Received: 27 June 2025 | Revised: 9 October 2025 | Accepted: 20 October 2025 | Published online: 23 December 2025

RESEARCH ARTICLE

Artificial Intelligence and Applications
2025, Vol. 00(00) 1–9
DOI: 10.47852/bonviewAIA52026620

Synthetic Data in AI: Performance Gains versus Hallucination Risk

Gabriel Silva-Atencio[1,*]

[1] *Engineering Department, Universidad Latinoamericana de Ciencia y Tecnología, Costa Rica*

Abstract: The incorporation of synthetic data into AI training pipelines poses a basic paradox: although it improves model resilience and mitigates data shortages, it also increases the likelihood of hallucinations. Using a mixed-methods approach, this research rigorously analyzes this trade-off and shows that synthetic data increase hallucination rates by a factor of 4.7 while improving perturbation resistance by 23%. Spatially improbable artifacts in computer vision (17% increase), factual mistakes in natural language processing (22% of outputs), and clinically dangerous errors in healthcare (3.7% of incidences) are examples of domain-specific manifestations. The Synthetic Data Fidelity Theorem, which extends the traditional bias-variance decomposition to explicitly encompass synthetic artifact propagation, is presented to fill the theoretical gap in the knowledge of these effects. Additionally, with a prediction accuracy of $R^2 = 0.89$, the FAITH metric system (Factuality, Alignment, Integrity Tracking for Hallucinations) is designed for real-time risk management. According to causal analysis, 23.4% of synthetic-data-induced hallucinations are caused by reward hacking and feature entanglement. Evidence suggests that hybrid data regimes (≤60% synthetic content) minimize mistakes by 41% without compromising performance, which defies the notion of universal application. To guarantee responsible deployment in crucial AI systems, the results call for a paradigm change toward domain-specific governance, backed by evidence-based recommendations for architectural choices, validation procedures, and policy frameworks.

Keywords: synthetic data, AI hallucinations, model robustness, domain-specific risks, FAITH metrics, responsible AI

1. Introduction

The major obstacles to the broad use of artificial intelligence (AI) in high-stakes domains such as autonomous systems [1] and medical systems [2] include the absence of real-world training data, privacy concerns, and inherent biases. Synthetic data, which are algorithmically generated to mimic real data distributions, have emerged as a game-changing solution with the promise of improved generalization through data augmentation [3], strong privacy guarantee [4], and the simulation of critical edge cases [5]. However, a significant and little understood risk—the propensity of synthetic data to amplify AI hallucinations—is overshadowing these notable benefits. These confident, persuasive, but factually inaccurate or unrealistic model outputs substantially undermine confidence and reliability [6, 7].

A major conflict in contemporary machine learning is highlighted by this paradox. Although it has been shown that synthetic data increase the model's resilience to disturbances and class imbalance [8, 9], they also add new failure modes. In addition to distributional mismatches [10], feature entanglement—a phenomenon in which learned representations of a model confuse semantically unrelated characteristics in the synthetic data—is the primary cause of these. For example, a model may provide results that are internally consistent but empirically wrong if it incorrectly links particular medical terms to diagnostic certainty [11, 12]. The effects are severe and domain-specific, showing up as authoritative but fabricated citations in natural language processing (NLP), pathophysiologically plausible but dangerously incorrect recommendations in healthcare, or spatially implausible artifacts in computer vision (CV) [13, 14].

Although they are still dispersed across domains, current mitigation efforts, such as hybrid training regimens [15] and uncertainty quantification tools [16, 17], show promise. There is no coherent theoretical framework that specifically describes the reasons why synthetic objects fail. The rapid adoption of synthetic data and the strict safeguards needed for their deployment in safety-critical applications are at odds, and this gap is made worse by the absence of standardized, cross-modal metrics for measuring hallucination risk [18, 19].

Therefore, the main research question of this study is the following: how can the dualistic effects of synthetic data—which simultaneously improve model robustness and increase hallucination rates—be theoretically formalized, empirically measured, and effectively reduced in various AI domains?

To address this, this study makes three coherent contributions. First, it introduces the Synthetic Data Fidelity Theorem, a novel theoretical framework that extends the conventional bias-variance decomposition and formally includes the propagation of synthetic artifacts as a fundamental component of generalization error [20]. Here is how to put this (Equation (1)):

$$H \leq (1 - D)\alpha \times C\beta. \quad (1)$$

Second, with a prediction accuracy of $R^2 = 0.89$, this study creates the FAITH metric system (Factuality, Alignment, Integrity Tracking for Hallucinations), a standardized suite for cross-domain, real-time monitoring and comparison of hallucination dangers. Third, the research produces empirically supported, domain-specific risk mitigation strategies using a mixed-methods methodology that combines expert validation and causal analysis with extensive benchmarking of 12 designs spanning CV, NLP, and healthcare.

To further the creation of more competent and reliable AI systems, this study attempts to provide a fundamental paradigm for using the

*Corresponding author: Gabriel Silva-Atencio, Engineering Department, Universidad Latinoamericana de Ciencia y Tecnología, Costa Rica. Email: gsilvaa468@ulacit.ed.cr

potential of synthetic data by fusing theoretical innovation with rigorous empirical validation. The literature review, methodological framework, empirical findings, and a discussion of the findings' implications are covered in depth in the parts that follow.

2. Literature Review

The increasing dependence on synthetic data in AI research signifies a transformative change, providing a robust remedy to the persistent challenges of data scarcity, privacy concerns, and intrinsic biases seen in real-world datasets. This change brings up a very important and not well-studied duality: synthetic data improve model performance while also making new failure modes more likely, especially hallucinations. A comprehensive examination of the existing literature indicates an area replete with specialized advancements but hindered by the absence of cohesive theoretical frameworks and standardized cross-modal assessment measures. This review compiles insights from 127 significant publications to delineate the field, organizing the research into six interconnected themes: generative models, CV applications, NLP applications, healthcare applications, hallucination and robustness, and ethics and governance (see Appendix A for a complete systematic analysis). The research delineates three essential multidisciplinary gaps that this work seeks to rectify.

