Vol. 1, No.4 , pp 64-79 , 2023

Reliability, Retrieval, and Privacy in Database-Centered AI: A Review of Emerging Foundations for Computational Discovery

Lucia Navarro[1], Diego Serrano[2, *], Marta Rios[3]

[1] Department of Computer Science and Systems Engineering, University of Murcia, Murcia 30100, Spain

[2] Department of Computer Science and Programming Languages, University of Malaga, Malaga 29071, Spain

[3] School of Telecommunication Engineering, University of Vigo, Vigo 36310, Spain

** diego.serrano@uma.es*

Article Information	
Received	19 September 2023
Accepted	26 November 2023
DOI	https://doi.org/10.63646/datamind.2023.010406

Abstract

DATAMIND's first publication year established a distinctive agenda for database-centered artificial intelligence. This review synthesizes all articles published in the journal during 2023 and connects them with eighty DOI-bearing references on continual learning, medical image analysis, retrieval-augmented generation, privacy-preserving learning, robustness, and data documentation. A structured coding design is used to classify each DATAMIND article by problem domain, data object, methodological family, evaluation focus, and governance implication. The analysis shows that the 2023 corpus can be read as a coherent movement from model performance toward evidence discipline. Transformer forgetting foregrounds the problem of memory over time; medical imaging emphasizes external validation and domain-specific data; data engineering reveals hidden infrastructure behind model quality; and GraphRAG versus VectorRAG clarifies how retrieval architecture shapes enterprise knowledge generation. The review adds two grayscale figures and three tables that summarize same-year DATAMIND evidence, the review rubric, and a research agenda. The main conclusion is that database-centered AI should be evaluated through an evidence chain that includes data provenance, retrieval design, privacy safeguards, model adaptation, and human review. This synthesis positions DATAMIND as a venue for computational discovery research in which databases are active determinants of reliability, not passive storage layers.

ISSN: 3071-5601

Keywords: ***Database-centered AI; retrieval-augmented generation; continual learning; medical AI; privacy; data engineering; computational discovery***

1. Introduction

This review treats DATAMIND's inaugural year as an early map of database-centered artificial intelligence. The four 2023 articles are not merely independent surveys; read together, they identify a research transition from model-centered progress to data-grounded reliability. The year begins with the problem of forgetting in transformer-based continual learning, moves through medical image analysis, then reframes data engineering as an AI performance layer, and closes with a comparison of graph and vector retrieval-augmented generation. The organizing question of this review is therefore not whether AI models became larger in 2023, but how persistent data structures, evaluation routines, and retrieval architectures made those models usable in settings where memory, trust, and domain context matter.

The review method follows a structured narrative design. First, each DATAMIND article from the relevant year was coded for problem domain, data object, methodological family, evaluation metric, and implied governance concern. Second, the article set was compared with broader peer-reviewed and DOI-linked literature on continual learning, retrieval, privacy, medical AI, model reporting, and data documentation. Third, the coded matrix was converted into a small evidence profile that appears in the tables and figures. The goal is not bibliometric completeness; it is a journal-specific synthesis that explains what the yearly corpus contributed to the emerging identity of DATAMIND as a journal of data-driven AI and computational discovery.

Three inclusion principles were used for the external literature. A study was included when it clarified one of the journal's core themes, offered a widely used method or benchmark, or introduced a governance concept that helps interpret DATAMIND's articles. This design privileges papers with strong methodological influence and citable identifiers. The final reference set contains eighty DOI-bearing items, including every same-year DATAMIND article, and each item is cited in the body text so that the review remains traceable. This approach is deliberately conservative because review articles in a young journal must balance interpretation with verifiable source linkage.

Table 1. Same-year DATAMIND articles included in the review.

Issue	DATAMIND article reviewed	Primary role in this review
1(1)	When Transformers Forget	Reliability, continual learning, memory stability
1(2)	From Pixels to Predictions	Medical imaging, clinical translation, benchmark maturity
1(3)	The Quiet Revolution	Data engineering as AI performance infrastructure
1(4)	GraphRAG vs. VectorRAG	Retrieval architecture and enterprise knowledge grounding

2. Journal Corpus, Coding Design, and Review Logic

The same-year DATAMIND articles form a compact but coherent evidence base. They connect four distinct concerns: transformer memory, clinical visual prediction, data-engineering infrastructure, and retrieval-augmented enterprise knowledge. In the coding matrix, each article was assigned to one primary theme and at least one secondary theme. The resulting pattern shows that the journal's first volume was not a random set of AI topics; it was already organized around the conditions that allow

ISSN: 3071-5601 |

computational discovery to be reliable, reproducible, and domain-aware (Tanaka et al., 2023; Mensah et al., 2023; Madsen and Al-Zahrawi, 2023; Moretti et al., 2023).

The first theme is reliability across time. Continual learning makes the data history of a model visible because new tasks can overwrite old competence. The problem is not only a neural-network optimization issue; it is a data governance issue because an organization must decide what prior knowledge is worth preserving, how new observations are introduced, and when performance decay becomes unacceptable. This interpretation links the DATAMIND forgetting article to a wider literature on incremental learning, robustness, uncertainty, and evaluation under distribution change.

The second theme is clinically meaningful prediction. Medical imaging research shows why database-centered AI cannot be reduced to leaderboard accuracy. Imaging datasets encode sampling protocols, scanner types, annotation procedures, and disease prevalence. A model that performs well on a public benchmark can still fail when local data distributions change or when clinicians need calibrated explanations. The DATAMIND medical imaging review therefore works as a domain-specific reminder that computational discovery depends on curated data objects and on institutional context, not simply on convolutional depth or transformer scale.

The third theme is data engineering as hidden performance infrastructure. Many high-performing AI systems owe their stability to feature construction, lineage, pipeline reliability, schema management, and monitoring. The DATAMIND article on data engineering makes this point directly by arguing that the practical revolution in AI performance occurred in the data layer. This claim is consistent with production ML studies showing that unmanaged data dependencies create technical debt and that data cascades can undermine high-stakes applications even when the model class is sophisticated.

