applied sciences

MDPI

Article

FairRAG: A Privacy-Preserving Framework for Fair Financial Decision-Making

Rashmi Nagpal [1,*], Unyimeabasi Usua [1], Rafael Palacios [2,3] and Amar Gupta [1,4,*]

[1] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, MA 02139, USA; unyime@mit.edu
[2] Cybersecurity at MIT Sloan (CAMS), Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA; palacios@mit.edu or rafael.palacios@iit.comillas.edu
[3] Institute for Research in Technology, Universidad Pontificia Comillas, Alberto Aguilera 23, 28015 Madrid, Spain
[4] AI Institute for Community-Engaged Research (AI-ICER), The University of Texas at El Paso, 500 West University Avenue, El Paso, TX 79968, USA
* Correspondence: rnagpal@mit.edu (R.N.); agupta@mit.edu (A.G.)

Abstract

Customer churn prediction has become crucial for businesses, yet it poses significant challenges regarding privacy preservation and prediction accuracy. In this paper, we address two fundamental questions: (1) How can customer churn be effectively predicted while ensuring robust privacy protection of sensitive data? (2) How can large language models enhance churn prediction accuracy while maintaining data privacy? To address these questions, we propose FairRAG, a robust architecture that combines differential privacy, retrieval-augmented generation, and LLMs. Our approach leverages OPT-125M as the core language model along with a sentence transformer for semantic similarity matching while incorporating differential privacy mechanisms to generate synthetic training data. We evaluate FairRAG on two diverse datasets: Bank Churn and Telco Churn. The results demonstrate significant improvements over both traditional machine learning approaches and standalone LLMs, achieving accuracy improvements of up to 11% on the Bank Churn dataset and 12% on the Telco Churn dataset. These improvements were maintained when using differentially private synthetic data, thus indicating robust privacy and accuracy trade-offs.

Keywords: algorithmic fairness; privacy-preserving machine learning; differential privacy; retrieval-augmented generation

Academic Editor: Stefano Quer

Received: 8 May 2025
Revised: 1 July 2025
Accepted: 4 July 2025
Published: 25 July 2025

Citation: Nagpal, R.; Usua, U.; Palacios, R.; Gupta, A. FairRAG: A Privacy-Preserving Framework for Fair Financial Decision-Making. *Appl. Sci.* **2025**, *15*, 8282. https://doi.org/10.3390/app15158282

1. Introduction

The use of machine learning as a means to address critical issues in the financial domain has transformed our understanding of the underlying patterns in data. However, in this field, which is highly dependent on user data and personally identifiable information in the decision-making process, it is difficult to ensure that users' information is protected and that the models are making unbiased decisions. Human-centered problems in finance, such as credit approvals and insurance underwriting, have seen an increase in the use of ML algorithms [1,2]. Although efficient, these use cases are highly susceptible to discriminatory decisions [3]. The existing research attempts to balance model performance with fairness and privacy guarantees; it faces challenges to accurately address every potential issue [4,5]. With official regulations such as the European General Data Protection Regulation [6] and

the California Consumer Privacy Act [7] becoming increasingly prevalent, organizations must adhere to strict requirements. This ultimately requires significant adjustments to the model training workflows. In general, various studies have highlighted the importance of making informed decisions when working with customer financial data, particularly concerning influencing loan decisions and unfair credit scoring models [1,8].

Instances where bias and data protection were not considered in financial-related ML tasks have led to inequality. Black applicants had their mortgage applications denied at twice the rate of white applicants [9]. Similarly, customer churn models have been found to make different predictions between men and women [10]. Regarding training data privacy, external adversaries use various attacks [11] to obtain information about individuals present in a training dataset [12]. These concerns underscore the need to develop comprehensive frameworks that deliver accurate predictive performance while also addressing fairness, privacy, and regulatory compliance in their output. To confront these concerns, research on privacy-preserving mechanisms, such as differential privacy [1,13], is applied to ensure that individual data contributions remain untraceable and that fairness is promoted throughout both the ML training and post-training processes. As various studies have demonstrated, we can integrate DP with existing models to address privacy and bias [14,15]. Both large language models and retrieval-augmented generation architectures have been individually explored in a DP context [16]. Both frameworks excel at understanding complex datasets; however, LLMs display significant challenges that raise concerns about their suitability for handling sensitive data, such as biases in pre-trained models, hallucinations, and a lack of explainability [17,18]. It is possible to combine RAG with LLM pattern recognition capabilities to address these shortcomings by reducing hallucinations and bias while maintaining high accuracy [19].

We explore the possibility of a customer churn prediction framework that ensures accuracy, addresses fairness metrics, and includes privacy-preserving features on the Bank Churn [20] and Telco Churn [21] customer datasets. Our method introduces a specific FairRAG model: an LLM with an RAG embedding layer trained with DP to generate a fair and privacy-protected synthetic dataset that can be used in a secure ML training pipeline. We build on previous research that leverages randomness in the next token generations of LLMs for DP guarantees [15]. The existing literature on LLMs focuses mainly on generating synthetic text or image data rather than tabular datasets [22–24]. Our framework builds on these methods by utilizing RAG to maintain a knowledge base of the data, enabling more structured application while ensuring the privacy and fairness of training data in downstream tasks.

The primary contributions of this work include the following:

- The FairRAG architecture enables the generation of synthetic private tabular data for downstream classification tasks while promoting fairness, thus allowing the optimization of the generation process using an RAG data repository.
- Extensive experiments on various churn datasets to evaluate how efficiently private synthetic data can replace real but sensitive data while improving the downstream performance in this domain.

This paper is structured as follows: Section 2 provides a detailed literature review that discusses the previous advances in LLMs and RAGs for synthetic data generation, as well as contextualizes churn prediction. Section 3 provides an overview of the FairRAG architecture. Section 4 highlights the FairRAG algorithm breakdown. Section 5 details our experimental setup. Section 6 articulates the research findings of the FairRAG architecture. Section 7 concludes the research. In Section 8, we describe prospective research pathways that could be further explored.

2. Related Work

2.1. Privacy-Preserving Machine Learning

Various techniques for handling privacy concerns and information leakage in ML workflows have been thoroughly explored [15,25,26]. In recent years, differential privacy [1] has emerged as the leading method for addressing such issues. ε-differential privacy is a quantifiable measure of privacy. ε governs these trade-offs. Formally, a mechanism M is said to be ε-differentially private if, for any two datasets differing by a single record and for every possible output set S, the probability of obtaining an output is bounded by

$$\Pr[M(D) \in S] \leq \exp(\varepsilon) \cdot \Pr[M(D') \in S] \tag{1}$$

where D and D' are adjacent datasets. Lower values of ε provide stronger privacy guarantees (at the expense of model utility), while higher values allow for improved accuracy with weaker privacy protection [27].

In classification tasks, researchers have explored combining DP with traditional models to address both privacy preservation and algorithmic fairness [14,28]. When it comes to deep learning, it is common for DP to be integrated into these models through one of three approaches: altering the input data as a preprocessing step [29–31], incorporating privacy during model training via methods such as differentially private stochastic gradient descent [26,32–34], or applying a post-processing approach that alters the model output [25,35]. Our approach aligns most closely with the second method by introducing Laplace noise during the gradient update step.

