INTERNATIONAL JOURNAL OF ADVANCED ARTIFICIAL INTELLIGENCE RESEARCH (IJAAIR)

eISSN: 3087-4270

Volume. 02, Issue. 08, pp. 09-15, August 2025

AI-AUGMENTED FRAMEWORKS FOR DATA QUALITY VALIDATION: INTEGRATING RULE-BASED ENGINES, SEMANTIC DEDUPLICATION, AND GOVERNANCE TOOLS FOR ROBUST LARGE-SCALE DATA PIPELINES

John M. Davenport
University of Edinburgh

Article Received: 01/08/2025, Article Accepted: 11/08/2025, Article Published: 18/08/2025

ABSTRACT

Background: The exponential growth of data generation, coupled with the proliferation of large language models (LLMs) and complex analytic systems, has elevated the importance of comprehensive, scalable, and explainable data quality validation. Traditional rule-based and statistical validation systems face challenges at web-scale data volumes, semantic duplication, and heterogeneous governance requirements (Apache Griffin, 2024; Deequ, 2024; Great Expectations, 2024). Recent work on semantic deduplication and LLM-assisted validation suggests hybrid frameworks that combine deterministic checks, probabilistic inference, and semantic reasoning can yield higher-quality, more actionable validation outcomes (Abbas et al., 2023; Achiam et al., 2023).
Methods: This article synthesizes design principles, operational architectures, and analytic methods into a unified, publication-ready research narrative. We construct a methodological taxonomy that integrates three principal components: (1) deterministic rule engines and metric-based validators drawn from industry-grade tools (Apache Griffin, Deequ, Great Expectations); (2) semantic deduplication and representation learning to reduce redundancy and improve downstream model training (Abbas et al., 2023); and (3) governance orchestration and qualitative-process integration for auditability and human-in-the-loop oversight (Qualitis, Nvivo, wenjuanxing). Each component is elaborated with procedural steps, expected outputs, failure modes, and interoperability constraints, building from both open-source tooling and contemporary academic research (Malviya & Parate, 2025; Wu et al., 2023).
Results: Through a detailed descriptive analysis, we identify how hybrid validation pipelines can achieve improvements in precision and recall of data error detection, reduce model degradation attributable to duplicated or low-quality samples, and enhance human interpretability. Specifically, semantic deduplication reduces redundant training exposures and dataset bloat, while rule-based validators ensure invariants and schema-level integrity (Abbas et al., 2023; Apache Griffin, 2024). Governance modules provide audit trails and decision rationales necessary for regulated domains such as insurance and healthcare (Malviya & Parate, 2025; Diaby et al., 2013).
Conclusions: An AI-augmented hybrid approach—anchored by robust rule engines, enriched by representation-aware deduplication, and governed through orchestration platforms—offers a promising direction for modern data quality validation. This framework balances computational efficiency, explainability, and adaptability, enabling institutions to manage the twin demands of scale and accountability in contemporary data ecosystems (Great Expectations, 2024; Deequ, 2024).

KEYWORDS

Data quality validation, semantic deduplication, rule engines, governance, large-scale data, AI-augmented validation

INTRODUCTION

The modern data landscape has evolved rapidly from closed, manageable datasets to sprawling, multi-source ecosystems. Enterprises and research institutions ingest billions of records daily from streams, user submissions, telemetry, and third-party vendors. This shift magnifies the importance of data quality validation, as low-quality inputs propagate through analytic systems and machine learning models, producing incorrect inferences, biased decisions, and, in regulated settings, legal or ethical violations (Apache Griffin, 2024; Deequ, 2024). Historically, practitioners relied on schema-checks, statistical profiling, and handcrafted rules to detect anomalies and preserve data integrity. While these approaches remain necessary, they are no longer sufficient when datasets are noisy, semantically redundant, or when subtle distributional shifts undermine model performance (Abbas et al., 2023; Wu et al., 2023).

This article constructs a comprehensive, theoretically grounded, and operationally detailed framework for AI-augmented data quality validation. We aim to provide an academically rigorous yet practically grounded account that preserves reproducibility and interpretability. The proposed framework synthesizes three strands of literature and tooling: (1) rule-based and metric-driven validators, exemplified by industry open-source projects that codify data expectations and compute metrics; (2) recent innovations in semantic deduplication and representation learning that address redundancy and data efficiency; and (3) governance and human-centered tools that provide qualitative validation, audit records, and stakeholder coordination. Each strand addresses a facet of the overall problem: correctness, efficiency, and accountability.

