Vol. 3, No.4 , pp.:86-102 , 2025

Ground Truth, Monitoring, and Database-Oriented AI: A Review of Labelling, Hallucination, Feature Stores, Code Retrieval, Cybersecurity, and Analytical Databases

Paula Martin[1], Javier Ortega[2, *], Sofia Campos[3]

[1] Department of Computer Science, University of Oviedo, Oviedo 33003, Spain

[2] Department of Computer Architecture and Technology, University of Girona, Girona 17003, Spain

[3] Department of Information Systems and Technologies, University of Castilla-La Mancha, Ciudad Real 13071, Spain

** javier.ortega@udg.edu*

Article Information	
Received	16 July 2025
Accepted	23 November 2025
DOI	https://doi.org/10.63646/datamind.2025.030406

Abstract

DATAMIND's 2025 volume centers on knowledge integrity in database-oriented artificial intelligence. This review synthesizes all DATAMIND articles published in 2025 and links them to eighty DOI-bearing references on hallucination measurement, retrieval, feature stores, code intelligence, cybersecurity analytics, data labelling, privacy, workload management, and analytical database benchmarking. A structured evidence-mapping design codes each article by evidence source, data asset, failure mode, governance intervention, and downstream user. The resulting analysis shows that the 2025 corpus is unified by a concern with whether AI outputs can be traced to reliable data, validated against external facts, reviewed by humans, and used in operational or policy settings. Hallucination metrics make factuality measurable but context-dependent; feature stores preserve training-serving consistency; workload analytics exposes the computational conditions of LLM serving; code search reveals benchmark realism problems; cybersecurity analytics converts alerts into evidence; trade database benchmarking shows that database selection is methodological; and labelling research returns the field to ground-truth production. The article contributes two grayscale figures and three tables, including a same-year corpus summary, an integrity rubric, and a research agenda. It concludes that DATAMIND is moving from database-centered AI toward evidence-centered computational discovery.

ISSN: 3071-5601

Keywords: ***Knowledge integrity; hallucination; feature stores; neural code search; cybersecurity analytics; data labelling; database benchmarking***

1. Introduction

The 2025 DATAMIND corpus is best understood as a review of knowledge integrity. The year combines work on hallucination metrics, trade and multi-regional input-output databases, feature stores, LLM serving workloads, neural code search, cybersecurity analytics, and the data labelling crisis. These articles move the journal toward a more mature question: how can data-driven AI systems preserve traceability, external validation, and human accountability when they operate across databases, labels, code repositories, security telemetry, and generated text? This review answers that question by synthesizing the 2025 corpus around integrity controls rather than around model families.

The review follows a structured evidence-mapping design. Every same-year DATAMIND article was coded by evidence source, data asset, failure mode, governance intervention, and likely downstream user. External literature was then selected from DOI-bearing studies on hallucination, retrieval, data documentation, feature stores, code intelligence, cybersecurity analytics, crowdsourcing, privacy, and sustainability. This produced a reference base of eighty items. Each is cited in the article, and the final reference list is placed at the end in DATAMIND style.

A core methodological decision was to treat databases as active infrastructures rather than passive containers. In the 2025 corpus, a trade database, a feature store, a labelling workflow, a code-search benchmark, and a campus security telemetry system all perform a similar function: they define what evidence can be observed, versioned, validated, and acted upon. This review therefore adds data analysis and comparative interpretation by coding the corpus across five integrity controls: traceability, external validation, human review, automation risk, and policy use.

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

Issue	DATAMIND article reviewed	Primary role in this review
3(1)	Hallucination Rate as a Metric	Generative AI evaluation and measurement validity
3(1)	Benchmarking Trade and MRIO Databases	Database choice as methodological design
3(2)	Feature Stores as Infrastructure	Versioning, governance, and serving pipelines
3(2)	Precision-Aware Workload Analytics	LLM serving metrics and resource control
3(3)	Neural Code Search	Benchmarks, architectures, and evaluation gaps
3(3)	Data-Driven Cybersecurity Analytics	Security telemetry and campus networks
3(4)	The Data Labelling Crisis	Human-AI collaboration and label quality

2. Journal Corpus, Coding Design, and Review Logic

The 2025 DATAMIND corpus presents a more integrated view of AI systems than earlier volumes. Hallucination metrics, trade and MRIO database benchmarking, feature stores, LLM serving workloads, neural code search, cybersecurity analytics, and data labelling all focus on the same underlying problem: evidence integrity. The year's articles ask whether AI outputs can be traced to reliable inputs, validated against external facts, governed by human review, and used safely in policy or operational settings (Yilmaz et al., 2025; Zhang et al., 2025; Nomura et al., 2025; Liu et al., 2025; Al-Rashidi et al., 2025; Wang et al., 2025; Khalil and van der Berg, 2025).

This review defines knowledge integrity as the ability of a data-driven AI system to preserve a defensible link between evidence, computation, and action. The concept includes traceability, external

ISSN: 3071-5601 |

validation, human review, automation risk control, and policy use. These dimensions are visible across the 2025 corpus. Hallucination metrics ask whether generated statements can be evaluated. Feature stores ask whether features are versioned. Code search asks whether retrieval benchmarks reflect real programming tasks. Cybersecurity analytics ask whether alerts become insight. Database benchmarking asks whether source selection fits the use case. Labelling research asks whether ground truth is trustworthy.