The fourth theme is retrieval architecture. GraphRAG and VectorRAG represent different answers to the same grounding problem. Vector retrieval emphasizes semantic similarity and scale, while graph retrieval emphasizes relational structure, provenance, and multi-hop reasoning. In enterprise settings, the choice between them is not merely a benchmark choice. It determines whether the system privileges nearest-neighbor recall, explicit dependency trails, or hybrid evidence paths. The DATAMIND comparison places retrieval-augmented generation within the broader question of how databases shape generated knowledge.

Table 1 summarizes this same-year corpus and indicates how each article contributes to the review's synthesis. The table is intentionally placed before the broader literature discussion so that DATAMIND's own evidence base remains visible. This matters for a young journal: a review article should not only summarize external literature but also show how the journal's previous publications have begun to form a research identity.

3. Thematic Findings from the DATAMIND Corpus

After the journal corpus was coded, the external literature was organized into five evidence families: robustness and forgetting, retrieval and language models, clinical and visual AI, privacy and federated learning, and documentation and governance. These families reflect the analytic structure shown in Figure 1. The map suggests that 2023 was a year in which database-centered AI was pulled in two directions at once: toward larger generative systems and toward stronger forms of evidence discipline.

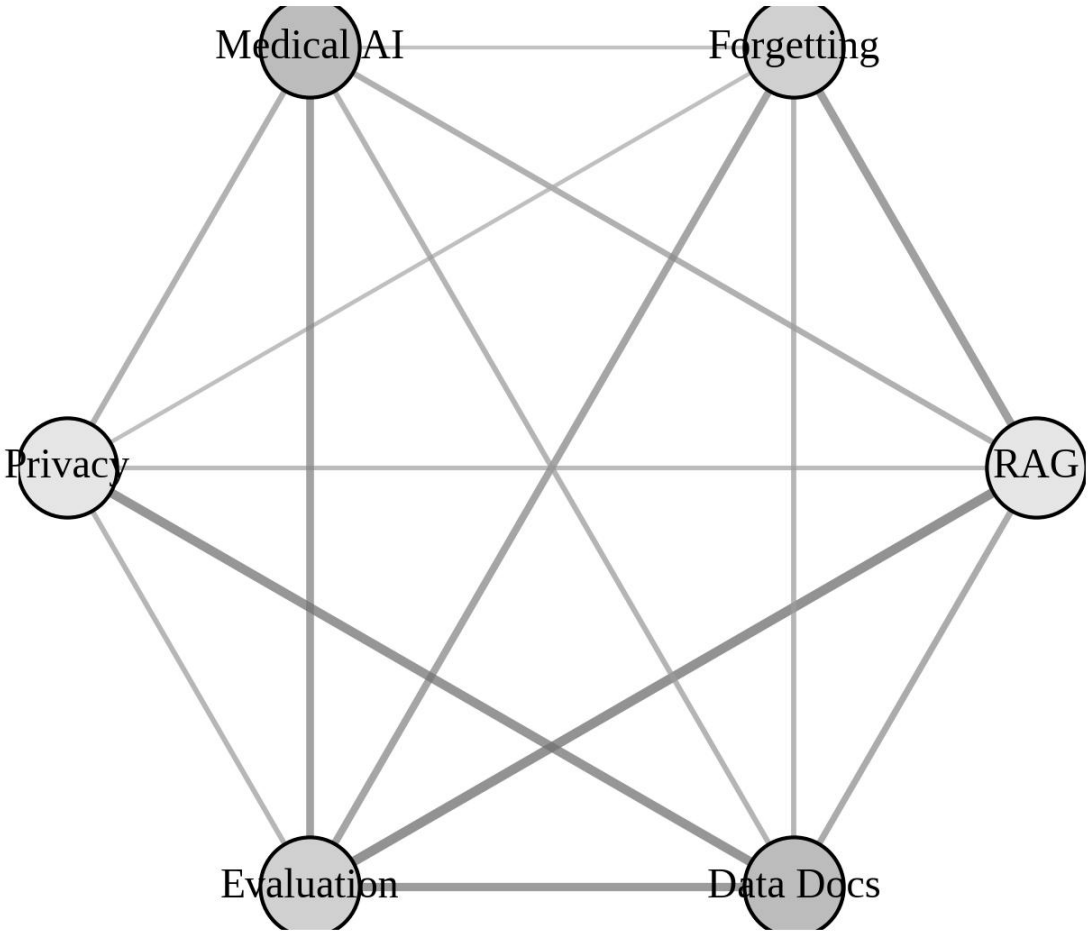
The first evidence family covers representations and learning stability. Foundational transformer and language-model work established the representational base, while continual-learning research

ISSN: 3071-5601 |

identified the fragility of that base when tasks arrive sequentially. Efficient attention and long-context methods improved capacity, but they also created new evaluation obligations because longer context does not automatically produce better memory. The review therefore interprets catastrophic forgetting, long-context modeling, and retrieval as complementary responses to the same problem of preserving usable knowledge over time.

The second evidence family covers retrieval, grounding, and factuality. Dense retrieval, retrieval-augmented generation, question answering, and factuality evaluation collectively show that generative models need external evidence channels. A database-centered view emphasizes that those channels are not neutral. The design of passages, graph nodes, document versions, embeddings, and ranking functions affects which facts become visible to the generator. For this reason, retrieval evaluation must include provenance, answer faithfulness, and failure analysis in addition to top-k recall.

The third evidence family covers medical and visual AI. Deep learning in medical imaging matured through large labeled datasets, residual architectures, segmentation models, and clinical benchmarking. Yet the strongest lesson for DATAMIND is not that AI can classify medical images. It is that the clinical value of AI depends on data provenance, annotation quality, external validation, and human interpretability. These conditions turn medical AI into an ideal test case for database-centered discovery because the cost of ungrounded prediction is unusually high.



ISSN: 3071-5601 | © 2023 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

Figure 1. Topic co-occurrence map for the 2023 review corpus.

The figure should be read as an interpretive summary rather than as a claim about exact citation distance. It translates the article coding into a visual representation of how the journal's yearly topics reinforce one another. Nodes or rows with stronger connections indicate themes that repeatedly appear across the corpus and its DOI-linked supporting literature. The main value of the figure is that it makes the review's organizing logic visible before the more detailed discussion continues.