The need for such synthesis arises from several observed gaps. First, many validation systems operate in isolation—performing schema validation but ignoring semantic redundancy; others perform de-duplication without preserving provenance or audit trails (Apache Griffin, 2024; Deequ, 2024). Second, large language models and representation-based methods introduce both opportunities and risks: they can suggest transformations, identify semantic similarities, and generate synthetic augmentations, but they can also amplify biases and fabricate plausible falsehoods if trained or validated on noisy corpora (Achiam et al., 2023; Abbas et al., 2023). Third, governance requirements in domains like insurance, healthcare, and public policy demand transparent, reproducible processes that document how data quality checks were performed and why specific remediation decisions were made (Malviya & Parate, 2025; Diaby et al., 2013).

This paper addresses these gaps by articulating a hybrid approach that combines deterministic rules, semantic deduplication, and governance orchestration. We do not present a single software artifact; instead, we produce a conceptual and methodological blueprint that can be operationalized using established tools and recent research advances. Our contribution is threefold. First, we offer a rigorous taxonomy of data validation components and their interrelations. Second, we trace failure modes and propose mitigations that are informed by both industrial practice and scholarly work. Third, we provide prescriptive guidance for implementing staged pipelines that balance automation with human oversight, emphasizing auditability, explainability, and adaptivity.

Throughout the manuscript, we ground claims in extant tools and studies: Apache Griffin, Deequ, and Great Expectations for rule-based validation and metrics (Apache Griffin, 2024; Deequ, 2024; Great Expectations, 2024); Semdedup and representation-learning work to address redundancy (Abbas et al., 2023); and governance/qualitative tools such as Qualitis, Nvivo, and wenjuanxing for orchestration and human-centered validation (Qualitis, 2024; Nvivo, 2024; wenjuanxing, 2024). We also reference survey and methodological texts to provide theoretical context for LLMs, benchmarking pitfalls, and the broader implications for recommendation and clinical systems (Wu et al., 2023; McIntosh et al., 2024; Agrawal et al., 2022).

METHODOLOGY

This section articulates a detailed, text-based methodology appropriate for academic and industrial adoption. The methodology is prescriptive: it describes the architecture, the components and their interactions, validation checks and metrics, remediation strategies, human-in-the-loop protocols, and governance recordkeeping. The design is intentionally modular so that organizations can adopt individual components according to maturity and resource constraints (Apache Griffin, 2024; Deequ, 2024).

Framework Overview

The proposed framework comprises three interlocking modules: Validator Layer, Semantic Layer, and Governance Layer.

1. Validator Layer (Deterministic checks and metrics): This layer executes schema validation, statistical profiling, constraint checks, and expectation assertions. It is modeled on tools such as Apache Griffin and Deequ, which provide scalable metric computation and rule evaluation across large datasets (Apache Griffin, 2024; Deequ, 2024). Validators are organized into categories: structural (schema adherence), content (value ranges, types, nullability), relational (foreign key consistency), distributional (expected distributions, drift detection), and custom domain rules (business logic).

2. Semantic Layer (Representation-aware deduplication and semantic checks): Leveraging representation learning, this layer computes embeddings and semantic similarity to identify duplicates, near-duplicates, and semantically equivalent records that syntactic validators miss. Semdedup demonstrates that semantically aware deduplication can dramatically reduce training set redundancy and improve data efficiency (Abbas et al., 2023). The Semantic Layer integrates deduplication with probabilistic scoring to avoid overzealous deletions.

3. Governance Layer (Orchestration, audit trails, human-in-the-loop): This final layer orchestrates validations, records provenance, facilitates human review, and enforces policy. Tools like Qualitis provide workflow and governance mechanisms, while qualitative software like Nvivo and survey platforms like wenjuanxing support stakeholder engagement and labeling processes (Qualitis, 2024; Nvivo, 2024; wenjuanxing, 2024). Governance also includes retention of supplemental materials and DOIs for reproducibility (Supplemental Materials, 2024).

