Geração de Dados Sintéticos com IA Generativa: Inovação Segura e Governança Alinhadas à LGPD e ao ECA Digital

Suerdo Flaubert C. de Lucena Júnior[1], Juliana Saraiva[1]

[1]Departamento de Ciências Exatas – Universidade Federal da Paraíba (UFPB)
Av. Santa Elizabeth, S/N, Centro, CEP 58297-000 - Rio Tinto - PB - Brasil

{suerdo.campos, julianajags}@dcx.ufpb.br

Abstract. *This paper presents a modular software pipeline for generating synthetic personal data using Generative AI to support testing, research, and AI experimentation, aligned with Brazil's LGPD and the ECA Digital. The approach combines fictitious data with GAN to produce coherent socio-demographic profiles, incorporating traceability, uniqueness checks, similarity controls, and semantic validation rules. The utility was assessed through statistical fidelity and relational consistency, while robustness was assessed using plausibility-based rules. The results indicate reduced exposure of real data subjects when synthetic data generation is treated as a risk-management strategy. Also, the code is available to ensure reproducibility.*

Resumo. *Este artigo apresenta um pipeline de software modular para geração de dados pessoais sintéticos com IA Generativa para testes, pesquisa e experimentação em IA, alinhado à LGPD e ao ECA Digital. A proposta combina identificadores fictícios determinísticos e uma GAN tabular para produzir perfis coerentes, com rastreabilidade, checagens de unicidade, controles de similaridade e validações semânticas. A utilidade foi avaliada por fidelidade estatística e consistência relacional, e a robustez por regras de plausibilidade. Os resultados sugerem redução de exposição de titulares quando a geração é tratada como gestão de risco e o código é disponibilizado publicamente para reprodutibilidade.*

1. Introdução

A crescente utilização de IA generativa no desenvolvimento de sistemas tem ampliado as possibilidades de automação, simulação e experimentação orientadas por dados. Ao mesmo tempo, a manipulação de informações pessoais em ambientes de desenvolvimento, testes e pesquisas tornou-se um ponto crítico na relação entre computação e sociedade. Vazamentos, acessos indevidos e reutilização inadequada de bases reais afetam não apenas a conformidade jurídica, mas também a confiança institucional, a segurança dos indivíduos e a legitimidade social de sistemas baseados em dados [Strzelecki and Rizun 2022].

Instituições públicas e privadas que desenvolvem, testam e comercializam softwares necessitam de dados realistas para validar funcionalidades, treinar modelos e conduzir experimentos. Essa necessidade cria uma tensão estrutural entre utilidade técnica e proteção de direitos fundamentais, especialmente quando se trata de dados pessoais. Nesse cenário, a geração de dados sintéticos por meio de IA generativa surge

como alternativa promissora para reduzir a exposição de titulares reais durante fases críticas do ciclo de vida dos sistemas.

Entretanto, é fundamental reconhecer que "dado sintético" não é automaticamente equivalente a "dado anonimizado". Modelos generativos podem memorizar padrões do conjunto de treinamento, permitir inferências por cruzamento com outras bases e reproduzir vieses sociais históricos, potencialmente ampliando desigualdades quando utilizados para simulação, treinamento ou validação. Assim, a geração de dados sintéticos deve ser compreendida como um problema sociotécnico, pois envolve não apenas arquitetura e desempenho técnico, mas também métricas de risco, mecanismos de controle, limites de uso, governança institucional e impactos sociais.

No Brasil, o debate regulatório sobre anonimização, pseudonimização e uso de dados sintéticos encontra-se em consolidação no âmbito da Lei Geral de Proteção de Dados (LGPD). A LGPD estabelece princípios como finalidade, adequação, necessidade, segurança, prevenção e responsabilização, que devem orientar todo o ciclo de vida do tratamento de dados — inclusive quando esses dados são artificialmente gerados. Além disso, a recente publicação do Estatuto Digital da Criança e do Adolescente (ECA Digital), com entrada em vigor iminente, eleva o padrão de cuidado exigido no ambiente digital, reforçando a centralidade da prevenção de danos e da proteção do melhor interesse de crianças e adolescentes. Esse novo marco normativo amplia as expectativas sociais e institucionais de que sistemas sejam concebidos com mecanismos concretos de mitigação de risco desde sua concepção.

Diante desse contexto, este artigo propõe uma solução modular e escalável para geração de dados sintéticos brasileiros por meio de IA generativa, orientada à inovação segura e à governança de privacidade. A proposta integra: (i) geração controlada de atributos estruturais e identificadores fictícios exclusivamente para ambientes de teste e homologação; (ii) utilização de redes neurais generativas (GANs – *Generative Adversarial Networks*) para compor perfis socioeconômicos e demográficos coerentes; e (iii) incorporação de salvaguardas verificáveis, como checagens de unicidade, controles de similaridade, testes de overfitting, validações estatísticas e regras semânticas.

Para reforçar a independência em relação a indivíduos reais e mitigar riscos residuais de associação ou inferência indevida, a solução incorpora mecanismos de rastreabilidade, controle de parâmetros e documentação do processo de geração. Além da dimensão técnica, o trabalho discute a governança do uso de dados sintéticos e propõe um protocolo mínimo de adoção responsável, alinhado aos princípios da LGPD e às diretrizes reforçadas de prevenção previstas no ECA Digital. Embora o pipeline possa ser parametrizado para diferentes faixas etárias, os experimentos reportados neste artigo foram conduzidos, por padrão, com perfis adultos, isto é, maiores de 18 anos, por controle experimental e comparabilidade de métricas. Assim, o alinhamento ao ECA Digital é discutido neste trabalho principalmente como diretriz de governança preventiva e como orientação para parametrizações futuras que envolvam crianças e adolescentes.

