BenchCouncil Transactions on Benchmarks, Standards and Evaluations 5 (2025) 100245

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/

Research article

Evaluatology-driven artificial intelligence

Guoxin Kang [b], Wanling Gao [a,b,*], Jianfeng Zhan [a,b,c]

[a] The International Open Benchmark Council, China
[b] ICT, Chinese Academy of Sciences, Beijing, China
[c] University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:
Evaluatology
Artificial intelligence

ABSTRACT

The prevailing data-driven paradigm in AI has largely neglected the generative nature of data. All data, whether observational or experimental, are produced under specific conditions, yet current approaches treat them as context-free artifacts. This neglect results in uneven data quality, limited interpretability, and fragility when models face novel scenarios. Evaluatology reframes evaluation as the process of inferring the influence of an evaluated object on the affected factors and attributing the evaluation outcome to specific ones. Among these factors, a minimal set of indispensable elements determines how changes in conditions propagate to outcomes. This essential set constitutes the evaluation conditions. Together, the evaluated object and its evaluation conditions form a self-contained evaluation system — a structured unit that anchors evaluation to its essential context.

We propose an evaluatology-based paradigm, which spans the entire AI lifecycle — from data generation to training and evaluation. Within each self-contained evaluation system, data are generated and distilled into their invariant informational structures. These distilled forms are abstracted into reusable causal-chain schemas, which can be instantiated as training examples. By explicitly situating every learning instance within such condition-aware systems, evaluation is transformed from a passive, post-hoc procedure into an active driver of model development. This evaluation-based paradigm enables the construction of causal training data that are interpretable, traceable, and reusable, while reducing reliance on large-scale, unstructured datasets. This paves the way toward scalable, transparent, and epistemically grounded AI.

1. Introduction

In artificial intelligence, the forms of data are endlessly diverse — spanning text, images, audio, and structured records — yet their underlying essence remains invariant [1]. This *nature of data* represents the stable informational structure that persists beneath superficial variations. However, the prevailing data-driven paradigm has largely overlooked this essence, instead pursuing a brute-force strategy of enumerating variations to cover possible conditions. Such an approach, even when exhaustive, remains brittle in the face of unseen scenarios, as it fails to address the causal factors that truly govern generalization [2].

However, The continued viability of data-driven AI is now being challenged by deep scarcity limitations. As models grow larger and more data-hungry, the supply of high-quality, human-authored training data has become a critical bottleneck [3]. Recent projections by Epoch AI warn that the stock of usable Internet-scale text data may be depleted by 2028 [4]. This looming *data ceiling* threatens not only the scalability of current systems but also undermines their epistemic reliability and generalization capabilities [5]. In response, many have turned to generative models to produce synthetic data in an attempt to overcome this limitation [6–10]. However, such data often deviates significantly from real-world distributions, introducing distributional shifts that compromise its effectiveness as training material. Coupled with growing concerns over bias, opacity, hallucination, and uncontrollable behaviors — problems exacerbated by the black-box nature of large models and their dependence on opaque training corpora — the limitations of the current paradigm are becoming increasingly apparent [11].

We argue that a fundamental shift is needed: from scaling data volume to capturing and leveraging the nature of data itself — its invariant informational essence beneath diverse forms. To this end, we propose an *evaluation-based paradigm*, in which learning is guided and constrained by well-defined evaluation conditions rather than by uncontrolled data accumulation [12,13]. Drawing from the emerging discipline of Evaluatology, we reimagine the AI lifecycle such that evaluation is not a terminal procedure but a core methodological principle, shaping data generation and governing training dynamics. This

* Corresponding author at: ICT, Chinese Academy of Sciences, Beijing, China.
E-mail address: gaowanling@ict.ac.cn (W. Gao).

https://doi.org/10.1016/j.tbench.2025.100245
Received 28 August 2025; Received in revised form 26 September 2025; Accepted 30 September 2025
Available online 15 October 2025

reconceptualization is both technical and epistemological, redefining how intelligence should be constructed, assessed, and trusted in the post-data-scaling era.

At the heart of this paradigm are Self-contained Evaluation Systems (SESs): self-contained units that produce data — both observational and experimental — under explicit evaluation conditions. Within each SES, data are produced under explicit evaluation conditions and then distilled into their essential informational structures, from which causal-chain schemas are abstracted and later instantiated into concrete training examples. This process guarantees that the resulting data are interpretable, causally grounded, and reusable across tasks.