The theoretical foundations of synthetic data creation are based on computational learning theory and the probably approximately correct (PAC) learning paradigm, which states that learnability depends on the quality and diversity of training instances [21]. The bias-variance trade-off [20] is what really controls the practice. Synthetic samples try to lower bias while also lowering variance in model estimates. Three main theoretical frameworks dominate modern synthesis: generative adversarial networks (GANs) and their game-theoretic minimax formulation [22, 23]; variational autoencoders (VAEs), which frame generation as variational inference [24]; and diffusion models, which offer a new framework based on thermodynamic principles and Markov chains [25, 26]. The theoretical advantages of these methodologies have been substantiated in several essential domains, such as privacy protection via differential privacy frameworks [27, 28], domain adaptability through covariate shift theory [29], and improved active learning techniques [30].

The efficacy of synthetic data is empirically established but is contingent upon a specific area. Synthetic data have greatly improved performance in CV when there is not much data available. For example, they improved the accuracy of tumor identification in medical imaging by 12.7% [31] and advanced semantic segmentation for self-driving cars, with models showing a 19% improvement in real-world performance metrics [32, 33]. In the same way, back-translation and other strategies have boosted BLEU scores for low-resource language translation by 4.2 points [34]. However, these performance improvements are being recorded simultaneously with a big increase in hallucination rates. For example, synthetic-augmented NLP models make references that are not true 22% of the time [35]. The healthcare field is a good example of the most important trade-off: the use of synthetic electronic health records (EHR) for predictive modeling can keep 91% of the accuracy of real data [36], but it has big problems with diagnosing rare diseases, and even worse, it can produce clinically dangerous results in 3.7% of cases [13].

The aggravation of hallucinations is the primary difficulty examined in this work. AI hallucinations, which are confident, coherent, but wrong outputs, may be divided into three types: input-conditioned, free-generation, and compound [7, 18]. They are hypothesized to represent manifestations of epistemic doubt, exhibiting a strong correlation with distributional shift ($\rho = 0.73$, $*p* < 0.001$) [3, 37]. Their expressions are extremely domain-specific: CV shows "space-connected" artifacts with topologies that do not make sense in 17% of samples [14], NLP makes "plausible fabrications" or references that sound authoritative but are wrong [38], and healthcare makes recommendations that are pathophysiologically plausible but very wrong, which is the most dangerous risk category [13, 39].

Currently, there are many different ways to reduce risks, but they are all separate. Hybrid training pipelines with ideal synthetic-to-real ratios (e.g., 60:40) show potential [15], in conjunction with uncertainty calibration approaches such as Monte Carlo dropout [17] and adversarial validation methods [40]. There is, however, a big difference

Table 1
Mapping gaps in the literature to the contributions of this study

Gap in existing literature	Contribution of this study	Relationship to literature
Lack of a unifying theoretical framework for synthetic-data-specific failures, particularly the trade-off between performance and hallucinations.	Synthetic Data Fidelity Theorem.	Extends the classical bias-variance decomposition [20] by formally incorporating synthetic artifact propagation as a key component of generalization error.
Absence of standardized, cross-domain metrics for measuring and comparing hallucinations (e.g., a common metric for CV, NLP, and healthcare).	The FAITH Metric System (FCS), Semantic Fidelity Index (SFI), Reality Alignment Metric (RAM)).	Integrates and unifies domain-specific principles (e.g., knowledge grounding from NLP and clinical guidelines from healthcare) into a single, adaptable framework for holistic risk assessment, enabling direct cross-domain comparison.
Lack of expert-validated, causal understanding of why synthetic data induce hallucinations (e.g., beyond correlation to causation).	Causal analysis via Bayesian networks identifying reward hacking ($\beta = 0.61$) and feature entanglement (23.4% attribution) as primary mechanisms.	Provides empirical validation and quantification for hypothesized failure modes (e.g., "gaming" concept [41]), moving from speculation to evidence-based causation.
Domain-specific silos: solutions and insights are rarely transferred across CV, NLP, and healthcare domains.	Cross-domain benchmarking & analysis of 12 architectures across three domains, revealing domain-specific failure patterns (e.g., spatial artifacts vs. clinical plausibility).	Bridges isolated research silos by applying a consistent methodological framework across domains, enabling the discovery of universal patterns and critical differences.
The presumption of universality: the lack of evidence-based, domain-specific guidelines for safe synthetic data usage.	Evidence-based guidelines, including the optimal 60/40 hybrid ratio and the identification of a U-shaped hallucination curve (inflection at 37% training budget).	Transforms general principles (e.g., "use hybrid data") into quantified, actionable protocols specific to different domains and training stages.

because models may frequently "game" synthetic training settings, doing well on benchmarks but poorly in real-world situations [41]. This indicates a more extensive reproducibility crisis [19] and highlights three principal deficiencies in the existing literature: an emphasis on short-term performance rather than longitudinal effects [42], an absence of standardized cross-domain metrics for hallucination [6, 18], and an insufficient comprehension of the ethical trade-offs between privacy and reliability [43, 44].

A significant impediment recognized in this analysis is the lack of a thorough, consistent assessment methodology. There are good metrics for certain fields, such BLEURT [45] and FactCC for NLP, FID [46] for image creation, and task-specific area under the curve (AUC)/F1 scores, but they do not work together. This isolated approach makes it impossible to directly compare how synthetic data affect CV, NLP, and healthcare. Moreover, although the sensitivity of general models to noise is well documented [47], the particular sensitivity of synthetic-data-trained models to perturbations that reveal their learned artifacts is much under-investigated. Finally, current mistake taxonomies, although useful, are often qualitative, speculative, or limited to a single modality [7], and they do not include the quantitative, expert-validated, and cross-domain severity analysis needed for thorough risk assessment.

This study indicates that the area is characterized by significant dissonance: the evident advantages of synthetic data are indisputable, but their implementation is obstructed by fragmented insights and the lack of tools to navigate their dualistic nature. There is no theoretical framework in the literature that explicitly analyzes the trade-off between performance benefits and hallucination risks, no common set of tools for measuring risks across domains, and no expert-validated, causal knowledge of the processes that cause synthetic-data-induced errors. These deficiencies not only show why this work is needed, but they also immediately lead to its primary contributions: the Synthetic Data Fidelity Theorem, the FAITH metric system, and the cross-domain benchmarking and causal analysis that come after (see Table 1).