O artigo está estruturado da seguinte forma, incluindo esta: a Seção 2 discute os fundamentos teóricos sobre dados sintéticos, IA generativa e governança à luz da LGPD

e do ECA Digital. A Seção 3 descreve a metodologia e o pipeline proposto. A Seção 4 analisa os resultados e seus limites, enquanto a Seção 5 examina os impactos regulatórios e as contribuições à proteção de dados. Já a Seção 6 explicita os trabalhos relacionados, e por fim, a Seção 7 apresenta as considerações finais.

2. Referencial Teórico

2.1. Anonimização e Pseudonimização: Fundamentos e Limites Jurídicos

A distinção entre anonimização e pseudonimização é central para enquadrar dados sintéticos na LGPD, especialmente em contextos de testes, homologação e pesquisa. A LGPD define dado anonimizado como aquele relativo a titular que não possa ser identificado, considerando "meios técnicos razoáveis e disponíveis" (art. 5º, XI), e dispõe que dados efetivamente anonimizados deixam de ser dados pessoais (art. 12) [Brasil 2018]. Já a pseudonimização reduz a identificação direta por substituição de atributos, mas mantém possibilidade de reidentificação, permanecendo sob proteção da LGPD (art. 13, §4º) [Brasil 2018].

Essas definições sustentam uma abordagem baseada em risco: anonimização deve ser tratada como processo de transformação e validação, dependente do contexto e do risco residual de associação, inclusive por cruzamento com outras bases. Documentos técnicos da ANPD reforçam a necessidade de avaliar meios técnicos, plausibilidade de reidentificação e evidências de mitigação [ANPD 2023a; ANPD 2023b].

Nesse cenário, dados sintéticos podem reduzir a dependência de bases reais, mas não são automaticamente "fora da LGPD". Dependendo do pipeline e do treinamento, podem reter padrões do conjunto de referência e permitir inferências em cenários adversariais. Assim, o uso responsável requer controles verificáveis, como critérios de desvinculação, mecanismos para reduzir memorização/replicação e documentação técnica compatível com os princípios de necessidade, segurança, prevenção e responsabilização (art. 6º) [Brasil 2018; ANPD 2023b].

2.2. IA Generativa e Modelos para Geração de Dados Sintéticos

Dados sintéticos são informações artificialmente geradas para preservar propriedades estatísticas e estruturais de um conjunto de referência, viabilizando testes e experimentação quando o uso de dados reais é indesejável ou arriscado [Patki et al. 2016]. A IA generativa ampliou esse potencial, especialmente em dados tabulares, por meio de modelos como as Redes Gerativas Adversariais (GANs), nas quais um gerador e um discriminador interagem de forma adversarial para produzir amostras progressivamente mais plausíveis [Goodfellow et al. 2014]. Modelos condicionais como o CTGAN tornaram-se amplamente utilizados por lidarem com variáveis categóricas e desbalanceamento [Xu et al. 2019].

Entretanto, a geração sintética envolve um trade-off entre privacidade e utilidade. Modelos generativos podem memorizar padrões do conjunto de treinamento e permitir inferências sobre registros reais, sobretudo em bases pequenas ou sensíveis [Hilprecht et

al. 2019; Chen et al. 2019]. Além disso, dados sintéticos podem reproduzir vieses históricos, mantendo padrões discriminatórios em aplicações posteriores.

Por isso, recomenda-se que pipelines de geração incorporem avaliação em dois eixos: (i) utilidade, verificando fidelidade estatística e preservação de relações entre variáveis; e (ii) risco, analisando overfitting, duplicação e similaridade excessiva. Abordagens com privacidade diferencial e métricas específicas de risco reforçam a maturidade do campo, embora impliquem maior complexidade [Xie et al. 2018; Jordon et al. 2019; Steier et al. 2025]. Assim, a adoção responsável de IA generativa para dados sintéticos depende de controles verificáveis, validações técnicas e documentação transparente do processo de geração.

2.3. Governança de Dados, Privacy by Design e Accountability

A geração de dados sintéticos com IA generativa deve ser compreendida como prática inserida em programas de governança de dados, e não como simples substituição técnica de bases reais. Governança envolve definição de finalidade, limites de uso, responsabilidades, controles técnicos e mecanismos de prestação de contas ao longo do ciclo de vida do dado. Na LGPD, princípios como necessidade, segurança, prevenção e responsabilização precisam ser traduzidos em requisitos implementáveis e auditáveis [Brasil 2018].

Sob a perspectiva de engenharia, isso se concretiza por meio de privacy by design e privacy by default, incorporando salvaguardas desde a concepção do sistema. Em pipelines de dados sintéticos, tais salvaguardas incluem separação de ambientes (teste vs. produção), parametrização reprodutível, validações automáticas, regras semânticas e mecanismos de rastreabilidade (logs e documentação de execução).

A accountability exige evidências técnicas de redução de risco e adequação à finalidade declarada. Esse cuidado é ainda mais relevante em contextos que envolvem crianças e adolescentes, dada a proteção reforçada prevista na LGPD e no ECA Digital [Brasil 2018; Brasil 2025]. Assim, a geração de dados sintéticos configura um problema sociotécnico no qual utilidade estatística deve ser acompanhada de governança verificável, rastreabilidade e mitigação estruturada de riscos.