When applied to large language model development, SESs enable the construction of causal training data — for instance, reasoning traces, grounding steps, or labeled decision points — that make the provenance of every learning instance explicit. In contrast to opaque, data-hungry pipelines, our approach yields a white-box training system in which condition changes can be precisely traced to output variations. This ensures interpretability, attribution, and transparency while significantly reducing dependence on large-scale unstructured datasets.

2. Related work and motivation

2.1. Related work

2.1.1. The existing AI paradigm

As shown in Fig. 1(a), early developments in data-driven artificial intelligence relied heavily on manually labeled data, which enabled supervised learning for tasks such as classification and regression. While effective in domains like computer vision [14] and natural language processing [15], labeled data incurs high annotation costs, suffers from limited scalability, and is often prone to label noise [16], limiting its applicability in large-scale or open-ended scenarios.

To overcome these limitations, researchers have increasingly turned to unlabeled data, leveraging unsupervised and self-supervised learning techniques to extract patterns and representations without explicit human annotations. Classical methods include clustering [17], dimensionality reduction [18], and association mining. Recent advances in self-supervised learning — such as masked language modeling [19], contrastive learning [20], and masked image modeling [21] — have demonstrated strong performance across modalities and become foundational for large-scale pretraining.

From the perspective of learning paradigms, supervised and unsupervised learning are often sufficient for solving simple tasks. However, as task complexity increases — e.g., involving dynamic environments, delayed feedback, or alignment with human intent — these paradigms are frequently integrated with reinforcement learning to enable adaptive, multi-stage training pipelines. This hybrid approach has become central to the development of general-purpose systems such as large language models, where unsupervised learning supports knowledge acquisition, supervised learning guides task-specific behavior, and reinforcement learning fine-tunes alignment with user preferences.

In this work, we focus on static data sources. While, reinforcement learning, which depends on interaction-driven data collection and reward-based optimization, follows a fundamentally different data paradigm and is therefore excluded from our data-centric analysis [22–24].

2.1.2. The basic concepts, theories, and methodologies in evaluatology

Evaluatology [12,13] conceptualizes evaluation as the process of inferring the effect of an evaluated object on the affected objects. An object naturally exerts influence on multiple others, which can be categorized as directly or indirectly affected. Based on this principle, Evaluatology stresses the need to specify evaluation conditions (ECs) and to identify the set of affected objects. Together, the evaluated object, the affected objects, and the evaluation conditions form a Self-contained Evaluation System (SES). For instance, in evaluating CPU performance, the CPU serves as the evaluated object, while stakeholders' concerns — such as running time — require consideration of affected objects including the dataset, algorithm implementation, programming framework, operating system, compiler, processor, memory, etc [25]. Each possible configuration of these affected objects constitutes a point in a vast evaluation condition space, where variations in conditions naturally lead to variations in running time results. This structured perspective highlights that evaluation is inseparable from the context in which data are generated.

Inspired by Evaluatology, we reflect on the current data-driven paradigm in AI. Regardless of whether observational or experimental, all data are inherently generated under specific conditions. However, prevailing AI training methods largely ignore these generative conditions and focus exclusively on the data themselves. Such a deficiency leads to uneven and difficult-to-evaluate data quality, constrains interpretability and the capacity for causal discovery, and renders models fragile in the face of novel scenarios.

Our research intuition is that explicitly incorporating both data and their generative conditions into the training process can substantially enhance the effectiveness and transparency of AI. Even under limited data availability, leveraging the interplay between data and conditions allows the discovery of deeper causal structures, enabling models to capture the invariant informational essence beneath data diversity. By grounding learning in condition-aware causal relationships, we move toward more robust, interpretable, and genuinely intelligent systems.

2.2. Motivation: The limitations of existing AI paradigm

The modern trajectory of data-driven artificial intelligence has been shaped by the belief that more data yields better models. This principle underlies the development of large language models (LLMs), whose performance scales predictably with training data volume, model size, and compute budget [26]. However, this scaling paradigm is increasingly constrained by a looming data bottleneck. As high-quality human-authored data becomes saturated and expensive to curate, synthetic data generation has emerged as a promising alternative.