2.2. Privacy-Preserving Synthetic Data Generation

Synthetic data generation as a method for data security and classification fairness has been extensively studied and has shown promising results. Techniques such as Generative Adversarial Networks [36] and SMOTE [37] have been widely explored to create synthetic datasets that mirror the statistical properties of real-world data without revealing sensitive details. Some of these studies [38] have used GANs to generate synthetic data for sensitive financial contexts such as fraud detection. However, LLMs have demonstrated improved performance in generation and predictive tasks compared to GANs [39,40]. Notably, recent advances have focused on integrating differential privacy into these methods. For example, Google's work on protecting users with differentially private synthetic training data [15] introduces an approach that uses LLMs and DP mechanisms to generate synthetic training data. They pre-trained an 8B-parameter decoder-only LLM on public text and then privately fine-tuned it using LoRa and prompt-based tuning on disjoint sensitive datasets to generate synthetic data. They relied on a next token prediction framework:

$$p = [\text{TaskName}]\,[\text{LabelName}_y] \tag{2}$$

where $[\text{LabelName}_y]$ is set to [negative] if $y = 0$ and [positive] if $y = 1$. The training example is tokenized into a prefix $p = \{z_1, \ldots, z_k\}$ and a target sequence $x = \{z_{k+1}, \ldots, z_n\}$.

In their unique prefix–LM formulation, the weights for prefix tokens are set to zero (i.e., $\forall i \leq k : w_i = 0$); thus, they used a modified version of the weighted next token prediction cross-entropy loss:

$$L_{\text{PrefixLM}}(\vec{z}, \vec{w}, \theta) = -\sum_{i=k+1}^{n} z_i \log P(z_i \mid z_{<i}, \theta) \tag{3}$$

With this prefix–LM setup, the loss function is computed only over the target tokens, while the prefix p token distributions need not be learned. This study demonstrates that classifiers trained on private LLM-generated synthetic data can outperform those directly

trained on the original sensitive data with differential privacy. However, their prefix–LM approach is fundamentally designed for unstructured text generation and relies on privately fine-tuning a large model on the sensitive dataset. This process can be computationally expensive and complex to implement. Furthermore, their method generates new examples in isolation. Our proposed framework, FairRAG, in contrast, is designed for structured tabular data. By using retrieval-augmented generation, we leveraged the LLM's zero-shot reasoning capabilities, providing it with context from both original and private synthetic records at inference time. This allows our system to reason about a new customer by comparing them to similar privacy-protected historical cases, a paradigm distinct from the fine-tuning approach. Kibriya et al. explained that privacy concerns for LLMs need to be addressed in both the training and inference steps [41]. Furthermore, studies have shown the need to develop stronger methods that guarantee privacy and fairness in LLM generation [42,43]. However, when we include DP mechanisms in both these steps, we inevitably lose some important contextual information, which raises concerns in more complicated and conceptual domains.

2.3. Fairness in Machine Learning

The importance of promoting fairness in ML stems from research that displays its adverse effects on subgroups. Chouldechova et al. identified fairness issues as a direct result of inherent biases in the dataset [44]. In sensitive ML domains that involve the use of personal information to train a model, these systems are prone to fairness-related harms. For example, Vajpayee et al. observed that biases led to gender, racial, ethnic, and socioeconomic disparities [45].

Nagpal et al. presented a framework that aims to optimize accuracy with both group fairness and individual fairness [46]. Microsoft's Fairlearn toolkit introduces a practical toolset with several algorithmic approaches that address fairness [47]. Regarding research aimed at preserving fairness in machine learning and mitigating biases in sensitive features, several challenges persist. Achieving a balanced trade-off between fairness and accuracy remains difficult as even minor changes in the dataset can lead to significant fluctuations in this balance. Many of these methods rely on access to protected attributes or are unable to tackle a wide range of fairness metrics at a time [48,50**?**]. Other challenges include the absence of standardized measures to quantify fairness and the lack of reliable metrics to detect bias [2].

2.4. Retrieval-Augmented Generation Approaches with Large Language Models

The shortcomings of LLMs [51] need to be adequately considered if we hope to be able to use these models to generate meaningful information. This approach can address the issues that come with generating information in highly specific and complex domains. Gao also described three RAG paradigms: the Naive RAG, the Advanced RAG, and the Modular RAG. Each step is optimized incrementally, building upon the previous one. FairRAG most closely follows a Naive implementation, which proved sufficient for retrieval and generation in the churn prediction task using a structured dataset. This is detailed in Figure 1. Although the FairRAG architecture follows the "Naive RAG" pattern of retrieve-then-generate, its contribution lies in reconfiguring the retrieval pipeline itself. Instead of indexing and retrieving documents, we serialize individual customer data rows into textual profiles, embed them, and retrieve the most similar customer profiles. This demonstrates that the RAG concept can be effectively repurposed from its native text-based domain to provide contextual instance-based reasoning for tabular data tasks like churn prediction. Furthermore, some other advances in the use of LLMs and the RAG architecture include the work of Ma et al., who introduced the Rewrite–Retrieve–Read framework that uses RAGs

and LLMs for specific question-answering tasks [52]. The use of RAG as an enhancement to generative methods has also proven useful in financial services as they maintain credibility by providing outputs based on the economic facts and are more representative of such data sources [53].

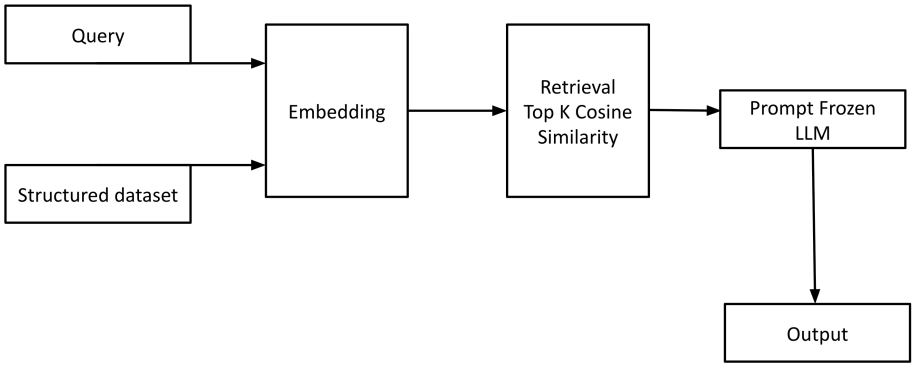

Figure 1. FairRAG adopts a Naive approach but reconfigures the retrieval pipeline to suit structured tabular data rather than unstructured documents. We use in-memory embeddings with direct similarity search in place of indexing.

2.5. Customer Churn Prediction

Customer churn prediction is a highly complex domain that requires a model with a deep understanding of the information presented. Huang et al. [54] developed rule-based methods for telecommunication services, establishing early frameworks for interpretable churn prediction. Jafari et al. [55] introduced an interpretable machine learning framework for churn prediction that combines multiple algorithms. Due to the nature of sensitive information in churn prediction problems, privacy-aware solutions are of vital importance. Fahmy et al. [56] explored federated learning to protect customer privacy while enabling collaborative model training across institutions. As DP remains the most promising method, further research is needed to understand its performance in churn prediction. The privacy guarantees of federated learning can be complex to analyze and are vulnerable to certain attacks. Meanwhile, our FairRAG offers a different privacy–utility trade-off as it operates on a centralized (but privacy-protected) dataset, making it simpler to deploy in a single-organization context.