3. Metodologia

3.1. Delineamento do Estudo

Este estudo adota a abordagem de ***Design Science Research*** **(DSR)**, orientada à proposição, desenvolvimento e avaliação de um artefato computacional com relevância prática e contribuição científica [Delport et al. 2024]. A DSR parte da identificação de um problema no contexto aplicado, avança para a definição de objetivos de solução, projeta e implementa um artefato, demonstra sua aplicabilidade e realiza sua avaliação à luz de critérios técnicos e contextuais.

O problema abordado na pesquisa decorre do conflito entre a necessidade de dados pessoais realistas para desenvolvimento, testes e experimentação em sistemas baseados em inteligência artificial e as exigências regulatórias impostas pela LGPD, recentemente reforçadas pelo ECA Digital. Nesse contexto, torna-se necessário projetar

soluções que possibilitem inovação tecnológica sem exposição indevida de titulares reais.

Com base nesse diagnóstico, definiu-se como objetivo o desenvolvimento de um *pipeline* modular para geração de dados pessoais sintéticos que preservasse a coerência relacional, reduzisse o risco de associação ou reidentificação e incorporasse mecanismos verificáveis de governança e rastreabilidade, em consonância com os princípios da LGPD, notadamente necessidade, segurança, prevenção e responsabilização.

3.2. Design e Implementação do *Pipeline* de Software

O artefato desenvolvido consiste em um *pipeline* de software estruturado em etapas sequenciais e integradas, concebido para garantir modularidade, controle de risco e reprodutibilidade. O pipeline possui dois componentes acoplados: (i) geração generativa de atributos tabulares, como idade, gênero e renda, por meio de uma GAN tabular; e (ii) geração programática de identificadores fictícios e atributos complementares, como CPF, CNH, RG, título de eleitor, telefone, nome e data de nascimento, seguida de validações, checagens de unicidade e exportação. A implementação foi realizada em Python 3.12 no ambiente Google Colab, utilizando bibliotecas amplamente reconhecidas para processamento de dados e modelagem generativa. O uso de sementes fixas e registro sistemático dos parâmetros assegurou reprodutibilidade experimental. Os experimentos foram executados em GPU NVIDIA T4. O código-fonte[1] completo do *pipeline*, incluindo *scripts* de geração, validação e documentação de execução, está disponível em repositório público no GitHub, permitindo reprodutibilidade dos experimentos e auditoria das etapas do processo.

Inicialmente, realiza-se a parametrização da população sintética, definindo-se atributos, domínios possíveis, regras de coerência e proporções-alvo. Essa etapa permite configurar o contexto de geração conforme o cenário de aplicação pretendido, mantendo controle explícito sobre as variáveis modeladas. A fase de geração é dividida em uma etapa determinística voltada à criação de identificadores fictícios sintaticamente válidos para testes e homologação. Esses dados são produzidos por meio de aleatoriedade controlada e cálculo de dígitos verificadores, garantindo compatibilidade com validações formais dos sistemas. Além disso, o pipeline implementa validações automáticas para aumentar a segurança e confiabilidade, incluindo rejeição de padrões triviais, verificação de unicidade para evitar colisões internas e registro de parâmetros e finalidade de uso, restringindo a geração a ambientes não operacionais.

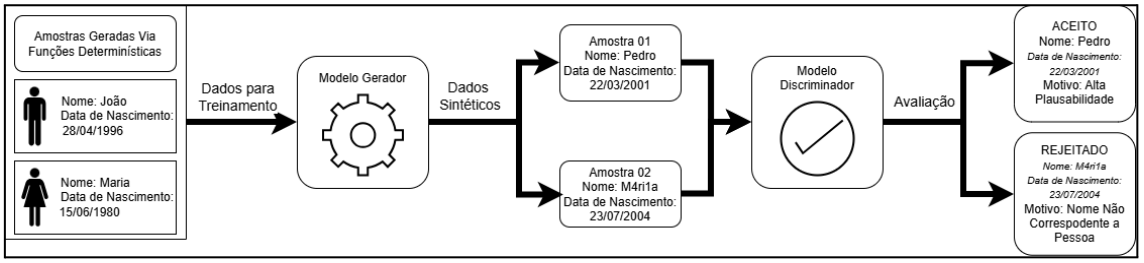

Figura 1. Visão geral do processo de geração e validação de dados sintéticos via GAN

[1] Disponível em: https://github.com/Suerdo/synthetic-br-profiles-gan

A segunda parte emprega um modelo generativo do tipo GAN tabular para compor perfis socioeconômicos e demográficos coerentes. O processo envolve pré-processamento dos dados de calibração, incluindo tratamento de valores ausentes, codificação de variáveis categóricas e padronização de variáveis numéricas. O treinamento ocorre em regime adversarial, no qual o gerador produz registros candidatos e o discriminador avalia sua plausibilidade estatística. O monitoramento de estabilidade e o controle por semente aleatória permitem maior reprodutibilidade e consistência. O objetivo dessa etapa é aproximar distribuições e relações entre atributos, preservando utilidade estatística sem reproduzir literalmente registros de treinamento.