Despite its scalability, synthetic data introduces a new layer of complexity [27–29]. Crucially, the quality of synthetic data is fundamentally limited by the generative models that produce it, which are often black-box architectures with little transparency or interpretability. This lack of visibility makes it difficult to trace the root causes of errors or biases in downstream models back to specific properties of the synthetic data. When performance deteriorates, it remains unclear whether the issue lies in data coverage, semantic consistency, or deeper representational flaws.

In practice, current synthetic data suffers from several well-documented issues: (1) generative models may fail to match the statistical distribution of real data, introducing biases that impair generalization. (2) synthetic samples often contain logical contradictions or distorted features that are difficult to detect but can corrupt pretraining (see Fig. 1). Low diversity and mode collapse: generators tend to produce samples with limited variation, leading to models that overfit narrow modes and underperform on real-world variability.

To improve the quality, reliability, and usefulness of synthetic data, it is imperative to enhance the interpretability and evaluation of generative models. Without understanding what a generator has learned — and what it systematically omits — scaling synthetic corpora becomes a blind process, susceptible to spurious correlations and misalignment.

These observations motivate a shift toward an evaluatology-driven AI paradigm, in which systematic attribution and interpretability are not afterthoughts but central components of the AI development cycle. By incorporating formal principles of evaluatology into the design, analysis, and deployment of generative models, we can better align synthetic data generation with downstream objectives, ensure quality control, and build AI systems that are not only larger, but measurably better.

