innovations; and Domain-Specific Data Science, emphasizing application modeling constrained by professional prior knowledge, exemplified by multi-source data fusion analysis in clinical medicine. This theoretical bifurcation marks data science's transition from an instrumental technology to a foundational discipline.

2.3 Comprehensive Integration and Deepening Period (2019–Present)

Since 2019, research in data science has experienced explosive growth, significantly catalyzed by the COVID-19 pandemic. Data-driven technologies were deployed during the pandemic for real-time infection tracing and vaccine allocation optimization [11]. Concurrently, the surge in remote work and online education generated large-scale user behavioral datasets, accelerating the development of behavioral prediction models [12]. In interdisciplinary research, sociology remained dominant, while applications in education witnessed the most substantial growth. As research scaled, discussions on data ethics—particularly regarding data justice—gained prominence, reflecting an urgent societal need for governance frameworks in this emerging discipline. By 2023, rapid advancements in natural language processing (NLP), especially iterative breakthroughs in LLMs such as GPT and Deepseek series, propelled a paradigm shift across data science, enabling deeper applications in specialized fields and facilitating autonomous model operations through LLMs.

3. Developmental Trajectory of Large Language Models in the Domain of Data Science

The development of large language models can be traced back to the 1950s. Early research on language models (LMs) was confined to language translation and grammatical analysis, relying on rule-based systems for text processing, which exhibited significant limitations in handling large-scale,

complex linguistic data. By the 1980s, advancements in computational power and the emergence of large-scale corpora propelled statistical methods to dominate natural language processing, with statistical machine translation (SMT) serving as a representative milestone.

Post-2010, deep learning achieved groundbreaking progress. Introducing attention mechanisms and the Transformer architecture enabled quantum leaps in performance for tasks such as text comprehension, generation, and machine translation. In recent years, the rise of reinforcement learning and multimodal fusion technologies has further extended capabilities to encompass collaborative understanding of image, audio, and video data, driving language models toward cross-modal cognitive evolution.

As application scenarios grew increasingly complex and model parameters continued to scale exponentially, large language models (LLMs) emerged. Leveraging massive parameters and exceptional learning capabilities, these models have achieved breakthroughs in natural language processing and furnished data science with novel methodological tools for interdisciplinary research. The following section delineates the developmental trajectory of LLM applications within the data science domain.

3.1 Foundational Construction Period (2017–2020)

Ashish Vaswani et al. introduced the Transformer model in 2017, a groundbreaking advancement in natural language processing (NLP) [13]. Entirely based on attention mechanisms, the Transformer architecture abandoned recurrent neural networks (RNNs) and convolutional neural networks (CNNs) previously prevalent in sequence transduction models. This framework gained widespread attention for its parallel processing capabilities and effective modeling of long-range dependencies, subsequently forming the foundation for numerous NLP models.

In 2018, Howard and Ruder proposed the ULMFiT (Universal Language Model Fine-tuning) transfer learning framework [14]. This work pioneered efficient fine-tuning of pretrained models for text classification tasks, establishing the groundwork for NLP task adaptation and providing a more effective approach for classifying structured data.

In 2019, Chen and colleagues constructed the first large-scale tabular reasoning dataset (TabFact) and designed a structure-aware attention mechanism [15]. This research enabled subsequent LLMs to perform automated data cleaning. The 2020 TaPas model adopted a weakly supervised approach, selecting table cells through applied aggregation operators for prediction.

While these studies demonstrated remarkable potential in individual tasks during this period, the absence of a cross-task collaborative framework hindered systematic problem-solving. As TaPas author Herzig aptly remarked in 2021: "We crafted exquisite screwdrivers, yet users needed a comprehensive toolkit." [16]

3.2 Process Integration Period (2021–2022)

Chen and colleagues introduced the groundbreaking contextual inheritance mechanism in 2021. By fine-tuning GPT-3 on 159GB of code data, this approach enabled models to parse historical operational states. Experiments demonstrated a 37% improvement in end-to-end code generation accuracy on Kaggle datasets, significantly outperforming isolated task models. This advancement propelled automated code generation in data science [17].

Jason Wei proposed the Chain-of-Thought (CoT) method in 2022 [18]. By decomposing complex tasks into inferential sub-steps, CoT enabled LLMs to exhibit human-like logical capabilities, establishing a cognitive science foundation for workflow integration in data science. This breakthrough marked a critical advancement in machine reasoning, providing an extensible task decomposition framework for subsequent tools like LangChain and AutoFeat. It also directly catalyzed the emergence of Dynamic CoT in 2023.

