LEVERAGING SYNTHETICALLY GENERATED DATA FOR REAL ESTATE DOCUMENT CLASSIFICATION

Tobias Deußer*[1,2], Gregor Ramien [3], Nico Weber [3], Maximilian Meidinger [3], Max Hahnbück [1,2], Christian Bauckhage [1,2], and Rafet Sifa [1,2]
[1]University of Bonn, Bonn, Germany
[2]Fraunhofer IAIS, Sankt Augustin, Germany
[3]Atruvia AG, Karlsruhe, Germany

ABSTRACT

Document classification in regulated domains like law, finance, or real estate is hindered by the scarcity of labeled data and strict privacy constraints. This paper presents a pipeline for synthetically generating training data for document classifiers using a combination of domain-specific templates, large language models, and data augmentation techniques. Focusing on two key document types relevant to real estate workflows, *Child Support Certificate* and *Refurbishment Roadmap*, we construct realistic multi-page documents and generate negative classes using LLM-generated distractors. We train a BERT-based classifier on this synthetic dataset and evaluate it on real-world OCR-extracted documents, achieving strong performance despite the absence of real documents in training. Our findings highlight the feasibility of using synthetic data to overcome annotation bottlenecks and pave the way for broader applications in privacy-sensitive industries.

Keywords Document Classification · Synthetic Data · Large Language Models · Natural Language Processing · Finance · Machine Learning

1 Introduction

Document recognition in regulated industries like real estate, banking, or law presents significant challenges due to limited access to annotated data, strict privacy policies, and a high degree of variation in document formats. Traditional supervised learning methods often struggle with such constraints.

Machine learning has proven to be a powerful tool in the finance and banking sectors, allowing applications such as fraud and contradiction detection [1, 2, 3, 4], credit risk scoring [5, 6, 7], anonymization [8], sentiment analysis [9, 10, 11], and document classification [12]. The adoption of supervised and unsupervised learning in these industries is accelerating, driven by the availability of computing resources and the increasing digitalization of services. However, the success of such models heavily depends on access to high-quality, labeled datasets, which remains a major challenge in regulated environments.

By artificially creating data that mimics the structure and variability of real datasets, synthetic data can enable financial institutions to explore machine learning solutions without exposing sensitive information. Some models, particularly in exploration and proof-of-concept phases, are already trained on partially synthetic datasets [13, 14, 15].

Ultimately, synthetic data generation presents an attractive research and development path. It could help meet both operational goals and legal constraints, but still requires careful evaluation, validation, and domain-specific adaptation.

While large language models such as GPT-4o show promising capabilities, preliminary testing revealed that generating real estate documents end-to-end with such models proved too brittle and lacked sufficient structural and visual variation. Additionally, we found that current image generation methods cannot yet produce document renderings with the fidelity

*tdeusser@uni-bonn.de, ORCID-ID: 0000-0003-4685-0847

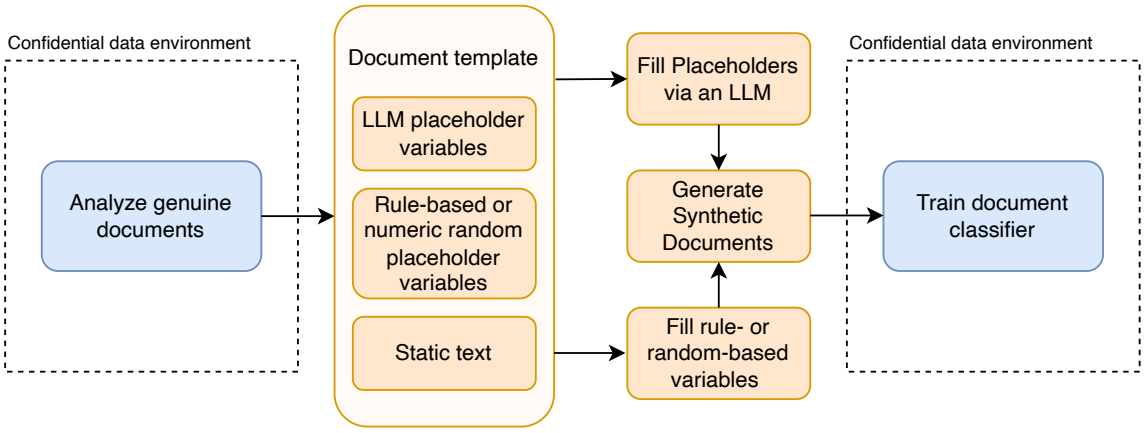

Figure 1: The complete data synthesizing and training pipeline for our document classifier.

required for realistic training data, further motivating our approach of using structured, text-only templates and a dedicated classifier.

Therefore, this work explores how domain-specific templates combined with modern text generation techniques can be leveraged to create rich datasets for training document classification models. Without such synthetic data generation techniques, training a document classification model would be near impossible due to either the lack or sensitivity of real documents, which can only partly be solved by anonymizing the data, as no anonymization is 100% accurate [8, 16]. Our focus for this proof-of-concept lies on two document types: *Nachweis Kindergeld* (German for "Child Support Certificate") and *Individueller Sanierungsfahrplan* (German for "Refurbishment Roadmap"). Both play a central role in German real estate and financial services workflows, as they are commonly required during the application process for traditional consumer mortgage financing.

