Why Do AI Projects Fail in Drug Development and Pharma? Insights from Multi-Year Pistoia Alliance Studies

Vladimir Makarov

vladimir.makarov@pistoiaalliance.org

Pistoia Alliance

Mario Sänger

AstraZeneca (Spain)

Roman Saiz

Roche (Germany)

Matthias Hartung

Boehringer Ingelheim (Germany)

Nick Lynch

Curlew Research

Research Article

Keywords: Artificial Intelligence, Drug Discovery, Organizational Maturity, Governance, Good Machine Learning Practices

Posted Date: June 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9871864/v1

Additional Declarations: Competing interest reported. • The funding for this work was provided by AstraZeneca, Boehringer-Ingelheim, and Roche • Mario Sänger is an employee of AstraZeneca • Roman Saiz is an employee of Roche • Matthias Hartung is an employee of Boehringer-Ingelheim • Vladimir Makarov receives consulting fees from the Pistoia Alliance

---

Abstract

Artificial intelligence (AI) is playing an increasingly central role in drug discovery and the pharmaceutical industry more broadly. However, despite some high-profile success stories, many technically successful pilots do not translate into sustained business value. In this study, we analyze the determinants of successful AI adoption in the life sciences through three complementary approaches: (i) brainstorming workshops with senior industry professionals, (ii) a global executive survey, and (iii) a statistical analysis of a collection of AI/ML use cases collected by the Pistoia Alliance. Across all methods, a consistent pattern emerges: high strategic importance and demand for AI contrast with low organizational maturity, resulting in limited success rates for production deployments. Statistical analysis reveals that business success correlates strongly with project maturity, business support, and developer support, i.e., dimensions of organizational capability maturity, while factors such as data sources or model types are not predictive on their own. The findings reinforce that AI projects follow the same success drivers as other complex engineering initiatives and that organizational readiness, change management, and alignment between stakeholders are critical. Furthermore, trustworthiness encompassing reproducibility, explainability, and governance, is identified as a limiting factor for adoption, particularly in high-risk domains. We synthesize these observations into a set of recommended practices that emphasize fitness-for-purpose method selection, integration of AI into human workflows, and development of shared standards.

Introduction

Conventional drug discovery and development is a lengthy, costly, and complex process, underscoring the need for more efficient approaches. Recent advances in artificial intelligence (AI), machine learning (ML) and natural language processing (NLP), e.g., AI-aided compound design using protein [1] or chemical language models [2] and ML-based methods for evaluating key molecular properties [3, 4], have the potential to accelerate and improve solutions to complex drug development challenges. These potential benefits extend beyond the core stages of drug design and development to virtually every step of the drug development pipeline, as well as to a wide range of applications across pharmaceutical and healthcare industries in general. For example, AI may support clinical trial design by improving patient cohort selection and recruitment strategies [5] and may also enhance public communication by accelerating the generation of plain-language summaries of scientific findings [6]. Alongside applications focused on a single task or problem setting, a recent trend has been the emergence of multi-purpose AI systems, exemplified by recent “AI co-scientist” systems [7–10] and specialized life science large language models developed by Amazon BioDiscovery [11] and OpenAI GPT Rosalind [12].

Despite the enormous potential of these applications, integrating AI- and ML-based approaches into real-world business processes remains challenging because of technical, organizational, regulatory, and operational barriers. Practitioners often face a broad range of challenges, including (but not limited to) data accessibility and quality issues, regulatory and compliance complexity, trust in and explainability of the applied models, integration of new approaches into existing workflows, and change management and workforce readiness. Hence, implementing AI in pharmaceutical business processes is not merely a matter of model performance, but often requires broader enterprise transformation. Due to their high complexity and the vast number of factors involved, AI engineering projects are often risky undertakings and frequently fail to achieve their intended objectives and business value. This raises the question of which factors are most dominant and influential in determining the success of AI engineering projects. Indeed, several industry trends can be identified that may provide a basis for improving the governance of AI and ML deployments, increasing the likelihood of successful application in drug discovery, and enhancing the return on investment in AI and ML.