Harrison Chase's team launched the open-source workflow orchestration framework LangChain in 2022, whose core innovation lies in engineering CoT principles into a scalable system. LangChain integrated 200+ data science tools (e.g., HuggingFace models and Scikit-learn) into its ecosystem [19]. It gained native support from platforms like Databricks and Airflow, establishing itself as the de facto integration platform for data science and LLMs. McKinsey reports indicate that frameworks

like LangChain are pivotal shifts from experimental AI prototypes to scalable enterprise solutions. By standardizing tool integration and workflow automation, these frameworks fundamentally lowered the barrier to deployment, evidenced by adoption within 30% of Fortune 500 tech companies within 12 months of release. This phase signifies LLMs' qualitative transformation from standalone tools to production pipelines, driving their applications from experimentation to operationalization, ultimately catalyzing the emergence of autonomous systems with dynamic workflow generation capabilities in 2023.

3.3 Autonomous Systems Period (2023–Present)

The Autonomous Systems Period, commencing in 2023, marks a qualitative leap in LLM applications within data science. While the preceding Process Integration Period achieved task-chain connectivity, it remained constrained by static workflow design. Geekan et al. introduced hierarchical graph-based task and code planning in 2024, enabling models to effectively manage complex inter-task dependencies and dynamically adapt to real-time data fluctuations in data science tasks [20]. This approach draws from hierarchical planning techniques in automated machine learning. It decomposes complex data science problems into manageable subtasks through layered structures and translates them into executable code actions for granular planning and implementation.

Furthermore, Data Interpreter pioneered tool integration and generation methodologies. Through tool recommendation and organization, tasks are classified based on descriptions to select the optimal toolkits. During execution, it dynamically embeds and adjusts tool parameters by leveraging structured information from tool parameter descriptions and documentation, tailoring solutions to specific task requirements. Crucially, Data Interpreter achieves self-evolution: it abstracts core functionalities from execution experiences to form reusable code snippets, integrating them into a tool function library. These functions reduce debugging frequency and enhance execution efficiency in future tasks.

Concurrently, Zhibin Gou proposed the CRITIC framework in 2023. This framework emulates human verification processes using external tools to validate and refine outputs, enabling inherently "black-box" LLMs to self-verify and iteratively correct their results [21]. CRITIC demonstrates LLMs’ capacity for runtime iterative improvement, transforming the human role in data science tasks from strategy formulator to strategy auditor. These technological advancements propel LLMs from auxiliary tools toward autonomous systems capable of self-directed task execution in data science.

4. Developmental Trajectory of Large Language Models in the Domain of Data Science

4.1 R&D-Agent

R&D-Agent, released in 2025 by Xu Yang et al., is a cross-domain data science agent framework designed to address persistent challenges of high labor intensity and cross-disciplinary knowledge demands in advanced data science tasks. This dual-agent architecture enables iterative exploration through specialized role allocation [22]. The authors contend that a successful machine learning engineering agent must achieve active learning and adaptability through iterative exploration—integrating domain insights, generating deep hypotheses, and continuously refining methodologies based on phased discoveries rather than relying on single solution paths.

Recognizing distinct strengths across foundation LLMs—such as o1 series' superiority in reasoning innovation and GPT-4.1's excellence in instruction execution—the framework strategically assigns model roles to form collaborative teams for optimal outcomes. It comprises two specialized agents: the Researcher generates innovative ideas using performance feedback, and the Developer agent optimizes code based on error feedback. R&D-Agent substantially narrows the performance gap between automated solutions and expert-level implementations by enabling mutual reinforcement across parallel exploration paths.

In the MLE-Bench benchmark, R&D-Agent was validated as the top-performing machine learning engineering agent, demonstrating significant potential for accelerating innovation and enhancing precision in cross-domain data science applications.

4.2 Agent Laboratory

Agent Laboratory, proposed by Samuel Schmidgall et al. in 2025, is an autonomous research framework based on large language models (LLMs) capable of executing the entire scientific research workflow. Within this framework, upon receiving a research proposal from humans, LLMs sequentially advance through three phases: literature review, experimental design, and report writing [23]. At each stage, specialized agents driven by LLMs collaborate to achieve distinct objectives, integrating external tools such as arXiv, Hugging Face, Python, and LaTeX to optimize outcomes. The final output comprises a complete deliverable, including code repositories and research reports, allowing users to provide stage-specific feedback for guidance.

During the literature review phase, a PhD-level agent employs an iterative optimization mechanism: it retrieves the top 20 most relevant paper abstracts via the arXiv API using agent-generated search queries, parses the full content of selected papers, and synthesizes curated summaries or full texts into a refined literature review.

In the experimental design phase, doctoral and postdoctoral agents collaboratively define machine learning models to implement, datasets to employ, and experimental procedures through conversational coordination.

For the report writing phase, doctoral and professor-level agents utilize a dedicated Paper Solver module to consolidate research findings into an academic report. Although the framework's generated papers have not yet achieved top-tier conference acceptance standards under human review, their evaluation scores surpass those of outputs from GPT-4o and o1-mini.

4.3 SPIO

SPIO is a framework to address persistent challenges in multi-agent systems—specifically, workflow rigidity and insufficient feedback integration—by leveraging LLMs to optimize the entire data analysis process through multi-agent collaboration [24]. This framework features two core innovations: modular design and an optimized multi-agent coordination mechanism.