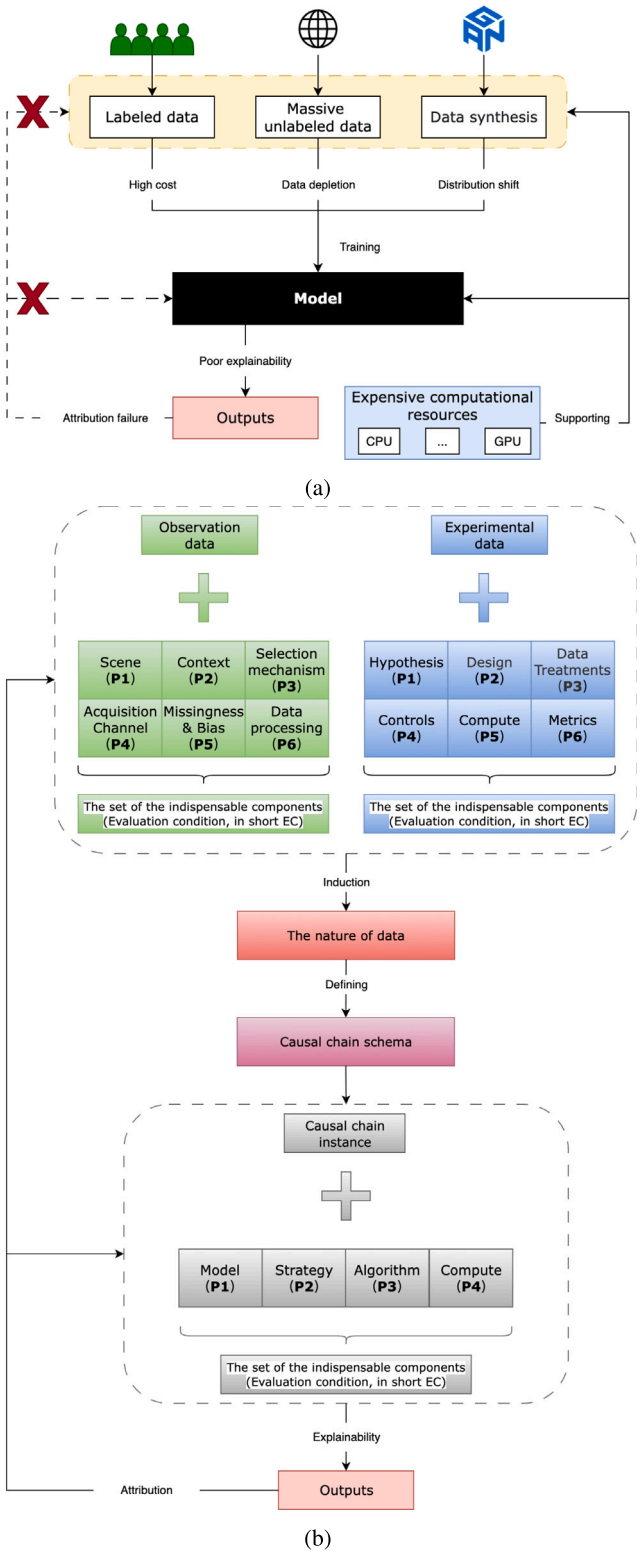

Fig. 1. Comparison of AI paradigms. (a) Data-driven: models are trained on massive, unstructured data and self-discover patterns, offering limited causal attribution and interpretability. (b) Evaluatology-driven: data are generated under explicit evaluation conditions within Self-contained Evaluation Systems and abstracted into causal-chain training instances, yielding interpretable, traceable, and attributable outputs.

3. The new AI paradigm based on evaluatology

3.1. Overview

Our methodology is grounded in *Evaluatology*, which reconceptualizes evaluation as the process of inferring the influence of an evaluated object on affected factors and attributing the observed outcomes to specific ones. The central methodological unit is the *Self-contained Evaluation System (SES)*, which anchors evaluation to its essential generative context.

Definition 3.1 (*Self-contained Evaluation System*). A Self-contained Evaluation System is defined as

$$SES = (E, C),$$

where E denotes the *evaluated object* and C denotes the *evaluation conditions*. The evaluation conditions C are composed of a minimal set of indispensable affected factors, which determine how variations in conditions propagate to outcomes.

Within this formalism, both data and models are situated in explicitly defined SESs. Data instances arise as functions of E and C, while induction over multiple condition configurations reveals their invariant informational essence. These essences can be abstracted into causal-chain schemas and instantiated into causal-chain instances, which serve as interpretable and reusable training data. Models themselves can also be represented as SESs, ensuring that their outputs are attributable to well-defined evaluation conditions. For example, in the task of video keyframe selection, event segmentation is one indispensable factor in C, and can be instantiated through visual-based shot boundary detection, audio-based segmentation via automatic speech recognition (ASR) pauses or speaker changes, or multimodal fusion of visual and auditory cues. The segmentation strategy determines how the video is partitioned and directly affects the structure of the output, illustrating how variations in condition configurations propagate to outcomes. Models themselves can also be represented as SESs, ensuring that their outputs are attributable to well-defined evaluation conditions. While defining an SES requires some domain-specific effort, it enforces strict evaluation conditions that enhance interpretability and reduce data dependency and long-term cost.

3.2. Self-contained evaluation systems for observational data

Observational data emerge from natural processes but are nonetheless shaped by indispensable generative factors as shown in Fig. 1(b). We formalize their essential affected factor set as

$$\mathcal{O} = \{S, T, M, A, B, P\},$$

where S is the *Scene*, T the *Context*, M the *Selection Mechanism*, A the *Acquisition Channel*, B the *Missingness & Bias*, and P the *Data Processing*.

Given an evaluated object E, an observational data instance d^{obs} is generated as

$$d^{obs} = f_{obs}(E, \mathcal{O}),$$

where f_{obs} denotes the observational generative mapping. By considering multiple admissible configurations $\mathcal{O}_i \in \Omega_{obs}$, we obtain the invariant informational essence of observational data:

$$\phi^{obs} = \text{Induction}\big(\{f_{obs}(E, \mathcal{O}_i) \mid \mathcal{O}_i \in \Omega_{obs}\}\big).$$

3.3. Self-contained evaluation systems for experimental data

Experimental data are deliberately generated under interventions and controls, guided by explicit causal inquiry. Their essential affected factor set is formalized as

$$\mathcal{X} = \{H, D, K, Q, R\},$$

where H is the *Hypothesis*, D the *Design*, K the *Controls*, Q the *Compute*, and R the *Metrics*.