This study aims to contribute to this question by drawing on insights from a multi-year Pistoia Alliance initiative. The Pistoia Alliance is a global, not-for-profit life sciences collaboration that brings together pharmaceutical companies, biotech, vendors, academics, and other stakeholders to solve shared R&D data and technology challenges through pre-competitive collaboration, standards, and innovation. It runs multiple sub-initiatives and working groups focused on different aspects of pharma and biotechnology, such as data and annotation standards, real world data mining, laboratory workflows, and scientific collaboration. This paper builds on the work and findings of the *LLM and NLP Use Case Database* project, which focuses on documenting successes and failures in AI, ML and NLP systems development across the pharmaceutical industry. Since 2022, we have collected 67 use cases spanning the full business landscape, including R&D, pharmacovigilance, manufacturing, and HR management. To our knowledge, this represents the largest collection of its kind in the industry. Moreover, we conducted i) community brainstorming workshops and ii) executive surveys interviews focusing on recent advances and challenges in engineering and production-level application of AI- and ML-based systems in pharmaceutical use cases. The information gathered provides unique and valuable insights for AI practitioners and decision-makers on key success factors. Our findings, particularly from the use case analysis, suggest that the success of AI/ML initiatives depends less on technical design and performance than on organizational maturity and execution capability. Future progress will therefore require stronger governance, business alignment, and workforce readiness to enable scalable and sustained impact.

In the following, we describe our three complementary approaches used to gather evidence from real-world pharmaceutical AI applications and senior experts, and then present the resulting findings. On this basis, we formulate ten best-practice recommendations derived from the analysis that we strongly believe support the successful implementation of AI projects in the pharmaceutical industry.

Methods

To gather information on the opportunities, challenges and risk factors associated with AI engineering projects in the pharmaceutical industry, we applied three distinct initiatives to collect diverse inputs and broaden the coverage of use cases and perspectives: i) collective brainstorming workshops, ii) collection of AI / ML / NLP use cases in a structured way and iii) executive surveys and interviews. Table 1 provides an overview of the applied approaches and their base characteristics and time spans. The applied initiatives are presented in greater detail in the following sections.

Table 1
Overview of the information collection methods applied. For each method we report the time span covered in the data and type of information (structured, free form or mixed).

Information Collection Method	Time Span	Information Type
Collective Brainstorming Workshops	2017–2026	Free Form
Collection and Analysis of AI / NLP Use Cases	2018–2025	Structured Format
Executive Surveys and Interviews	2025	Mixed
Collective Brainstorming Workshops		

Collective brainstorming workshops in teams composed of senior professionals ranging from senior scientist to executive senior directors from biotech and pharmaceutical industries. We conducted such events both in-person and online multiple times during each year since 2017 and accumulated a large body of knowledge. It was in part formalized in the 2021 and 2024 papers on Good Machine Learning Practices [13, 14]. This paper builds on top of these earlier findings. Here we report the most up-to-date results of the community brainstorming events in 2025 and the first half of 2026. Our typical event aggregates the inputs from 80 or more individuals.

Collection and Analysis of AI / NLP Use Cases

Since 2022, the Pistoia Alliance *LLM and NLP Use Case Database* project has been collecting AI and ML use cases from participating organizations. Initially, the project focused on documenting the workflows and protocols used in complex NLP system development initiatives. However, recent advances in large language models, generative AI, agentic AI, and the integration of text-based and multimodal data have led to a significant expansion of the project's scope. It now covers a broad range of AI use cases, spanning everything from molecular property modeling and scientific literature mining to HR applications. As of now, the database consists of 67 use cases in total. Initially we gathered 47 properties for each use case across the following dimensions:

- the use case name, description, and domain context
- the target audience together with the business objectives
- success or failure for business
- success or failure for developers
- project completeness (e.g., finished or not, and proof-of-concept (POC) or production stage)
- deployment type (e.g., on-premises vs. cloud-based)
- organizational capability maturity dimensions (e.g., project maturity, business support, and developer support)

- the training data and taxonomies used, along with associated data quality issues and metrics
- the AI, ML, and NLP methods applied
- the evaluation data and performance metrics used

The complete list of properties is provided in the Supplementary Material. The structured schema for collecting use cases was defined at the start of the project by all participating organizations and iteratively refined throughout its execution to reflect changes in the AI technology landscape. Except for the use case name and description, controlled vocabularies were used to capture the information, enabling better integration, consistency, and analysis across use cases. To protect the reputation and the intellectual property of the organizations that submitted use cases to the collection only an anonymized and restricted version of the gathered use case information could be made available. The restricted use case collection is provided in Section B of the Supplementary Material.