Detailed Procedural Steps

Below we present concrete steps to implement the framework. Each step frames expected inputs, outputs, computational considerations, and failure modes with suggested citations to underpin the rationale.

1.Ingestion and Initial Profiling

○Objective: Capture raw data, metadata, and provenance. Compute initial metrics: row counts, null rates, type distributions, cardinalities.

○Tools & Rationale: Use Deequ-style metric computation for scalability and expressiveness (Deequ, 2024). Apache Griffin offers pipeline-level integration and lineage capture that is beneficial at enterprise scale (Apache Griffin, 2024).

○Failure Modes & Mitigations: Incomplete metadata leads to ambiguous rule application; therefore embed automated metadata extraction and require explicit provenance fields during ingestion (Apache Griffin, 2024).

2.Schema and Constraint Validation

○Objective: Enforce schema constraints (types, ranges), identify structural anomalies, and flag violations.

○Procedure: Implement expectations (e.g., value ranges, regex patterns) similar to Great Expectations' human-friendly assertions (Great Expectations, 2024). Prioritize deterministic checks for invariants that are safety-critical or regulatory (e.g., patient IDs, financial transaction formats).

○Notes on Design: Maintain parametrized constraints to adapt to changing upstream formats without rewriting rules entirely (Great Expectations, 2024).

3.Statistical and Distributional Checks

○Objective: Identify shifts in distributions, detect missing modes, and surface anomalies based on historical baselines.

○Method: Compute statistical distance metrics and establish rolling baselines. Use adaptive thresholds calibrated to domain tolerance for variance. Document assumptions and thresholds in governance records (Apache Griffin, 2024).

○Caveat: Statistical tests can be sensitive to sample size; therefore combine statistical alarms with semantic checks to reduce false positives (Deequ, 2024).

4.Semantic Deduplication and Representation-Based Validation

○Objective: Remove or annotate semantically redundant records to improve model training efficiency and reduce bias propagating from duplicated content.

○Method: Compute embeddings (textual or multimodal) and use clustering or approximate nearest neighbor search to identify near-duplicates. Apply semantic deduplication thresholds informed by Semdedup methodology to choose which records to remove or downsample (Abbas et al., 2023).

○Key Considerations: Preserve provenance and the capacity to recover removed records. Use probabilistic scoring to avoid deterministic deletions when semantic similarity is ambiguous (Abbas et al., 2023).

5.Hybrid Anomaly Scoring and Prioritization

○Objective: Combine deterministic rule-violations and semantic anomaly scores to derive a composite risk score per record or batch.

○Method: Construct a composite scoring function that weights constraint violations more heavily for safety-critical features and semantic anomalies more heavily for model-training concerns (Malviya & Parate, 2025). Tune weights based on domain impact analyses (Diaby et al., 2013).

○Governance: Record the weighting logic and maintain versioning so decisions are auditable.

6.Human-in-the-Loop Review and Remediation

○Objective: Route prioritized anomalies for expert review. Utilize qualitative and survey tools to collect contextual judgments.

○Implementation: Integrate platforms such as Nvivo for qualitative analysis of flagged records and wenjuanxing for structured reviewer feedback (Nvivo, 2024; wenjuanxing, 2024). Use review outcomes to refine rules and retrain representation models.

Design Principle: Avoid overwhelming reviewers; present aggregated summaries, representative exemplars, and clear remediation options.

7.Governance, Documentation, and Reproducibility

○Objective: Ensure every validation action is logged, rationales recorded, and supplemental materials archived.

○Procedure: Use governance platforms like Qualitis to track workflows, approvals, and policy enforcement (Qualitis, 2024). Archive datasets, validation reports, and scripts using persistent identifiers (Supplemental Materials, 2024).

○Regulatory Fit: Tailor the governance artifacts to the compliance regime (e.g., insurance, healthcare) and maintain data lineage for auditability (Malviya & Parate, 2025).

Interoperability and Implementation Notes

The framework is intentionally tool-agnostic. Apache Griffin, Deequ, and Great Expectations provide complementary capabilities; decisions as to which to use depend on organizational constraints. For example, Deequ offers tight integration with JVM-based ETL frameworks and excels at large-scale metric computation (Deequ, 2024), whereas Great Expectations promotes human-friendly expectation suites and integrates well with Python-centric ML stacks (Great Expectations, 2024). Apache Griffin is optimized for pipeline-level governance and distributed lineage capture (Apache Griffin, 2024). Semantic deduplication pipelines can be implemented using embedding libraries and ANN indices, but must be carefully designed to preserve provenance and to avoid introducing selection biases (Abbas et al., 2023).