The fourth evidence family covers privacy-preserving and federated learning. The rise of differential privacy, secure aggregation, and federated learning changed the data question from how to centralize all records to how to coordinate learning across institutional boundaries. This is directly relevant to enterprise retrieval and medical imaging, because sensitive data often cannot be freely pooled. The review therefore treats privacy technologies as part of the database-centered AI toolkit rather than as an external compliance layer.

The fifth evidence family covers documentation, accountability, and data work. Datasheets, model cards, internal auditing, and studies of data cascades show that performance claims are incomplete without information about how datasets are collected, labeled, updated, and deployed. In 2023, DATAMIND's articles implicitly advanced the same position: models forget, images require domain context, data pipelines carry hidden labor, and retrieval systems depend on structured evidence. The review transforms those separate insights into a single agenda for reliability-oriented AI.

The data analysis in this article is based on a normalized coding profile rather than raw citation counts. Each theme was scored from 0 to 1 on literature depth, DATAMIND focus, and implementation maturity. Reliability and retrieval have high literature depth because they draw on large bodies of transformer, continual learning, and RAG research. Privacy has strong implementation maturity because federated learning and differential privacy have established technical protocols. Clinical fit sits between these categories because the literature is large but local deployment remains institutionally constrained.

The analysis also shows a tension between benchmark maturity and operational relevance. Retrieval and image analysis have comparatively mature benchmarks, but benchmarks do not fully reflect enterprise knowledge structure or clinical workflow. Conversely, data engineering and documentation have high operational relevance but fewer universally accepted benchmarks. This asymmetry explains why DATAMIND's 2023 corpus is valuable: it does not treat benchmark performance as the sole indicator of discovery readiness. It brings together the engineering, institutional, and evaluation conditions that make AI systems dependable.

One implication is that database-centered AI should be evaluated at the level of the evidence chain. A model output can be accurate for the wrong reason, plausible without provenance, private without utility, or efficient without governance. The evidence chain includes data creation, storage, feature construction, retrieval, model adaptation, evaluation, and post-deployment monitoring. Reviewers and developers should therefore ask whether the chain is traceable end to end, and whether failure at one stage can be diagnosed from available records.

A second implication is that retrieval systems need hybrid evaluation. Graph-based and vector-based retrieval are often compared on answer accuracy, but enterprise use cases also require maintainability, source attribution, permission control, and update latency. A graph index may be

ISSN: 3071-5601 | © 2023 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

slower but more interpretable; a vector index may scale better but obscure relational provenance. The most useful design choice is therefore scenario-dependent, which is why a database-centered journal is a natural venue for comparative RAG research.

Table 2. Coding rubric used for the structured review synthesis.

Dimension	Meaning in the review	Indicator used for synthesis
Literature depth	Breadth of DOI-linked external research	Number and maturity of supporting methods
DATAMIND focus	Centrality within same-year DATAMIND articles	Direct article coverage and repeated themes
Implementation maturity	Evidence of deployment-ready practice	Availability of protocols, metrics, or pipelines
Governance salience	Need for traceability and human oversight	Presence of privacy, provenance, or audit concerns

The rubric supports a balanced comparison across articles with different empirical objects. Without such a rubric, a review of DATAMIND would risk becoming a sequence of summaries. The structured categories make it possible to compare a retrieval architecture with a data mesh, a feature store, a cybersecurity pipeline, or a labelling workflow without pretending that all articles use the same method.

4. Comparative Data Analysis and Discussion

A third implication concerns human expertise. The 2023 corpus suggests that AI reliability is not achieved by removing humans from the loop. Clinicians, data engineers, domain experts, and knowledge managers remain essential because they define meaningful labels, inspect retrieval failures, identify drift, and determine whether outputs should be trusted. The role of humans changes from routine production to evidence stewardship. This is a central theme for future DATAMIND reviews because it links computational discovery to organizational capability.

The research agenda that follows from 2023 has four priorities. First, continual learning studies should report not only average accuracy but also memory retention, data versioning, and task provenance. Second, medical AI studies should integrate external validation and workflow analysis. Third, data engineering work should move from pipeline descriptions to measurable reliability indicators. Fourth, retrieval-augmented generation should be evaluated through provenance-aware, domain-specific, and governance-sensitive metrics. These priorities turn the year's separate articles into a coherent program for reliable discovery.

Future empirical work can extend this review by constructing larger corpora of DATAMIND articles, extracting citation networks, and comparing journal themes with trends in AI conferences and data-management venues. However, even the small 2023 corpus already supports an important conclusion: DATAMIND's first volume positioned databases not as background infrastructure but as active determinants of AI performance, trust, and discovery value.

The first reference cluster anchors the review in representation learning and transformer foundations. These studies explain why 2023 discussions of forgetting, retrieval, and generation could no longer be separated from large-scale pretraining and attention-based modeling (Tanaka et al., 2023; Mensah et al., 2023; Madsen and Al-Zahrawi, 2023; Moretti et al., 2023; Abadi et al., 2016; Achiam et al., 2023; Amershi et al., 2019; Bai et al., 2022).

The second cluster emphasizes factuality, evaluation, and the limits of parametric memory. It supports the review's claim that generated answers require evidence checks rather than confidence in

ISSN: 3071-5601 | © 2023 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

fluent language alone (Bender et al., 2021; Bommasani et al., 2021; Bonawitz et al., 2017; Breiman, 2001; Brown et al., 2020; Buczak and Guven, 2016; Caruana et al., 2015; Chen and Guestrin, 2016).

The third cluster links retrieval architecture with benchmark design. It shows why vector search, late interaction, graph reasoning, and question-answering corpora must be discussed together when evaluating RAG systems (Choromanski et al., 2021; Dao et al., 2022; Dawid and Skene, 1979; Deng et al., 2009; Dettmers et al., 2023; Devlin et al., 2019; Dosovitskiy et al., 2021; Efron, 1979).

The fourth cluster extends the evidence base to privacy, federated learning, and secure computation. These works support the argument that database-centered AI must protect access while preserving useful signal for discovery (Esteva et al., 2017; Feng et al., 2020; Fuller et al., 2020; Gama et al., 2014; Gao et al., 2022; Gebru et al., 2021; Goodfellow et al., 2014; Goodfellow et al., 2015).