Each use case was classified as a success or failure based on the business perception of the organization conducting it. In other words, the assessment was made not on technical criteria, e.g., whether a predefined performance threshold was achieved or not, but on whether the use case delivered the desired business value. Projects and pilots that were successful technically but did not deliver business value were classified as failures (if complete) or excluded from analysis (if unfinished). To the best of our knowledge, this use case analysis represents the largest and most methodologically rigorous study of its kind to-date.

Executive Surveys and Interviews

In 2025 the Pistoia Alliance in collaboration with the consulting company Zuhlke Engineering AG organized a global survey of 54 senior leaders across pharmaceutical, biotech, CRO, and technology organizations, followed by in-depth interviews with 9 executives [15]. (Zuhlke Engineering AG is a member company of the Pistoia Alliance). The Pistoia Alliance / Zuhlke survey reported on the quantitative landscape: where demand is highest, where maturity lags, and where collaboration potential is strongest. The difference between this survey research method and community brainstorming works in the prior paragraph is in the autonomy of the participants: surveys are designed and guided, and leave less room for creativity of the participants, whereas brainstorming puts an emphasis on generation and discussion of new provocative ideas.

Results and Discussion

In the following section, we present the results obtained through the three information-gathering and analysis methods. The results are presented in the same order as the methods described in the previous section. These findings serve as the empirical foundation for the recommendations outlined in the final subsection of this section.

Brainstorming and Pistoia Alliance Surveys

Former results of the expert panels and community meetups have been documented in two publications in 2021 and 2024 [13, 14]. Here in total 23 common work challenges for four different business personas, i.e., data scientists, architects, model users and executives. Common challenges across personas included the ethical use of AI (e.g., with respect to privacy protection) and the compliant implementation and deployment of AI systems within governmental and regulatory frameworks.

In the most recent rounds of community brainstorming conducted in 2025 and early 2026, we identified the following findings:

- Leading organizations across the industry report a shift from incremental to transformative AI adoption. Rather than merely accelerating existing business processes and improving efficiency, increasing attention is being directed toward redesigning processes and workflows around AI capabilities. This magnifies the potential benefits of the technology and the return on investment, while increasing the risks of losses in case of failure at the same time.
- Self-reported maturity of AI governance is low: over half of the respondents describe it as “ad-hoc” or “inconsistent”. Further, the observed low level of organizational maturity appears to correlate with the low success rates of AI production deployments relative to pilot projects, as reported in independent studies [16].
- Self-reported AI literacy among R&D staff also appears to be low, with fewer than 50% of respondents reporting having received any AI training. At the same time, specialists demonstrate awareness of emerging technology trends, including agentic AI and AI “co-scientist” systems.
- AI interoperability standards are regarded as highly important by the majority of participants, with more than 60% identifying them as a priority.

Overall, the findings suggest that while organizations are increasingly pursuing transformative AI adoption and recognize the strategic importance of AI interoperability, this ambition is constrained by low governance maturity and limited AI literacy, both of which may hinder successful large-scale deployment.

Analysis of the Pistoia Alliance NLP, AI, and ML use case collection

Figure 1 illustrates the distribution of key properties captured in the dataset.

We conducted a statistical analysis of project characteristics that may correlate with, and therefore potentially predict, project success. We used the Chi-square test (at 0.05 significance level) to establish the relationships between the project success and the properties listed in the project description. The project properties that statistically correlate with the business success of AI deployments were *project maturity, business support, and developer support*. On the other hand, we did not find a statistically significant dependence between project success and any other properties including data preparation methods, encountered data quality issues, training data category, customer size, deployment type or usage of standardized taxonomies. Although these aspects intuitively influence projects, they did not statistically predict successful outcomes within this dataset, suggesting that their impact might be mitigated by other compensating project factors.

The three project properties that statistically correlate with the business success of AI deployments, i.e., project maturity, business support, and developer support, are collectively dimensions of organizational capability maturity [17]. These are the same properties that are known in the general management literature to correlate with the successes or failures of other complex engineering projects [18]. Therefore, established best practices for the planning and execution of IT projects are also applicable to AI projects. In this respect, AI projects do not appear to differ fundamentally from other IT initiatives.