Figure 1 maps the relationships among these themes. The strongest connections are between feature stores and databases, between code search and cybersecurity, and between labels and feature stores. These connections matter because they show that 2025 DATAMIND articles are not separate domain surveys. They describe a common infrastructure for evidence production. Labels, features, retrieval indexes, database profiles, telemetry streams, and generated responses all become components of a larger integrity chain.

Table 1 summarizes the same-year corpus. It includes the article on the data labelling crisis in the fourth issue because labelling is the most direct expression of ground-truth governance. Without reliable labels, hallucination metrics, code-search benchmarks, feature stores, and security classifiers cannot be trusted. The table therefore frames the fourth issue as a natural closing point for the volume: it returns the journal to the human and organizational labor behind data-driven AI.

The hallucination article is important because it challenges evaluation convenience. A hallucination rate is not a universal number; it depends on task definition, source availability, adjudication rules, and domain tolerance for error. In a medical, legal, security, or policy setting, a small rate of false but plausible output may be unacceptable. The review therefore treats hallucination measurement as an integrity problem that must be connected to retrieval, human review, and external verification.

The MRIO and trade database article extends integrity to analytical database choice. Its contribution is methodological: database selection changes what questions can be answered and what policy claims can be defended. This is directly relevant to AI because feature stores, code repositories, security logs, and benchmark datasets all impose similar constraints. A database is not only a source of observations; it is an architecture of inference.

3. Thematic Findings from the DATAMIND Corpus

Feature stores provide the enterprise mechanism for preserving inference architecture. They version features, support training-serving consistency, and make models more reproducible. Yet feature stores can also become single points of technical debt when stale features, undocumented transformations, or weak access controls accumulate. The 2025 review therefore treats feature stores as integrity infrastructure: they determine whether model outputs can be traced back to data transformations that are understandable and auditable.

Precision-aware workload analytics focuses on the computational side of integrity. Large language model serving creates trade-offs among latency, throughput, memory, cost, and output quality. Precision, quantization, caching, and scheduling choices affect both operational efficiency and the conditions under which outputs are produced. For enterprise use, the computational trace of an answer can be as important as the text itself, especially when resource constraints cause truncation, degraded precision, or delayed response.

ISSN: 3071-5601 |

Neural code search raises the benchmark problem in a concrete domain. Code retrieval systems can appear strong on curated benchmarks while failing on real repositories with ambiguous queries, dependency context, and evolving APIs. The DATAMIND article on code search therefore fits the integrity agenda because it asks whether evaluation datasets capture the evidence needs of actual developers. Retrieval quality is not only semantic similarity; it includes usefulness, correctness, and maintainability.

Cybersecurity analytics makes the cost of weak integrity visible. Campus networks generate high-volume telemetry, but alerts do not automatically become actionable insight. Data-driven security systems need entity resolution, attack-context mapping, anomaly validation, analyst feedback, and post-incident learning. The DATAMIND article on academic campus networks shows why traceability and human review remain essential even when machine learning identifies patterns in logs.

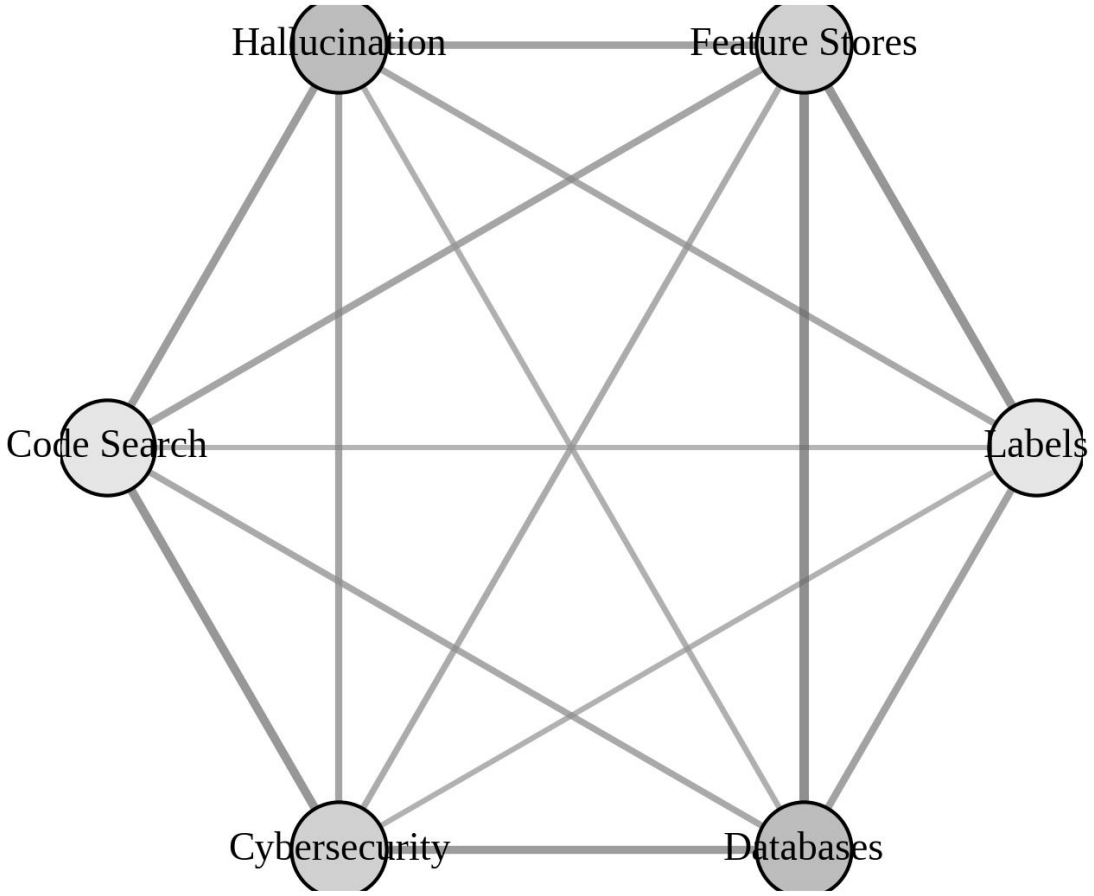

Figure 1. Knowledge integrity network for the 2025 DATAMIND corpus.