RESULTS

Given the nature of this paper—a methodological and synthesis-focused contribution—we present descriptive, theory-driven results rather than empirical datasets. The results synthesize how the proposed hybrid pipeline addresses failure modes, improves certain measurable outcomes, and clarifies trade-offs. Each subsection discusses expected measurable impacts, the mechanisms by which these impacts arise, and the limitations of those expectations.

Impact on Detection Precision and Recall

By combining deterministic rules with semantic checks, an organization can expect an increased precision in anomaly detection and improved recall for semantically anomalous cases that syntactic checks miss. Deterministic validators reliably capture schema and constraint violations, ensuring high precision for structural errors (Apache Griffin, 2024; Great Expectations, 2024). Semantic deduplication increases recall for semantic problems—such as paraphrased spam, duplicated content, or near-duplicate reports—that would otherwise pass syntactic validation (Abbas et al., 2023). The composite scoring approach reduces false positives by cross-validating statistical anomalies with semantic context; anomalies that are statistical but semantically benign (e.g., a natural distributional shift due to a new campaign) can be deprioritized for remediation, thus reducing unnecessary human effort (Deequ, 2024).

Effect on Model Training Efficiency and Robustness

Semantic deduplication has been shown to reduce effective dataset size without sacrificing representation diversity, which can improve training efficiency and reduce overfitting to repeated data artifacts (Abbas et al., 2023). When models are trained on deduplicated corpora, they are exposed to more unique content per epoch, often leading to better generalization and reduced memorization of rare idiosyncratic patterns. However, excessive deduplication risks removing legitimate repeated signals (e.g., multiple independent reports of the same event) which can be important for estimating prevalence. To mitigate this, deduplication must retain provenance metadata and allow domain-specific retention policies (Abbas et al., 2023).

Governance and Auditability Outcomes

Embedding governance at the orchestration level—tracked via workflow platforms—yields significant improvements in reproducibility, policy compliance, and risk management. When validation processes are versioned, policy changes are recorded, and human review outcomes are stored, organizations can reconstruct decisions and satisfy auditors or regulators (Qualitis, 2024; Supplemental Materials, 2024). In regulated domains this capacity is often a legal requirement, and the framework provides a structured approach to meeting those obligations (Malviya & Parate, 2025).

Human Workload and Cognitive Load

By prioritizing anomalies through composite scoring and grouping similar issues, the framework reduces human workload. Presenting aggregated clusters of semantically similar anomalies allows reviewers to adjudicate categories rather than individual records, increasing throughput (Nvivo, 2024). The framework's hybrid approach mitigates the cognitive overload that arises when reviewers see many low-impact, noisy alerts.

Qualitative Benefits: Trust and Explainability

The explicit separation between deterministic rules and AI-driven semantic checks preserves explainability. Rule violations are inherently interpretable (e.g., a missing required field), while semantic decisions include probabilistic scores and representative exemplars to justify automated actions. This design fosters trust among stakeholders who require transparent rationales for remediation or data rejection (Great Expectations, 2024; Deequ, 2024).

Limitations of the Results

These results are descriptive and derived from synthesis of tooling and prior research rather than from controlled experiments. The magnitude of improvements will vary according to domain characteristics (e.g., text vs. tabular data), available compute, and the quality of existing metadata. Further empirical validation is necessary to quantify effects across diverse organizational contexts and data regimes (McIntosh et al., 2024).

DISCUSSION

The primary contribution of this paper is conceptual: articulating a practical, theoretically grounded hybrid framework for data quality validation that is robust to modern challenges such as semantic redundancy and scale. Below we explore theoretical implications, operational trade-offs, counter-arguments, and future research directions.

Theoretical Implications

The framework situates data validation within a layered epistemology. Deterministic checks correspond to the syntactic and ontological layers of knowledge—what the data must be to be considered valid—while semantic validation addresses the pragmatic and referential layers—what the data means and what it represents in the world. This layered approach aligns with philosophical accounts of meaning and representation and provides a basis for formalizing validation as a multi-level inference process (Abbas et al., 2023).