Another observation that we made in this study is that certain specialized NLP algorithms had higher success rates when aligned with specific tasks. For example, multitask BERT excelled in Information Extraction task (100% success) while Lingo4G performed not perfectly, but well in the Text Classification and Topic Modeling task (~ 67% success). This observation correlates with the established Good Machine Learning Practice that calls for the *educated selection of the technologies to meet the business goals (method selection and fitness-for-purpose)* [13], that in turn demands tight collaboration between the AI professionals and the subject matter experts [14]. Other practical examples in this area are exploitation of the generative capabilities of Large Language Models in tasks such as generation of novel molecular structures or protein sequences, while limiting their use in tasks where hallucinations are critically undesirable or where a deterministic API query is sufficient [14, 19]. Use of multi-agentic AI systems that allow AI agents to perform quality control on each other's outputs ("LLM-as-a-judge") and thus drive the complex processes that require feedback and error correction is another example of the same principle [14]. In relation to this, other authors observe that agentic and autonomous AI systems may need tight control with minimal necessary agency and zero-trust design to contain business risks associated with these non-deterministic systems [20–23].

The Pistoia Alliance / Zuhlke Survey

A particularly notable finding is the consistent pattern observed across the survey responses: although internal demand for AI and machine learning (ML) is high, and their perceived strategic importance is likewise substantial, organizational maturity remains comparatively low. This pattern is consistent with the broader findings reported above and may indicate a gap between strategic intent and organizational preparedness. Such a gap may, at least in part, contribute to the relatively low success rates of production-grade AI and ML projects. The Pistoia Alliance / Zuhlke survey observed an emergent pattern of successful pilots failing to translate into production systems with ongoing impact. These cases illustrate business failure against the background of formal technical success. This is consistent with other reports of the troubling high failure rates of AI projects [16].

In contrast to the dimensions of project maturity, business support, and developer support identified in our analysis of NLP, AI, and ML use cases, respondents in the Zühlke study placed greater emphasis on the other three components of organizational maturity, namely organizational fragmentation, misaligned incentives, and insufficient change management, as underlying contributors to project failure. This divergence suggests that the barriers to successful AI implementation extend beyond project-level execution and technical support, encompassing broader organizational conditions that shape adoption and scale-up. Notably, these challenges are not unique to AI, but are well documented in earlier waves of digital transformation; however, the present findings suggest that they have yet to be adequately addressed in contemporary AI initiatives [18, 24].

This suggests that improving the success rate of AI projects requires sustained attention to the foundational best practices that strengthen organizational capability maturity. At a high level, these include business process analysis, measurement of improvement, and change management. Accordingly, the focus of AI professionals should extend beyond marginal gains in model accuracy to the broader organizational and operational factors that enable successful implementation and scaling.

Any improvements in business performance should be evaluated at the level of the human team executing a given process with the support of AI tools. In contrast to the frequently invoked notion of the "human in the loop," which can imply a subordinate role for human judgment, we advocate an "AI in the loop" perspective in which AI functions in support of human decision-making and action. From this perspective, responsibility for the outcomes of AI-supported processes remains with the relevant human operators and organizations, rather than being displaced onto the technology itself. No kind of legal sophistry can be used to avoid such responsibility. This principle is particularly important for high-risk AI systems, especially those used in clinical settings [23, 25]. Taken together, these considerations suggest that trust, rather than benchmark performance alone, may now constitute a principal limiting factor in the adoption of AI [21, 26, 27].

The second most significant observation concerns the well-established finding that most failures in AI and ML system performance originate in the data rather than in the model architecture [14, 28–32]. Robust modelling therefore depends on the availability of minimal metadata standards, consistent experimental annotation, and harmonized data engineering pipelines. In this context, AI-ready data should be understood as related to, but not synonymous with, FAIR (Findable, Accessible, Interoperable, and Reusable) data. Whereas FAIR data primarily emphasizes federated human and machine discovery and access, AI-ready data additionally requires standardized, high-throughput engineering pipelines, structurally normalized inputs, and contextual metadata that explicitly describe dataset limitations. Relevant enabling artifacts may include interoperable schemas, experimental documentation standards, imaging ingestion patterns, and dataset cards.

The final observation arising from the survey is that the current pace of technological change may exceed the ability of organizations to adequately assess fitness for purpose. This not only complicates the method selection discussed in the previous section but also highlights important opportunities for the development of shared evaluation benchmarks, model comparison frameworks, and guidance for technology lifecycle management. More broadly, the findings point to a need for common practices supporting reproducibility, explainability, traceability, and bias assessment. In this context, close collaboration between AI professionals and domain experts in the relevant application areas appears to be essential for achieving business success [33].