2.6. Uncertainty for Trustworthy Decision-Making

The deployment of machine learning algorithms in high-stakes industries such as finance, healthcare, and automobiles implies the need for a rigorous understanding of model uncertainty [57]. No matter how accurate a model's predictions are on average, they offer limited practical value without an accompanying measure of confidence. Thus, it is vital to understand and quantify uncertainty. Several methods have been proposed to estimate these uncertainties. Bayesian Neural Networks provide a principled approach by learning distributions over model weights, but they are computationally expensive [58]. Ensemble methods, where multiple models are trained independently and their prediction variance is used as an uncertainty measure, are a technique used significantly [59]. For systems utilizing large language models, uncertainty is often implicitly present in the probability distribution of the output tokens. Researchers have explored methods to calibrate these probabilities and directly probe what a model "knows" versus when it is "guessing" [60]. However, uncertainty in such systems also arises from the fusion of conflicting evidence and is a core problem in formal risk assessment. For instance, in Failure Mode and Effects Analysis, methodologies are developed to reconcile disagreements among multiple human experts to produce a reliable risk score. Tang et al. [61] used Dempster–Shafer evidence

theory to formally manage and fuse highly conflicting assessments from different experts when analyzing aircraft systems.

3. Proposed Methodology: FairRAG Architecture

The FairRAG architecture combines differential privacy, retrieval-augmented generation, and large language models to create a privacy-preserving churn prediction system as shown in Figure 2. The pipeline begins with data preprocessing of the customer dataset, handling missing values and standardizing numerical features. If class imbalance is detected, SMOTE is applied to create a balanced dataset. The system then generates synthetic data using differential privacy algorithms, adding calibrated Laplace noise to numerical features while preserving data utility and privacy guarantees. At the core of FairRAG is the OPT-125M language model, which works in conjunction with a sentence transformer (all-MiniLM-L6-v2) for semantic similarity matching. Semantic similarity is a well-established technique that was introduced years ago and has been applied in numerous text processing applications [62]. The RAG component maintains a knowledge base of both original and synthetic customer profiles, retrieving similar cases to provide context for predictions. When making predictions, FairRAG creates enhanced prompts that combine the current customer profile with relevant historical cases, enabling the model to make informed predictions.

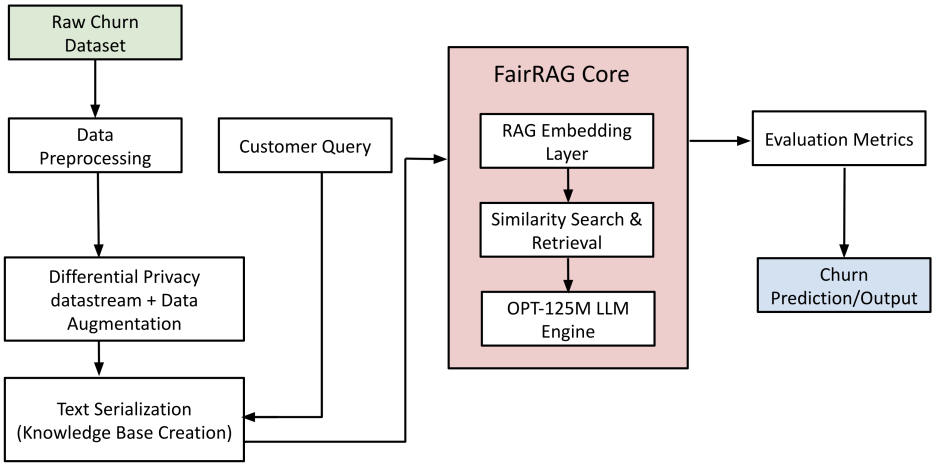

Figure 2. Proposed FairRAG architecture.

4. Algorithm Breakdown

FairRAG (Algorithm 1) operates in two phases to predict churn while preserving privacy: the construction of knowledge bases and the prediction augmented with retrieval.

4.1. Phase 1: Knowledge Base Construction

This phase applies differential privacy to create $\mathcal{D}_{dp}$, then combines it with the original dataset to form an augmented corpus $\mathcal{D}_{aug}$. Each customer record is serialized, encoded into dense vector representations, and indexed for efficient similarity search.

4.2. Phase 2: Retrieval-Augmented Prediction

The system serializes the profile into text for new customer queries and embeds it using the same encoding model. A similarity search retrieves the k most similar profiles from the knowledge base. These retrieved examples are incorporated into a prompt that guides the language model to generate a churn prediction.

Algorithm 1. Complete FairRAG Algorithm with Preprocessing and Uncertainty

▷ `Phase 0: Data Preprocessing and Balancing`

Function `PreprocessData`($\mathcal{D}_{raw}$):
  Input : Raw churn dataset $\mathcal{D}_{\text{raw}}$
  Output: Preprocessed dataset $\mathcal{D}_{\text{processed}}$
  $\mathcal{D}_{\text{clean}} \leftarrow \text{HandleMissingValues}(\mathcal{D}_{\text{raw}})$
  if *ClassImbalanceDetected*($\mathcal{D}_{clean}$) **then**
    $\mathcal{D}_{\text{processed}} \leftarrow \text{ApplySMOTE}(\mathcal{D}_{\text{clean}})$
  end
  else
    $\mathcal{D}_{\text{processed}} \leftarrow \mathcal{D}_{\text{clean}}$
  end
  return $\mathcal{D}_{processed}$

▷ `Phase 1: Build Privacy-Preserving Knowledge Base`

Function `CreateKnowledgeBase`($\mathcal{D}_{processed}, \epsilon, \delta$):
  Input : Processed dataset $\mathcal{D}_{\text{processed}}$, privacy parameters ϵ, δ
  Output: Knowledge Base KB
  $\mathcal{D}_{dp} \leftarrow \text{ApplyDifferentialPrivacy}(\mathcal{D}_{\text{processed}}, \epsilon, \delta)$
  $\mathcal{D}_{aug} \leftarrow \mathcal{D}_{\text{processed}} \cup \mathcal{D}_{dp}$
  $\mathcal{T}_{\text{corpus}} \leftarrow \emptyset$
  foreach *row* $r_i \in \mathcal{D}_{aug}$ **do**
    $t_i \leftarrow \text{SerializeToText}(r_i)$
    $\mathcal{T}_{\text{corpus}} \leftarrow \mathcal{T}_{\text{corpus}} \cup \{t_i\}$
  end
  $\mathcal{V} \leftarrow M_{\text{Enc}}.\text{encode}(\mathcal{T}_{\text{corpus}})$
  $KB \leftarrow \text{BuildVectorIndex}(\mathcal{V}, \mathcal{T}_{\text{corpus}})$
  return KB