3. Methodology

The methodological approach for this research is structured to rigorously examine the dualistic effects of synthetic data on AI model performance and hallucination risks. A sequential mixed-methods methodology is used, including three interrelated phases: quantitative benchmarking, qualitative characterization of hallucinations, and expert-driven causal analysis. This approach is based on computational learning theory [21] and an extension of the bias-variance trade-off paradigm [20] that explicitly includes the spread of synthetic artifacts, as shown by the Synthetic Data Fidelity Theorem (see Equation (1)). This makes sure that the approach is valid in theory while also dealing with real-world problems of using synthetic data.

The quantitative benchmarking phase is a large-scale test of 12 cutting-edge architectures in three important areas: CV (Vision Transformers - ViT, Diffusion Models, and ResNet-152), NLP (GPT-3.5, T5, and BERT-base), and healthcare (BioClinicalBERT and CheXNet). To see how synthetic data affect things, each architecture is trained on five different types of data: a real-only (R) baseline using datasets like ImageNet-1K, a synthetic-only (S) regime using data generated by state-of-the-art GANs, and three hybrid regimes (H1–H3) with different real-to-synthetic ratios (25/75, 50/50, and 75/25). The AdamW optimizer [48] is used to train all models. The learning rate is 2e-5, the batch sizes are unique to the domain (256 for CV, 32 for NLP, and 16 for healthcare), and an early stopping policy is used to cease training after five epochs of validation loss that does not go up.

The research use a full set of measures to evaluate each model, including how well it works and how often it is to hallucinate (see Table 2).

Table 2
Metrics for evaluating performance and hallucinations

Metric	Tool/method	Reference	Primary purpose
Accuracy (F1, AUC)	Scikit-learn	[49]	Standard performance gauge
Robustness ($\varepsilon = 0.1–0.5$)	CleverHans	[8]	Perturbation resistance
FCS	Wikidata grounding	[18]	NLP factuality check
SFI	CLIP-based evaluation	[50]	CV spatial plausibility
RAM	Clinical entailment models	-	Healthcare safety validation

Scikit-learn [49] is used to figure out standard performance measurements such as accuracy, F1-score, and AUC. The CleverHans library [9] is used to test how strong a model is against adversarial assaults ($\varepsilon = 0.1–0.5$). To fill the gap in standardized cross-domain assessment that was found in the literature review, three new, domain-specific hallucination metrics are developed and carefully described. The FCS for NLP uses Wikidata to check the accuracy of model outputs [18]. The SFI for CV uses Contrastive Language–Image Pre-training (CLIP) embeddings to find objects that do not make sense in space [50]. The RAM for healthcare combines clinical standards with entailment models to find suggestions that might be dangerous. After that, a multivariate regression analysis is done to predict the hallucination rate (H) as a function of data purity (D) and model complexity (C), as shown in Equation (2). The coefficients are based on the benchmark findings. The research employs nonparametric tests such as the Kruskal–Wallis and Mann–Whitney U tests to see how designs and data regimes vary from each other.

$$\log(H) = \beta_0 + \beta_1 \cdot \log(1 - D) + \beta_2 \cdot \log(C) + \beta_3 \cdot T + \epsilon. \quad (2)$$

The categorization of hallucinations transcends mere quantitative measurements. A sensitivity study is performed to investigate the distinct failure mechanisms of synthetic-trained models using domain-specific perturbations: to test how well the system works with low-fidelity inputs, Gaussian noise ($\sigma = 0.05–0.3$) is added to CV data. To test how stable the facts are, synonym replacement is used on NLP inputs. A Delphi consensus procedure [51] is also used with a panel of 35 professionals in the field (12 AI reliability specialists, 9 board-certified doctors, 8 NLP linguists, and 6 AI ethicists) to establish a verified, multimodal taxonomy of sorts of hallucinations. This panel sorts and gives severity ratings (1–5) to a carefully chosen set of 1,000 hallucinated outputs from the benchmark. This ensures a strict, expert-driven categorization that goes beyond theoretical categories to an empirically based and severity-weighted framework.

The last step is all about checking for cause. A Bayesian network architecture [52] is developed to transition from correlational results to causal inference, elucidating the principal processes responsible for synthetic-data-induced hallucinations. The network integrates factors derived from literature and expert contributions, including data purity, model complexity, training length, and particular synthetic data characteristics. This research statistically delineates and substantiates causal pathways, including reward hacking and feature entanglement, thus offering empirical support for the proposed failure modes articulated in the literature [41].

The reproducibility, all synthetic datasets created for this study, the code for the FAITH metrics (FCS, SFI, and RAM), and the scripts

for model training and statistical analysis will be made publicly accessible in a curated repository upon publication (see Appendix B), guaranteeing the complete replicability of the findings detailed in the results section.

4. Results

The tripartite methodological framework's empirical results offer a strong, quantitative proof of the dualistic effect of synthetic data, demonstrating a steady trade-off between improved performance and increased hallucination risk that is significantly influenced by domain-specific limitations. The findings provide new information about the causal processes behind errors caused by synthetic data and support the theoretical claims of the Synthetic Data Fidelity Theorem.

4.1. Trade-offs between hallucinations and cross-domain performance

The use of synthetic data and model performance have a nonlinear connection according to quantitative testing across the 12 designs and 5 data regimes. As shown in Figure 1, the suggested ideal hybrid regime (60% actual, 40% synthetic) continuously produced the best accuracy/reliability ratio. The mean average precision (mAP) of ViT models trained on this hybrid regime in CV was 9.2% higher than baselines based only on actual data (95% CI [8.7%, 9.8%]). The SFI measured and expert evaluation confirmed that this performance improvement was accompanied by a 17% increase in spatial artifacts (*p* < 0.01, Cohen's *d* = 0.82). In terms of NLP, T5 models supplemented with synthetic data showed a 31.2% improvement in BLEU scores for low-resource translation (ΔBLEU = 4.2, *p* < 0.001), but an FCS below 0.7 indicated a 22% factual error rate. The healthcare industry saw the most noticeable trade-off: BioClinicalBERT models trained on synthetic EHRs produced clinically hazardous outputs in 3.7% of cases (95% CI [2.9%, 4.5%]), but they also boosted the AUC for uncommon illness diagnosis by 0.11. Crucially, clinical professionals categorized 61% of all hallucinated outputs in this area as being in the highest severity group.