Common artifacts in this area may include validation protocols, bias assessment methods, audit checklists [34], and model documentation templates. Organizations may build on existing standards and guidance relating to AI governance [21–23], validation and external testing [35], and reporting [35–37]. In addition, we note that the implementation of FAIR principles and good documentation practices in AI requires dedicated model management systems capable of tracking all relevant components, including training and testing data, inputs, model architecture, and model parameters, in order to support backup, restoration, disaster recovery, business continuity, and the archival and retention of AI models [14]. The key common artifacts identified in our study are summarized in Table 2.

Table 2
The most significant focus areas and key common artifacts.

Strategic Pillar	Core Focus Areas	Key Common Artifacts
I. Organizational Maturity	• Managing AI initiatives with IT engineering rigor • Broad upskilling and baseline AI literacy • Aligning cross-functional stakeholder incentives	• Capability Maturity Frameworks • Standardized Change Management Playbooks • Shared Resource Allocation Models
II. Human-Centric Architecture	• Shifting from "Human-in-the-loop" to "AI-in-the-loop" • End-to-end business workflow redesign • Definite human accountability for outputs	• Redesigned Workflow Schemas • Human Team Performance Trackers • Human Responsibility Protocols
III. Trust, Governance & Lifecycle	• Fitness-for-purpose method and algorithm selection • Shared benchmarks & lifecycle management • Rigorous data stewardship and validation	• Validation Protocols & Checklists • Dataset Data Cards • Model Management Systems

Recommendation of Best Practices

This study indicates that the success of AI and ML projects in drug discovery is shaped less by algorithmic sophistication than by organizational capability and execution discipline. Across brainstorming exercises, survey data, and the empirical analysis of use cases, a consistent set of recommended best practices emerges:

1. **Prioritize organizational capability maturity**: The strongest predictors of business success—project maturity, business support, and developer support—reflect the broader concept of organizational readiness. AI initiatives should be managed with the same rigor as other complex IT and engineering projects, incorporating structured governance, clear ownership, and sustained resourcing.
2. **Focus on business process redesign rather than isolated models**: Transformative value arises when workflows are redesigned around AI capabilities, not when AI is incrementally inserted into existing processes. Success should be measured at the level of end-to-end business outcomes delivered by human teams augmented with AI.
3. **Adopt “AI-in-the-loop” rather than “human-in-the-loop”**: AI systems should support human decision-making, with accountability and responsibility remaining firmly with human operators. This principle is essential for maintaining trust and for managing risk in high-impact applications.
4. **Ensure strong alignment between stakeholders**: Close collaboration between AI practitioners, domain experts, and business leaders is critical for selecting fit-for-purpose methods and ensuring that solutions address real needs. Misalignment, organizational fragmentation, and weak change management are major causes of failure.
5. **Select methods based on fitness-for-purpose**: The choice of algorithms must be driven by the specific task and business objective. Matching methods to use cases—while understanding their strengths, limitations, and failure modes—is more important than pursuing state-of-the-art performance in isolation.
6. **Invest in trustworthiness and governance**: Trust, rather than raw model performance, is the primary limiting factor for adoption. Organizations should implement practices for reproducibility, explainability, traceability, bias evaluation, and auditability, supported by standardized documentation and validation protocols.
7. **Strengthen evaluation and lifecycle management**: Rapid technological change necessitates robust frameworks for benchmarking, model comparison, and lifecycle management. Shared evaluation standards and external validation practices are key enablers of reliable deployment.
8. **Improve AI literacy and organizational readiness**: Broad-based training and upskilling are required to bridge the gap between high strategic ambition and low operational maturity. Enhancing AI literacy across R&D organizations supports more effective adoption and reduces resistance to change.
9. **Adopt disciplined data and model management practices**: Although not statistically predictive of project success in isolation, data quality, metadata standards, and FAIR principles remain foundational. Proper data stewardship and model management systems are essential to ensure reproducibility, maintainability, and long-term value.
10. **Embrace iterative development with early validation**: Rapid experimentation, including early identification of unviable approaches, reduces risk and prevents the persistence of misleading models. Baseline comparisons and realistic performance assessment should be standard practice.

In summary, AI does not appear to differ fundamentally from other enterprise technologies with respect to the factors that determine implementation success. Organizations that prioritize capability maturity, human-centered deployment, and trustworthy practices are likely to be best positioned to realize the transformative potential of AI in drug discovery.

Conclusion

This study examined why AI projects in drug discovery and the broader pharmaceutical context fail to translate from technical promise into sustained business value. Drawing on a multi-year Pistoia Alliance initiative, we combined evidence from community brainstorming workshops, a global executive survey and interviews, and a statistical analysis of 67 AI/ML use cases. Across these complementary sources, a consistent picture emerged.