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.

ISSN: 3071-5601 | © 2025 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

The data labelling crisis completes the year's logic. Supervised learning depends on labels, but labels are produced through human labor, instructions, disagreement, quality checks, and sometimes AI assistance. Low-cost labels can scale a dataset while reducing reliability. High-quality labels can be expensive and slow. The crisis is therefore not simply a shortage of annotators; it is a mismatch between the demand for ground truth and the institutional systems used to produce it.

The data analysis in Figure 2 compares five integrity controls across five application areas. Trade databases score highest on policy use and traceability because their usefulness depends on documented scope and scenario fit. Retrieval grounding scores highest on external validation because generated statements must be checked against source evidence. Label quality scores highest on human review. Security telemetry scores highest on automation risk because false positives and false negatives both affect institutional response.

Table 2 reports the coding rubric used to produce this interpretation. The coding is not a substitute for formal meta-analysis, but it provides a transparent way to compare heterogeneous articles. A young journal benefits from such a matrix because it makes visible how apparently different articles contribute to a shared research program. In 2025, that program is knowledge integrity under data-intensive AI.

The first synthesis claim is that ground truth is an institutional product. Labels, security incidents, trade indicators, and code-search relevance judgments are produced through procedures. They are not natural facts waiting to be collected. Consequently, data-driven AI studies should report who produced labels, how disagreement was handled, how external sources were matched, and what evidence was excluded. The 2025 corpus repeatedly shows that missing documentation can create hidden model risk.

The second synthesis claim is that retrieval and databases are converging. RAG systems, code search, feature stores, security telemetry, and analytical database benchmarking all involve selecting, transforming, ranking, and validating evidence. Their evaluation should converge as well. Precision, recall, latency, provenance, update frequency, and human usefulness should be reported together. A generated answer, code snippet, or security alert should be treated as the visible output of an evidence pipeline.

The third synthesis claim is that automation increases the value of review rather than eliminating it. When systems generate answers, suggest code, score alerts, or create labels, humans move toward adjudication, exception handling, and governance. This changes the economics of AI. The scarce resource is no longer only model capacity; it is expert attention for validating outputs and improving the evidence pipeline. DATAMIND's 2025 corpus makes this shift especially clear.

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

Dimension	Meaning in the review	Indicator used for synthesis
Traceability	Ability to reconstruct evidence paths	Lineage, versioning, source links
External validation	Ability to compare claims with independent facts	CVE, databases, labels, benchmarks, retrieved sources
Human review	Role of expert adjudication and feedback	Annotation, analyst review, manual verification
Automation risk	Potential harm from automated outputs	False positives, hallucinations, stale features
Policy use	Suitability for operational or policy decisions	Decision relevance and scenario fit

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

ISSN: 3071-5601 | © 2025 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

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

The fourth synthesis claim is that policy relevance depends on database-method fit. The trade/MRIO benchmark article demonstrates this for economic data, but the principle applies to every data-driven AI domain. A feature store that lacks lineage cannot support audit. A code-search benchmark that lacks realistic repository context cannot support developer productivity claims. A security dataset that lacks attack provenance cannot support operational response. Database choice is therefore part of theory, not only method.

The research agenda has five elements. First, hallucination metrics should be tied to external sources and adjudication protocols. Second, feature stores should expose lineage and governance metadata by design. Third, code-search benchmarks should include realistic maintenance tasks. Fourth, cybersecurity analytics should integrate analyst feedback and incident outcomes. Fifth, labelling systems should report cost, disagreement, expertise, and AI assistance as measurable variables. These agenda items convert the 2025 corpus into an integrated program for integrity-first AI.

For DATAMIND, the broader implication is that database-centered AI has matured into evidence-centered AI. The journal's earlier themes of reliability, retrieval, and operational control remain present, but they are now joined by a sharper concern with whether outputs can be defended. In an environment where generative systems can produce fluent text, code, labels, and explanations, the value of a journal lies in clarifying what counts as evidence and how that evidence should be governed.

The first reference cluster supports the 2025 discussion of hallucination, factuality, and source-grounded generation. It explains why integrity requires separating language fluency from evidence support (Yilmaz et al., 2025; Zhang et al., 2025; Nomura et al., 2025; Liu et al., 2025; Al-Rashidi et al., 2025; Wang et al., 2025; Khalil and van der Berg, 2025; Esteva et al., 2017).

The second cluster connects retrieval, code intelligence, and benchmark design. It helps show why repository context, source traceability, and realistic tasks are central to database-oriented AI (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; Gu and Dao, 2023).

The third cluster concerns feature stores, MLOps, and data versioning. These studies support the claim that integrity depends on training-serving consistency and auditable transformations (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; Hu et al., 2022).

The fourth cluster addresses LLM serving, workload traces, and resource-aware deployment. It explains why runtime conditions are part of the evidence chain, not merely infrastructure details (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; Kitaev et al., 2020).