Moreover, the integration of representation learning into validation challenges conventional boundaries between data cleaning and modeling. Traditionally, data cleaning precedes modeling and remains model-agnostic. The Semantic Layer introduces model-aware operations—using learned representations to shape the dataset—thereby blurring the separation of concerns. This raises important philosophical and practical questions about whether dataset curation should be informed by models that will later be trained on that dataset, and what safeguards are required to prevent circular reasoning or confirmation bias (Achiam et al., 2023; Abbas et al., 2023).

Operational Trade-offs and Design Choices

Several trade-offs emerge when operationalizing the framework:

- Aggressiveness of Deduplication vs. Retention of Signal: Aggressive deduplication reduces redundancy but risks discarding legitimate repeated signals. The framework recommends probabilistic scoring, human review for borderline cases, and configurable retention policies sector-specific retention thresholds (Abbas et al., 2023).

- Automation Depth vs. Explainability: Greater automation reduces human burden but may decrease transparency. The framework suggests hybrid automation: automatic handling of low-risk violations and human adjudication for high-impact cases (Qualitis, 2024).

- Computational Cost vs. Benefit: Embedding computations and ANN searches introduce additional compute overhead. Cost–benefit analyses should be conducted to ensure representation-based checks are justified by downstream model improvements or governance needs (Abbas et al., 2023; Deequ, 2024).

- Versioning and Policy Stability: As rules, representation models, and thresholds change, governance must track versions and rationales. This requirement increases process complexity but is indispensable for auditability and reproducibility (Supplemental Materials, 2024).

Counter-Arguments and Responses

1.Counter-Argument: Representation-based deduplication introduces model bias and can homogenize datasets, harming minority representation.

Response: This is a valid concern. The framework requires monitoring of class-wise and subgroup-wise retention rates post-deduplication. If deduplication disproportionately affects underrepresented groups, this indicates a need to adjust similarity thresholds or adopt subgroup-preserving strategies (Abbas et al., 2023).

2.Counter-Argument: Adding layers increases operational complexity and slows pipelines.

Response: While complexity increases, the framework proposes staged adoption. Organizations can begin with deterministic validators and incrementally add semantic checks and governance modules as needed. Prioritizing checks based on impact reduces upfront complexity (Apache Griffin, 2024).

3.Counter-Argument: Human-in-the-loop processes are expensive.

Response: The framework reduces human costs by prioritizing and clustering anomalies, enabling batch review and targeted human intervention where most valuable. Investment in human review often pays dividends in regulatory compliance and risk reduction (Nvivo, 2024; Qualitis, 2024).

Limitations of the Proposed Framework

Although comprehensive, the framework is not a silver bullet. It relies on adequate metadata, reasonable compute budgets, and domain expertise for rule specification. In data-sparse environments or where metadata is missing, its efficacy is reduced. Additionally, the framework assumes the availability of embedding models suitable for the domain; domain adaptation may be needed for niche technical or multilingual datasets (Abbas et al., 2023; Wu et al., 2023).

Future Research Directions

We identify several promising areas for future research:

- Empirical Benchmarks: Standardized benchmarks to measure the marginal benefits of semantic deduplication across tasks and domains are needed. Benchmarks should capture model performance, data efficiency, and fairness metrics (McIntosh et al., 2024).

- Causal Impacts of Deduplication: Investigate how deduplication affects causal inference tasks where repeated observations of the same unit may be informative.

- Adaptive Governance Policies: Develop meta-policies that adapt thresholds and rules based on observed downstream impacts, closing the loop between model outcomes and validation logic.

- Human-Computer Interaction (HCI) for Review: Design interfaces and workflows that reduce reviewer fatigue, improve judgment calibration, and integrate crowd or expert feedback effectively (Nvivo, 2024).

- Explainable Representation Models: Advance embedding methods that provide interpretable semantic features to facilitate governance and human understanding (Abbas et al., 2023; Achiam et al., 2023).

CONCLUSION

The contemporary data environment demands validation frameworks that are scalable, explainable, and sensitive to semantic meaning. This paper presents a hybrid architecture that combines deterministic rule-based validators, semantic deduplication informed by representation learning, and governance orchestration to meet these demands. By integrating these components, organizations can improve anomaly detection, enhance model training efficiency, and satisfy regulatory and audit requirements.

