# SynEval: A Multidimensional Framework for Evaluating Synthetic Data on Fidelity, Utility, Diversity, and Privacy

Yefeng Yuan
*Santa Clara University*
Santa Clara, USA
yyuan4@scu.edu

Liang Cheng
*eBay Inc.*
San Jose, USA
liacheng@ebay.com

Zhan Shi
*Santa Clara University*
Santa Clara, USA
ashi2@scu.edu

Tevin Atwal
*Santa Clara University*
Santa Clara, USA
tatwal@scu.edu

Chan Nam Tieu
*Santa Clara University*
Santa Clara, USA
ctieu@scu.edu

Yuhong Liu*
*Santa Clara University*
Santa Clara, USA
yhliu@scu.edu

***Abstract*—Synthetic data is increasingly recognized as a practical solution for privacy-sensitive and data-scarce domains, where reliance on real data is limited by cost, imbalance, or regulatory constraints. Large language models (LLMs) and other generative approaches have shown strong potential for producing such data. Yet, despite this promise, the community still lacks a comprehensive, standardized framework for evaluating the quality and safety of multimodal synthetic datasets. We present SynEval[1], an open-source, metadata-driven evaluation framework that systematically assesses synthetic data across four critical dimensions: fidelity, utility, diversity, and privacy, through an extensive suite of quantitative metrics spanning both structured and unstructured fields. To substantiate cross-domain generalizability, we validate SynEval across three distinct, high-stakes domains: e-commerce (Amazon Reviews), healthcare (Medical Transcriptions), and finance (CFPB Consumer Complaints) using state-of-the-art LLMs (ChatGPT, Claude, Lilium). Crucially, SynEval reveals non-obvious phenomena, such as the "predictability trap"—where poor-quality generators artificially inflate downstream utility scores—and cross-modal illusions that isolated metrics fail to catch. By supplementing baseline privacy scores with empirical adversarial models (e.g., Membership Inference Attacks), SynEval provides actionable, robust diagnostics for researchers and practitioners, making it a practical and extensible tool for advancing responsible synthetic data deployment.**

***Index Terms*—Synthetic data, Evaluation framework, Large language models, Privacy, Diversity, Utility, Fidelity**

## I. Introduction

While high-quality data drives recent AI advances, its collection and labeling remain costly and slow, and its sharing and reuse are increasingly limited by strict privacy regulations [1]. Synthetic data offers a scalable and controllable substitute for real data, capable of preserving essential statistical properties while supporting targeted utility [2], [3]. Recent advances in generative modeling, from GANs [4] and VAEs [5] to diffusion models [6] and LLMs [7], have substantially improved the realism and feasibility of synthetic data, leading to their accelerated adoptions. For example, Gartner predicts that synthetic data will underpin the majority of AI model development by 2030, and 89% of technology executives identify it as critical to maintaining competitiveness [8].

Despite this momentum, evaluation remains a bottleneck. First, existing frameworks typically assess only one or two dimensions (e.g., fidelity or privacy) rather than integrating fidelity, utility, diversity, and privacy within a unified evaluation, offering limited guidance regarding the quantitative tradeoff among different dimensions. Second, existing evaluation often overlook complex synthetic data with heterogeneous data types, such as categorical, numerical, temporal, and free-text fields. As modern datasets increasingly exhibit tight dependencies between different data types, such as text (e.g., reviews, notes) and structured attributes (e.g., ratings, IDs, timestamps), the ignorance of cross-type correlations fails to reflect the realistically of synthetic data. Third, limited attention has been paid to evaluating synthetic data diversity [9], even though biased or under-represented subpopulations can adversely affect downstream tasks such as anomaly and fraud detection, amplifying distributional skew and increasing false positives and negatives. Recent work on bias-aware generation is a step forward [10], but a comprehensive, multi-dimensional evaluation that jointly considers fidelity, diversity, privacy, and downstream utility remains largely absent.