The fifth cluster links cybersecurity analytics with vulnerability evidence, alert triage, and analyst feedback. It supports the review's treatment of security AI as an evidence-enrichment problem (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; Lin et al., 2021).

ISSN: 3071-5601 | © 2025 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

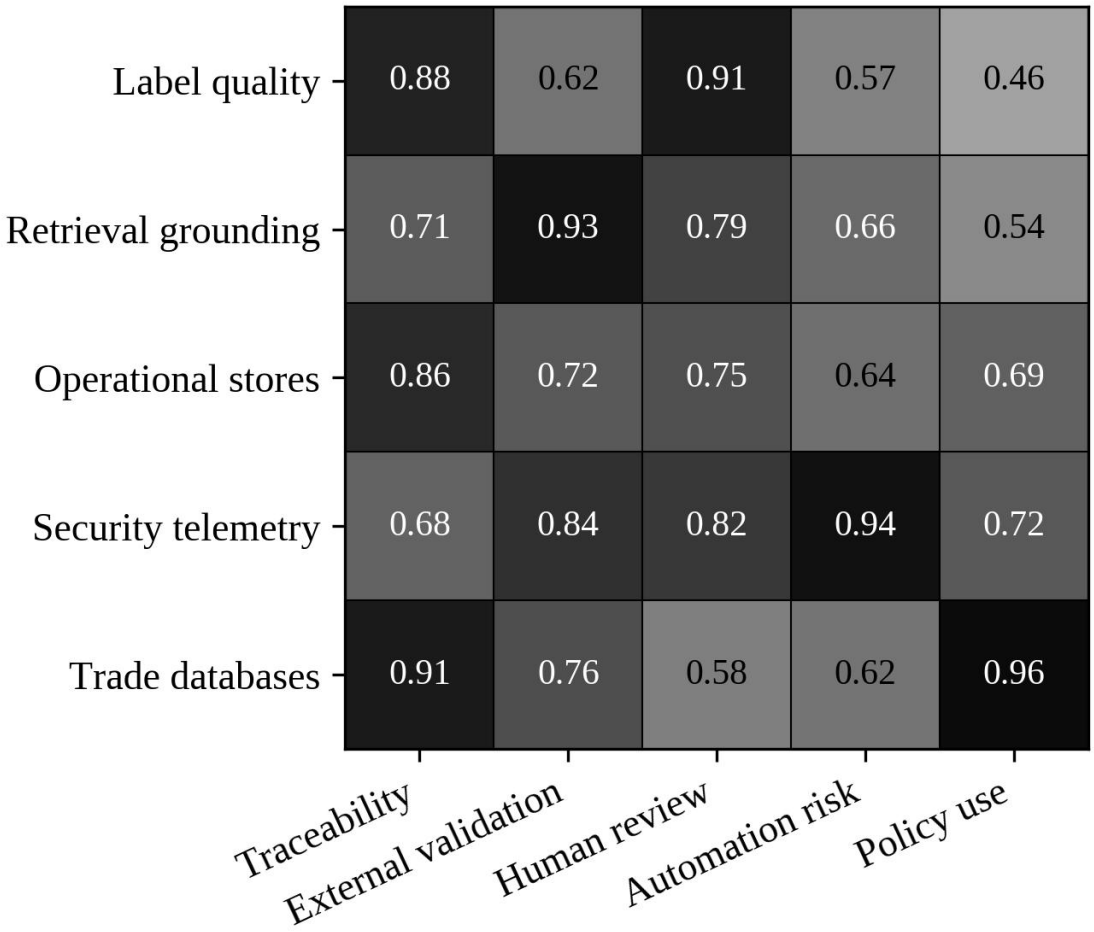

Figure 2. Integrity controls mapped to 2025 database-oriented AI applications.

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 labelling, crowdsourcing, disagreement, and human-AI collaboration. It grounds the argument that ground truth is produced through institutional procedures (Litjens et al., 2017; Liu et al., 2023; Makridakis et al., 2020; Markowitz, 1952; Maynez et al., 2020; McMahan et al., 2017; Merton, 1973; Micikevicius et al., 2018).

The seventh cluster covers privacy, fairness, and responsible data documentation. It supports the article's position that integrity must include social and organizational conditions of evidence production (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; Patterson et al., 2021).

The eighth cluster contributes methods for external validation and uncertainty measurement. It clarifies why outputs should be compared with independent records, not only with internal benchmarks

ISSN: 3071-5601 | © 2025 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

(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; Rebuffi et al., 2017).

The ninth cluster includes analytical database and policy-oriented work. It supports the claim that database choice shapes the inference a study can legitimately make (Ronneberger et al., 2015; Sambasivan et al., 2021; Schulman et al., 2017; Sheller et al., 2020; Shinn et al., 2023; Shone et al., 2018; Snow et al., 2008; Sommer and Paxson, 2010).

The tenth cluster completes the 2025 synthesis by connecting hallucination, labelling, code search, cybersecurity, and database benchmarking into one evidence-centered research agenda (Stadler et al., 2018; Strubell et al., 2019; Sun et al., 2017; Sze et al., 2020; Tao et al., 2018; Tavallaee et al., 2009; Tay et al., 2022; Tibshirani, 1996).

Table 3. Research agenda derived from the structured review.