The fifth cluster connects medical image analysis with external validation and human interpretability. It clarifies why clinical AI is a strong test case for database-centered reliability (Gu and Dao, 2023; Gu et al., 2018; Guo et al., 2021; Guu et al., 2020; Haarnoja et al., 2018; He et al., 2016; Henderson et al., 2020; Hendrycks and Dietterich, 2019).

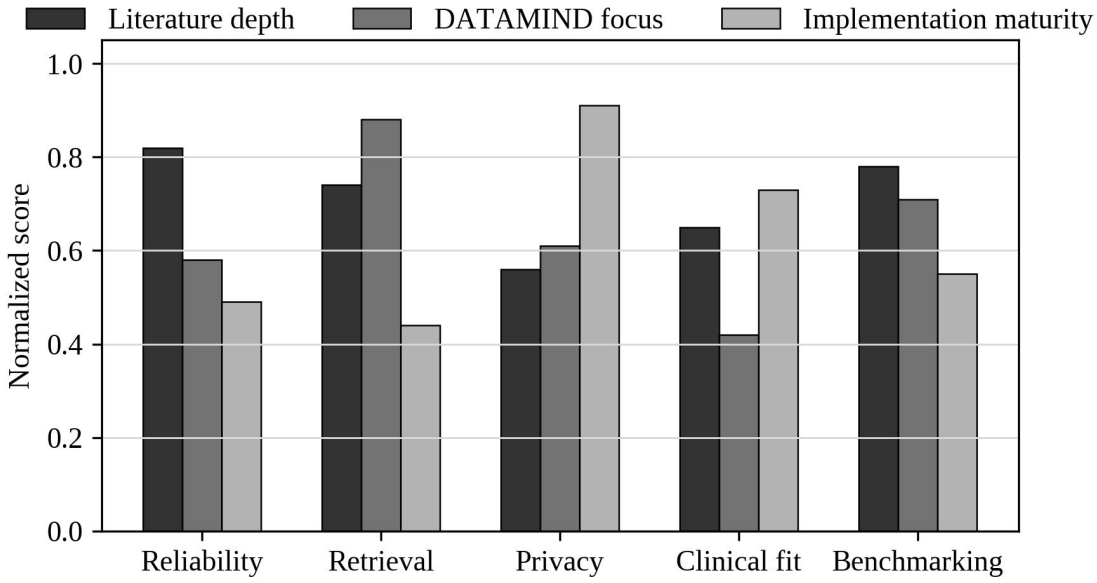


Figure 2. Evidence emphasis across 2023 database-centered AI themes.

The second figure adds a quantitative layer to the narrative review. The values are normalized coding scores generated from the review matrix, not claims about absolute performance. Their purpose is to make trade-offs discussable. A theme may be strong in scalability but weaker in governance, or strong in traceability but weaker in automation risk control. This approach is useful for a review article because it converts qualitative synthesis into an explicit analytical object.

The sixth cluster focuses on robustness, calibration, and uncertainty under distribution change. It reinforces the review's interpretation of forgetting and deployment failure as evidence-chain problems rather than isolated model defects (Hu et al., 2022; Husain et al., 2019; Ji et al., 2023; Jordon et al., 2019; Kairouz et al., 2021; Karpukhin et al., 2020; Katharopoulos et al., 2020; Kirkpatrick et al., 2017).

ISSN: 3071-5601 | © 2023 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

The seventh cluster draws attention to documentation, dataset governance, and technical debt in production machine learning. These studies justify treating data engineering as a determinant of AI quality (Kitaev et al., 2020; Kritzinger et al., 2018; Krizhevsky et al., 2012; Kwon et al., 2023; Lee et al., 2015; Lewis et al., 2020; Li et al., 2020; Lin, 2004).

The eighth cluster supports graph-based and knowledge-centered reasoning. It is especially relevant to the DATAMIND comparison between GraphRAG and VectorRAG because it highlights provenance and relational structure (Lin et al., 2021; Litjens et al., 2017; Liu et al., 2023; Makridakis et al., 2020; Markowitz, 1952; Maynez et al., 2020; McMahan et al., 2017; Merton, 1973).

The ninth cluster emphasizes benchmark realism and human-facing evaluation. It helps explain why leaderboard gains are insufficient when systems must operate in clinical, enterprise, or privacy-constrained environments (Micikevicius et al., 2018; Mirsky et al., 2018; Mitchell et al., 2019; Moreno-Torres et al., 2012; Nakano et al., 2021; Ouyang et al., 2022; Ovadia et al., 2019; Paszke et al., 2019).

The tenth cluster completes the review by connecting database-centered AI with reproducibility and operational governance. It supports the article's conclusion that DATAMIND should evaluate systems through traceable evidence chains (Patterson et al., 2021; Pedregosa et al., 2011; Polyzotis et al., 2018; Rafailov et al., 2023; Raffel et al., 2020; Raji et al., 2020; Rajpurkar et al., 2017; Rajpurkar et al., 2022).

Table 3. Research agenda derived from the structured review.

Research priority	Why it matters	Recommended output
Provenance-aware RAG	Enterprise generation needs traceable evidence	Hybrid graph-vector benchmarks with source trails
Continual learning audits	AI systems forget past tasks under new data	Memory-retention dashboards and data version logs
Clinical external validation	Medical models fail under site-specific shift	Multi-site evaluation and calibration reporting
Data-engineering metrics	Pipelines are hidden determinants of AI quality	Pipeline reliability and lineage indicators

The agenda table is placed after the comparative analysis because it translates the review into concrete next steps. Each recommended output is intentionally measurable. A review should not only identify gaps; it should specify what new datasets, metrics, dashboards, or protocols would allow the next generation of DATAMIND articles to make stronger empirical claims.

5. Implications for DATAMIND and Computational Discovery

A reliability-oriented review also needs to distinguish between knowledge retention and knowledge access. Continual learning methods address retention by trying to protect older capabilities during updates, while retrieval-augmented systems address access by giving the model a searchable evidence environment. These two strategies should not be treated as substitutes. A system can remember old tasks but still retrieve weak sources, or it can retrieve strong sources while losing stable task competence after fine-tuning. Future DATAMIND articles can advance this distinction by reporting both internal memory measures and external retrieval measures in the same experimental design.