To address these gaps, this paper presents SynEval, an open-source, metadata-driven, and command-line configurable framework for evaluating synthetic tabular datasets across fidelity, utility, diversity, and privacy. We focus on tabular data because its inherently heterogeneous structure, including categorical, numerical, temporal, and often free-text attributes, creates complex inter-field dependencies that make both generation and evaluation uniquely challenging. Specifically, SynEval implements more than 150 quantitative metrics spanning both structured and unstructured fields and is

*Corresponding author

[1] [https://github.com/privacy-enhancing-technologies/SynEval)

designed to be resource-aware for edge settings. The major contributions of this work are as follows:

- **Unified evaluation framework.** We introduce SynEval, an open-source, configurable system that consolidates synthetic tabular data evaluation across four quality dimensions into a single, standardized framework to facilitate flexible evaluation on tradeoff among different dimensions.
- **Heterogeneous data evaluation.** SynEval jointly evaluates structured tabular attributes and unstructured text, capturing dependencies across modalities that existing frameworks typically ignore.
- **Diversity module.** We propose a dedicated diversity module for synthetic tabular data that separates *what is present* (support coverage) from *how it is distributed* (distributional similarity) for tabular data, with structure-aware signals for text.
- **Privacy suite.** We propose a layered, modality-aware privacy suite that consolidates leakage detection, re-identifiability, and authorship risk into baseline-normalized scores, making privacy–utility trade-offs explicit.

By filling the evaluation gaps, SynEval provides a standardized and extensible toolkit that enables researchers and practitioners to rigorously assess the trustworthiness of synthetic data. The framework supports responsible deployment across sensitive domains such as e-commerce, healthcare, and finance, where data quality and privacy are critical.

## II. Experiments and Key Findings

To ensure broad empirical validity and address the limitations of single-domain testing, we evaluated SynEval across three highly regulated, multimodal datasets: Amazon Product Reviews (e-commerce), Kaggle Medical Transcriptions (healthcare), and the CFPB Consumer Complaint Database (finance). We synthesized data across these domains using three LLMs: Claude 3.7, ChatGPT-4o, and Lilium-8B. Our evaluation yielded several counter-intuitive findings:

*a) The Predictability Trap (Utility vs. Realism).:* Prior assumptions often suggest that higher downstream utility correlates with higher data quality. SynEval exposes this as a fallacy. In our classification tasks, Lilium-8B achieved the highest downstream utility (Accuracy 0.91), vastly outperforming the original human dataset (Accuracy 0.55). However, SynEval's fidelity metrics revealed that this high utility was actually a symptom of severe mode collapse: the model generated overly brief, rigidly templated text that artificially perfectly aligned with tabular labels. SynEval demonstrates that near-perfect downstream predictability is often a red flag for a lack of natural human variance.

*b) The Cross-Modal Illusion.:* Across the medical and financial datasets, isolated NLP metrics (like BERTScore) and standard tabular evaluators assigned high fidelity scores to generated records. However, SynEval's joint text-tabular alignment metrics uncovered severe semantic contradictions—such as a CFPB complaint narrative describing a "mortgage foreclosure" paired with a tabular product category of "student loan." SynEval proved that high independent modal quality does not guarantee joint dataset coherence.

*c) Adversarial Privacy vs. Diversity.:* Moving beyond intuitive baseline metrics, we subjected the generated data to Membership Inference Attacks (MIA) to simulate realistic adversaries. When we applied evaluation-guided prompt optimization to successfully increase semantic diversity (expanding minimum spanning tree edge length by $\sim 12\times$), SynEval's adversarial suite detected a disproportionate spike in vulnerability. The richer semantic generation inadvertently hallucinated highly specific nominals and entities, increasing the MIA success rate. This empirically proves that diversity optimization directly scales real-world adversarial risk, requiring explicit PII redaction pipelines.

## III. Conclusion

SynEval is a multi-dimensional framework implementing 156 metrics to assess heterogeneous synthetic data. By validating the framework across e-commerce, healthcare, and financial datasets, our experiments demonstrate its robust cross-domain applicability. SynEval successfully moves beyond surface-level evaluation to expose counter-intuitive phenomena—such as the predictability trap, cross-modal logical illusions, and the exact adversarial cost of optimizing for diversity. By integrating empirical adversarial proxies like Membership Inference Attacks alongside standard statistical measures, SynEval provides a comprehensive, plug-and-play solution for trustworthy synthetic data assessment in high-stakes, regulated environments.

## References


[1] Data Privacy Manager, "Meta hit with record €1.2b gdpr fine – data privacy manager," https://dataprivacymanager.net/meta-hit-with-record-e1-2b-gdpr-fine/, 2023, accessed: 2024-04-20.

[2] N. Savage, "Synthetic data could be better than real data," Nature, 2023, published in April.