The remainder of this paper is structured as follows: In Section 2, we discuss related work on synthetic data generation and its growing role in regulated domains such as finance and real estate. Thereafter, we detail our pipeline for generating synthetic documents, including template design, LLM integration, and dataset composition. Section 4 presents experimental results across multiple classification scenarios, highlighting the robustness of our approach. Finally, we close this paper by summarizing our findings and outlining directions for future research.

2 Related Work

Synthetic data has gained traction in regulated domains like finance and real estate, where data privacy and limited access to annotated documents pose major challenges [13, 17]. Previous work has demonstrated that synthetic datasets can effectively train models for tasks like fraud detection and invoice parsing without exposing sensitive information [18].

More recently, large language models (LLMs) have emerged as powerful tools for synthetic data creation in NLP. Studies have shown that prompting LLMs can yield synthetic text suitable for training classifiers, especially in low-resource settings [19, 20]. While LLM-generated data may introduce distributional shifts, techniques such as iterative refinement and hybrid pipelines have been proposed to mitigate this [21].

In the broader field of financial document analysis, [22, 23, 24, 25] studied how language models can be leveraged to extract information from financial reports, and [26, 27, 28] used machine learning methods to improve the financial auditing process. Moreover, [29, 30, 31] studied approaches to automatize the financial document classification process.

Our work builds on these insights by combining template-driven document synthesis with LLM-based content generation, targeting document classification in real estate—an area with similar data scarcity and privacy constraints as finance. We show that carefully designed synthetic datasets can serve as viable stand-ins for real data in training classification models.

3 Methodology

In this section, we describe how we synthetically generate real estate documents and structure our dataset for training a document classifier. Additionally, we briefly touch upon the training process itself. The overall structure of this pipeline is illustrated in Figure 1, which provides an overview of the document generation and classification workflow.

3.1 Document Analysis and Template Design

To understand the structure and content patterns of our target documents, we perform a manual analysis of a range of real samples. Across all collected examples, we observe a high degree of consistency in layout and formatting. In particular, documents of the type *Refurbishment Roadmap* follow a common structure, primarily due to their generation by a limited number of planning software tools based on shared templates. While we do not assume this uniformity across all document types, the consistent styles observed in our dataset enable us to generate synthetic documents that closely resemble authentic examples in both structure and appearance.

Based on this analysis, we manually construct Jinja-based templates for two representative document types, using real samples as references. These templates consist of static text sections and placeholders for dynamic content. The variable components are annotated and categorized into two groups. Simple variables are replaced using the Faker library [32], which generates synthetic names, dates, locations, and similar data. Complex variables are handled using LLM-generated content, tailored to preserve the semantic integrity of the documents.

For multi-page documents, we create one template per page to accurately capture the layout across pages. This modular design enables flexible content assembly while maintaining visual and structural consistency throughout the document.

3.2 Synthetic Dataset Generation

We design modular pipelines capable of generating large volumes of synthetic documents. These pipelines support dynamic document composition and include built-in mechanisms for class balancing, ensuring alignment with dataset distribution goals. To introduce variation, documents can be restructured through controlled reordering or rephrased using language models. The pipeline also supports optional document degradation (e.g., OCR-induced noise) to simulate the low fidelity often found in scanned real-world documents. Additional augmentation strategies are available to support both classification and layout-related tasks.

While the pipeline facilitates rapid generation of new documents, creating new templates still requires manual effort.

3.3 Negative classes

In addition to the two primary document types we analyzed and templated, *Refurbishment Roadmap* and *Child Support Certificate*, we introduce a third document class to serve as a contrasting category in our classification task called *Sonstige* (German for "other"). This class consists of *other*, miscellaneous documents that do not belong to the first two categories. We do not define several templates but instead prompt an LLM to generate these documents for us. To enrich this class, we define a variety of synthetic document types, each annotated with relevant fields such as names, dates, transaction IDs, addresses, or lists, depending on its nature, where applicable. Examples include:

- **Personal Communication and Notices:** Personal letter, Postcard, Handwritten note, Community notice, Wedding invitation
- **Financial and Transactional Documents:** Receipt, Restaurant bill, Utility bill (e.g., electricity or water), Insurance policy, Warranty certificate, Travel booking confirmation, Gym membership documents
- **Informational and Reference Materials:** User manual, Magazine article, Employment reference letter, Contact list
- **Casual or Everyday Texts:** Shopping list, Cooking recipe, Advertisement

This dataset is designed to simulate a realistic mix of document types that may appear when performing document classification and filtering for specific target classes. In addition to the main categories, we include a class labeled *empty pages*, representing pages that are empty or contain only a few characters. While such documents are not necessarily common, they do occur in practical settings and are not well represented by the other classes.