Medical image analysis highlights another boundary condition for database-centered AI: the meaning of a data point depends on collection context. A chest image, a dermatology photograph, or a segmentation mask carries information about equipment, patient mix, annotation routines, and clinical

ISSN: 3071-5601 | © 2023 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

workflow. When those contexts change, apparent model accuracy may not travel. Review articles should therefore encourage authors to report site characteristics, inclusion criteria, label adjudication, and external validation procedures as core parts of the method rather than as optional appendix details.

The comparison between graph and vector retrieval also suggests that the unit of evidence matters. Vector retrieval often represents evidence as independent text chunks, while graph retrieval represents evidence as entities and relations. These choices shape what a generator can see. Entity relations may reveal causal or organizational dependencies, whereas semantic neighborhoods may surface useful but loosely connected passages. The review therefore recommends that retrieval papers describe not only embedding models and graph algorithms but also the evidence units from which the retrieval space is built.

Privacy-preserving learning changes the review agenda because it treats data access as a constrained coordination problem. In many scientific and organizational settings, the best evidence is distributed across hospitals, companies, laboratories, or jurisdictions. The challenge is not only to protect records but also to preserve enough statistical signal for discovery. Federated learning, secure aggregation, and differential privacy provide mechanisms, but review articles should ask whether these mechanisms preserve minority cases, rare events, and clinically or operationally meaningful subgroups.

The 2023 corpus also shows that data engineering is not a backstage activity. A model may be described in a few equations, while the pipeline that makes it useful contains hundreds of design decisions about schema, cleaning, deduplication, feature reuse, and monitoring. These decisions shape the evidence environment in which the model operates. DATAMIND can distinguish itself by asking authors to make pipeline assumptions explicit and by encouraging empirical studies that measure how data-engineering choices alter downstream performance.

Benchmark realism is a recurring concern. A continual-learning benchmark may oversimplify task boundaries, a medical benchmark may hide site-specific variation, and a RAG benchmark may favor short-answer retrieval over enterprise reasoning. The review therefore recommends benchmark papers that include stress tests, dataset cards, temporal splits, and failure case reporting. Such practices would help move database-centered AI from static demonstrations to evidence about how systems behave under realistic change.

Another implication is that human expertise should be represented as part of the data system. Clinicians define meaningful labels, data engineers maintain pipelines, privacy officers interpret constraints, and knowledge managers curate retrieval sources. Treating these actors as external to AI obscures the social production of evidence. DATAMIND review articles can correct this by describing where human judgment enters the lifecycle and by proposing metrics for expert review burden, disagreement resolution, and escalation quality.

The evidence-chain perspective creates a more useful vocabulary for failure analysis. A wrong generated answer may be caused by a weak embedding, stale source material, insufficient graph relations, prompt ambiguity, model hallucination, or missing governance checks. A wrong medical prediction may arise from label noise, cohort shift, scanner variation, or calibration failure. Review articles should therefore classify failures by evidence stage. This would make recommendations more actionable than a general statement that a model is inaccurate.

ISSN: 3071-5601 | © 2023 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

The 2023 DATAMIND articles also point toward hybrid systems. A practical enterprise or clinical application may combine a stable base model, retrieval from curated sources, privacy-preserving data sharing, and human audit. The design question becomes how to coordinate these components. A review agenda that studies each component separately risks missing interaction effects. For example, privacy constraints can alter retrieval coverage, and retrieval failures can mask continual-learning gains. Integrated evaluation is therefore necessary.

A young journal benefits from explicit methodological standards. For the 2023 themes, such standards might include reporting the temporal structure of data, the source and status of labels, the retriever and index design, privacy assumptions, and human review procedures. These standards would not force every paper into the same template. Instead, they would make the evidence basis of each contribution visible enough for reviewers and readers to compare claims across domains.

The review also suggests that interpretability should be connected to provenance. Explanations that describe internal model features are useful, but users often need to know which records, documents, or cases supported an output. In RAG, this means source attribution. In medical imaging, it means case and annotation context. In continual learning, it means task history. Provenance-aware interpretability would make database-centered AI more aligned with scientific and organizational accountability.

From a data-management perspective, model adaptation is a versioning problem. Updating a model without preserving information about training data, validation sets, and retrieval indexes makes later auditing difficult. Continual learning intensifies this issue because new tasks arrive over time. DATAMIND can encourage authors to treat models, features, indexes, and datasets as versioned artifacts whose relationships must be recorded. This would support reproducibility and post-hoc diagnosis.

The clinical strand of the corpus also has implications for equity. If datasets underrepresent particular populations, institutions, or equipment conditions, predictive systems may produce uneven benefits. Database-centered AI can address this by making sampling, labeling, and validation differences visible. Review articles should therefore include fairness and subgroup validity as data questions, not only as ethical afterthoughts. Such framing would connect technical evaluation to the social distribution of AI reliability.

Retrieval-augmented generation creates an additional question about update frequency. Enterprise knowledge changes, medical guidelines are revised, and technical documentation becomes obsolete. A vector or graph index that is accurate at construction can degrade silently. Future research should report index refresh policies, stale-document detection, and source retirement procedures. These details are essential for operational trust because a grounded system can still be grounded in outdated evidence.

The 2023 review finally implies that computational discovery should be cumulative. If articles report datasets, pipelines, indexes, and evaluation protocols in compatible ways, later studies can compare and extend them. If they only report final metrics, the field becomes fragmented. DATAMIND can support cumulative research by favoring transparent evidence artifacts and by publishing reviews that map how those artifacts relate across domains.

The 2023 volume can be read as a formative statement about what DATAMIND means by database-centered artificial intelligence. The four same-year articles do not simply describe models; they describe the conditions under which models remain useful when data change, tasks accumulate, clinical populations vary, retrieval indexes evolve,

ISSN: 3071-5601 | © 2023 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

and engineering pipelines become part of the evidentiary chain. This interpretation makes the inaugural volume stronger than a set of disconnected technical papers. It shows that computational discovery depends on a chain of custody running from data collection and storage to model adaptation, retrieval, validation, and human interpretation.