The principal finding is that the success of AI and ML initiatives depends less on algorithmic sophistication than on organizational capability maturity and execution discipline. In the use case analysis, project maturity, business support, and developer support were the factors most strongly associated with business success, whereas technical characteristics such as model type, deployment type, and data category were not independently predictive. Survey and workshop findings reinforced this pattern, showing high strategic ambition and demand for AI alongside comparatively low maturity in governance, change management, and workforce readiness, which may help explain why many technically successful pilots fail to scale.

Overall, the findings indicate that AI should be treated not as an exceptional technology, but as a form of enterprise innovation whose success depends on organizational readiness, stakeholder alignment, and disciplined implementation. Organizations that invest in capability maturity, human-centered deployment, and trustworthy governance will be best positioned to realize the transformative potential of AI in drug discovery and pharma.

Declarations

• The funding for this work was provided by AstraZeneca, Boehringer-Ingelheim, and Roche

• Mario Sänger is an employee of AstraZeneca

Roman Saiz

• Roman Saiz is an employee of Roche

Competing Interests

• The funding for this work was provided by AstraZeneca, Boehringer-Ingelheim, and Roche• Mario Sänger is an employee of AstraZeneca• Roman Saiz is an employee of Roche• Matthias Hartung is an

employee of Boehringer-Ingelheim• Vladimir Makarov receives consulting fees from the Pistoia Alliance

Author Contribution

VM planned and coordinated the research, secured funding, wrote the text of the paper; MS wrote the text of the paper, created Figure 1 and Table 1; MH wrote the text of the paper; NL wrote the text of the paper and created Table 2; all authors reviewed the manuscript

Acknowledgements

The funding for this work was provided by AstraZeneca, Boehringer-Ingelheim, and Roche (Pistoia Alliance member companies), and is gratefully acknowledged.

References

1. Madani A, Krause B, Greene ER et al (2023) Large language models generate functional protein sequences across diverse families. Nat Biotechnol. https://doi.org/10.1038/s41587-022-01618-2
2. Grisoni F (2023) Chemical language models for de novo drug design: Challenges and opportunities. Current Opinion in Structural Biology. https://doi.org/10.1016/j.sbi.2023.102527
3. Obrezanova O (2023) Artificial intelligence for compound pharmacokinetics prediction. Curr Opin Struct Biol. https://doi.org/10.1016/j.sbi.2023.102546
4. Cáceres EL, Tudor M, Cheng AC (2020) Deep learning approaches in predicting ADMET properties. Future Med Chem. https://doi.org/10.4155/fmc-2020-0259
5. Harrer S, Shah P, Antony B, Hu J (2019) Artificial Intelligence for Clinical Trial Design. Trends Pharmacol Sci 40:577–591. https://doi.org/10.1016/j.tips.2019.05.005
6. Shyr C, Grout RW, Kennedy N, Akdas Y, Tischbein M, Milford J et al (2024) Leveraging artificial intelligence to summarize abstracts in lay language for increasing research accessibility and transparency. J Am Med Inf Assoc 31:2294–2303. https://doi.org/10.1093/jamia/ocae186
7. Gottweis J, Weng WH, Daryin A et al (2026) Accelerating scientific discovery with Co-Scientist. Nature. https://doi.org/10.1038/s41586-026-10644-y
8. Huang K, Zhang S, Wang H, Qu Y, Lu Y, Roohani Y et al (2025) Biomni: A general-purpose biomedical AI agent. bioRxiv. https://doi.org/10.1101/2025.05.30.656746
9. Ghareeb AE, Chang B, Mitchener L et al (2026) A multi-agent system for automating scientific discovery. Nature. https://doi.org/10.1038/s41586-026-10652-y
10. Aygün E, Belyaeva A, Comanici G et al (2026) An AI system to help scientists write expert-level empirical software. Nature. https://doi.org/10.1038/s41586-026-10658-6
11. Amazon Web Services (2026) AWS for bio discovery and AI drug research. https://www.aboutamazon.com/news/aws/aws-amazon-bio-discovery-ai-drug-research. Accessed