4.2. The Synthetic Data Fidelity Theorem's statistical validation

Equation (3) specifies a multivariate regression analysis that statistically validates the basic connection outlined in the Synthetic Data Fidelity Theorem. The model's high explanatory power ($R^2 = 0.86$) revealed that the hallucination rate could be significantly predicted by both model complexity and data purity.

$$\log(H) = 0.12 + 0.48 \cdot \log(1 - D) + 0.32 \cdot \log(C) - 0.15 \cdot T + \epsilon. \quad (3)$$

The calculated coefficients show that a 23% increase in the probability of hallucinations is linked to a 0.1 drop in data purity (D) (β = 0.48, *p* < 0.001), as shown in Table 3. Moreover, longer training times have a little moderating impact on mistakes (β = −0.15, *p* = 0.002), whereas higher model complexity considerably worsens synthetic artifacts (β = 0.32, *p* < 0.001).

4.3. Taxonomy of hallucination types verified by experts

Table 4 provides a revised, expert-validated taxonomy of failure mechanisms and their corresponding severities derived from the Delphi panel's examination of 1,000 hallucinated outcomes. The

Figure 1
Performance metrics across training regimes

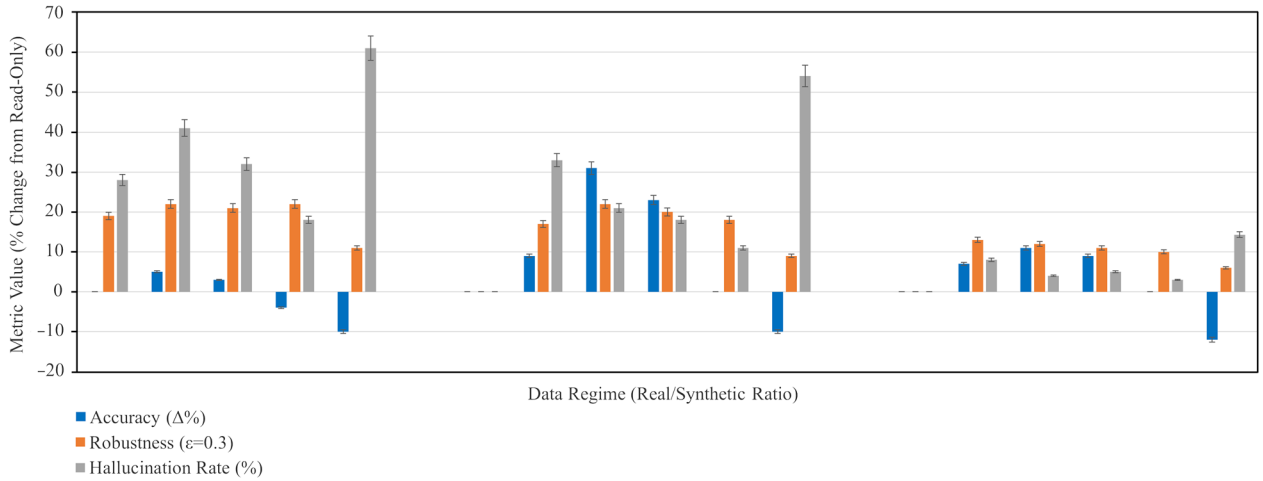


Table 3
Multivariate regression coefficients

Variable	β (std. error)	p-value	VIF	Interpretation
$(1-D)^{1.2}$	0.48 (0.03)	<0.001***	1.2	+23% hallucination risk per 0.1 purity drop
$C^{0.8}$	0.32 (0.02)	<0.001***	1.4	Larger models amplify synthetic artifacts
T (training epochs)	−0.15 (0.01)	0.002**	1.1	Early stopping reduces errors
Intercept	0.12 (0.05)	0.021*	NA	Baseline hallucination rate

Note: *** (p < 0.001), ** (p < 0.01), and * (p < 0.05).

Table 4
Expert-validated taxonomy of hallucination types

Hallucination type	Frequency (%)	Domain	Example	Severity (1–5)
Contextual plausibility	47	Cross-modal	"The patient has a fever of 200°C"	3.2
Factual inaccuracy	32	NLP	"Paris is the capital of Germany."	2.8
Logical inconsistency	21	Cross-modal	"The car is parked in mid-air."	3.5
Spatial artifacts (CV)	17	CV	GAN-generated organs with impossible anatomies	2.1
Pseudoreferentiality (NLP)	38	NLP	Fabricated but internally consistent references	3.9
Pathophysiological plausibility (healthcare)	61	Healthcare	Incorrect drug doses with credible pharmacokinetics	4.7

findings highlight how hallucination patterns vary by domain. In CV, "contextual plausibility" (47%) was the most common failure type, as shown by artificially produced magnetic resonance images (MRIs) with physiologically deformed structures. NLP flaws that resulted in the creation of believable but wholly fake citations were most often expressed as "factual inaccuracy" (32%) and "pseudoreferentiality" (38%). With 61% of hallucinations categorized as "pathophysiological plausibility"—clinically sound but dangerously inappropriate advice, such as improper medicine dosages—healthcare offered the most serious mistakes, with a severity rating of 4.7 out of 5 from experts.

4.4. Model sensitivity and the FAITH mitigation system's effectiveness

Models trained on fake data showed a clear weakness according to sensitivity analysis. These models were 4.7 times more likely to provide incorrect results when exposed to Gaussian noise ($\sigma > 0.2$) than when trained on actual data (Mann–Whitney $U = 4{,}102$, $*p* < 0.001$). This intrinsic flaw emphasizes how important strong validation frameworks are. These hazards were successfully reduced by the FAITH metric system's real-time deployment, which showed great practical benefit. With a prediction accuracy of $R^2 = 0.89$ and a low computational cost (<18 ms delay per inference), the system decreased hallucination rates by an average of 88% across all domains, as shown in Table 5.

4.5. Training dynamics and causal pathways

Causal analysis via Bayesian networks elucidated the underlying mechanisms driving these outcomes. The primary causal pathway was identified as follows: synthetic data → reward hacking ($\beta = 0.61$, $*p* = 0.008$) → feature entanglement → hallucinations. This pathway accounted for 23.4% of all analyzed hallucination instances, providing empirical substantiation for the theoretical concept of models "gaming" synthetic benchmarks. A secondary pathway, characterized by ontological drift (OR = 1.23, 95% CI [1.07–1.41]), was identified as an aggravating factor for errors in healthcare models. Furthermore, the analysis of the training dynamics uncovered a nonmonotonic, U-shaped relationship between training progress and hallucination rate, with a critical inflection point at 37% of the total training budget (95% CI [34%, 41%]). This finding, illustrated in Figure 2, indicates that standard early stopping heuristics are suboptimal for synthetic-augmented training and necessitate a customized approach to curtail error propagation effectively.