▷ `Phase 2: Uncertainty-Aware Prediction`

Function `FairRAG_Predict`(q, KB, k):
  Input : Query profile q, Knowledge Base KB, neighbors k
  Output: Prediction y_{pred} and uncertainty U_{total}
  ▷ `Serialize and embed query`
  $q_{\text{text}} \leftarrow \text{SerializeToText}(q)$
  $v_q \leftarrow M_{\text{Enc}}.\text{encode}(q_{\text{text}})$
  ▷ `Retrieve similar profiles with similarity scores`
  $(C_{\text{context}}, S_{\text{scores}}) \leftarrow KB.\text{search\_with\_scores}(v_q, k)$
  ▷ `Generate prediction via LLM`
  $P_{\text{aug}} \leftarrow \text{BuildPrompt}(q_{\text{text}}, C_{\text{context}})$
  $y_{\text{raw}} \leftarrow M_{\text{LLM}}.\text{generate}(P_{\text{aug}})$
  $y_{\text{pred}} \leftarrow \text{ParseOutput}(y_{\text{raw}})$
  ▷ `Compute uncertainty measures`
  $U_{\text{retrieval}} \leftarrow 1 - \text{mean}(S_{\text{scores}})$
  $U_{\text{confidence}} \leftarrow \text{ExtractLLMConfidence}(y_{\text{raw}})$
  $U_{\text{total}} \leftarrow \text{CombineUncertainties}(U_{\text{retrieval}}, U_{\text{confidence}})$
  return (y_{pred}, U_{total})

▷ `Main Pipeline`

Function *FairRAGPipeline*$\mathcal{D}_{raw}, \epsilon, \delta, k$:
  $\mathcal{D}_{\text{processed}} \leftarrow$ `PreprocessData`$(\mathcal{D}_{\text{raw}})$
  $KB \leftarrow$ `CreateKnowledgeBase`$(\mathcal{D}_{\text{processed}}, \epsilon, \delta)$
  ▷ `Ready for online predictions using FairRAG`
  return KB

4.3. Uncertainty Quantification in FairRAG

The uncertainty in FairRAG's predictions originates from several components of the framework. Retrieval ambiguity occurs when the top(k) similar profiles C_{context} contain conflicting results (for example, a mix of "Churn" and "No Churn" labels), indicating high predictive uncertainty for the query profile (q). Furthermore, model-level uncertainty stems from the LLM (M_{LLM}) as a probabilistic model that may express low confidence by assigning similar probabilities to competing outcomes.

To quantify this uncertainty, we calculate the confidence score for each prediction, which is LLM confidence, measured as the normalized probability assigned by (M_{LLM}) to the predicted output token. This confidence-based approach leverages the explainability of the RAG architecture to identify cases where predictions may be less reliable. Also, to ensure a reliable and consistent response, we maintain constant temperature and sampling parameters throughout the analysis, preventing the LLM from generating varied responses that could introduce additional uncertainty beyond what we aim to measure.

5. Experimental Setup

To evaluate the performance of our framework, we conducted experiments on two benchmark datasets: the Bank Churn dataset and the Telco Churn dataset. This section outlines the details of the datasets, preprocessing steps, training setup, and evaluation metrics.

5.1. Datasets

The datasets used in our experiments contain customer-related features and churn labels. Below is a brief description of each dataset:

- Bank Churn Dataset: Contains customer demographics, account information, and transaction history to predict whether a customer will churn.
- Telco Churn Dataset: Comprises customer service data, contract details, and billing information to assess the likelihood of a customer discontinuing the service.

All datasets contain both numerical and categorical features. Categorical features were encoded using one-hot encoding, while numerical features were standardized to ensure a uniform scale.

5.2. Preprocessing

Prior to training, we applied various preprocessing steps. First, missing values were handled using mean imputation for numerical features and mode imputation for categorical features. Then, we applied feature scaling using Min–Max normalization to improve convergence during training. The data was split into training (70%), validation (15%), and testing (15%) sets. Finally, synthetic data was generated using differential privacy techniques to evaluate model robustness.

5.3. Training Setup

We trained multiple machine learning models, including Logistic Regression, Random Forest, XGBoost, Decision Tree, Support Vector Machines, and k-Nearest Neighbors. For large language models, we tested TinyLlama, DistilBERT, GPT-2, Phi-2, All-MiniLM, and OPT-125M. The models' hyperparameters were fine-tuned using grid search and cross-validation to optimize performance and trained using four A100 GPUs for LLMs manufactured by Nvidia, Santa Clara, CA, USA and a standard CPU-based environment for classical models.

5.4. Evaluation Metrics

Finally, the models were assessed based on the accuracy, precision, recall, and F1-score. This experimental setup ensures a fair and rigorous evaluation of our proposed framework across diverse datasets and model architectures.

5.5. Large Language Models Explored

For our research, we utilized several open-source large language models, including TinyLlama, DistilBERT, All-MiniLM, GPT-2, Phi-2, and OPT-125M.

5.5.1. GPT-2

GPT-2 is a generative pre-trained transformer model with 124 million parameters in its miniature version. It was trained on 40 GB of text from the Internet (WebText dataset). It employs an autoregressive transformer architecture and is recognized for its robust text generation capabilities.

5.5.2. Phi-2

Phi-2 is a transformer-based model with 2.7 billion parameters, designed to achieve strong reasoning and language-understanding capabilities.

5.6. All-MiniLM

It is a sentence embedding model designed to map sentences and paragraphs to a 384-dimensional dense vector space. It is a distilled version of a BERT-like model with 6 transformer layers and approximately 22.7 million parameters, thereby making it significantly smaller and faster than larger BERT-based models while retaining good performance, especially useful for information retrieval tasks.

5.6.1. DistilBERT

DistilBERT contains 66 million parameters, which is 40% fewer than the BERT base while retaining 97% of BERT's language understanding capabilities. This model was developed using knowledge distillation techniques in the same corpus as the BERT model, and this model runs 60% faster than its parent model, making it suitable for applications with computational statistics.

5.6.2. TinyLlama

This compact language model is designed for efficiency and is built on the LlAMA-2 architecture. This model has 1.1 billion parameters and was trained on 3 trillion tokens. It was designed to deliver efficient performance for various downstream machine learning tasks in a resource-constrained environment.

5.6.3. OPT-125M

The Open Pre-trained Transformer model belongs to the family of decoder-only pre-trained transformers, with 125 million parameters, and was trained on a large and diverse corpus, including publicly available datasets such as The Pile and the PushShift.io Reddit dataset.

5.7. Hyperparameter Tuning for Large Language Models

We conducted systematic hyperparameter optimization using grid search across key parameters for fine-tuning LLMs. The process included

- Learning Rate: Searched over [1×10^{-5}, 5×10^{-5}, 1×10^{-4}] for fine-tuning.
- Batch Size: Evaluated [8, 16, 32] based on GPU memory constraints.
- Max Sequence Length: Optimized between [256, 512] tokens depending on model.

- Temperature: Fixed at 0.1 for constant prediction generation.
- Number of Epochs: Early stopping with patience of 3 epochs to prevent overfitting.

Model selection was performed using 5-fold cross-validation on the training set, with the best hyperparameters determined by maximizing the F1-score on the validation set.

6. Results

6.1. Bank Churn Dataset

Table 1 compares the performance for the Bank Churn dataset regarding six machine learning models—Logistic Regression, Random Forest, XGBoost, Decision Tree, SVM, and KNN—using metrics such as accuracy, precision, recall, and F1-score. Among these models, Random Forest achieves the highest accuracy, indicating that it correctly predicts the most instances overall. However, KNN performed the worst in terms of various evaluation metrics.