26 May 2026

12. OpenAI (2026) Introducing GPT Rosalind. https://openai.com/index/introducing-gpt-rosalind/. Accessed 26 May 2026
13. Makarov VA, Stouch T, Allgood B, Willis CD, Lynch N (2021) Best practices for artificial intelligence in life sciences research. Drug Discov Today 26:1107–1110. https://doi.org/10.1016/j.drudis.2021.01.017
14. Makarov V, Chabbert C, Koletou E, Psomopoulos F, Kurbatova N, Ramirez S et al (2024) Good machine learning practices: Learnings from the modern pharmaceutical discovery enterprise. Comput Biol Med 177:108632. https://doi.org/10.1016/j.compbiomed.2024.108632
15. Kharde S, Ramos dos Santos T, Gill R, Maskell C, Upton B (2025) Advancing AI and ML in life sciences R&D. Pistoia Alliance, Zühlke. https://marketing.pistoiaalliance.org/hubfs/Pistoia-Zuhlke%20whitepaper%202025.pdf
16. Project NANDA (2025) State of AI in business report. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf
17. Curley M (2008) Introducing an IT Capability Maturity Framework. In: Filipe J, Cordeiro J, Cardoso J (eds) Enterprise Information Systems. Springer, Berlin. https://doi.org/10.1007/978-3-540-88710-2_6
18. Fetais A, Abdella GM, AlKhalifa KN, Hamouda AM (2022) Business Process Re-Engineering: A Literature Review-Based Analysis of Implementation Measures. Information 13:185. https://doi.org/10.3390/info13040185
19. Daugherty PR, Wilson HJ (2018) Human + machine: reimagining work in the age of AI. Harvard Business Review, Boston
20. Steidl M, Felderer M, Ramler R (2023) The pipeline for the continuous development of artificial intelligence models—Current state of research and practice. J Syst Softw 199:111615. https://doi.org/10.1016/j.jss.2023.111615
21. Consumer Technology Association (2020) The use of artificial intelligence in healthcare: trustworthiness. ANSI/CTA 2090
22. Consumer Technology Association (2021) Guidelines for developing trustworthy artificial intelligence systems
23. US FDA, Health Canada MHRA (2021) Good machine learning practice for medical device development. https://www.fda.gov/media/153486/download Accessed May 29, 2026
24. Kotter JP, Schlesinger LA (2008) Choosing strategies for change. Harv Bus Rev 86:130–139
25. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P et al (2020) Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 368:l6927. https://doi.org/10.1136/bmj.l6927
26. Cabitza F, Zeitoun JD (2019) The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. Ann Transl Med 7:161. https://doi.org/10.21037/atm.2019.04.07

27. Char DS, Abràmoff MD, Feudtner C (2020) Identifying Ethical Considerations for Machine Learning Healthcare Applications. Am J Bioeth 20:7–17. https://doi.org/10.1080/15265161.2020.1819469
28. Beck B, Seeliger D, Kriegl JM (2015) The impact of data integrity on decision making in early lead discovery. J Comput Aided Mol Des 29:911–921. https://doi.org/10.1007/s10822-015-9871-2
29. Wise J, Grebe de Barron A, Splendiani A, Balali-Mood B, Vasant D, Little E et al (2019) Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discov Today 24:933–938. https://doi.org/10.1016/j.drudis.2019.01.008
30. European Commission (2018) Cost of not having FAIR research data. https://doi.org/10.2777/02999 Accessed May 29, 2026
31. Schneider G (2018) Automating drug discovery. Nat Rev Drug Discov 17:97–113. https://doi.org/10.1038/nrd.2017.232
32. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G et al (2019) Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18:463–477. https://doi.org/10.1038/s41573-019-0024-5
33. Littmann M, Selig K, Cohen-Lavi L et al (2020) Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. Nat Mach Intell 2:18–24. https://doi.org/10.1038/s42256-019-0139-8
34. Eid FE, Elmarakeby HA, Chan YA et al (2021) Systematic auditing is essential to debiasing machine learning in biology. Commun Biol 4:183. https://doi.org/10.1038/s42003-021-01674-5
35. Walsh I, Fishman D, Garcia-Gasulla D et al (2021) DOME: recommendations for supervised machine learning validation in biology. Nat Methods 18:1122–1127. https://doi.org/10.1038/s41592-021-01205-4
36. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH (2020) MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. J Am Med Inf Assoc 27:2011–2015. https://doi.org/10.1093/jamia/ocaa088
37. Liu X, Cruz Rivera S, Moher D et al (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 26:1364–1374. https://doi.org/10.1038/s41591-020-1034-x