A Figura 1 ilustra, de forma simplificada, o funcionamento do pipeline proposto e a lógica adversarial de uma GAN no contexto de geração de perfis sintéticos. A partir de uma base sintética inicial composta por atributos tabulares de calibração, como idade, gênero e renda, o modelo gerador produz registros sintéticos candidatos. Esses registros podem ser plausíveis ou incoerentes quanto aos atributos tabulares modelados. Em seguida, o discriminador avalia essas amostras e estima sua plausibilidade estatística conforme os padrões aprendidos durante o treinamento. Esse mecanismo atua como um filtro de qualidade e, no treinamento, fornece um sinal de erro que orienta o gerador a reduzir inconsistências ao longo das iterações, aumentando progressivamente a qualidade e a coerência dos atributos gerados. Após essa etapa, os registros aceitos ou selecionados são enriquecidos por funções programáticas com nomes, datas de nascimento, identificadores fictícios e contatos, sendo então submetidos a validações estruturais e checagens de unicidade.

3.3. Teste e Avaliação

Os testes do *pipeline* consistiram na geração de conjuntos populacionais sintéticos destinados a cenários típicos de testes de sistemas e experimentação em IA. A avaliação foi conduzida sob duas perspectivas complementares: utilidade e risco. No eixo da utilidade, analisaram-se distribuições univariadas e relações multivariadas entre atributos, com o intuito de verificar a preservação de estruturas estatísticas relevantes para simulações e validações funcionais. Foram examinadas correlações e associações entre variáveis demográficas e socioeconômicas, buscando evidências de coerência relacional compatível com o cenário brasileiro.

No eixo do risco, implementaram-se checagens de unicidade e detecção de duplicações, bem como análises de similaridade excessiva que pudessem indicar *overfitting* ou proximidade indevida com padrões de calibração. Adicionalmente, após a etapa generativa, os registros foram enriquecidos com atributos complementares gerados programaticamente, incluindo documentos fictícios e contatos. Em seguida, aplicaram-se regras de coerência semântica, validações estruturais e checagens de unicidade para identificar combinações inconsistentes, formatos inválidos ou colisões internas, reforçando o caráter iterativo do ciclo de design e avaliação característico da DSR.

3.4. Coleta e Análise de Dados

Para viabilizar a rastreabilidade e a coleta de evidências experimentais, o pipeline foi implementado para exportar, juntamente com a base sintética gerada, um relatório de

execução (relatorio_execucao.json) contendo parâmetros relevantes (por exemplo, *seed*, *latent_dim*, tamanho do lote e limiar do discriminador), tempo de geração e estatísticas resumidas. Essa documentação permite a replicação do experimento e auditoria básica das etapas executadas.

A análise dos dados foi conduzida de forma predominantemente quantitativa. Inicialmente, os registros foram gerados em ambiente controlado (Google Colab com GPU), com semente fixa (seed = 41). Em seguida, foram computadas métricas operacionais (número de candidatos, número de registros selecionados, tempo total e vazão em registros por segundo) e estatísticas descritivas univariadas para variáveis-chave (idade, renda e proporções por gênero, antes do pós-processamento). Por fim, foram aplicadas validações estruturais no conjunto final (N = 1.000), incluindo checagem de formato para CPF/RG/telefone e verificação de unicidade (ausência de colisões) para CPF, CNH, RG, título de eleitor e telefone, produzindo uma taxa de conformidade para os critérios avaliados.

Além disso, a plausibilidade semântica foi tratada como requisito de governança por meio de restrições e pós-processamento no pipeline (por exemplo, coerência entre gênero e nomes gerados e limitação de faixa etária), complementada por inspeção amostral do resultado final para identificação de inconsistências residuais. Os achados foram discutidos à luz de utilidade para testes e do papel de validações e rastreabilidade como mecanismos de controle de risco no uso de dados sintéticos.

4. Discussão de Resultados

Esta seção apresenta e discute os resultados decorrentes da implementação e avaliação do pipeline de software para geração de dados pessoais sintéticos com IA generativa. A análise integra evidências de utilidade técnica (plausibilidade estatística e adequação para testes), consistência estrutural e desempenho computacional com dimensões de governança, risco regulatório e proteção diferenciada de grupos vulneráveis. Ao articular resultados quantitativos e considerações sociotécnicas, busca-se demonstrar que a proposta não apenas produz dados plausíveis, mas operacionaliza um modelo de inovação mais segura alinhado à LGPD e ao ECA Digital.

4.1. Qualidade e Evidências de Utilidade Técnica

A utilidade dos dados sintéticos foi avaliada a partir da preservação de padrões estatísticos relevantes para cenários de testes, simulações e experimentação em inteligência artificial. A Figura 2 apresenta um recorte ilustrativo de perfis gerados, evidenciando diversidade de combinações e consistência estrutural entre atributos. A avaliação de utilidade foi organizada em uma análise univariada.