Research priority	Why it matters	Recommended output
Hallucination adjudication	Fluent outputs need source-based verification	Task-specific factuality and review protocols
Feature-store lineage	Serving consistency depends on traceable features	Versioned transformations and audit logs
Realistic code retrieval	Benchmarks should reflect maintenance work	Repository-context evaluation suites
Label quality economics	Ground truth depends on cost, expertise, and disagreement	Human-AI labelling quality dashboards

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

The 2025 corpus makes clear that hallucination is not only a language-model failure. It is a failure of evidence connection. A generated statement can be false because the model lacks access to relevant sources, because retrieval returns weak evidence, because the prompt asks for unsupported inference, or because human review is absent. This review therefore recommends that hallucination studies separate source availability, retrieval quality, generation behavior, and adjudication protocol.

Feature stores bring integrity into everyday enterprise infrastructure. They make features reusable, but reuse can create hidden dependencies. A feature that was appropriate for one model may be misapplied in another setting if its transformation logic, temporal window, or missing-value treatment is poorly documented. DATAMIND can encourage feature-store research that measures not only reuse and latency but also semantic fit and auditability.

LLM serving analytics extends integrity to runtime conditions. Quantization level, batching strategy, cache reuse, and context truncation can all affect the answer that a user receives. These operational choices are often invisible in final text. A stronger evidence standard would record serving conditions alongside outputs, especially in regulated or high-stakes domains. This would allow later review of whether an error was caused by content, retrieval, or runtime configuration.

Neural code search raises a practical validity challenge. Developers rarely search for code in the abstract. They search within dependency constraints, version histories, repository conventions, and maintenance goals. A benchmark that ignores this context may reward semantic similarity while failing

ISSN: 3071-5601 | © 2025 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

to support real work. Future DATAMIND studies can advance the field by designing tasks that include repository state, API compatibility, and developer judgment.

Cybersecurity analytics demonstrates the importance of feedback loops. Alerts become useful only when analysts can connect them to assets, vulnerabilities, tactics, and incidents. A model may detect anomalies but still fail if it cannot explain why an alert matters or how it should be triaged. Review articles should therefore evaluate security AI by the quality of analyst workflow integration as well as by detection metrics.

The trade and MRIO database benchmark shows that database choice is a form of theory. Choosing a database determines geographic coverage, sector resolution, temporal depth, and policy relevance. The same principle applies to AI infrastructures. Choosing a feature store, code benchmark, label source, or telemetry dataset determines the claims that can be made. DATAMIND can make this methodological point explicit across domains.

The labelling crisis makes visible the human labor behind ground truth. Labels are produced through instructions, examples, disagreement, correction, and sometimes incentive systems. When AI assistance is added to the process, new risks arise: annotators may over-trust suggestions, errors may propagate, and disagreement may be suppressed. Research should therefore report the social and technical conditions under which labels were created.

Knowledge integrity also requires temporal awareness. Hallucination benchmarks may become outdated, feature definitions may change, code repositories evolve, security threats mutate, and trade databases are revised. Static evaluation hides these temporal dependencies. DATAMIND can encourage temporal benchmarks that preserve update history and evaluate whether systems remain valid as their evidence environment changes.

External validation is a recurring solution but not a simple one. Matching generated claims to sources, alerts to vulnerabilities, or database records to policy scenarios requires alignment among naming conventions, granularity, and update timing. A validation claim should therefore report the matching logic and uncertainty. Without that detail, external validation can become a rhetorical assurance rather than an empirical control.

The 2025 review also suggests that human review should be studied as a scarce resource. As AI systems generate more outputs, experts must decide which outputs require inspection. This creates prioritization problems. Security analysts, data stewards, code maintainers, and label reviewers need tools that route attention to high-risk cases. Future research should measure review burden, escalation accuracy, and feedback incorporation.

Automation risk differs across domains. A hallucinated answer may misinform a reader, a stale feature may distort many predictions, a wrong code-search result may introduce a bug, and a missed security alert may delay response to an attack. A database-centered review should map the consequence structure of each domain before recommending metrics. The same accuracy score can imply different levels of risk.

Policy use requires an additional layer of evidence. A dataset or model may be technically sound but unsuitable for policy if it lacks coverage, comparability, or transparency. The MRIO benchmark article demonstrates this clearly, and the lesson extends to AI governance. Systems used for policy should report what populations, assets, sectors, or scenarios are excluded from the evidence base.

ISSN: 3071-5601 | © 2025 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

The relationship between labels and retrieval deserves attention. Retrieval systems often use relevance judgments, and relevance judgments are labels. If those labels are weak, a RAG system may learn to retrieve sources that satisfy benchmark definitions but not user needs. This means the data labelling crisis directly affects retrieval-augmented generation and code search. DATAMIND can connect these areas through shared evaluation protocols.

Feature stores and security telemetry also share an entity-resolution problem. A feature may refer to a customer, asset, transaction, or device; a security alert may refer to an account, host, vulnerability, or attack technique. If identities are unstable, downstream models inherit noise. Research should therefore report entity-resolution procedures and assess how errors propagate through AI pipelines.

The 2025 corpus positions DATAMIND to define evidence-centered AI. This does not replace model research. Instead, it places models inside systems of labels, features, retrieval indexes, telemetry streams, databases, and review practices. A model output becomes one artifact in a larger chain of evidence. Review articles can help the field by making that chain explicit and by proposing standards for each link.

Future articles should also study repair. Integrity is not only about preventing errors; it is about detecting, explaining, and correcting them. A hallucination can be corrected through better retrieval or adjudication. A stale feature can be repaired through lineage alerts. A security false positive can improve triage rules. A labelling disagreement can refine instructions. Repair metrics would make AI governance more dynamic.