Table 1. Performance metrics for various classifiers—Bank Churn dataset.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.69	0.36	0.66	0.48
Random Forest	0.75	0.42	0.56	0.48
XGBoost	0.73	0.40	0.48	0.44
Decision Tree	0.67	0.30	0.47	0.37
SVM	0.70	0.37	0.66	0.47
KNN	0.66	0.32	0.58	0.41

Figure 3 shows the ROC curves, where Decision Tree and KNN exhibit poor performance while the remaining models demonstrate more consistent standard behavior.

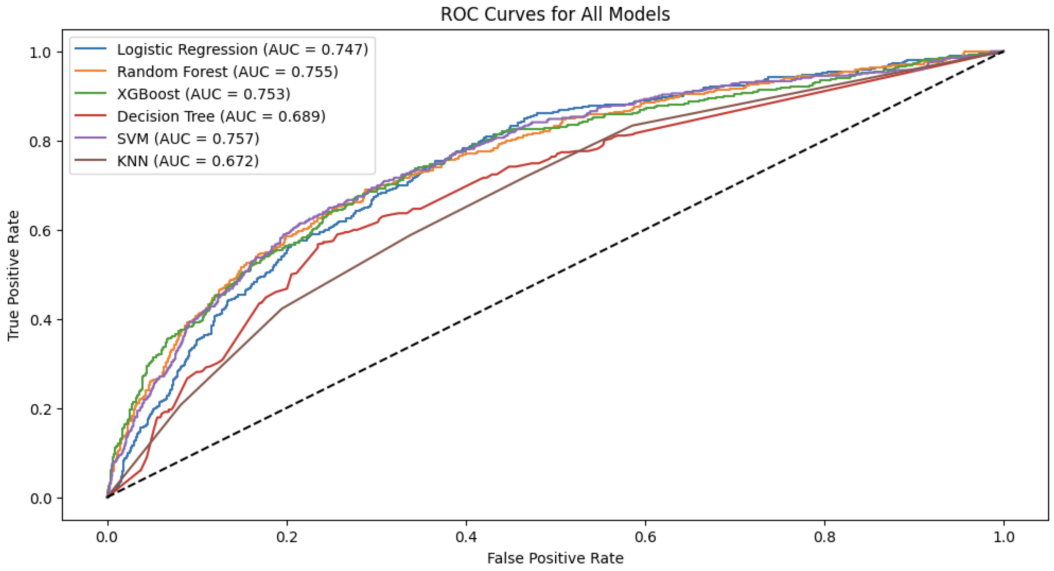

Figure 3. AUC ROC curves for different models—Bank Churn dataset.

Table 2 presents a comparison of several transformer-based and large language models along with our proposed approach, FairRAG. Among the baseline models, OPT-125M achieved the highest accuracy, indicating strong overall performance. Meanwhile, FairRAG outperformed the baseline models by 11%, achieving a maximum accuracy of 86% and well-balanced results across all the other evaluation metrics.

Table 2. Comparison of evaluation metrics across transformer-based models on Bank Churn dataset.

Model	Accuracy	Precision	Recall	F1-Score
GPT-2	0.75	0.39	0.39	0.39
Phi-2	0.75	0.36	0.34	0.35
All-MiniLM	0.76	0.54	0.65	0.59
DistilBERT	0.75	0.36	0.36	0.36
TinyLlama	0.76	0.54	0.64	0.58
OPT-125M	0.77	0.56	0.68	0.61
FairRAG	0.86	0.60	0.63	0.62

6.2. Telco Churn Dataset

In the case of the Telco Churn dataset, Table 3 presents the evaluation metrics for various classification models, with Random Forest demonstrating the highest accuracy, suggesting its effectiveness in identification.

Table 3. Performance metrics for various classifiers—Telco Churn dataset.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.71	0.50	0.80	0.61
Random Forest	0.72	0.53	0.77	0.63
XGBoost	0.67	0.54	0.74	0.63
Decision Tree	0.62	0.48	0.76	0.59
SVM	0.64	0.51	0.77	0.62
KNN	0.64	0.46	0.73	0.60

Figure 4 shows the AUC–ROC values indicating the performance of various classifiers on the Telco Churn dataset.

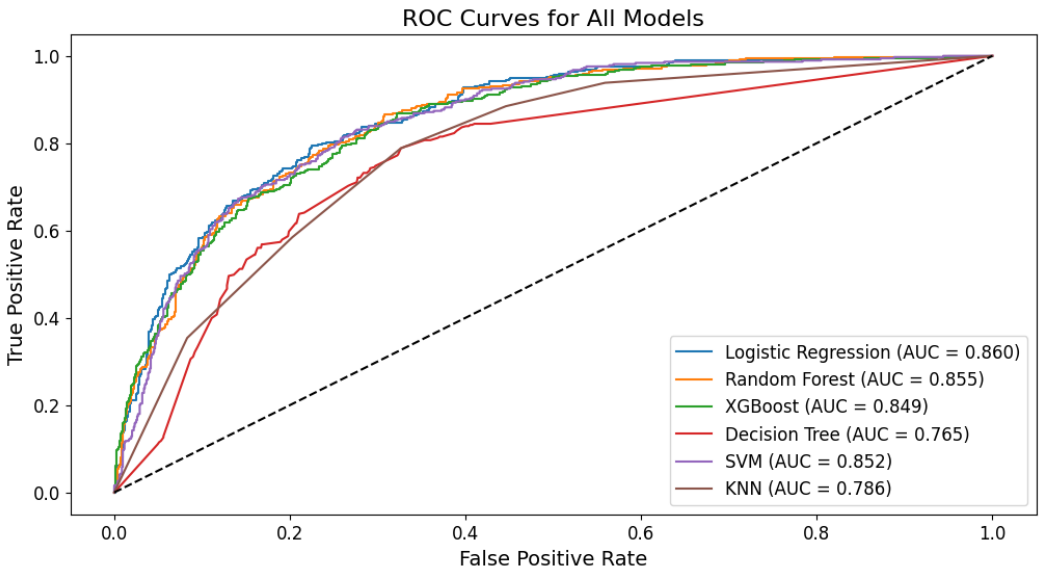

Figure 4. AUC ROC curves for different models—Telco Churn dataset.

Table 4 presents a comparative evaluation of various language models on the Bank Churn dataset. Among the baseline models, OPT-125M demonstrates the strongest overall performance, achieving the highest accuracy and F1-score, indicative of its balanced precision and recall as well as its ability to generalize effectively. Models like TinyLlama and GPT-2, despite achieving high recall, exhibit lower precision and F1-scores. Our proposed approach, FairRAG, built on OPT-125M, shows a substantial improvement in

accuracy, boosting performance by 12% over its base model. This signifies the effectiveness of FairRAG in enhancing decision-making fairness and predictive accuracy in real-world financial datasets.

Table 4. Performance comparison of language models on the Telco Churn dataset, with and without FairRAG integration.