[3] C. Dilmegani, "Synthetic data vs real data: Benefits, challenges in 2023," https://research.aimultiple.com/synthetic-data-vs-real-data/, 2023, accessed: 2024-04-20.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[5] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[6] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. pmlr, 2015, pp. 2256–2265.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] VentureBeat, "89% of tech execs see synthetic data as a key to staying ahead," https://venturebeat.com/ai/89-of-tech-execs-see-synthetic-data-as-a-key-to-staying-ahead, 2021, accessed: 2024-04-20.

[9] A. D. Lautrup, T. Hyrup, A. Zimek, and P. Schneider-Kamp, "Syntheval: a framework for detailed utility and privacy evaluation of tabular synthetic data," *Data Mining and Knowledge Discovery*, vol. 39, no. 1, p. 6, 2025.

[10] E. Barbierato, M. L. D. Vedova, D. Tessera, D. Toti, and N. Vanoli, "A methodology for controlling bias and fairness in synthetic data generation," *Applied Sciences*, vol. 12, no. 9, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/9/4619

# SynEval: A Multidimensional Framework for Evaluating Synthetic Data on Fidelity, Utility, Diversity, and Privacy


Yefeng Yuan∗, Zhan Shi∗, Liang Cheng†, Tevin Atwal∗, Chan Nam Tieu∗, Yuhong Liu∗

∗Santa Clara University, Santa Clara, CA, USA
†eBay Inc., San Jose, USA


## Problem Statement

Generative AI is driving the adoption of multimodal synthetic datasets (structured tables + unstructured text) in high-stakes domains like healthcare and finance. However, our ability to generate this complex data has outpaced our ability to safely and rigorously evaluate it.

## The Gap: The "Stitching Fallacy"

Current evaluation frameworks are trapped in modular isolation:

- **Tabular isolation:** Tools like SDV excel at numerical data but cannot process natural language.
- **Text isolation:** NLP metrics (BLEU, BERTScore) evaluate text in a vacuum, blind to tabular parameters.
- **The Result:** Isolated evaluators miss critical cross-modal logical contradictions (e.g., a tabular diagnosis of "Diabetes" paired with a text summary of a "Fractured femur").

## Proposed Scheme

**SynEval** is an open-source framework that explicitly models the joint distribution of multimodal synthetic data.

( 💻 github.com/privacy-enhancing-technologies/SynEval )

- **Mathematical Unification:** Maps discrete tabular features and continuous text embeddings into a singular, joint geometric space.
- **Overcoming Dimensionality:** Leverages Probabilistic Graphical Models (PGMs) to marginalize the intractable global joint distribution into highly tractable, low-dimensional clique marginals.
- **Active Optimization:** Outputs rigorous scalar metrics that can serve as reward functions (RLAIF) to help generative models dynamically self-correct cross-modal hallucinations.

## Core Contributions

- **Unified Evaluation Framework:** Consolidates assessment across four critical quality dimensions into a single system featuring **156 quantitative metrics**.
- **Heterogeneous Data Evaluation:** Jointly evaluates structured attributes and unstructured text, capturing cross-modal dependencies that existing frameworks ignore.
- **Diversity Module:** Separates support coverage (what is present) from distributional similarity (how it is distributed) for tabular data, adding structure-aware signals for text.
- **Layered Privacy Suite:** A modality-aware suite that consolidates leakage detection, re-identifiability, and authorship risk (stylistic outliers) into baseline-normalized scores, exposing explicit privacy–utility trade-offs.