	Nome	Gênero	Data_Nascimento	CPF	CNH	RG	Titulo_Eleitor	Telefone	Renda
0	Sofia Andrade	Feminino	12/06/1966	725.559.180-94	98321971447	75.235.407-7	9998 8987 26 42	(35) 96815-9605	8171.27
1	Valentina Souza	Feminino	13/09/1964	699.788.819-48	65407539835	92.189.613-0	2846 2567 26 00	(18) 96616-1819	8469.04
2	Milena Pastor	Feminino	21/02/1971	742.150.869-49	76339461005	48.286.386-4	4922 2100 25 55	(19) 90941-7802	8938.85
3	Maria Fernanda Pimenta	Feminino	13/06/1967	577.429.880-97	48331989901	47.563.917-9	9138 9940 11 93	(21) 97411-7314	9037.37
4	Cecilia Barbosa	Feminino	02/03/1971	975.527.928-84	9419053292	40.021.787-6	8298 8426 11 90	(31) 95396-2127	9604.47
5	Isadora Vieira	Feminino	15/11/1966	066.123.612-98	8164899948	25.720.401-8	1181 3350 15 07	(12) 95427-3433	7926.98
6	Anna Liz da Luz	Feminino	11/01/1970	365.875.971-22	61734873655	73.651.562-3	8315 2630 02 03	(19) 92455-6552	7896.13
7	Maria Eduarda Costela	Feminino	19/09/1965	129.855.282-67	49969688747	72.249.826-5	2221 3624 23 82	(48) 96177-2825	7003.65
8	Júlia Pires	Feminino	02/09/1968	100.209.557-31	22630865887	52.756.174-5	1685 8801 22 80	(37) 98930-7116	9831.87
9	Maria Luísa Melo	Feminino	06/05/1969	733.306.079-98	45105472727	20.300.962-5	5508 5039 18 94	(55) 97481-4477	8980.40
10	Lívia Brito	Feminino	02/12/1968	267.708.026-51	15111936692	37.754.664-2	6176 6027 05 03	(37) 91505-1265	8321.00

Figura 2. Exemplo de Perfis Sintéticos Gerados.

Na análise univariada, foram examinadas distribuições de variáveis contínuas e categóricas (por exemplo, idade, renda e gênero codificado em {0,1}), com foco em padrões plausíveis e coerência do domínio. Observou-se assimetria positiva na variável renda e maior concentração em faixas inferiores, comportamento compatível com distribuições socioeconômicas tipicamente assimétricas. No experimento controlado, antes do pós-processamento, registrou-se idade média de 29,64 (DP = 2,21), renda média de R$ 3.115,74 (mediana = R$ 2.970,40, DP = R$ 911,64) e proporção de "gênero = 1" de 27,7%. Esses valores isoladamente não comprovam utilidade, mas oferecem evidência descritiva de que o modelo produz amostras em faixas plausíveis e com variação mensurável, subsidiando calibração e ajustes futuros (por exemplo, ampliar diversidade e reduzir concentração).

Para quantificar desempenho e explicitar o procedimento de geração, foi executada uma rodada onde o modelo gerou inicialmente um lote de 2.048 candidatos (batch_gen), valor escolhido por conveniência computacional (processamento em batch na GPU e redução de chamadas sucessivas ao gerador - que pode ser ajustado). Em seguida, 1.000 registros foram selecionados para compor o conjunto final e os 1.048 restantes não foram utilizados apenas por já se ter atingido o tamanho-alvo (subamostragem por tamanho desejado), e não por reprovação. Com limiar do discriminador $\tau = 0{,}50$ e regras de domínio (idade 18–65, renda 1.200–25.000 e gênero $\in \{0,1\}$), não houve rejeições por violação de domínio nesta execução. A geração do lote final demandou 0,598 s, resultando em vazão de aproximadamente 1.673 registros, tendo como base no cálculo do tempo dos 1.000 registros selecionados.

Por fim, embora parte da consistência de campos como documentos e telefone decorra de rotinas determinísticas (máscaras e verificadores), a contribuição central do artefato não é apenas gerar valores válidos, e sim viabilizar perfis completos e coerentes para testes, com rastreabilidade e governança, sem depender de bases reais, reduzindo a probabilidade de reutilização indevida de dados pessoais e o risco de associação indevida. Essa utilidade é sustentada por decisões de projeto e evidências operacionais: (i) execução em ambiente controlado e sem ingestão de registros reais, (ii) parametrização e reprodutibilidade por semente, (iii) validações estruturais e controles de unicidade no lote final, e (iv) etapa generativa que produz combinações de atributos adequadas para simulação e testes funcionais. Assim, a contribuição da solução reside no cruzamento consistente de atributos com governança e rastreabilidade, oferecendo

um caminho pragmático para experimentação orientada por dados com menor exposição e menor risco de "contaminação" por dados de pessoas reais.

4.2. Robustez Semântica e Controle de Risco Residual

Além da fidelidade estatística, a qualidade dos dados foi analisada sob a perspectiva de coerência estrutural e risco residual de inconsistência, por meio de regras objetivas. Para o conjunto final com N = 1.000 registros, verificou-se: (i) validade de formato de campos críticos (CPF, RG e telefone) e (ii) unicidade/ausência de colisões internas para identificadores e contatos (CPF, CNH, RG, título de eleitor e telefone). Os resultados indicaram 0 ocorrências de formato inválido e 0 ocorrências de duplicidade em todos os campos avaliados, totalizando 0 erros no lote e conformidade estrutural de 100% para esses critérios. Esses testes evidenciam qualidade sintática e ausência de colisões internas, requisitos importantes para uso em ambientes de desenvolvimento e homologação.

É importante, contudo, distinguir a consistência estrutural de garantias de privacidade. Validar formato e unicidade demonstra que o conjunto é tecnicamente viável, mas não é, por si só, a comprovação de inexistência de risco de associação com registros reais. No presente trabalho, o controle de risco foi tratado como requisito de arquitetura: o pipeline não utiliza registros reais como entrada nem como saída, opera com parâmetros auditáveis (*seed e logs*) e impõe validações e pós-processamento antes da exportação. Como extensão natural de avaliação (a ser reportada em versões futuras do artefato), recomenda-se incorporar métricas de proximidade e testes adversariais (por exemplo, *distance to closest record* e ataques de *membership inference*), conforme discutido na literatura [Hilprecht et al. 2019; Chen et al. 2019], para estimar risco residual sob diferentes cenários de ameaça.