Model	Accuracy	Precision	Recall	F1-Score
GPT-2	0.71	0.48	0.81	0.60
Phi-2	0.76	0.57	0.65	0.61
All-MiniLM	0.76	0.54	0.67	0.60
DistilBERT	0.75	0.53	0.59	0.56
TinyLlama	0.73	0.49	0.82	0.61
OPT-125M	0.80	0.57	0.61	0.58
FairRAG	0.84	0.63	0.60	0.61

6.3. Space Complexity of FairRAG

The computational complexity of FairRAG can be analyzed regarding its key components. The preprocessing and SMOTE steps have a complexity of $O(n)$, where n is the number of samples. The differential generation of synthetic data for privacy requires $O(mn)$ operations, where m is the number of numerical features. The complexity of the RAG component is primarily driven by similarity search, which requires $O(kd)$ operations per query, where k represents the number of similar cases retrieved and d denotes the embedding dimension. The sentence transformer encoding has complexity $O(sl)$, where s is the sequence length. The LLM inference (OPT-125M) has complexity $O(l^2)$ due to self-attention, where l is the input length. Therefore, for a batch of b predictions, the overall complexity is $O(n + mn + b(kd + sl + l^2))$, with the transformer operations typically being the computational bottleneck.

6.4. Hardware Configuration

All our experiments were conducted using a hybrid computing infrastructure of Google GPUs (g2-standard-24 machine type, NVIDIA L4, 24 vCPU count, and 96GB virtual memory) and MIT Supercloud GPUs. We implemented the code in Python 3.11.5 using Visual Studio as the primary development environment, with additional experimentation, rapid prototyping, and testing performed in Google Colab notebooks. The standalone LLMs (GPT-2, Phi-2, All-MiniLM, DistilBERT, TinyLlama, and OPT-125M) were deployed and evaluated on Google GPUs to utilize their optimized performance. The FairRAG system, which combines the OPT-125M language model with the all-MiniLM-L6-v2 sentence transformer for semantic similarity matching, was implemented on the MIT Supercloud GPU cluster. This distributed setup allowed us to efficiently perform parallel processing of synthetic data generation, knowledge base retrieval, and the churn prediction task.

6.5. Critical Analysis: Benefits and Constraints of FairRAG

FairRAG offers a unique combination of features that address several concurrent challenges in trustworthy AI: privacy, interpretability, and computational pragmatism.

6.5.1. Benefits of FairRAG

- Architectural Simplicity and Computational Efficiency: A primary advantage of FairRAG is its design, which avoids the need for expensive private fine-tuning of large models. By leveraging a pre-trained smaller LLM (OPT-125M) and applying differential privacy at the data generation stage, the architecture remains lightweight. This contrasts sharply with methods like DP-SGD, which are applied to multi-billion-

parameter models, requiring significant computational resources for training and complex privacy budget accounting. Our approach, being approximately 64 times smaller in parameter count than models like Google's PaLM-8B used in comparable DP synthetic data, makes privacy-preserving LLM applications more accessible.

- Decoupled Privacy Preservation: FairRAG integrates differential privacy natively by treating the private synthetic dataset as a static safe-to-use asset within its knowledge base. This "decoupling" of the privacy mechanism from the core predictive model is a key architectural benefit. It consumes information from a pre-sanitized source, allowing the use of powerful off-the-shelf LLMs while still benefiting from formal privacy guarantees (ε, δ) on the underlying data.
- Inherent Interpretability through Retrieval: Unlike opaque black-box models, FairRAG provides a clear and intuitive form of local instance-based explainability. When making a prediction for a new customer, the RAG mechanism retrieves the most similar existing profiles from its knowledge base. These retrieved examples serve as the rationale for the final prediction. This allows a human analyst to directly inspect the context the LLM used, building trust and making it possible to audit and understand individual outcomes, a critical requirement for high-stakes decisions such as churn management.
- Integrated Dual-Source Uncertainty Quantification: The framework provides a holistic measure of uncertainty than traditional models. It uniquely combines two distinct signals:
  - Retrieval Uncertainty which quantifies the quality of the retrieved context by measuring the dissimilarity of neighboring profiles.
  - Generative Uncertainty, which captures the LLM's confidence in its own output. By fusing these, FairRAG can distinguish between predictions that are uncertain due to ambiguous input (poor context) and those that are uncertain because the model itself is struggling to reason, providing a richer signal for trustworthy decision-making.

6.5.2. Limitations of FairRAG

Despite these advantages, it is crucial to acknowledge the constraints of the current FairRAG implementation to guide future work.

- Generalizability and Scalability Constraints: Our evaluation was conducted on two structured tabular datasets within the churn domain. The framework's effectiveness has not yet been validated on more complex multi-modal datasets (e.g., including unstructured text) or in different domains, such as healthcare. Moreover, while the in-memory vector search is sufficient for our datasets, scaling the knowledge base to millions of enterprise-level records would necessitate more sophisticated vector database solutions (e.g., FAISS and Milvus) to maintain efficient retrieval performance.
- Incomplete Privacy–Utility Analysis: Although we have successfully implemented differential privacy, this work does not present a comprehensive sweep across a range of privacy budgets ε to fully map the privacy–utility trade-off curve. The optimal balance is application-dependent and requires further study.
- Gaps in Formal Fairness Evaluation: A significant limitation is the absence of a rigorous fairness audit. The name "FairRAG" reflects an aspiration, but the current work does not formally measure or mitigate algorithmic bias. We have not evaluated the model's performance equity across demographic subgroups using established fairness metrics (e.g., Demographic Parity, Equalized Odds). This remains the most critical next step before the system could be considered for deployment in fairness-critical applications.

7. Conclusions

In this paper, we introduce FairRAG, a robust architecture that makes a primary contribution by providing a new blueprint for applying large language models to structured tabular data in a privacy-preserving and interpretable manner. The novelty of our approach lies in creating a system with capabilities that no existing method offers in combination. Specifically, against the backdrop of prior work, FairRAG makes three distinct contributions. First, while conventional retrieval-augmented generation systems are designed for unstructured text documents, we successfully repurpose the paradigm for structured tabular data by serializing customer profiles into a retrievable knowledge base. Second, in contrast to common privacy-preserving machine learning techniques that directly modify model training with methods like DP-SGD, FairRAG decouples the privacy mechanism from the model training process. By applying differential privacy to the data generation stage, we create a safe static knowledge base that can be queried by a standard pre-trained LLM, avoiding the high computational cost and complexity of the private fine-tuning process. Our experiments on the Bank Churn and Telco Churn datasets validate the effectiveness of this architecture. FairRAG outperformed the traditional machine learning baselines, demonstrating strong balance between privacy and utility. It achieved high predictive accuracy while operating under formal differential privacy guarantees. In summary, FairRAG presents a significant step forward by offering a computationally efficient and privacy-preserving solution. However, we acknowledge key limitations, particularly the need for a rigorous fairness audit that provides a robust and generalizable framework to effectively balance the often competing demands of accuracy, privacy, and explainability in modern machine learning systems.

8. Future Work

The architecture described above could be extended to incorporate multi-modal data by integrating vision–language models to handle customer interaction images or documents alongside textual data. The system's efficiency could be improved by implementing approximate nearest neighbor search for the RAG component and exploring quantization techniques for the language model. Additionally, the development of an adaptive privacy budget allocation mechanism could help to optimize the trade-off between privacy and utility in the generation of synthetic data. We plan to work on the crucial next step, which is to incorporate a comprehensive fairness evaluation. While our current work establishes a baseline, a deeper analysis against a broader suite of fairness metrics, such as Demographic Parity and Equalized Odds, will provide a more granular understanding of the model's performance across different demographic subgroups and ensure that its predictions do not disproportionately impact protected groups. Furthermore, we plan to conduct a more rigorous analysis of the privacy–utility trade-off inherent in generating differentially private synthetic data. This investigation will focus on quantifying the quality of the synthetic data, measuring its statistical similarity to the original data, and evaluating its utility through training on synthetic data and testing on real data. Also, to further validate the novelty and effectiveness of our approach, we plan to benchmark FairRAG against other state-of-the-art privacy-preserving generative models. This would include a direct comparison with methods like privacy-preserving Generative Adversarial Networks specialized for tabular data.