For an evaluated object E, an experimental data instance d^{exp} is generated as

$$d^{exp} = f_{exp}(E, \mathcal{X}),$$

with f_{exp} representing the experimental generative mapping. The invariant informational essence of experimental data is induced as

$$\phi^{exp} = \text{Induction}\big(\{f_{exp}(E, \mathcal{X}_j) \mid \mathcal{X}_j \in \Omega_{exp}\}\big).$$

3.4. Induction and causal chain construction

Both ϕ^{obs} and ϕ^{exp} serve as abstract representations of the invariant essence of data. These are formalized as *causal-chain schemas* that describe how changes in evaluation conditions propagate to outcomes. When instantiated under concrete configurations of $\mathcal{O}$ or $\mathcal{X}$, these schemas yield *causal-chain instances*, which serve as interpretable, reusable, and condition-aware training examples.

3.5. Self-contained evaluation systems for models

Models themselves operate within SESs. The essential affected factor set for models can be denoted as

$$\mathcal{M} = \{T, A, Q, I\},$$

where T denotes the *Strategy*, A the *Algorithm*, Q the *Compute*, and I the *Implementation*.

The output of a model y can then be expressed as

$$y = f_{model}(E, \mathcal{M}),$$

with f_{model} representing the mapping from the evaluated object E and model-level evaluation conditions $\mathcal{M}$ to outputs. Since y is generated under explicit evaluation conditions, it is inherently explainable and attributable to specific elements of $\mathcal{M}$.

3.6. Case study: applying self-contained evaluation systems to video retrieval

Our experiments show that a user's query of vector database is aligned with only about **1.95%** of the keyframes in long-form videos. Motivated by an evaluation-based perspective, we introduce a *essential factor set* to defining the SES for keyframe selection — *event segmentation*, *textual signals*, *temporal context*, *scene/quality constraints*, *selection mechanism*, and *de-redundancy*. This essential factor set, grounded in a self-contained evaluation system, not only enables fine-grained attribution of retrieval performance to individual design factors, but also exemplifies the evaluatology-driven methodology underlying automatic database design.

In the Self-contained Evaluation System (SES), the **evaluation object** E is the target keyframe. The **evaluation conditions** C refer to a essential set of indispensable affected factors (see Definition 3.1):

1. Segment each video into *clips* using shot boundaries together with automatic speech recognition (ASR) pauses and speaker changes;
2. Derive a short *title/summary* per clip from subtitles/ASR as textual features;
3. Rank candidate frames by *textual match* (best matching 25 or embedding similarity) and by *temporal proximity* to query-relevant timestamps (closer is better);
4. Apply *scene/quality filters* (avoid blur and heavy motion; prefer stable frames shortly after shot transitions);
5. Within each clip, keep *1–3* frames and *merge near-duplicates* (retain a single frame for adjacent time spans);
6. Enforce a *keep-rate* of $\kappa \approx 2\%$–3% via a gating threshold, relaxing to $\sim 5\%$ when textual evidence is sparse.

By restricting both training and inference to keyframes identified by the SES, the model concentrates on *causally relevant semantics*, achieving retrieval quality comparable to full-frame pipelines while *substantially reducing* training data and inference cost.

3.7. Summary

In summary, our methodology situates both data and models within explicit Self-contained Evaluation Systems, each defined by an evaluated object E and its indispensable evaluation conditions C. By formalizing observational and experimental data generation as functions of (E, C), and by extending the same logic to models, we establish a unified framework in which invariant informational essences can be induced, abstracted into causal-chain schemas, and instantiated into training data. This paradigm transforms evaluation into an active methodological principle that spans the entire AI lifecycle, enabling interpretability, traceability, and epistemic grounding in artificial intelligence. Notably, this paradigm provides a theoretical foundation for automatic database design, by enabling the isolation of external confounding factors and attributing performance outcomes directly to the database design itself.

4. Conclusion

This work advances an evaluatology-based paradigm that spans data generation, training, and assessment. Its central construct — the Self-contained Evaluation System — couples an evaluated object with evaluation conditions constituted by a essential set of indispensable affected factors, thereby fixing the context required for coherent causal attribution. Within SESs, data are generated under explicit conditions, distilled into invariant informational structures, abstracted as causal-chain schemas, and instantiated as training examples; consequently, evaluation becomes a generative principle rather than a post-hoc procedure. The paradigm yields condition-aware learning with interpretable, traceable outputs and reduces dependence on massive, unstructured corpora. Taken together, these elements provide a scalable, unifying foundation for transparent, robust, and epistemically grounded AI, and furnish a precise basis on which future theory, benchmarks, and systems can be systematically developed.