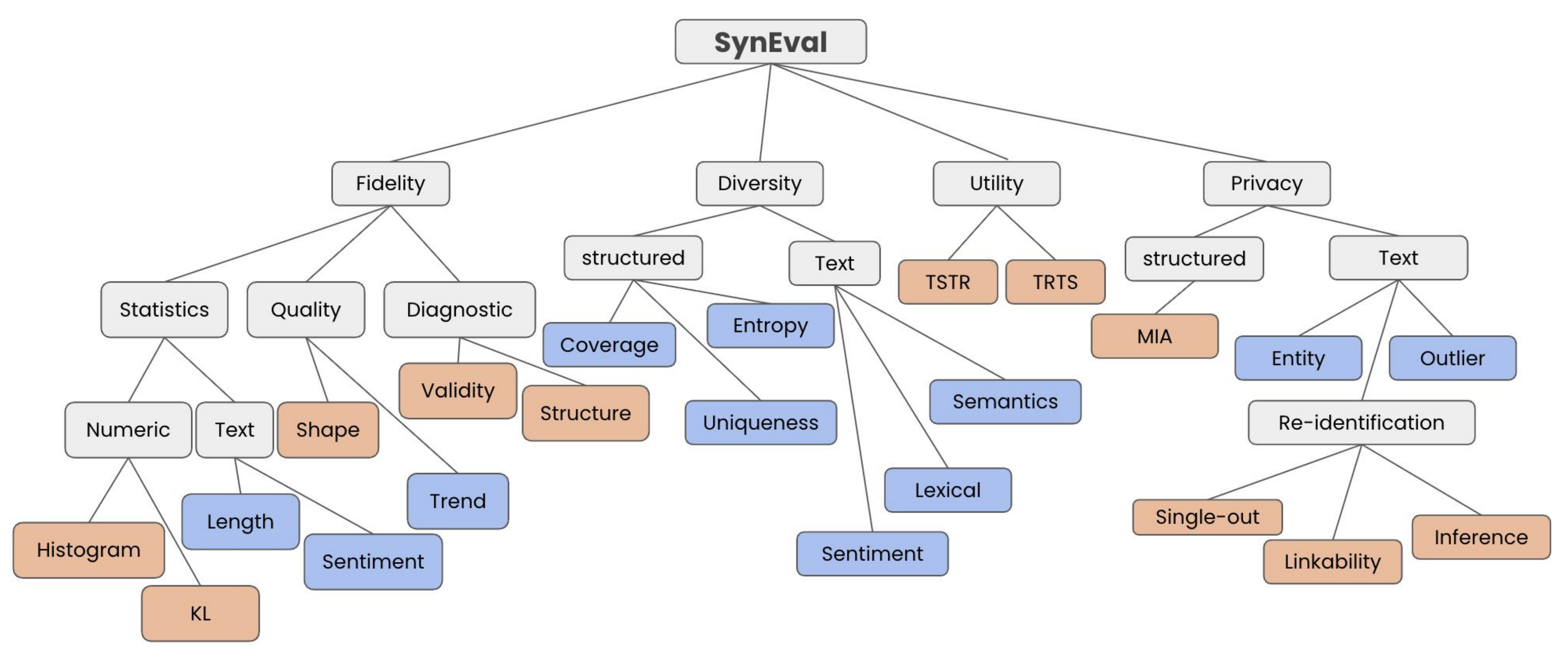


**Figure 1: SynEval Architecture**

This figure maps our pipeline, illustrating how SynEval integrates standard baselines (orange blocks) with our novel, multimodal joint-evaluation metrics (blue blocks) to create a comprehensive diagnostic tool.

## Text Fidelity Evaluation

TABLE I: Text fidelity summary by dataset. Sentiment shown as *neg/neu/pos* in percent.

| Data Source | Avg. Length (words) | Sentiment (neg/neu/pos %) | Top 3 Keywords | Top 3 Positive Words (+) | Top 3 Negative Words (-) |
|---|---|---|---|---|---|
| Original | 23.76 | 9.0 / 23.0 / 68.0 | love, dress, like | good, beautiful, great | return, product, like |
| Lilium-8B Synthetic | 5.75 | 10.6 / 22.5 / 66.9 | perfect, excellent, awesome | perfect, excellent, awesome | terrible, horrible, material |
| Claude 3.7 Synthetic | 25.12 | 12.8 / 24.8 / 62.3 | quality, perfect, just | perfect, beautifully, excellent | terrible, disappointed, money |
| ChatGPT-4o Synthetic | 10.61 | 18.0 / 21.0 / 61.0 | quality, broke, love | perfect, great, amazing | disappointed, horrible, terrible |

**Key observations:** Claude 3.7 successfully mirrors human sentiment and verbosity. In contrast, Lilium-8B suffers from severe mode collapse, generating unnaturally short (**5.75** words), and the top 3 keywords are the same as top 3 positive words.