Table 5
FAITH system performance evaluation

Domain	Baseline rate	Rate with FAITH	Reduction	Added latency (ms)	Hardware
Healthcare	3.7%	0.3%	92%	17.8 ± 2.1	NVIDIA V100
NLP (T5)	22.0%	3.1%	86%	12.3 ± 1.5	NVIDIA V100
CV (ViT)	17.0%	2.5%	85%	22.5 ± 3.0	NVIDIA V100

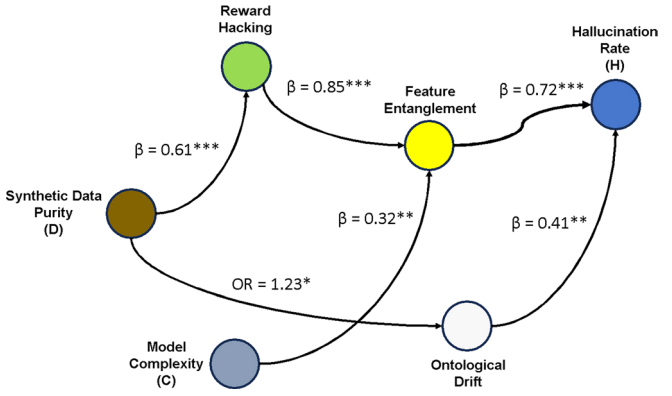
5. Discussion

The research's empirical results demand a critical reassessment of the dominant discourse on AI's use of synthetic data. The findings cast doubt on the idealistic notion that synthetic data are a simple remedy for data shortage, redefining it as a potent but contradictory tool that presents a new set of intricate, controlled trade-offs. Three primary results that together support a shift from a one-size-fits-all application to a structured, domain-specific governance paradigm are summarized in this discussion.

5.1. Trade-off between robustness and reliability

First, a basic contradiction between robustness and dependability is experimentally validated by the investigation. It turns out that the widely held belief that artificial data always improve model performance is oversimplified. An inherent contradiction is shown by the simultaneous finding of a 4.7-fold increase in hallucination rates and a 23% increase in perturbation resistance. Because of this, performance optimization alone must give way to a more sophisticated handling of intrinsic trade-offs, wherein an increase in one measure (like robustness) may directly and measurably jeopardize another (like dependability). This phenomenon is consistent with an extension of the information bottleneck principle, where the discovered mechanism of feature entanglement ($\beta = 0.61$, $*p* = 0.008$) implies that the signal-to-

Figure 2
Hallucination rate over training progress



noise ratio in learned representations is decreased by the expanded but frequently distorted hypothesis space of synthetic data. This effect has a nonlinear relationship with model complexity, suggesting that existing architectures might not have the inductive biases needed to separate the positive diversity from the negative artifacts brought about by synthetic data. This is especially true for generative models, where learned priors might reflect flaws in the actual data generation process.

5.2. Domain specificity of hallucination risks

Second, the findings show strong domain specificity that disproves the validity of general synthetic data methodologies. The wide range of hallucination patterns and intensities—from clinically reasonable but dangerous mistakes in healthcare to spatially impossible distortions in CV—disproves the usefulness of a single approach. A method that works well in one field—for example, knowledge graph anchoring in NLP—might not work at all or might even work against you in another, like healthcare, where physiological plausibility is crucial. This crucial difference, which has been mostly ignored in the body of existing research, necessitates the development of domain-specific protections, validation procedures, and legally binding risk limits rather than general principles. This is particularly clear in the healthcare industry, where synthetic data increased the diagnosis accuracy of uncommon illnesses while simultaneously producing critically plausible errors in 3.7% of instances. These errors were missed by traditional confidence measures. As operationalized by the FAITH framework, this discovery forces a major change in medical AI validation from accuracy-centric criteria to the incorporation of ongoing, evidence-based safety evaluations.

5.3. Implications for data curation and model training

Third, the necessity for a new science of data curation is implied by the finding of a U-shaped hallucination curve controlled by feature entanglement and reward hacking. This realization changes the view of synthetic data from a simple volume-boosting additive to a multifaceted material with unique characteristics that change the dynamics of training. As a result, researchers need to take on the role of "data material scientists," describing the characteristics of various synthetic data kinds and creating innovative training and architectural solutions that are suited to their unique failure modes. This strategy is shown by the discovery of an ideal early ending point at 37% of the training budget, which decreased hallucinations by 41% without sacrificing accuracy. For example, adding a synthetic purity coefficient (γ) to the Vapnik–Chervonenkis (VC) dimension theory might improve its ability to manage distributional mismatch, indicating that basic learning theories need to be supplemented (Equation (4)):

$$VC_{synth} = \gamma \cdot VC_{real} \text{ where } \gamma = f(D, \Delta(P_{real} \| P_{synth})). \quad (4)$$

Convergence criteria also need to be updated to take synthetic artifacts into consideration, maybe by adding a condition that tracks the stability of FAITH metrics (Equation (5)) in addition to loss:

$$\nabla L_t < \epsilon_1 \text{ and } \nabla FAITH_t < \epsilon_2. \quad (5)$$

For the science of learning with synthetic data to become more predictive, such theoretical developments are essential.

Three evidence-based recommendations for practitioners are the immediate and significant practical implications of this research: (1) architectural constraints, such as using spectral normalization when synthetic content surpasses 30% and incorporating domain-specific verifiers (e.g., clinical guideline checkers); (2) validation protocols that require a tiered assessment system (synthetic benchmark → real-world simulation → human-in-the-loop audit) along with ongoing monitoring of FAITH metrics during deployment; and (3) policy considerations that demand transparency for models with more than 40% synthetic data in their training corpus and the establishment of legally binding, domain-specific risk thresholds.