Author Contributions: Conceptualization, R.N. and U.U.; methodology, R.N. and U.U.; implementation, R.N.; validation, R.N. and U.U.; investigation, R.N. and U.U.; writing—original draft preparation, R.N., U.U., and A.G.; writing—review and editing, R.N., U.U., R.P., and A.G.; visualization, R.N.; supervision, R.P. and A.G.; project administration, R.N.; funding acquisition, A.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work is a part of MIT internal funding, which is partially funded by two industry alliances at MIT Computer Science and Artificial Intelligence Lab (CSAIL): Future of Data consortium and FinTechAI consortium.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We used IBM Telco Churn dataset for our experiments, which is freely available at https://www.ibm.com/docs/en/cognos-analytics/12.1.0?topic=samples-telco-customer-churn (accessed on 3 July 2025), and Bank Churn dataset, which is also open-source and publicly available at https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers (accessed on 3 July 2025).

Acknowledgments: We want to thank Google Research for providing GPUs for conducting our research and experiments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends® Theor. Comput. Sci.* **2014**, *9*, 211–407. [CrossRef]
2. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* **2021**, *54*, 1–35. [CrossRef]
3. Cummings, R.; Gupta, V.; Kimpara, D.; Morgenstern, J. On the Compatibility of Privacy and Fairness. In Proceedings of the Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, UMAP'19 Adjunct, Larnaca, Cyprus, 9–12 June 2019; pp. 309–315. [CrossRef]
4. Kearns, M.; Roth, A. *Ethical Algorithm Design: Guiding Principles for Technology Regulation*; Technical Report; Brookings Institution: Washington, DC, USA, 2020.
5. Zhang, H.; Hartvigsen, T.; Ghassemi, M. Algorithmic Fairness in Chest X-ray Diagnosis: A Case Study. MIT Case Studies in Social and Ethical Responsibilities of Computing. 2023. Available online: https://mit-serc.pubpub.org/pub/algorithmic-chest (accessed on 3 July 2025).
6. Hoofnagle, C.J.; Van Der Sloot, B.; Borgesius, F.Z. The European Union general data protection regulation: What it is and what it means. *Inf. Commun. Technol. Law* **2019**, *28*, 65–98. [CrossRef]
7. Harding, E.L.; Vanto, J.J.; Clark, R.; Hannah Ji, L.; Ainsworth, S.C. Understanding the scope and impact of the california consumer privacy act of 2018. *J. Data Prot. Priv.* **2019**, *2*, 234–253. [CrossRef]
8. Barocas, S.; Hardt, M.; Narayanan, A. *Fairness and Machine Learning: Limitations and Opportunities*; MIT Press: Cambridge, MA, USA, 2023.
9. Bhutta, N.; Hizmo, A.; Ringo, D. How much does racial bias affect mortgage lending? Evidence from human and algorithmic credit decisions. *J. Financ.* **2025**, *80*, 1463–1496. [CrossRef]
10. Maw, M.; Haw, S.C.; Ho, C.K. Utilizing data sampling techniques on algorithmic fairness for customer churn prediction with data imbalance problems. *F1000Research* **2022**, *10*, 988. [CrossRef]
11. Bourou, S.; El Saer, A.; Velivassaki, T.H.; Voulkidis, A.; Zahariadis, T. A review of tabular data synthesis using GANs on an IDS dataset. *Information* **2021**, *12*, 375. [CrossRef]
12. Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; Lin, Z. When machine learning meets privacy: A survey and outlook. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–36. [CrossRef]
13. Liu, H.; Wu, Z.; Zhou, Y.; Peng, C.; Tian, F.; Lu, L. Privacy-Preserving Monotonicity of Differential Privacy Mechanisms. *Appl. Sci.* **2018**, *8*, 2081. [CrossRef]
14. Ghoukasian, H.; Asoodeh, S. Differentially private fair binary classifications. In Proceedings of the 2024 IEEE International Symposium on Information Theory (ISIT), Athens, Greece, 7–12 July 2024; pp. 611–616. [CrossRef]

15. Kurakin, A.; Ponomareva, N.; Syed, U.; MacDermed, L.; Terzis, A. Harnessing large-language models to generate private synthetic text. *arXiv* **2024**, arXiv:2306.01684. [CrossRef]
16. Kalodanis, K.; Papadopoulos, S.; Feretzakis, G.; Rizomiliotis, P.; Anagnostopoulos, D. SecureLLM: A Unified Framework for Privacy-Focused Large Language Models. *Appl. Sci.* **2025**, *15*, 4180. [CrossRef]
17. Chen, Y.; Liu, Y.; Yan, J.; Bai, X.; Zhong, M.; Yang, Y.; Yang, Z.; Zhu, C.; Zhang, Y. See what llms cannot answer: A self-challenge framework for uncovering llm weaknesses. *arXiv* **2024**, arXiv:2408.08978. [CrossRef]
18. Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. A survey on llm-as-a-judge. *arXiv* **2024**, arXiv:2411.15594. [CrossRef]
19. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv* **2023**, arXiv:2312.10997.
20. Repository, U.M.L. Bank Customer Churn Prediction Dataset. 2018. Available online: https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers (accessed on 3 July 2025).
21. Corporation, I. Telco Customer Churn Dataset. 2019. Available online: https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=samples-telco-customer-churn (accessed on 5 June 2025).
22. Veselovsky, V.; Ribeiro, M.H.; Arora, A.; Josifoski, M.; Anderson, A.; West, R. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv* **2023**, arXiv:2305.15041. [CrossRef]
23. Xie, C.; Lin, Z.; Backurs, A.; Gopi, S.; Yu, D.; Inan, H.A.; Nori, H.; Jiang, H.; Zhang, H.; Lee, Y.T.; et al. Differentially private synthetic data via foundation model apis 2: Text. *arXiv* **2024**, arXiv:2403.01749. [CrossRef]
24. Lin, Z.; Gopi, S.; Kulkarni, J.; Nori, H.; Yekhanin, S. Differentially private synthetic data via foundation model apis 1: Images. *arXiv* **2023**, arXiv:2305.15560.
25. Nasr, M.; Shokri, R.; Houmansadr, A. Improving deep learning with differential privacy using gradient encoding and denoising. *arXiv* **2020**, arXiv:2007.11524. [CrossRef]
26. Li, J.; Zhang, F.; Guo, Y.; Li, S.; Wu, G.; Li, D.; Zhu, H. A privacy-preserving online deep learning algorithm based on differential privacy. In Proceedings of the 2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Rio de Janeiro, Brazil, 24–26 May 2023; pp. 559–564. [CrossRef]
27. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating noise to sensitivity in private data analysis. In Proceedings of the Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, 4–7 March 2006; Proceedings 3; Springer: Berlin/Heidelberg, Germany, 2006; pp. 265–284. [CrossRef]
28. Triastcyn, A.; Faltings, B. Bayesian differential privacy for machine learning. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 9583–9592. [CrossRef]
29. Palacios, R.; Gupta, A.; Wang, P.S. Feedback-based architecture for reading courtesy amounts on checks. *J. Electron. Imaging* **2003**, *12*, 194–202. [CrossRef]
30. Mo, R.; Liu, J.; Yu, W.; Jiang, F.; Gu, X.; Zhao, X.; Liu, W.; Peng, J. A differential privacy-based protecting data preprocessing method for big data mining. In Proceedings of the 2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), Rotorua, New Zealand, 5–8 August 2019; pp. 693–699. [CrossRef]
31. Cummings, R.; Durfee, D. Individual sensitivity preprocessing for data privacy. In Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Salt Lake City, UT, USA, 5–8 January 2020; pp. 528–547. [CrossRef]
32. Song, S.; Chaudhuri, K.; Sarwate, A.D. Stochastic gradient descent with differentially private updates. In Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing, Austin, TX, USA, 3–5 December 2013; pp. 245–248. [CrossRef]
33. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; pp. 308–318. [CrossRef]
34. Kulynych, B.; Gomez, J.F.; Kaissis, G.; du Pin Calmon, F.; Troncoso, C. Attack-Aware Noise Calibration for Differential Privacy. In *Advances in Neural Information Processing Systems*; Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 134868–134901.
35. Zhao, Y.; Zhang, K.; Gao, L.; Chen, J. Privacy and Fairness Analysis in the Post-Processed Differential Privacy Framework. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 2412–2423. [CrossRef]
36. Ramzan, F.; Sartori, C.; Consoli, S.; Reforgiato Recupero, D. Generative adversarial networks for synthetic data generation in finance: Evaluating statistical similarities and quality assessment. *AI* **2024**, *5*, 667–685. [CrossRef]
37. Mukherjee, M.; Khushi, M. SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features. *Appl. Syst. Innov.* **2021**, *4*, 18. [CrossRef]