No nível semântico, o pipeline incorpora restrições e pós-processamento para reduzir combinações implausíveis (por exemplo, limitação explícita de faixa etária e geração de nomes coerentes com o gênero). Em inspeções amostrais, observou-se que essas camadas contribuem para reduzir inconsistências contextuais e reforçam a interpretação de que geração sintética é um processo governado por validações, e não uma "anonimização automática" por definição. Assim, o controle de risco é parte integrante da arquitetura proposta, consistente com a natureza iterativa do ciclo de ***Design Science Research***.

4.3. Governança Incorporada ao Design: *Privacy by Design* e Rastreabilidade

Um dos principais diferenciais da proposta reside na incorporação de mecanismos explícitos de governança ao design do *pipeline*. Diferentemente de abordagens que tratam dados sintéticos como automaticamente anonimizados, o sistema foi concebido sob uma perspectiva baseada em risco, alinhada a *privacy by design* e *privacy by default*. A geração determinística de identificadores fictícios, as checagens de unicidade, as validações estruturais e o registro sistemático de parâmetros, versões e sementes compõem um conjunto de salvaguardas que viabilizam rastreabilidade e auditoria. Do ponto de vista regulatório, a aderência não decorre da simples afirmação de que os dados são *sintéticos*, mas da existência de evidências técnicas verificáveis e mecanismos

de controle estruturados, relevantes para contextos organizacionais que demandam justificativa formal do tratamento de dados.

5. Implicações Regulatórias e Contribuições à Governança de Dados à luz da LGPD e do ECA Digital

Os resultados técnicos apresentados fornecem evidências sobre a sustentação das implicações regulatórias da arquitetura proposta. A geração estruturada de dados pessoais sintéticos por meio de técnicas determinísticas e modelos generativos não apenas reduz a dependência de bases reais, mas configura um modelo operacional de inovação segura alinhado aos princípios da LGPD e às expectativas reforçadas pelo ECA Digital.

Sob a perspectiva da LGPD, a adoção de dados sintéticos conecta-se diretamente ao princípio da necessidade, ao evitar o tratamento de dados identificáveis quando estes não são indispensáveis para atividades de desenvolvimento, teste e validação de sistemas. A proposta também materializa os princípios da segurança e da prevenção, ao reduzir a superfície de exposição em fases historicamente sensíveis do ciclo de vida de sistemas — como desenvolvimento, homologação e demonstração — nas quais incidentes de vazamento e acessos indevidos são mais recorrentes.

Contudo, é fundamental destacar que dados sintéticos não devem ser automaticamente considerados “anonimizados” por definição. A incidência da LGPD depende de avaliação técnica do risco residual de associação ou reidentificação, especialmente em cenários que envolvem treinamento de modelos, cruzamento de bases ou inferências indiretas. Nesse sentido, o diferencial da proposta reside na incorporação de controles verificáveis — como checagens de unicidade, validações semânticas, mecanismos de prevenção de replicação e rastreabilidade completa da execução — que permitem justificar tecnicamente o uso dos conjuntos gerados e demonstrar mitigação estruturada de riscos.

A solução aqui proposta, portanto, não apenas produz dados sintéticos, mas operacionaliza princípios de *privacy by design* e rastreabilidade, transformando exigências normativas em parâmetros de engenharia. Essa conversão de princípios jurídicos em requisitos técnicos verificáveis representa contribuição relevante ao debate sobre governança de dados e uso responsável de IA, especialmente em ambientes organizacionais que demandam evidências formais de conformidade. A disponibilização pública do código-fonte do *pipeline* reforça a transparência e a auditabilidade da solução, permitindo inspeção independente da implementação das salvaguardas técnicas.

Além disso, a proposta dialoga com agendas regulatórias contemporâneas que articulam segurança da informação, uso de inteligência artificial e técnicas de anonimização e pseudonimização. Ao propor diretrizes técnicas mínimas para adoção de dados sintéticos — incluindo documentação de finalidade e contexto de uso, avaliação de utilidade e risco, mecanismos de rastreabilidade e critérios de governança para compartilhamento — o trabalho contribui para consolidar boas práticas replicáveis em diferentes contextos institucionais. Em síntese, o *pipeline* de software proposto não apenas reduz riscos operacionais, mas estrutura um modelo de governança técnica

alinhado à LGPD e ao ECA Digital, no qual inovação e conformidade são concebidas como dimensões complementares do design de sistemas orientados por dados.

6. Trabalhos Relacionados

A literatura sobre geração de dados sintéticos converge no entendimento de que produzir dados "realistas" não é suficiente: é necessário demonstrar, de forma mensurável, tanto a utilidade estatística quanto o nível de proteção à privacidade. Revisões recentes organizam o campo em abordagens estatísticas tradicionais e modelos generativos baseados em GANs, VAEs e técnicas de IA generativa, destacando o persistente desafio do equilíbrio entre privacidade e utilidade, bem como a ausência de padronização nas métricas de avaliação [Goyal and Mahmoud 2024].

Além dos modelos generativos, há ferramentas amplamente utilizadas para geração de dados fictícios e massas de teste, como Faker, Mockaroo e bibliotecas de mock de dados. Essas soluções são úteis em cenários de desenvolvimento, testes funcionais, preenchimento de formulários e validação de fluxos de sistemas, pois permitem gerar nomes, endereços, contatos e outros campos sintaticamente plausíveis de forma simples e rápida. No entanto, em geral, essas abordagens são mais orientadas à geração aleatória ou baseada em regras, oferecendo suporte limitado à preservação de relações estatísticas entre variáveis, à avaliação de risco residual e à rastreabilidade do processo de geração. Assim, a proposta deste trabalho não substitui tais ferramentas, mas as complementa ao combinar geração programática de identificadores fictícios, modelagem generativa de atributos tabulares, validações estruturais, checagens de unicidade e documentação de execução em um pipeline orientado à governança.