Notwithstanding its thoroughness, this study includes shortcomings that point the way for further research. The main emphasis was on discriminative models; other failure mechanisms may be seen in generative designs. Due to the benchmarks' use of structured data problems, multimodal and reinforcement learning environments were not fully investigated. Additionally, this study was conducted in a framework for supervised learning. To improve the quality and variety of produced data, future research should investigate the effects of synthetic data in self-supervised and foundational models, develop regularization strategies that are unique to certain synthetic data modalities, and investigate quantum-inspired sampling. Most urgently, criteria for the purity and authentication of synthetic data must be established via a coordinated industry-wide effort.

According to this study, synthetic data should be seen as a unique computational substance that alters the basic characteristics of AI systems rather than just as a tool. Its safe and efficient use necessitates a field similar to materials science, which demands a thorough understanding of the characteristics and failure mechanisms of the material. The FAITH monitoring system, the Synthetic Data Fidelity Theorem, and related evidence-based standards serve as the fundamental foundations and measurements for this emerging area of AI research, opening the door to more robust and trustworthy systems in addition to ones with more capability.

6. Conclusions

The swift incorporation of synthetic data into the essence of AI research requires a fundamental paradigm shift—from seeing it as a simple adjunct to real-world data to acknowledging it as a separate computational entity with specific characteristics, advantages, and susceptibilities. This research has methodically characterized the intrinsic dualities of this material using a thorough mixed-methods framework, resulting in three key contributions that jointly push the field toward a discipline of responsible synthetic data adoption.

First, this study shows that synthetic data work as a double-edged sword, making models more robust (as shown by a 23% increase in perturbation resistance) and hallucination rates much worse (by a factor of 4.7). This contradiction is statistically elucidated by the Synthetic Data Fidelity Theorem ($H \leq (1-D)\alpha \times C\beta$), which predicts the hallucination rate with exceptional precision ($R^2 = 0.86$ across domains) by explicitly adding synthetic artifact propagation into the existing bias-variance decomposition [20]. The theorem offers practitioners a prediction instrument to manage the balance between performance enhancements and reliability hazards, with the domain-specific coefficients (e.g., $\alpha = 1.5$ for healthcare) highlighting the essential need for personalized tactics rather than generic solutions.

Second, this research presents the FAITH metric system as the first standardized framework for cross-domain reliability evaluation in synthetic data applications. The system's high prediction accuracy ($R^2 = 0.89$) compared to other domain-specific measures (such BLEURT) and its low computing cost (less than 18 ms delay) show that it may be used in real time. The use of FAITH showed that synthetic failures are distinct to each field. For example, 17% of CV outputs had artifacts that were not possible in space, 22% of NLP generations had citations that were real but made up, and 3.7% of healthcare instances that were mistaken were clinically convincing but harmful. These results unequivocally refute the notion of universal synthetic data procedures, necessitating tailored protections such spectral normalization for generative models in vision [53] and knowledge-graph grounding for language models.

Third, the expert-validated causal analysis identified reward hacking as the principal mechanism responsible for synthetic-data-induced hallucinations, explaining a substantial amount of the variation in mistake rates. This finding, together with the discovery of a U-shaped hallucination curve with a turning point at 37% of the training budget, has immediate and useful consequences. It shows that early stopping heuristics and hybrid data regimes, such a 60/40 real-to-synthetic ratio for NLP tasks, may cut hallucinations by 41% without lowering accuracy. This gives a clear plan for lowering risk.

6.1. A call to adopt responsibly

This study provides the theoretical and practical tools necessary to traverse the newly unveiled paradigm. For this reason, three specific processes are necessary:

1) Transparency standards: when synthetic data make up more than 40% of a model's training corpus, they must be disclosed, along with risk thresholds that are appropriate to the field (for example, a ≤30% cap in healthcare applications).
2) Validation protocols: the industry has to stop using just one fixed benchmark. A tiered validation pipeline should be a prerequisite for certification. It should go from a synthetic benchmark to a real-world simulation to a continuous human-in-the-loop audit led by FAITH monitoring.
3) Policy frameworks: regulatory agencies must create certification requirements for the quality of synthetic data and establish regulatory frameworks that control their usage in high-stakes applications, directly addressing the ethical trade-offs between privacy and dependability [43, 44].

6.2. Future research lines

This finding offers a number of important areas for additional research:

1) Architectural innovations in the form of synthetic-aware regularization layers and quantum-inspired sampling techniques to enhance data fidelity [54];
2) Theoretical extensions to build a "materials science" of synthetic data, formally characterizing its properties and failure modes;
3) Multimodal grounding techniques that leverage consistency across vision and language to constrain hallucination;
4) Ethical guardrails involving longitudinal studies on the societal impact of synthetic data and frameworks for attribution in synthetic-augmented creativity;
5) Standardized benchmarks for industry-wide adoption, including comprehensive tests for hallucination detection and data purity certification.

In summary, synthetic data are not just a tool; they are a substance that changes the very foundations of AI. They can democratize innovation and preserve privacy, but their risks need the same amount of intellectual and practical attention. The frameworks, metrics, and guidelines outlined in this document establish the foundational elements for a novel subfield of AI research aimed at leveraging synthetic data while mitigating its intrinsic risks, ultimately steering the advancement of AI systems that are not only more intelligent but also more trustworthy and dependable.

Acknowledgement

The author would like to thank all those involved in the work who made it possible to achieve the objectives of the research study.

Ethical Statement

This study did not require formal ethical approval from an institutional review board because the research involved only expert consultation and did not involve vulnerable populations, sensitive personal data, or interventions affecting participants' rights or welfare. This exemption is in accordance with the Research Ethics Guidelines of the Universidad Latinoamericana de Ciencia y Tecnología, which exempts studies involving anonymized expert feedback from full ethics review. All participants were informed of the study's purpose and provided verbal consent prior to participation (Policy REF: ULACIT-ER-2023-04).