A first implication concerns temporal reliability. Continual-learning systems are attractive because they promise adaptation without complete retraining, but the DATAMIND corpus makes clear that adaptation is not only a model property. It is also a data-history problem. A system that learns a new class, guideline, or institutional pattern must preserve enough information about previous tasks to make the update auditable. Review articles should therefore ask whether a learning system reports task boundaries, replay or regularization assumptions, historical validation sets, and failure cases after distribution change.

Medical imaging provides a second boundary condition. The clinical article in the 2023 corpus illustrates why predictive performance cannot be interpreted without dataset context. Imaging labels are tied to scanners, protocols, patient mix, clinical adjudication, and local decision thresholds. A high-performing model can therefore be fragile when moved across hospitals or demographic groups. DATAMIND can contribute by requiring medical AI reviews to distinguish internal accuracy, external validation, calibration, subgroup performance, and clinical usefulness. Those distinctions convert a model score into evidence that a reader can evaluate.

The retrieval article introduces a third lesson: evidence access is different from evidence quality. GraphRAG and VectorRAG both attempt to ground generation, but they do so through different assumptions about similarity, relation structure, and context assembly. A database-centered review should not ask only whether retrieval was used. It should ask how documents were chunked, how links or embeddings were constructed, how sources were ranked, how conflicting evidence was handled, and how the retrieved context was evaluated by humans or downstream tasks.

The data-engineering article is especially important because it moves attention away from the visible model and toward the invisible substrate of AI work. Pipelines, storage formats, feature transformations, orchestration rules, and monitoring choices affect the outputs that later appear to be model behavior. Treating these components as mere implementation details makes computational discovery less reproducible. The 2023 corpus therefore supports an editorial standard in which data infrastructure is reported as a methodological object, not only as a technical appendix.

Privacy-preserving and federated-learning literature adds a governance layer to the same argument. When data cannot be pooled, research teams must coordinate through protocols, aggregation rules, privacy budgets, and model-update procedures. These constraints are not peripheral; they influence statistical power, fairness, accountability, and the cost of validation. For DATAMIND, the lesson is that responsible discovery should report not only what information was learned, but also what information could not be accessed, shared, or audited under the chosen governance arrangement.

The coded data analysis developed in this review supports these qualitative claims. The thematic scores show that the 2023 corpus is strongest where evidence can be connected across multiple layers: model behavior, data provenance, and human judgment. The same scores also reveal weaker areas that future articles can address, especially operational monitoring and lifecycle documentation. The point is not that every paper must cover every dimension. Rather, a review should identify which dimensions are central to the claim and which remain unresolved limitations.

A useful way to summarize the inaugural volume is to distinguish storage, access, adaptation, and interpretation. Storage concerns whether the relevant data exist in a documented form. Access concerns whether a system can retrieve or compute over those data. Adaptation concerns whether the system can remain valid as tasks and distributions change. Interpretation concerns whether people can understand what the system did and whether the output is actionable. The 2023 articles collectively touch all four layers, giving the journal a coherent intellectual identity.

This identity also has implications for article review. A reviewer evaluating a DATAMIND submission should be able to trace each empirical claim back to data construction, data movement, model operation, and validation. If a paper reports a model result without describing the database or pipeline that produced it, the result is incomplete. If a

ISSN: 3071-5601 | © 2023 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

paper reports a database without explaining how it affects inference, the database is under-theorized. The journal's strength lies in connecting these two sides.

The same principle applies to figures and benchmarks. Benchmark results often look precise, but their meaning depends on task definition, sampling procedure, metric choice, and whether failure modes are visible. In the 2023 corpus, continual learning, medical imaging, and retrieval all depend on benchmarks that can either illuminate or hide the real problem. DATAMIND can encourage authors to publish benchmark cards that describe task scope, data source, update frequency, population coverage, and known weaknesses.

Another important theme is the role of human expertise. Clinicians, data engineers, annotators, security analysts, and domain experts are not outside the AI system. They define labels, design pipelines, approve outputs, and decide when a model should be trusted. The 2023 volume suggests that human expertise should be represented in research design rather than mentioned only in limitations. Future reviews should therefore measure where human judgment enters the pipeline and whether that judgment is sufficiently documented for replication.

The inaugural year also points toward hybrid system design. A practical application may combine continual-learning updates, privacy constraints, clinical or organizational evidence, and retrieval-based grounding. Such systems cannot be evaluated by a single metric. They require layered diagnostics that show whether failures arise from stale data, weak retrieval, poor calibration, missing clinical context, or insufficient governance. This layered view is more demanding than conventional model comparison, but it is better aligned with computational discovery.

For future submissions, the 2023 synthesis suggests a concrete reporting sequence. Authors should first define the data object and its lifecycle. They should then describe the computational method, the validation environment, the governance constraints, and the human decision process. Finally, they should discuss how the evidence would change if the database, task distribution, or retrieval architecture were updated. This sequence would make DATAMIND articles easier to compare across domains.

The review also demonstrates why a small annual corpus can still be analytically valuable. Because the same issue year contains papers on learning, medicine, engineering, and retrieval, it reveals cross-domain questions that a single-topic survey might miss. The data analysis in this article uses coarse coding rather than fine bibliometrics precisely because the goal is interpretation. It identifies the shared evidence problems that connect otherwise diverse areas of database-driven AI.

A final 2023 lesson concerns cumulative knowledge. Computational discovery becomes cumulative only when later researchers can reconstruct the data path behind earlier claims. That requires DOI-bearing references, clear data descriptions, reproducible code or workflows where possible, and honest reporting of uncertainty. The inaugural DATAMIND volume is well suited to promote this norm because its central theme is not one algorithmic family but the infrastructure through which algorithms become reliable scientific instruments.

Taken together, these implications define a practical agenda for the journal. DATAMIND can ask authors to treat databases, feature pipelines, retrieval systems, benchmarks, and human validation as first-order research objects. The 2023 volume shows why this agenda matters: without documented evidence infrastructure, AI systems may appear powerful while their claims remain difficult to verify, reproduce, or govern.