Em dados tabulares, modelos como o CTGAN tornaram-se referência por lidar com variáveis categóricas e distribuições desbalanceadas [Xu et al. 2019]. Contudo, essas soluções são predominantemente orientadas à modelagem estatística e, em geral, não incorporam mecanismos explícitos de governança, rastreabilidade e prestação de contas. Paralelamente, abordagens que integram privacidade diferencial (como DP-GAN e PATE-GAN) buscam reforçar garantias contra vazamento do treinamento, ainda que com maior complexidade técnica [Xie et al. 2018; Jordon et al. 2019]. Estudos sobre ataques de inferência reforçam a necessidade de métricas específicas para avaliação de risco em dados sintéticos [Hilprecht et al. 2019; Chen et al. 2019; Steier et al. 2025].

No contexto brasileiro, trabalhos aplicados indicam a importância da adequação sociocultural, especialmente quanto a padrões documentais e nomes locais [Corral 2021]. Diferentemente de geradores baseados apenas em regras, listas ou aleatoriedade, o presente estudo propõe um pipeline integrado que combina modelagem generativa de atributos tabulares, geração determinística de identificadores fictícios compatíveis com o contexto brasileiro, validações estatísticas e semânticas e mecanismos de rastreabilidade. Ao incorporar governança e controle de risco desde o design, a proposta avança da simples geração de dados realistas para um modelo de inovação segura e auditável.

7. Considerações Finais

A integração entre inteligência artificial generativa e proteção de dados exige que soluções técnicas sejam concebidas desde sua origem com critérios explícitos de governança e mitigação de risco. Neste trabalho, foi proposta e avaliada uma arquitetura modular para geração de dados sintéticos populacionais que combina identificadores fictícios determinísticos e modelagem adversarial para produção de atributos relacionais coerentes, incorporando salvaguardas estruturadas e mecanismos de rastreabilidade.

Os resultados indicam que o *pipeline* de software produz conjuntos com fidelidade estatística e consistência interna adequadas a cenários de desenvolvimento, homologação e experimentação inicial em projetos de IA, reduzindo a necessidade de utilização de bases reais. Mais do que viabilizar utilidade técnica, a solução demonstra que princípios como necessidade, segurança, prevenção e responsabilização — previstos na LGPD — podem ser operacionalizados no próprio design do sistema, por meio de controles verificáveis e documentação de execução.

O estudo também reforça que dados sintéticos não devem ser automaticamente classificados como anonimizados. A avaliação de risco residual e a adoção de medidas técnicas de desvinculação são elementos centrais para justificar seu uso em conformidade regulatória. Ao estruturar mecanismos de validação estatística, coerência semântica e auditoria do processo, a arquitetura contribui para um modelo de governança incorporado à Engenharia de Software.

Como perspectivas futuras, destacam-se o refinamento de regras condicionais para maior precisão semântica, especialmente em parametrizações voltadas a perfis infantojuvenis, considerando as diretrizes reforçadas pelo ECA Digital. Embora a avaliação experimental deste artigo tenha adotado, por padrão, perfis adultos, a arquitetura pode ser evoluída para incorporar seleções de atributos, validações e restrições específicas conforme a faixa etária e a finalidade de uso.

Ainda assim, apesar de inicial, a pesquisa demonstra que a geração de dados sintéticos com IA generativa pode constituir instrumento concreto de inovação segura quando acompanhada de governança técnica estruturada. Com isso, contribui para aproximar engenharia, inteligência artificial e proteção de dados, demonstrando que a geração de dados sintéticos pode ser reformulada, deixando de ser uma técnica puramente estatística para se tornar uma prática de engenharia orientada à governança.

Agradecimentos

Este trabalho é parcialmente apoiado pelo INES.IA (www.ines.org.br), CNPq processo 408817/2024-0. Este trabalho é parcialmente fomentado pelo Centro de Excelência em Tecnologias Sociais (NEES), afiliado ao Instituto de Computação (IC) da Universidade Federal de Alagoas (UFAL).

Uso de Inteligência Artificial

Ferramentas de Inteligência Artificial generativa foram utilizadas como apoio à geração de ilustrações conceituais, revisão gramatical, clareza textual e organização da exposição, sem substituir a análise técnica dos autores.

Referências

AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS (ANPD). Estudo técnico sobre a anonimização de dados na LGPD: análise jurídica (Versão 1.0). Brasília, DF: ANPD, 2023. Disponível em: https://www.gov.br/anpd/pt-br/centrais-de-conteudo/documentos-tecnicos-orientativos/estudo_tecnico_sobre_anonimizacao_de_dados_na_lgpd___analise_juridica.pdf. Acesso em: 1 mar. 2026.

AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS (ANPD). Estudo técnico sobre a anonimização de dados na LGPD: uma visão de processo baseado em risco e técnicas computacionais (Versão 1.0). Brasília, DF: ANPD, 2023. Disponível em: https://www.gov.br/anpd/pt-br/centrais-de-conteudo/documentos-tecnicos-orientativos/estudo_tecnico_sobre_anonimizacao_de_dados_na_lgpd_uma_visao_de_processo_baseado_em_risco_e_tecnicas_computacionais.pdf. Acesso em: 1 mar. 2026.

AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS (ANPD). Resolução nº 23, de 9 de dezembro de 2024. Aprova a Agenda Regulatória para o biênio 2025–2026. Diário Oficial da União: Seção 1, Brasília, DF, 11 dez. 2024. Disponível em: https://www.in.gov.br/en/web/dou/-/resolucao-n-23-de-9-de-dezembro-de-2024-601118310. Acesso em: 1 mar. 2026.

AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS (ANPD). Radar Tecnológico nº 3: Inteligência Artificial Generativa. Brasília, DF: ANPD, nov. 2024. Disponível em: https://www.gov.br/anpd/pt-br/documentos-e-publicacoes/documentos-de-publicacoes/radar_tecnologico_ia_generativa_anpd.pdf. Acesso em: 1 mar. 2026.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União: Seção 1, Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 1 mar. 2026.

BRASIL. Lei nº 15.211, de 17 de setembro de 2025. Dispõe sobre a proteção de crianças e adolescentes em ambientes digitais (Estatuto Digital da Criança e do Adolescente). Diário Oficial da União: Seção 1, Brasília, DF, 18 set. 2025. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/lei/L15211.htm. Acesso em: 1 mar. 2026.

CHEN, D.; YU, N.; ZHANG, Y.; FRITZ, M. GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. arXiv:1909.03935, 2019. Disponível em: https://arxiv.org/abs/1909.03935. Acesso em: 1 mar. 2026.

CORRAL, Vitor Curiel Trentin. Gerador de dados sintéticos para testes de rotinas de record linkage para o contexto brasileiro. 2021. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) — Universidade Federal do Rio de Janeiro, Instituto de Matemática, Rio de Janeiro, 2021. Disponível em: https://pantheon.ufrj.br/bitstream/11422/14769/1/VCTCorral.pdf. Acesso em: 1 mar. 2026.

DELPORT, P. M. J.; VON SOLMS, R.; GERBER, M. Methodological Guidelines for Design Science Research. Procedia Computer Science, v. 237, p. 195–203, 2024. DOI: 10.1016/j.procs.2024.05.096. Disponível em: https://doi.org/10.1016/j.procs.2024.05.096. Acesso em: 1 mar. 2026.

GOODFELLOW, I. J.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative Adversarial Nets. In: Advances in Neural Information Processing Systems (NeurIPS 2014). 2014. Disponível em: https://papers.nips.cc/paper/5423-generative-adversarial-nets. Acesso em: 1 mar. 2026.

GOYAL, M.; MAHMOUD, Q. H. A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. Electronics, v. 13, n. 17, art. 3509, 2024. DOI: 10.3390/electronics13173509. Disponível em: https://www.mdpi.com/2079-9292/13/17/3509. Acesso em: 1 mar. 2026.

GUO, X.; CHEN, Y. Generative AI for Synthetic Data Generation: Methods, Challenges and the Future. arXiv:2403.04190, 2024. Disponível em: https://arxiv.org/abs/2403.04190. Acesso em: 1 mar. 2026.

HILPRECHT, B.; HÄRTERICH, M.; BERNAU, D. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. Proceedings on Privacy Enhancing Technologies, v. 2019, n. 4, p. 232–249, 2019. DOI: 10.2478/popets-2019-0067. Disponível em: https://petsymposium.org/popets/2019/popets-2019-0067.php. Acesso em: 1 mar. 2026.

JORDON, J.; YOON, J.; VAN DER SCHAAR, M. PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. In: International Conference on Learning Representations (ICLR 2019). 2019. Disponível em: https://openreview.net/forum?id=S1zk9iRqF7. Acesso em: 1 mar. 2026.

PATKI, N.; WEDGE, R.; VEERAMACHANENI, K. The Synthetic Data Vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). 2016. p. 399–410. DOI: 10.1109/DSAA.2016.49. Disponível em: https://ieeexplore.ieee.org/document/7796926. Acesso em: 1 mar. 2026.

STEIER, A.; RAMASWAMY, L.; MANOEL, A.; HAUSHALTER, A. Synthetic Data Privacy Metrics. arXiv:2501.03941, 2025. Disponível em: https://arxiv.org/abs/2501.03941. Acesso em: 1 mar. 2026.

STRZELECKI, A.; RIZUN, M. Consumers' Change in Trust and Security after a Personal Data Breach in Online Shopping. Sustainability, v. 14, n. 10, art. 5866, 2022. DOI: 10.3390/su14105866. Disponível em: https://www.mdpi.com/2071-1050/14/10/5866. Acesso em: 1 mar. 2026.

XIE, L.; LIN, K.; WANG, S.; WANG, F.; ZHOU, J. Differentially Private Generative Adversarial Network. arXiv:1802.06739, 2018. Disponível em: https://arxiv.org/abs/1802.06739. Acesso em: 1 mar. 2026.

XU, L.; SKOULARIDOU, M.; CUESTA-INFANTE, A.; VEERAMACHANENI, K. Modeling Tabular Data Using Conditional GAN. arXiv:1907.00503, 2019. Disponível em: https://arxiv.org/abs/1907.00503. Acesso em: 1 mar. 2026.

ZHOU, Y.; MALIN, B.; KANTARCIOGLU, M. SMOTE-DP: Improving Privacy-Utility Tradeoff with Synthetic Data. arXiv:2506.01907, 2025. Disponível em: https://arxiv.org/abs/2506.01907. Acesso em: 1 mar. 2026.