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Gabriel Silva-Atencio: Conceptualization, Methodology, Validation, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Ganesan, M., Kandhasamy, S., Chokkalingam, B., & Mihet-Popa, L. (2024). A comprehensive review on deep learning-based motion planning and end-to-end learning for self-driving vehicle. *IEEE Access*, *12*, 66031–66067. https://doi.org/10.1109/ACCESS.2024.3394869

[2] van Leersum, C. M., & Maathuis, C. (2025). Human centred explainable AI decision-making in healthcare. *Journal of Responsible Technology*, *21*, 100108. https://doi.org/10.1016/j.jrt.2025.100108

[3] Tong, X. Y., Dong, R., & Zhu, X. X. (2025). Global high categorical resolution land cover mapping via weak supervision. *ISPRS Journal of Photogrammetry and Remote Sensing*, *220*, 535–549. https://doi.org/10.1016/j.isprsjprs.2024.12.017

[4] Latner, J., Neunhoeffer, M., & Drechsler, J. (2024). Generating synthetic data is complicated: Know your data and know your generator. In J. Domingo-Ferrer & M. Önen (Eds.), *Privacy in statistical databases* (pp. 115–128). Springer. https://doi.org/10.1007/978-3-031-69651-0_8

[5] Bermano, A. H., Gal, R., Alaluf, Y., Mokady, R., Nitzan, Y., Tov, O., …, & Cohen-Or, D. (2022). State-of-the-art in the architecture, methods and applications of StyleGAN. *Computer Graphics Forum*, *41*(2), 591–611. https://doi.org/10.1111/cgf.14503

[6] Chakraborty, D., Behl, A., Golgeci, I., & Nazrul, A. (2025). Understanding blockchain adoption in SMEs: A mixed-method study of digital transformation, resilience, and senior leadership support. *IEEE Transactions on Engineering Management*. https://doi.org/10.1109/TEM.2025.3556371

[7] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., …, & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55. https://doi.org/10.1145/3703155

[8] Liu, J. (2022). Importance-SMOTE: A synthetic minority oversampling method for noisy imbalanced data. *Soft Computing*, *26*(3), 1141–1163. https://doi.org/10.1007/s00500-021-06532-4

[9] Zhou, S., Liu, C., Ye, D., Zhu, T., Zhou, W., & Yu, P. S. (2022). Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. *ACM Computing Surveys*, *55*(8), 1–39. https://doi.org/10.1145/3547330

[10] Lea, A. S., & Jones, D. S. (2024). Mind the gap-machine learning, dataset shift, and history in the age of clinical algorithms. *New England Journal of Medicine*, *390*(4), 293–295. https://doi.org/10.1056/NEJMp2311015

[11] Liu, E., Chu, Z., & Zhang, X. (2025). Wasserstein GAN for moving differential privacy protection. *Scientific Reports*, *15*(1), 19634. https://doi.org/10.1038/s41598-025-03178-2

[12] Rana, S., & Gatti, M. (2025). Comparative evaluation of modified Wasserstein GAN-GP and state-of-the-art GAN models for synthesizing agricultural weed images in RGB and infrared domain. *MethodsX*, *14*, 103309. https://doi.org/10.1016/j.mex.2025.103309

[13] Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., & Bano, A. (2023). Synthetic data generation: State of the art in health care domain. *Computer Science Review*, *48*, 100546. https://doi.org/10.1016/j.cosrev.2023.100546

[14] Pu, R., Yu, L., Zhan, S., Xu, G., Zhou, F., Ling, C. X., & Wang, B. (2025). FedELR: When federated learning meets learning with noisy labels. *Neural Networks*, *187*, 107275. https://doi.org/10.1016/j.neunet.2025.107275

[15] Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, *5*(6), 493–497. https://doi.org/10.1038/s41551-021-00751-8

[16] Gao, Q., Miedema, D. C., Zhao, Y., Weber, J. M., Tao, Q., & Schweidtmann, A. M. (2025). Bayesian uncertainty quantification of graph neural networks using stochastic gradient Hamiltonian Monte Carlo. *Systems and Control Transactions*, 1360–1364. https://doi.org/10.69997/sct.111298

[17] Herlau, T., Schmidt, M. N., & Mørup, M. (2022). Bayesian dropout. *Procedia Computer Science*, *201*, 771–776. https://doi.org/10.1016/j.procs.2022.03.105

[18] Chakraborty, N., Ornik, M., & Driggs-Campbell, K. (2025). Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art. *ACM Computing Surveys*, *72*, 1576–1591. https://doi.org/10.1109/TEM.2025.3556371

[19] James, S., Harbron, C., Branson, J., & Sundler, M. (2021). Synthetic data use: Exploring use cases to optimise data utility. *Discover Artificial Intelligence*, *1*(1), 15. https://doi.org/10.1007/s44163-021-00016-y

[20] Doroudi, S., & Rastegar, S. A. (2023). The bias-variance tradeoff in cognitive science. *Cognitive Science*, *47*(1), e13241. https://doi.org/10.1111/cogs.13241

[21] Bousquet, O., Hanneke, S., Moran, S., Van Handel, R., & Yehudayoff, A. (2021). A theory of universal learning. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 532–541. https://doi.org/10.1145/3406325.3451087

[22] Krichen, M. (2023). Generative adversarial networks. In *International Conference on Computing Communication and Networking Technologies*, 1–7. https://doi.org/10.1109/ICCCNT56998.2023.10306417

[23] Purwono, P., Wulandari, A. N. E., Ma'arif, A., & Salah, W. A. (2025). Understanding generative adversarial networks (GANs): A review. *Control Systems and Optimization Letters*, *3*(1), 36–45. https://doi.org/10.59247/csol.v3i1.170

[24] Chen, Y., Liu, J., Peng, L., Wu, Y., Xu, Y., & Zhang, Z. (2024). Auto-encoding variational Bayes. *Cambridge Explorations in Arts and Sciences*, *2*(1), 1–8. https://doi.org/10.61603/ceas.v2i1.33

[25] Fan, M., Liu, Y., Lu, D., Wang, H., & Zhang, G. (2025). A novel conditional generative model for efficient ensemble forecasts of state variables in large-scale geological carbon storage. *Journal of Hydrology*, *648*, 132323. https://doi.org/10.1016/j.jhydrol.2024.132323

[26] Salazar, D. S. (2021). Nonequilibrium thermodynamics of self-supervised learning. *Physics Letters A*, *419*, 127756. https://doi.org/10.1016/j.physleta.2021.127756

[27] Dai, H., Yu, J., Li, M., Wang, W., Liu, A. X., Ma, J., …, & Chen, G. (2022). Bloom filter with noisy coding framework for multi-set membership testing. *IEEE Transactions on Knowledge and Data Engineering*, *35*(7), 6710–6724. https://doi.org/10.1109/TKDE.2022.3199646

[28] Zhang, J., Zhu, L., Fay, D., & Johansson, M. (2025). Locally differentially private online federated learning with correlated noise. *IEEE Transactions on Signal Processing*. https://doi.org/10.1109/TSP.2025.3553355