Figures

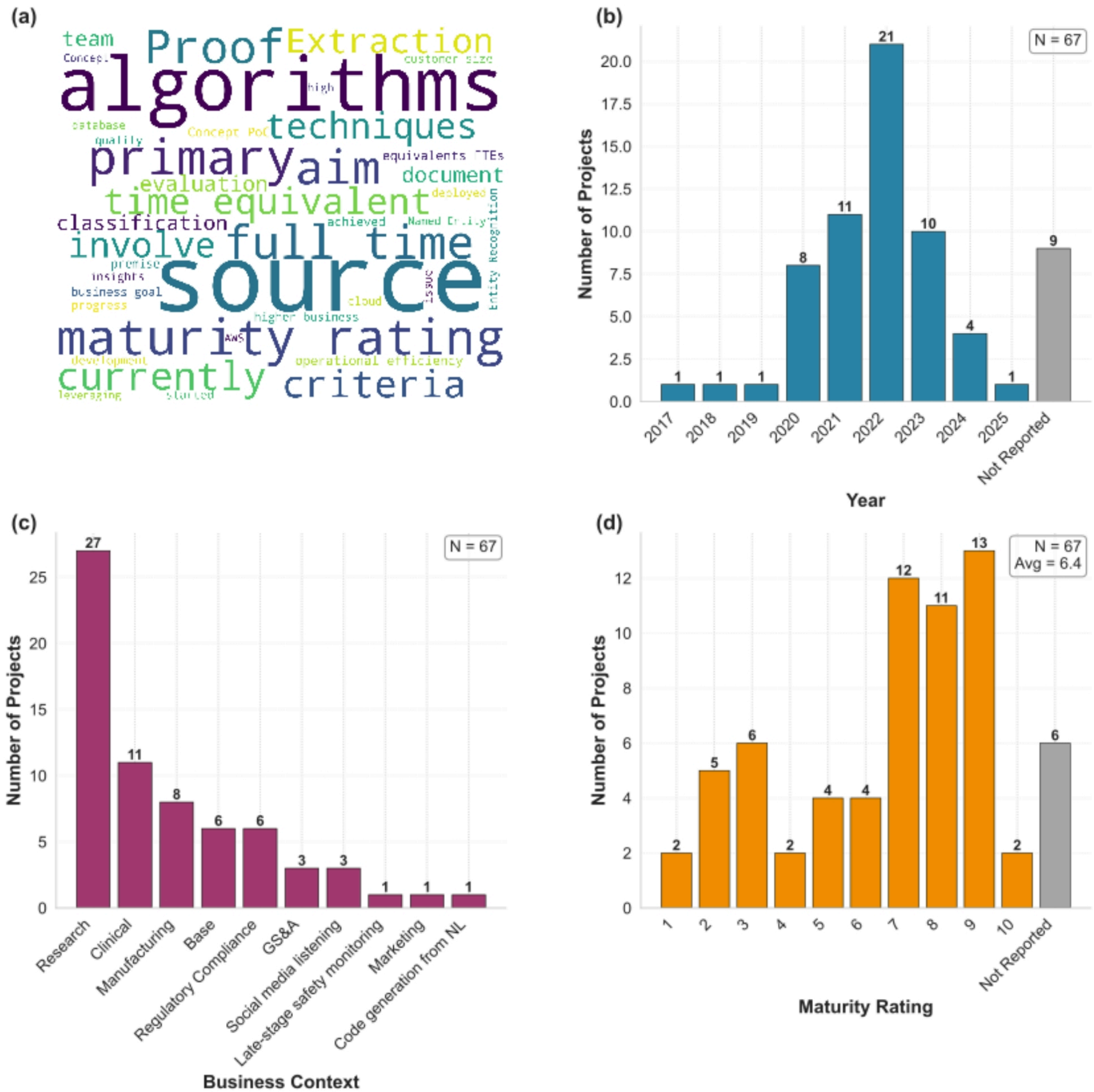

Figure 1

Insights into the 67 use cases included in the public, anonymized version of the database: a) a word cloud based on project titles and descriptions, highlighting key concepts that recur across use cases; b) an overview of the number of projects initiated each year; c) the distribution of use cases across different business contexts and application areas; and d) the maturity levels of the use cases, as reported by the conducting organizations. On average, a maturity level of 6.4 was reported.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S1.PistoiaAllianceNLPandLLMusecasecollectiondimensions.xlsx
- S2.PistoiaAllianceNLPandLLMusecasecollection.xlsx