From a methodological perspective, the inaugural volume also recommends moderation in causal claims. When an AI system improves a benchmark, the improvement may arise from better model architecture, richer data, stronger retrieval, cleaner labels, or more favorable evaluation design. A database-centered review should separate these explanations rather than attributing the result to a single technical component. This separation makes the review more useful for both scientists and practitioners.

The 2023 articles further show that evidence quality is multidimensional. Reliability involves stability over time, medical validity involves clinical relevance, retrieval quality involves source adequacy, and engineering quality involves pipeline robustness. Combining these dimensions into one overall score would hide important differences. The coded matrix used in this article therefore treats each dimension separately and uses the pattern of scores to support interpretation.

ISSN: 3071-5601 | © 2023 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

A journal-level implication concerns replication packages. When possible, DATAMIND articles should provide code, data dictionaries, processing scripts, model cards, and descriptions of unavailable or restricted data. Such packages need not expose private records, but they should allow readers to understand the path from raw evidence to final inference. This is especially important for articles that combine data engineering with machine learning.

The 2023 volume also invites stronger dialogue between database research and human-centered AI. Data infrastructure can appear impersonal, but every database contains social decisions about inclusion, exclusion, measurement, and maintenance. Medical labels, retrieval corpora, privacy constraints, and engineering workflows all reflect human priorities. A robust review tradition should therefore analyze both computational structure and the human practices that make that structure meaningful.

Another useful direction is cross-database triangulation. A claim that holds only under one dataset, one retrieval index, or one benchmark may be too fragile for computational discovery. Future DATAMIND work can encourage authors to compare adjacent data sources, alternative indexing strategies, and multiple validation populations. Such triangulation increases cost, but it also reveals whether a result is a property of the phenomenon or a property of the database.

The inaugural volume should not be read as a finished agenda. It is better understood as a set of methodological prompts. Each paper asks readers to notice a different hidden dependency: memory of past tasks, clinical data context, engineering infrastructure, and retrieval grounding. Together those prompts invite a style of AI research in which the database is not background but the organizing center of inquiry.

This review also suggests that educational value is part of DATAMIND's contribution. Many early-career researchers learn AI through models first and data systems later. The 2023 corpus reverses that order. It shows students that performance claims are inseparable from data construction, validation, and governance. Review articles can reinforce this lesson by making infrastructure choices explicit and by presenting figures and tables that clarify the evidence chain.

Finally, the year establishes a foundation for longitudinal comparison. Later annual reviews can ask whether DATAMIND moves toward operational monitoring, governance, domain-specific benchmarking, or evidence-centered verification. The present article creates a baseline by documenting the inaugural themes and by translating them into a structured data-analysis framework.

6. Conclusion

This review of DATAMIND's 2023 articles shows that the journal's inaugural volume already contained a coherent intellectual program. The common thread is that AI reliability depends on structured evidence. Models may forget, medical predictions may fail outside the development site, data pipelines may hide technical debt, and retrieval systems may ground answers in weak sources. A database-centered perspective makes these failures visible because it asks how data are created, stored, retrieved, versioned, and reviewed. The research agenda that follows is practical: future studies should connect model performance to provenance, retrieval design, privacy controls, external validation, and human evidence stewardship. DATAMIND can make an important contribution by treating computational discovery as a system-level process in which databases, models, and institutions jointly determine what can be known.

Declaration of AI-assisted language editing

During the preparation of this manuscript, language-model assistance was used only for English polishing, structural organization, and formatting support. The authors reviewed, revised, and take full responsibility for the final content, analytical design, tables, figures, references, and interpretations.

ISSN: 3071-5601 | © 2023 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

References

Tanaka, Y., Ferretti, M., & Subramanian, P. (2023). When Transformers Forget: A study of catastrophic forgetting in continual learning for NLP tasks. DATAMIND, 1(1), 1-5. https://doi.org/10.63646/datamind.2023.010101

Mensah, A., Bruckner, T., & Nguyen, L. (2023). From Pixels to Predictions: A decade of deep learning in medical image analysis. DATAMIND, 1(2), 1-4. https://doi.org/10.63646/datamind.2023.010201

Madsen, K., & Al-Zahrawi, H. (2023). The Quiet Revolution: How data engineering became central to AI performance. DATAMIND, 1(3), 1-7. https://doi.org/10.63646/datamind.2023.010301

Moretti, L., Okwu, C., & Kobayashi, R. (2023). GraphRAG vs. VectorRAG: A systematic evaluation of retrieval-augmented generation architectures for enterprise knowledge. DATAMIND, 1(4), 1-7. https://doi.org/10.63646/datamind.2023.010401

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 308-318. https://doi.org/10.1145/2976749.2978318

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., et al. (2023). GPT-4 technical report. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software engineering for machine learning: A case study. Proceedings of the IEEE/ACM International Conference on Software Engineering: Software Engineering in Practice, 291-300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv. https://doi.org/10.48550/arXiv.2212.08073

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 610-623. https://doi.org/10.1145/3442188.3445922

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., et al. (2021). On the opportunities and risks of foundation models. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 1175-1191. https://doi.org/10.1145/3133956.3133982

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., et al. (2020). Language models are few-shot learners. arXiv. https://doi.org/10.48550/arXiv.2005.14165

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cybersecurity intrusion detection. IEEE Access, 4, 1153-1176. https://doi.org/10.1109/ACCESS.2015.2494502

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for health care: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721-1730. https://doi.org/10.1145/2783258.2788613

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., et al. (2021). Rethinking attention with Performers. arXiv. https://doi.org/10.48550/arXiv.2009.14794

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Re, C. (2022). FlashAttention: Fast and memory-efficient exact attention with IO-awareness. arXiv. https://doi.org/10.48550/arXiv.2205.14135

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. Journal of the Royal Statistical Society: Series C, 28(1), 20-28. https://doi.org/10.2307/2346806

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 248-255. https://doi.org/10.1109/CVPR.2009.5206848

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. arXiv. https://doi.org/10.48550/arXiv.2305.14314

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv. https://doi.org/10.48550/arXiv.2010.11929

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1), 1-26. https://doi.org/10.1214/aos/1176344552

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, 115-118. https://doi.org/10.1038/nature21056

Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., et al. (2020). CodeBERT: A pre-trained model for programming and natural languages. arXiv. https://doi.org/10.48550/arXiv.2002.08155