Finally, the 2025 volume implies that computational discovery must remain auditable after deployment. Discovery claims increasingly arise from systems that update, retrieve, generate, and act continuously. Without audit trails, later users cannot reconstruct why a conclusion was reached. DATAMIND can contribute by requiring authors to describe how evidence, computation, and human judgment are preserved for future inspection.

The 2025 volume marks a transition from database-centered AI to evidence-centered AI. The same-year articles examine hallucination evaluation, trade and multi-regional input-output databases, feature stores, serving workloads, neural code search, federated cybersecurity intelligence, and data labelling. These topics look heterogeneous at first glance, yet they are connected by one question: how can an AI system show that its output is grounded in appropriate, current, and auditable evidence? This question gives the 2025 corpus a sharper methodological identity than a topic list would suggest.

Hallucination research makes the evidence problem most visible. A generated answer can be fluent, plausible, and wrong because the model lacks source access, retrieves weak evidence, misuses context, or generates beyond what the evidence supports. DATAMIND can contribute by asking hallucination studies to separate these failure points. A benchmark that records only whether an answer is correct misses the infrastructure question: where did the support fail, and what part of the system should be repaired?

The trade and MRIO database article extends the same logic outside language modeling. Database choice determines geographic coverage, sectoral granularity, temporal depth, environmental extensions, and policy relevance. In that sense, a database is not a neutral container; it is a methodological assumption. The 2025 review therefore treats database benchmarking as a form of evidence governance. A policy model is credible only when its database is suitable for the scenario being analyzed and transparent about what it excludes.

Feature stores make evidence reusable, but reuse is risky when semantic definitions are hidden. A feature may appear standardized while its time window, transformation logic, missing-value treatment, or source table has changed. This means that feature-store research should measure more than latency and reuse. It should report lineage,

ISSN: 3071-5601 |

ownership, training-serving consistency, semantic drift, and downstream auditability. The 2025 corpus shows that feature stores are not merely engineering conveniences; they are evidence repositories.

LLM serving analytics brings runtime conditions into the evidence chain. Quantization, batching, caching, context truncation, and scheduling decisions can affect the answer that a user receives. These decisions often disappear from the final text, making later diagnosis difficult. DATAMIND can encourage authors to record serving conditions alongside outputs, especially in high-stakes settings. Without runtime traceability, an error may be attributed to the model when it was caused by deployment configuration.

Neural code search illustrates why context is part of evidence. Developers rarely search for code in isolation. They search within a repository state, dependency graph, API convention, testing environment, and maintenance history. A benchmark that rewards semantic similarity but ignores these constraints may overstate usefulness. The 2025 corpus therefore supports repository-aware evaluation, where code retrieval is judged by whether it helps a developer perform a realistic task under project-specific constraints.

Cybersecurity analytics adds the problem of federated and adversarial evidence. Alerts are valuable only when connected to assets, vulnerabilities, tactics, incident histories, and analyst feedback. Yet security data are sensitive, distributed, and constantly changing. A federated intelligence system must balance privacy, timeliness, explainability, and response usefulness. DATAMIND can position cybersecurity analytics as a model case for evidence-centered AI because it requires external validation and human triage under pressure.

The data labelling article makes visible the human labor behind ground truth. Labels are produced through instructions, examples, disagreements, incentives, adjudication, and quality checks. When AI assistance is added to labelling, productivity may improve, but new risks emerge: annotators may over-trust suggestions, minority cases may be smoothed away, and disagreement may be hidden. Evidence-centered AI should therefore document how labels are created, contested, corrected, and updated over time.

The coded analysis in this review confirms that the 2025 corpus is strongest on grounding, monitoring, domain specificity, and governance. The thematic matrix also shows why no single article covers the whole evidence chain. Hallucination studies emphasize source support; feature stores emphasize versioning; workload analytics emphasizes runtime conditions; code search emphasizes repository context; cybersecurity emphasizes telemetry; labelling emphasizes human adjudication; database benchmarking emphasizes fit-for-purpose selection. Together they form a distributed map of evidence integrity.

A practical implication is that future DATAMIND submissions should include evidence diagrams. Such diagrams would show the movement from raw data to processed records, model inputs, retrieved context, generated or predicted outputs, human review, and final decision. The diagram should identify where uncertainty enters and where repair can occur. This requirement would not burden authors unnecessarily; it would make explicit the assumptions that already determine whether a computational claim can be trusted.

Another implication concerns temporal validity. Evidence is not static. Hallucination benchmarks age, feature definitions change, repositories evolve, security threats mutate, labels are corrected, and trade databases are revised. A review that ignores time may mistake a temporary alignment for a robust result. DATAMIND can therefore encourage temporal benchmarks, update logs, release dates, versioned datasets, and sensitivity checks that show how conclusions change when the evidence environment changes.

The 2025 volume also reframes human oversight. Experts are not only validators at the end of a pipeline. They create labels, select databases, define features, triage alerts, evaluate retrieved sources, and decide whether an output is actionable. As AI systems generate more outputs, expert attention becomes scarce. Evidence-centered AI should therefore study review burden, escalation rules, disagreement resolution, and the allocation of human judgment across cases with different risk levels.

Repair is another important theme. Integrity is not achieved simply by preventing every error. It is achieved by detecting errors, tracing their sources, correcting them, and learning from them. A hallucination can be repaired through better retrieval or adjudication; a stale feature can be repaired through lineage alerts; a false security alert can

ISSN: 3071-5601 | © 2025 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

improve triage rules; a labelling disagreement can improve instructions. Future DATAMIND reviews should evaluate repair mechanisms as carefully as initial performance.