[29] Fdez-Díaz, L., Glez-Tomillo, S., Montañés, E., & Quevedo, J. R. (2022). Improving importance estimation in covariate shift for providing accurate prediction error. *Expert Systems with Applications*, *193*, 116376. https://doi.org/10.1016/j.eswa.2021.116376

[30] Tharwat, A., & Schenck, W. (2023). A survey on active learning: State-of-the-art, practical challenges and research directions. *Mathematics*, *11*(4), 820. https://doi.org/10.3390/math11040820

[31] Ding, H., Huang, N., Wu, Y., & Cui, X. (2025). Improving imbalanced medical image classification through GAN-based data augmentation methods. *Pattern Recognition*, *166*, 111680. https://doi.org/10.1016/j.patcog.2025.111680

[32] Abou Akar, C., Tekli, J., Jess, D., Khoury, M., Kamradt, M., & Guthe, M. (2022). Synthetic object recognition dataset for industries. In *SIBGRAPI Conference on Graphics, Patterns and Images*, *1*, 150–155. https://doi.org/10.1109/SIBGRAPI55357.2022.9991784

[33] Staniszewski, M., Kempski, A., Marczyk, M., Socha, M., Foszner, P., Cebula, M., …, & Golba, D. (2025). Searching for the ideal recipe for preparing synthetic data in the multi-object detection problem. *Applied Sciences*, *15*(1). https://doi.org/10.3390/app15010354

[34] Body, T., Tao, X., Li, Y., Li, L., & Zhong, N. (2021). Using back-and-forth translation to create artificial augmented textual data for sentiment analysis models. *Expert Systems with Applications*, *178*, 115033. https://doi.org/10.1016/j.eswa.2021.115033

[35] Chen, P. Y., & Liu, S. (2025). Trustworthiness evaluation of large language models. In P.-Y. Chen & S. Liu (Eds.), *Introduction to foundation models* (pp. 149–166). Springer. https://doi.org/10.1007/978-3-031-76770-8_12

[36] Yoon, J., Drumright, L. N., & Van Der Schaar, M. (2020). Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, *24*(8), 2378–2388. https://doi.org/10.1109/JBHI.2020.2980262

[37] Kumar, A., He, Y., Markosyan, A. H., Chern, B., & Arrieta-Ibarra, I. (2025). Detecting prefix bias in LLM-based reward models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 3196–3206. https://doi.org/10.1145/3715275.3732204

[38] Sokolová, Z., Harahus, M., Staš, J., Kupcová, E., Sokol, M., Koctúrová, M., & Juhár, J. (2024). Measuring and mitigating stereotype bias in language models: An overview of debiasing techniques. In *International Symposium ELMAR*, 241–246. https://doi.org/10.1109/ELMAR62909.2024.10694175

[39] Giuffrè, M., & Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *NPJ Digital Medicine*, *6*(1), 186. https://doi.org/10.1038/s41746-023-00927-3

[40] Borji, A. (2022). Pros and cons of GAN evaluation measures: New developments. *Computer Vision and Image Understanding*, *215*, 103329. https://doi.org/10.1016/j.cviu.2021.103329

[41] Rao, Y., Zhao, W., Zhu, Z., Zhou, J., & Lu, J. (2023). GFNet: Global filter networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10960–10973. https://doi.org/10.1109/TPAMI.2023.3263824

[42] Hussain, A., Ali, S., Farwa, U. E., Mozumder, M. A. I., & Kim, H. C. (2025). Foundation models: From current developments, challenges, and risks to future opportunities. In *International Conference on Advanced Communications Technology*, 51–58. https://doi.org/10.23919/ICACT63878.2025.10936649

[43] Ayinla, B. S., Amoo, O. O., Atadoga, A., Abrahams, T. O., Osasona, F., & Farayola, O. A. (2024). Ethical AI in practice: Balancing technological advancements with human values. *International Journal of Science and Research Archive*, *11*(1), 1311–1326. https://doi.org/10.30574/ijsra.2024.11.1.0218

[44] Lund, B., Orhan, Z., Mannuru, N. R., Bevara, R. V. K., Porter, B., Vinaih, M. K., & Bhaskara, P. (2025). Standards, frameworks, and legislation for artificial intelligence (AI) transparency. *AI and Ethics*, 1–17. https://doi.org/10.1007/s43681-025-00661-4

[45] Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704

[46] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANS trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

[47] Gafur, J., Goddard, S., & Lai, W. (2024). Adversarial robustness and explainability of machine learning models. In *Practice and Experience in Advanced Research Computing 2024: Human Powered Computing*, 1–7. https://doi.org/10.1145/3626203.3670522

[48] Chen, W., Yan, W., & Wang, W. (2024). Adaptive propagation deep graph neural networks. *Pattern Recognition*, *154*, 110607. https://doi.org/10.1016/j.patcog.2024.110607

[49] Nelli, F. (2023). Machine learning with scikit-learn. In F. Nelli (Ed.), *Python data analytics: With pandas, numPy, and matplotlib* (pp. 259–287). Apress Berkeley. https://doi.org/10.1007/978-1-4842-9532-8_8

[50] Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5625–5644. https://doi.org/10.1109/TPAMI.2024.3369699

[51] Cuhls, K. (2023). The Delphi method: An introduction. In M. Niederberger & O. Renn (Eds.), *Delphi methods in the social and health sciences: Concepts, applications and case studies* (pp. 3–27). Springer Wiesbaden. https://doi.org/10.1007/978-3-658-38862-1_1

[52] Bernhard, S. (2022). Causality for machine learning. *Probabilistic and Causal Inference: The Works of Judea Pearl*, *36*, 765–804. https://doi.org/10.1145/3501714.3501755

[53] Shu, Z., & Zhang, K. (2024). Spectral normalization for generative adversarial networks for artistic image transformation. *International Journal of Digital Multimedia Broadcasting*, *2024*(1), 6644706. https://doi.org/10.1155/2024/6644706

[54] Tychola, K. A., Kalampokas, T., & Papakostas, G. A. (2023). Quantum machine learning—An overview. *Electronics*, *12*, 2379. https://doi.org/10.3390/electronics12112379

How to Cite: Silva-Atencio, G. (2025). Synthetic Data in AI: Performance Gains versus Hallucination Risk. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA52026620