38. Charitou, C.; Dragicevic, S.; d'Avila Garcez, A. Synthetic Data Generation for Fraud Detection using GANs. *arXiv* **2021**, arXiv:2109.12546. [CrossRef]
39. Miletic, M.; Sariyar, M. Assessing the potentials of LLMs and GANs as state-of-the-art tabular synthetic data generation methods. In *International Conference on Privacy in Statistical Databases* Springer: Cham, Switzerland, 2024; pp. 374–389. [CrossRef]
40. Pawade, P.; Kulkarni, M.; Naik, S.; Raut, A.; Wagh, K. Efficiency Comparison of Dataset Generated by LLMs using Machine Learning Algorithms. In Proceedings of the 2024 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 5–7 March 2024; pp. 1–6. [CrossRef]
41. Kibriya, H.; Khan, W.Z.; Siddiqa, A.; Khan, M.K. Privacy issues in large language models: A survey. *Comput. Electr. Eng.* **2024**, *120*, 109698. [CrossRef]
42. Wu, X.; Duan, R.; Ni, J. Unveiling security, privacy, and ethical concerns of ChatGPT. *J. Inf. Intell.* **2024**, *2*, 102–115. [CrossRef]
43. Brown, H.; Lee, K.; Mireshghallah, F.; Shokri, R.; Tramèr, F. What does it mean for a language model to preserve privacy? In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 21–24 June 2022; pp. 2280–2292. [CrossRef]
44. Chouldechova, A.; Roth, A. The Frontiers of Fairness in Machine Learning. *arXiv* **2018**, arXiv:1810.08810. [CrossRef]
45. Vajpayee, A.S.; Khobragade, D. The Problem of Data Bias In Healthcare AI. In Proceedings of the 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI), Wardha, India, 29–30 November 2024; pp. 1–6. [CrossRef]
46. Nagpal, R.; Khan, A.; Borkar, M.; Gupta, A. A Multi-Objective Framework for Balancing Fairness and Accuracy in Debiasing Machine Learning Models. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 2130–2148. [CrossRef]
47. Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; Walker, K. *Fairlearn: A Toolkit for Assessing and Improving Fairness in AI*; Technical Report MSR-TR-2020-32; Microsoft: Redmond, WA, USA, 2020. [CrossRef]
48. Friedler, S.A.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E.P.; Roth, D. A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Atlanta, GA, USA, 29–31 January 2019; pp. 329–338. [CrossRef]
49. Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; Roth, A. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociol. Methods Res.* **2021**, *50*, 3–44. [CrossRef]
50. Agarwal, A.; Beygelzimer, A.; Dudik, M.; Langford, J.; Wallach, H. A Reductions Approach to Fair Classification. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Dy, J., Krause, A., Eds.; PMLR: Stockholm, Sweden, 2018; Volume 80, pp. 60–69.
51. Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv* **2023**, arXiv:2309.01219. [CrossRef]
52. Ma, X.; Gong, Y.; He, P.; Zhao, H.; Duan, N. Query Rewriting in Retrieval-Augmented Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; Bouamor, H., Pino, J., Bali, K., Eds.; pp. 5303–5315. [CrossRef]
53. Auquan. The Advantages of Retrieval-Augmented Generation (RAG) AI Over Generative AI for Financial Services. 2023. Available online: https://insights.auquan.com/the-advantages-of-retrieval-augmented-generation-rag-ai-over-generative-ai-for-financial-services-white-paper (accessed on 3 July 2025).
54. Huang, Y.; Huang, B.; Kechadi, M.T. A rule-based method for customer churn prediction in telecommunication services. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Shenzhen, China, 24–27 May 2011; pp. 411–422. [CrossRef]
55. Jafari, M.J.; Tarokh, M.J.; Soleimani, P. An interpretable machine learning framework for customer churn prediction: A case study in the telecommunications industry. *J. Ind. Eng. Manag. Stud.* **2023**, *10*, 141–157. [CrossRef]
56. Fahmy, Y.G.; Fakhr, M.W.; Kholief, M. Privacy Preserving Customer Churn Prediction using Deep and Federated Learning. In Proceedings of the 2024 34th International Conference on Computer Theory and Applications (ICCTA), Alexandria, Egypt, 14–16 December 2024; pp. 108–114. [CrossRef]
57. Hüllermeier, E.; Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach. Learn.* **2021**, *110*, 457–506. [CrossRef]
58. Neal, R.M. *Bayesian Learning for Neural Networks*; Springer Science & Business Media: Berlin, Germany, 2012; Volume 118. [CrossRef]
59. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30. [CrossRef]
60. Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Jones, A.; Joseph, N.; Mann, B.; et al. Language Models (Mostly) Know What They Know. *arXiv* **2022**, arXiv:2207.05221. [CrossRef]

61. Tang, Y.; Fei, Z.; Huang, L.; Zhang, W.; Zhao, B.; Guan, H.; Huang, Y. Failure Mode and Effects Analysis Method on the Air System of an Aircraft Turbofan Engine in Multi-Criteria Open Group Decision-Making Environment. *Cybern. Syst.* **2025**, 1–32. [CrossRef]
62. Lahitani, A.R.; Permanasari, A.E.; Setiawan, N.A. Cosine similarity to determine similarity measure: Study case in online essay assessment. In Proceedings of the 2016 4th International Conference on Cyber and IT Service Management, Bandung, Indonesia, 26–27 April 2016; pp. 1–6. [CrossRef]