In its modular architecture, the framework divides the workflow into four distinct phases: data preprocessing, feature engineering, modeling, and parameter tuning. This structure enables autonomous agents to generate multiple candidate strategies independently, forming cascading exploration pathways. SPIO has two functionally optimized variants: SPIO-S and SPIO-E. SPIO-S employs LLMs to select and execute a single optimal path. At the same time, SPIO-E integrates multiple top-ranked paths (using soft voting for classification tasks and averaging for regression tasks) to enhance robustness during data analysis.

Evaluated against mainstream methods (e.g., Agent K, Data Interpreter) across 12 Kaggle and OpenML datasets spanning classification and regression tasks, SPIO-S and SPIO-E achieved superior performance in most scenarios, with peak average prediction accuracy improvements of up to 11%. This research pioneered the integration of multi-path exploration and ensemble learning into automated data analysis workflows, overcoming limitations of single-path approaches. Its dynamic strategy generation mechanism significantly improves adaptability to complex tasks while reducing dependence on manual intervention.

The framework offers scalable solutions for high-precision prediction scenarios such as business intelligence and scientific research automation.

5. Conclusion

The recent transformation of LLMs from assistive tools to autonomous systems marks a pivotal shift in the trajectory of data science. Frameworks such as R&D-Agent, SPIO, and Agent Laboratory

demonstrate that LLMs can now conduct end-to-end scientific workflows with minimal human oversight. However, several critical challenges remain. First, the general-purpose nature of most LLM applications leads to insufficient performance in domain-specific tasks, highlighting the need for tailored training and fine-tuning. Second, the absence of standardized benchmarks limits the comparability and evaluation of LLM-based agent systems across disciplines. Lastly, despite advancements in interpretability mechanisms such as CRITIC, LLMs still exhibit black-box behavior and occasional hallucinations, which hinder trust and adoption in high-stakes scenarios. Future research must address these limitations through the development of domain-customized models, unified evaluation standards, and mechanisms that balance efficiency with transparency. By overcoming these barriers, LLMs are poised to become central agents in the evolution of data science, enabling more robust, autonomous, and interpretable workflows across diverse fields.

References

[1] Naur, P. (1974). Concise survey of computer methods. (No Title).

[2] Chambers, J. M. (1993). Greater or lesser statistics: a choice for future research. Statistics and Computing, 3(4), 182-184.

[3] Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. International statistical review, 69(1), 21-26.

[4] Simon, S. (2000). Brewer's cap theorem. CS341 Distributed Information Systems, University of Basel (HS2012).

[5] Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META group research note, 6(70), 1.

[6] Humby, C. (2006). Data is the new oil. Proc. ANA Sr. Marketer's Summit. Evanston, IL, USA, 1.

[7] Taylor, D. (2016). Battle of the data science Venn diagrams. KDNuggets News.

[8] McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

[9] Rubin, V., & Lukoianova, T. (2013). Veracity roadmap: Is big data objective, truthful and credible?. Advances in Classification Research Online, 24(1), 4.

[10] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[11] Goldstein, J. R., & Lee, R. D. (2020). Demographic perspectives on the mortality of COVID-19 and other epidemics. Proceedings of the national academy of sciences, 117(36), 22035-22041.

[12] Davies, H. C., Eynon, R., & Salveson, C. (2021). The mobilisation of AI in education: A Bourdieusean field analysis. Sociology, 55(3), 539-560.

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[14] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

[15] Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., ... & Wang, W. Y. (2019). Tabfact: A large-scale dataset for table-based fact verification. arXiv preprint arXiv:1909.02164.

[16] Herzig, J., Müller, T., Krichene, S., & Eisenschlos, J. M. (2021). Open domain question answering over tables via dense retrieval. arXiv preprint arXiv:2103.12011.

[17] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., ... & Zaremba, W. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

[18] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.

[19] Chase, H. (2022, October). LangChain.

[20] Hong, S., Lin, Y., Liu, B., Liu, B., Wu, B., Zhang, C., ... & Wu, C. (2024). Data interpreter: An llm agent for data science. arXiv preprint arXiv:2402.18679.

[21] Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., & Chen, W. (2023). Critic: Large language models can self-correct with tool-interactive critiquing. arXiv preprint arXiv:2305.11738.

[22] Yang, X., Yang, X., Fang, S., Xian, B., Li, Y., Wang, J., ... & Bian, J. (2025). R&d-agent: Automating data-driven ai solution building through llm-powered automated research, development, and evolution. arXiv preprint arXiv:2505.14738.

[23] Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., ... & Barsoum, E. (2025). Agent laboratory: Using llm agents as research assistants. arXiv preprint arXiv:2501.04227.

[24] Seo, W., Lee, J., & Bu, Y. (2025). SPIO: Ensemble and Selective Strategies via LLM-Based Multi-Agent Planning in Automated Data Science. arXiv preprint arXiv:2503.23314.