## Utility Trade-offs (TSTR)

TABLE II: Classification results predicting *rating* from *review* sentiment for each dataset.

| Dataset | Accuracy | Macro-F1 | Weighted-F1 |
|---|---|---|---|
| Original | 0.5550 | 0.2608 | 0.4986 |
| Lilium-8B Synthetic | 0.9126 | 0.8691 | 0.9114 |
| ChatGPT-4o Synthetic | 0.5517 | 0.4929 | 0.5442 |
| Claude 3.7 Synthetic | 0.4400 | 0.4092 | 0.4252 |

**Key observations:** High performance doesn't equal realism. Lilium-8B's near-perfect accuracy (**0.9126**) indicates an overly deterministic, robotic generation. ChatGPT-4o (**0.5517**) successfully replicates the messy variability of the Original human data (**0.5550**).

## Evaluation-Guided Optimization

TABLE III: Evaluation metrics comparing model generations and original data.

| Evaluation Metrics | Original Data | Claude L1 | Claude L2 | ChatGPT L1 | ChatGPT L2 |
|---|---|---|---|---|---|
| Unigram ($L_r/H_n$) | 0.0042 / 0.6176 | 0.1672 / 0.8474 | 0.1480 / 0.8847 | 0.1419 / 0.8653 | 0.2386 / 0.8813 |
| Bigram | 0.1696 / 0.8419 | 0.5184 / 0.9439 | 0.6657 / 0.9692 | 0.6768 / 0.9721 | 0.7942 / 0.9786 |
| Trigram | 0.6588 / 0.9619 | 0.7114 / 0.9780 | 0.8739 / 0.9904 | 0.9335 / 0.9957 | 0.9592 / 0.9967 |
| Quadrigram | 0.9133 / 0.9937 | 0.7939 / 0.9876 | 0.9488 / 0.9970 | 0.9857 / 0.9993 | 0.9929 / 0.9996 |
| Pentagram | 0.9743 / 0.9987 | 0.8372 / 0.9915 | 0.9777 / 0.9988 | 0.9991 / 1.0000 | 0.9993 / 1.0000 |
| Avg MST Edge Length | 0.0642 | 0.000043 | 0.000018 | 0.000078 | 0.000914 |
| Semantic Ratio | 0.9572 | 0.8740 | 1.0000 | 1.0000 | 0.9980 |
| $D_{sen}$ Score | 0.9036 | 0.8999 | 0.8938 | 0.9211 | 0.9055 |
| Mean entity count | 0.7106 | 0.2040 | 0.5260 | 0.1636 | 0.2619 |
| Mean nominal count | 7.6358 | 2.6180 | 5.8930 | 1.6534 | 2.4506 |
| Mean entity density | 0.0269 | 0.0177 | 0.0206 | 0.0187 | 0.0272 |
| Mean nominal density | 0.2569 | 0.2151 | 0.2295 | 0.1849 | 0.2063 |
| 1st percentile $d_i^{nn}$ | 0.1847 | $1.49\times10^{-4}$ | $3.46\times10^{-5}$ | $2.89\times10^{-4}$ | 0.0064 |

**Key observations:** Using SynEval to guide prompt optimization (**Level 2**) massively boosts semantic diversity (MST Edge Length expanded **12x**). However, SynEval successfully flagged that this richer prompting slightly increased privacy leakage (entity counts rose), proving its value in monitoring the privacy-utility trade-off.

## Privacy Risks in Text (Contextual Identifiers)

TABLE IV: Examples of reviews with high entity/nominal density. Entities are underlined; nominal mentions are highlighted.

| Review Text | Entity Density | Nominal Density |
|---|---|---|
| "I love this baseball cap. I graduated from the University of Hawaii with my Bachelor's degree...and I love advertising Hawaii on the top of my head! The many years I lived in Hawaii ~ it was/is absolutely gorgeous, calm, safe, friendly and multi-ethnic. Great memories...thus, a happy cap to bring back happy memories." | 0.086 | 0.293 |
| "I have plantar fasciitis and have been trying and using various compression socks and sleeves... I ordered the large/extra-large because I take a 9 to 9.5 shoe... I'm passing them off to my boyfriend... These are not only 'fun' but they are medically helpful for my plantar fasciitis..." | 0.025 | 0.284 |
| "Bought this in XL for my 11yo who is 5'8 and 110." | 0.250 | 0.417 |
| "My granddaughter loves these!" | 0.000 | 0.750 |

**Key observations:** Even without explicit PII, high "Nominal/Entity Density" creates severe re-identification risks. For example, a single review mapped a nominal density of **0.293**, exposing a highly identifiable personal narrative.