The framework is concrete, modular, and amenable to staged implementation. It acknowledges trade-offs—between automation and oversight, cost and benefit, and deduplication aggressiveness and signal retention—and offers prescriptive mitigations. Importantly, the framework situates data validation not merely as a preprocessing chore but as a central, model-aware governance function. Future work should empirically quantify the gains of hybrid validation across domains, refine HCI for human review, and develop benchmarks to standardize evaluations.

This synthesis draws upon a diverse set of tools and studies, ranging from Apache Griffin, Deequ, and Great Expectations to Semdedup and governance platforms. Practitioners and researchers adopting this framework will need to tailor thresholds and policies to their specific domain constraints while preserving the core principles of provenance, explainability, and adaptive governance.

REFERENCES

1. Apache Griffin. 2024. https://griffin.apache.org/.

2. Deequ. 2024. https://github.com/awslabs/deequ.git.

3. Great Expectations. 2024. https://github.com/great-expectations/great_expectations.

4. Nvivo qualitative software. 2024. https://lumivero.com/products/nvivo/.

5. Qualitis. 2024. https://github.com/WeBankFinTech/Qualitis.

6. Supplemental Materials. 2024. https://doi.org/10.6084/m9.figshare.25928863.

7. wenjuanxing software. 2024. https://www.wjx.cn.

8. Abbas, Amro; Tirumala, Kushal; Simig, Dániel; Ganguli, Surya; Morcos, Ari S. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. arXiv preprint arXiv:2303.09540.

9. Achiam, Josh; Adler, Steven; Agarwal, Sandhini; Ahmad, Lama; Akkaya, Ilge; Aleman, Florencia Leoni; Almeida, Diogo; Altenschmidt, Janko; Altman, Sam; Anadkat, Shyamal; et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

10. Kachris, C. 2024. A Survey on Hardware Accelerators for Large Language Models. arXiv:2401.09890.

11. Wu, L.; Zheng, Z.; Qiu, Z.; Wang, H.; Gu, H.; Shen, T.; Qin, C.; Zhu, C.; Zhu, H.; Liu, Q.; et al. 2023. A Survey on Large Language Models for Recommendation. arXiv:2305.19860.

12. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.;

Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A Survey of Large Language Models. arXiv:2303.18223.

13. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open Foundation and Fine-tuned Chat Models. arXiv:2307.09288.

14. Agrawal, M.; Hegselmann, S.; Lang, H.; Kim, Y.; Sontag, D. 2022. Large Language Models Are Few-shot Clinical Information Extractors. arXiv:2205.12689.

15. Roussinov, D.; Conkie, A.; Patterson, A.; Sainsbury, C. 2022. Predicting Clinical Events Based on Raw Text: From Bag-of-Words to Attention-based Transformers. Frontiers in Digital Health, 3, 810260.

16. Malviya, S.; Parate, V. 2025. AI-Augmented Data Quality Validation in P&C Insurance: A Hybrid Framework Using Large Language Models and Rule-Based Agents. International Journal of Computational and Experimental Science and Engineering, 11(3). https://doi.org/10.22399/ijcesen.3613

17. Ollitrault, P.J.; Loipersberger, M.; Parrish, R.M.; Erhard, A.; Maier, C.; Sommer, C.; Ulmanis, J.; Monz, T.; Gogolin, C.; Tautermann, C.S.; et al. 2023. Estimation of Electrostatic Interaction Energies on a Trapped-ion Quantum Computer. arXiv:2312.14739.

18. Diaby, V.; Campbell, K.; Goeree, R. 2013. Multi-criteria decision analysis (MCDA) in health care: A bibliometric analysis. Operations Research for Health Care, 2, 20–24.

19. McIntosh, T.R.; Susnjak, T.; Liu, T.; Watters, P.; Halgamuge, M.N. 2024. Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence. arXiv:2402.09880.

20. Kuo, T. 2017. A modified TOPSIS with a different ranking index. European Journal of Operational Research, 260, 152–160.

21. Tang, R.; Han, X.; Jiang, X.; Hu, X. 2023. Does Synthetic Data Generation of LLMs Help Clinical Text Mining? arXiv:2303.04360.