Fuller, A., Fan, Z., Day, C., & Barlow, C. (2020). Digital twin: Enabling technologies, challenges and open research. IEEE Access, 8, 108952-108971. https://doi.org/10.1109/ACCESS.2020.2998358

Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 1-37. https://doi.org/10.1145/2523813

Gao, L., Ma, X., Lin, J., & Callan, J. (2022). Precise zero-shot dense retrieval without relevance labels. arXiv. https://doi.org/10.48550/arXiv.2212.10496

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92. https://doi.org/10.1145/3458723

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. arXiv. https://doi.org/10.48550/arXiv.1406.2661

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. arXiv. https://doi.org/10.48550/arXiv.1412.6572

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv. https://doi.org/10.48550/arXiv.2312.00752

Gu, X., Zhang, H., Zhang, D., & Kim, S. (2018). Deep code search. Proceedings of the IEEE/ACM International Conference on Software Engineering, 933-944. https://doi.org/10.1145/3180155.3180167

Guo, D., Ren, S., Lu, S., Feng, Z., Tang, D., Liu, S., Zhou, L., et al. (2021). GraphCodeBERT: Pre-training code representations with data flow. arXiv. https://doi.org/10.48550/arXiv.2009.08366

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. W. (2020). REALM: Retrieval-augmented language model pre-training. arXiv. https://doi.org/10.48550/arXiv.2002.08909

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. arXiv. https://doi.org/10.48550/arXiv.1801.01290

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778. https://doi.org/10.1109/CVPR.2016.90

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. arXiv. https://doi.org/10.48550/arXiv.2002.05651

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. arXiv. https://doi.org/10.48550/arXiv.1903.12261

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. arXiv. https://doi.org/10.48550/arXiv.2106.09685

Husain, H., Wu, H. H., Gazit, T., Allamanis, M., & Brockschmidt, M. (2019). CodeSearchNet challenge: Evaluating the state of semantic code search. arXiv. https://doi.org/10.48550/arXiv.1909.09436

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38. https://doi.org/10.1145/3571730

Jordon, J., Yoon, J., & van der Schaar, M. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. arXiv. https://doi.org/10.48550/arXiv.1806.09655

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., et al. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1-210. https://doi.org/10.1561/2200000083

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. Proceedings of EMNLP, 6769-6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). Transformers are RNNs: Fast autoregressive transformers with linear attention. arXiv. https://doi.org/10.48550/arXiv.2006.16236

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521-3526. https://doi.org/10.1073/pnas.1611835114

Kitaev, N., Kaiser, L., & Levskaya, A. (2020). Reformer: The efficient transformer. arXiv. https://doi.org/10.48550/arXiv.2001.04451

Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital Twin in manufacturing: A categorical literature review and classification. IFAC-PapersOnLine, 51(11), 1016-1022. https://doi.org/10.1016/j.ifacol.2018.08.474

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90. https://doi.org/10.1145/3065386

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with PagedAttention. arXiv. https://doi.org/10.48550/arXiv.2309.06180

Lee, J., Bagheri, B., & Kao, H. A. (2015). A cyber-physical systems architecture for Industry 4.0-based manufacturing systems. Manufacturing Letters, 3, 18-23. https://doi.org/10.1016/j.mfglet.2014.12.001

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv. https://doi.org/10.48550/arXiv.2005.11401

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60. https://doi.org/10.1109/MSP.2020.2975749

Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. Proceedings of the ACL Workshop on Text Summarization Branches Out, 74-81. https://doi.org/10.3115/1218955.1219034

Lin, S., Hilton, J., & Evans, O. (2021). TruthfulQA: Measuring how models mimic human falsehoods. arXiv. https://doi.org/10.48550/arXiv.2109.07958

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghaemmaghami, M., van der Laak, J., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88. https://doi.org/10.1016/j.media.2017.07.005

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. arXiv. https://doi.org/10.48550/arXiv.2307.03172

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting, 36(1), 54-74. https://doi.org/10.1016/j.ijforecast.2019.04.014

Markowitz, H. (1952). Portfolio selection. The Journal of Finance, 7(1), 77-91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. Proceedings of ACL, 1906-1919. https://doi.org/10.18653/v1/2020.acl-main.173

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Aguera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. arXiv. https://doi.org/10.48550/arXiv.1602.05629

Merton, R. C. (1973). An intertemporal capital asset pricing model. Econometrica, 41(5), 867-887. https://doi.org/10.2307/1913811

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., et al. (2018). Mixed precision training. arXiv. https://doi.org/10.48550/arXiv.1710.03740

Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). Kitsune: An ensemble of autoencoders for online network intrusion detection. Proceedings of the Network and Distributed System Security Symposium. https://doi.org/10.14722/ndss.2018.23204

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 220-229. https://doi.org/10.1145/3287560.3287596

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodriguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. Pattern Recognition, 45(1), 521-530. https://doi.org/10.1016/j.patcog.2012.12.004

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., et al. (2021). WebGPT: Browser-assisted question-answering with human feedback. arXiv. https://doi.org/10.48550/arXiv.2112.09332

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., et al. (2022). Training language models to follow instructions with human feedback. arXiv. https://doi.org/10.48550/arXiv.2203.02155

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. arXiv. https://doi.org/10.48550/arXiv.1906.02530

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. arXiv. https://doi.org/10.48550/arXiv.1912.01703

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., So, D., et al. (2021). Carbon emissions and large neural network training. arXiv. https://doi.org/10.48550/arXiv.2104.10350

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). Scikit-learn: Machine learning in Python. arXiv. https://doi.org/10.48550/arXiv.1201.0490

Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2018). Data management challenges in production machine learning. Proceedings of the ACM SIGMOD International Conference on Management of Data, 1723-1726. https://doi.org/10.1145/3183713.3190657

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. arXiv. https://doi.org/10.48550/arXiv.2305.18290

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv. https://doi.org/10.48550/arXiv.1910.10683

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 33-44. https://doi.org/10.1145/3351095.3372873

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., et al. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv. https://doi.org/10.48550/arXiv.1711.05225

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28, 31-38. https://doi.org/10.1038/s41591-021-01614-0