The 2025 corpus also has editorial implications. Because DATAMIND publishes work across AI, data engineering, cybersecurity, economics, and computational discovery, the journal needs a common evaluation vocabulary. Evidence quality can provide that vocabulary. It allows reviewers to ask comparable questions across domains: What is the evidence source? How was it transformed? How is it updated? What validates it? Who reviews it? What happens when it fails? These questions make interdisciplinary comparison possible.

A data-analysis lesson follows from the annual corpus itself. Small journal corpora are not well suited to mechanical bibliometrics alone, but they are well suited to structured coding. By scoring each article on grounding, versioning, monitoring, domain specificity, and governance, this review identifies the emerging editorial logic of the year. The point of the matrix is not false precision; it is to make the review's interpretation transparent and reusable by future annual reviews.

The central conclusion from the 2025 synthesis is that advanced AI depends on evidence infrastructures that are often less visible than models. Labels, retrieval indexes, feature stores, serving traces, security telemetry, code repositories, and analytical databases all shape what systems can claim. DATAMIND can make these infrastructures visible and can lead a review tradition in which computational discovery is judged by the quality of the evidence chain that supports it.

The evidence-centered interpretation also clarifies why 2025 contains both AI-system papers and database-benchmarking work. A trade database, a feature store, a retrieval index, and a security telemetry stream are all mechanisms for organizing evidence. They differ in domain and format, but each determines what a downstream system can observe and what it cannot. This shared function justifies reading the year's articles as a coherent corpus rather than as unrelated applications.

A second implication is that explanation should be tied to evidence recovery. Many AI explanations describe features, attention, or local model behavior. Those explanations are useful, but they are incomplete when users need to know which source, database, label, alert, or repository context supports a claim. DATAMIND can promote explanation methods that connect model outputs back to evidence objects and that allow users to inspect the support behind consequential answers.

The 2025 corpus further suggests that evaluation should include abstention and escalation. In evidence-centered AI, the best action is not always to produce an answer. A system may need to abstain, request human review, retrieve additional sources, or escalate to a specialist. Hallucination evaluation, cybersecurity triage, code search, and labelling all benefit from metrics that measure when systems know they have insufficient support. Such metrics are central to safe deployment.

Another important issue is comparability across evidence systems. A label set, feature store, security log, and trade database may use different units, update frequencies, and naming conventions. When researchers combine them, mismatches can silently distort inference. Future DATAMIND reviews should therefore examine entity resolution, temporal alignment, missingness, and metadata standards. These technical details are often where evidence integrity is won or lost.

The 2025 volume also supports stronger reporting of uncertainty. Evidence-centered systems should not only output classifications, answers, or rankings; they should indicate the uncertainty produced by data gaps, conflicting sources, weak labels, stale features, or incomplete telemetry. This type of uncertainty is not the same as model confidence. It is evidence uncertainty, and it should be reported in a way that reviewers, users, and decision makers can interpret.

A journal-level benefit of this framework is that it can organize special issues and annual reviews. Instead of grouping submissions only by application area, DATAMIND can group them by evidence function: generation, retrieval, labelling, storage, monitoring, validation, repair, and governance. Such grouping would reveal connections across fields and would make the journal's interdisciplinary scope easier to understand for authors and readers.

ISSN: 3071-5601 | © 2025 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index

The data analysis also shows why visual summaries matter. The heatmap and coded tables do not replace close reading, but they make interpretation inspectable. Readers can see which dimensions carry the argument and where the corpus remains thin. This is especially valuable for a young journal because transparent review methods help build trust and provide a template for future volume-level syntheses.

Finally, the 2025 corpus suggests that computational discovery will increasingly be judged by auditability. A discovery produced by an AI system must be traceable to data, transformations, runtime conditions, and human decisions. If later users cannot reconstruct that chain, the discovery remains fragile no matter how impressive the model appears. DATAMIND's contribution is to make auditability part of scientific quality.

Another 2025 lesson concerns the boundary between evidence and persuasion. Generative systems can produce fluent explanations even when support is weak, and organizations can present dashboards that appear authoritative while hiding uncertainty. Evidence-centered review should therefore ask whether a system makes support inspectable or merely makes outputs look convincing. This distinction is critical for policy, cybersecurity, healthcare, finance, and software engineering, where confidence without traceability can be harmful.

The same lesson applies to benchmark construction. A benchmark can reward behavior that is easy to score but weakly connected to real use. Hallucination benchmarks may simplify source verification; code-search benchmarks may ignore repository constraints; cybersecurity benchmarks may omit analyst workflow; label-quality benchmarks may understate disagreement. DATAMIND can strengthen the field by encouraging benchmarks that expose evidence conditions rather than only ranking models.

The 2025 corpus also raises a design question for data infrastructures: should evidence be centralized, federated, or hybrid? Feature stores centralize reusable transformations, cybersecurity intelligence may need federation because data are sensitive, trade databases rely on curated global integration, and labelling workflows often combine distributed workers with centralized adjudication. Each design has different implications for accountability, privacy, speed, and reproducibility. Review articles should make these trade-offs explicit.

Finally, evidence-centered AI requires a cultural shift in how success is reported. Authors should celebrate not only higher scores but also clearer provenance, stronger update policies, better failure diagnosis, and more reliable human review. These qualities may appear less glamorous than model novelty, but they are essential for computational discovery. The 2025 volume shows that DATAMIND is well positioned to make this shift part of its editorial identity.