CRediT authorship contribution statement

Guoxin Kang: Writing – original draft. **Wanling Gao:** Writing – review & editing. **Jianfeng Zhan:** Writing – review & editing, Conceptualization.

Funding

This paper is supported by the Strategic Research Special Funding of the Bureau of Development and Planning, Chinese Academy of Sciences (GHJ-ZLZX-2024-34).

Declaration of competing interest

The authors declare no competing interests.

References

[1] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.

[2] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F.A. Wichmann, Shortcut learning in deep neural networks, Nat. Mach. Intell. 2 (11) (2020) 665–673.

[3] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D.d.L. Casas, L.A. Hendricks, J. Welbl, A. Clark, et al., Training compute-optimal large language models, 2022, arXiv preprint arXiv:2203.15556.

[4] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim, M. Hobbhahn, Will we run out of data? Limits of LLM scaling based on human-generated data, 2024, Epoch AI Blog, https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data. (Accessed 28 August 2025).

[5] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, R. Anderson, The curse of recursion: Training on generated data makes models forget, 2023, arXiv preprint arXiv:2305.17493.

[6] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357.

[7] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, Domain randomization for transferring deep neural networks from simulation to the real world, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2017, pp. 23–30.

[8] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling tabular data using conditional gan, Adv. Neural Inf. Process. Syst. 32 (2019).

[9] J. Jordon, J. Yoon, M. Van Der Schaar, PATE-GAN: Generating synthetic data with differential privacy guarantees, in: International Conference on Learning Representations, 2018.

[10] S.R. Richter, V. Vineet, S. Roth, V. Koltun, Playing for data: Ground truth from computer games, in: European Conference on Computer Vision, Springer, 2016, pp. 102–118.

[11] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Trans. Inf. Syst. 43 (2) (2025) 1–55.

[12] J. Zhan, L. Wang, W. Gao, H. Li, C. Wang, Y. Huang, Y. Li, Z. Yang, G. Kang, C. Luo, et al., Evaluatology: The science and engineering of evaluation, BenchCouncil Trans. Benchmarks, Stand. Eval. 4 (1) (2024) 100162.

[13] J. Zhan, Fundamental concepts and methodologies in evaluatology, BenchCouncil Trans. Benchmarks, Stand. Eval. 4 (3) (2024) 100188.

[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012).

[15] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (7) (2011).

[16] D. Rolnick, A. Veit, S. Belongie, N. Shavit, Deep learning is robust to massive label noise, 2017, arXiv preprint arXiv:1705.10694.

[17] S. Lloyd, Least squares quantization in PCM, IEEE Trans. Inform. Theory 28 (2) (1982) 129–137.

[18] I. Jolliffe, Principal component analysis, in: International Encyclopedia of Statistical Science, Springer, 2011, pp. 1094–1096.

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[20] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PmLR, 2020, pp. 1597–1607.

[21] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[22] R.S. Sutton, A.G. Barto, et al., Reinforcement Learning: An Introduction, vol. 1, no. 1, MIT press Cambridge, 1998.

[23] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, et al., Dapo: An open-source llm reinforcement learning system at scale, 2025, arXiv preprint arXiv:2503.14476.

[24] C. Sun, S. Huang, D. Pompili, Llm-based multi-agent reinforcement learning: Current and future directions, 2024, arXiv preprint arXiv:2405.11106.

[25] C. Wang, L. Wang, W. Gao, Y. Yang, Y. Zhou, J. Zhan, Achieving consistent and comparable CPU evaluation outcomes, 2024, arXiv preprint arXiv:2411.08494.

[26] J. Kaplan, S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, 2020, arXiv preprint arXiv:2001.08361.

[27] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, Y. Gal, AI models collapse when trained on recursively generated data, Nature 631 (8022) (2024) 755–759.

[28] A. Mumuni, F. Mumuni, N.K. Gerrar, A survey of synthetic data augmentation methods in computer vision, 2024, arXiv preprint arXiv:2403.10075.

[29] A. Bauer, S. Trapp, M. Stenger, R. Leppich, S. Kounev, M. Leznik, K. Chard, I. Foster, Comprehensive exploration of synthetic data generation: A survey, 2024, arXiv preprint arXiv:2401.02524.