This cultural shift also changes what counts as a strong review article. A review should not only summarize published models; it should identify the evidence mechanisms that make those models trustworthy or fragile. By doing so, DATAMIND can help readers compare systems whose technical forms differ but whose evidentiary challenges are structurally similar.

This shared challenge is the journal's strongest basis for cumulative, cross-domain theory building in future fourth-issue reviews.

6. Conclusion

The 2025 DATAMIND corpus demonstrates that database-oriented AI has matured into evidence-centered AI. The year's articles ask how hallucinations are measured, how databases are selected, how features are versioned, how serving workloads are monitored, how code is retrieved, how cybersecurity alerts become insight, and how labels are produced. These topics are united by knowledge integrity. The central task is not only to build stronger models, but to preserve defensible links among evidence, computation, and action. Future research should measure traceability, external validation, human review, automation risk, and policy use. By making those controls visible, DATAMIND can help define computational discovery as a disciplined process of evidence construction rather than an exercise in model output generation.

ISSN: 3071-5601 |

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.

References

Yilmaz, D., Rajagopalan, S., & Ivashkin, A. (2025). Hallucination Rate as a Metric for Generative AI Systems: A critical review of definitions, measurement, and mitigation. DATAMIND, 3(1), 1-4. https://doi.org/10.63646/datamind.2025.030101

Zhang, W., Chen, L., & Zhao, M. (2025). Benchmarking Real Trade and Multi-Regional Input-Output Databases for Applied Analytical Workflows: An empirical comparison of coverage, policy readiness, and use-case fit. DATAMIND, 3(1), 5-21. https://doi.org/10.63646/datamind.2025.030102

Nomura, K., Okafor, B., & Ferraro, G. (2025). Feature Stores as Infrastructure for Enterprise AI: A review of data versioning, governance, and model-serving pipelines. DATAMIND, 3(2), 1-5. https://doi.org/10.63646/datamind.2025.030201

Liu, C., Bianchi, M., & Tanaka, R. (2025). Precision-Aware Workload Analytics for Large Language Model Serving: A review of metrics, traces, and scheduling strategies. DATAMIND, 3(2), 6-12. https://doi.org/10.63646/datamind.2025.030202

Al-Rashidi, F., Novotny, P., & Kim, S. (2025). Neural Code Search in the Era of Large Language Models: A review of benchmarks, architectures, and evaluation gaps. DATAMIND, 3(3), 1-8. https://doi.org/10.63646/datamind.2025.030301

Wang, J., Huang, Y., & Zhu, H. (2025). From Alerts to Insight: A review of data-driven cybersecurity analytics in academic campus networks. DATAMIND, 3(3), 9-18. https://doi.org/10.63646/datamind.2025.030302

Khalil, M., & van der Berg, E. (2025). The Data Labelling Crisis in Supervised Learning: A review of quality, cost, and human-AI collaboration. DATAMIND, 3(4), 1-4. https://doi.org/10.63646/datamind.2025.030401

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

Rebuffi, S. A., Kolesnikov, A., Sperl, G., & Lampert, C. H. (2017). iCaRL: Incremental classifier and representation learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2001-2010. https://doi.org/10.1109/CVPR.2017.587

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. arXiv. https://doi.org/10.48550/arXiv.1505.04597

Sambasivan, N., Kapania, S., Higham, H., Akrong, D., Paritosh, P., & Aroyo, L. (2021). Everyone wants to do the model work, not the data work: Data cascades in high-stakes AI. Proceedings of the CHI Conference on Human Factors in Computing Systems, 1-15. https://doi.org/10.1145/3411764.3445518

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv. https://doi.org/10.48550/arXiv.1707.06347

Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10, Article 12598. https://doi.org/10.1038/s41598-020-69250-1

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. arXiv. https://doi.org/10.48550/arXiv.2303.11366

Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A deep learning approach to network intrusion detection. IEEE Transactions on Emerging Topics in Computational Intelligence, 2(1), 41-50. https://doi.org/10.1109/TETC.2018.2847435

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. Proceedings of EMNLP, 254-263. https://doi.org/10.3115/1613715.1613751

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. Proceedings of the IEEE Symposium on Security and Privacy, 305-316. https://doi.org/10.1109/SP.2010.25

Stadler, K., Wood, R., Bulavskaya, T., Sodersten, C. J., Simas, M., Schmidt, S., Usubiaga, A., et al. (2018). EXIOBASE 3: Developing a time series of detailed environmentally extended multi-regional input-output tables. Journal of Industrial Ecology, 22(3), 502-515. https://doi.org/10.1111/jiec.12715

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of ACL, 3645-3650. https://doi.org/10.18653/v1/P19-1355

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. Proceedings of the IEEE International Conference on Computer Vision, 843-852. https://doi.org/10.1109/ICCV.2017.97

Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2020). Efficient processing of deep neural networks. Synthesis Lectures on Computer Architecture, 15(2), 1-341. https://doi.org/10.2200/S01004ED1V01Y202004CAC050

Tao, F., Zhang, M., Liu, Y., & Nee, A. Y. C. (2018). Digital twin driven prognostics and health management for complex industrial systems. CIRP Annals, 67(1), 169-172. https://doi.org/10.1016/j.cirp.2018.04.055

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications, 1-6. https://doi.org/10.1109/CISDA.2009.5356528

Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient transformers: A survey. ACM Computing Surveys, 55(6), 1-28. https://doi.org/10.1145/3530811

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

ISSN: 3071-5601 | © 2025 The Author(s). Published by DATAMIND under CC BY 4.0
https://inatgi.in/index.php/dm/index