Table 1: Document statistics by category of the real-world test set.

Category	Total Pages	# Documents
Empty Page	4	2
Child Support Certificate	53	32
Refurbishment Roadmap	120	12
Other	63	6
Total	**240**	**52**

3.4 Training and Testing Pipeline

We train a supervised text-based classification model, specifically a BERT-based encoder fine-tuned with a linear classification head [33], using synthetic datasets specifically formatted to align with our existing training infrastructure. The amount of synthetic training data varies by experiment, ranging from 1,300 to 2,800 pages. To evaluate generalization performance, we build a separate testing pipeline within a confidential data environment. This pipeline assesses model performance on real-world documents extracted via OCR. A manually annotated test set of 52 real estate documents, totaling 240 pages, was compiled by domain experts. The distribution of this evaluation dataset is shown in Table 1.

4 Experiments and Results

To formalize our classification setup, we define the task as follows. Let $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ be a set of OCR-extracted pages, where each d_i is a text sequence. The goal is to learn a function $f : d_i \mapsto y_i$ mapping pages to document type labels $y_i \in \mathcal{Y}$.

In the basic 3-class setting, we set

$$\mathcal{Y} = \{\text{CS Certificate}, \text{Refurbishment Roadmap}, \text{Other}\}, \tag{1}$$

plus an additional *empty page* class.

In the fine-grained 15-class setting, each page template corresponds to a distinct label, i.e., $\mathcal{Y} = \{y_1, \ldots, y_{15}\}$, where multiple labels belong to the same document family but reflect different page structures. This formulation enables us to test whether the classifier generalizes beyond simple document-level signals by recognizing individual page types.

Therefore, we conduct two series of page-wise classification experiments to evaluate the effectiveness of our approach. First, in the basic series, we assign the same label to every page in a document. Second, in the fine-grained series, we treat each document page as a separate class to test the scalability of our method. This fine-grained classification setup allows us to evaluate model performance in a more granular, 15-class setting. Finally, acknowledging that template creation is a manual and time-consuming process, we investigate how model performance is affected when training data is generated from only a subset of available templates. Specifically, we use just 75% of the templates from the *Refurbishment Roadmap*, dropping 25% of them at random. By excluding certain page types, we aim to evaluate the model's ability to generalize from the remaining pages and their content, recognizing unseen page types based on shared structural or semantic patterns. This setup allows us to assess whether similar classification accuracy can be achieved with fewer templates, thereby reducing manual effort.

4.1 Training Results

In the basic classification setup (standard 3-class configuration trained for 3 epochs), the model achieves a page-wise average precision (AP) of 88.3% and a document-wise AP of 33.4%.

For the fine-grained classification experiment, where each page is treated as a distinct class (15-class configuration), the model demonstrates scalability and achieves a page-wise AP of 81.4%.

In the reduced training set scenario, where only a subset of pages is used for training, the model still maintains competitive performance with a page-wise AP of 80.8%.

These results indicate that our method remains robust even under more demanding conditions and that effective models can be trained with partially labeled or synthetic datasets, thereby reducing the need for exhaustive manual template creation.

5 Conclusion and Future Work

In this work, we present a practical approach to synthetic data generation for real-world document recognition tasks in the real estate and banking industry. Our key contributions include a modular pipeline for creating high-quality synthetic datasets based on real document structures, integration of LLMs for balanced static and dynamic content generation, and empirical validation on real-world data confirming the generalization capabilities of models trained on synthetic data. The implementation of this was partly developed by Sopra Steria Custom Software Solutions. We achieve a page-wise average precision of 88.3% on our real-world, manually annotated test dataset. Our results demonstrate that carefully designed synthetic datasets can effectively replace or augment real data in industrial machine learning pipelines, reducing both cost and privacy risk.

Nonetheless, important limitations remain. First, while our synthetic data captures many structural and semantic properties of real documents, it may not fully reflect the long-tail variability observed in practice, which can reduce robustness. Second, template creation is still a manual and time-consuming process, which limits scalability across diverse document types. Finally, our work is currently limited to classification. Extending the pipeline to information extraction, layout parsing, and other document understanding tasks would be an important next step.

Future work will aim at fully automating template generation using structure-aware LLMs, improving diversity in underrepresented classes, and incorporating layout-aware models to further enhance document understanding. Additionally, an obvious next step is extending the synthetic data generation to more than two classes and transitioning to a production-ready system, including all classes in the classifier. We also plan to explore the application of this approach across broader document types and industries.

Acknowledgments

This research has been partially funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence.

References

[1] Patricia Craja, Alisa Kim, and Stefan Lessmann. Deep learning for detecting financial statement fraud. *Decision Support Systems*, 139, 2020.

[2] Tobias Deußer, Maren Pielka, Lisa Pucknat, Basil Jacob, Tim Dilmaghani, Mahdis Nourimand, Bernd Kliem, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. Contradiction detection in financial reports. In *Proc. NLDL*, 2023.

[3] Tobias Deußer, David Leonhard, Lars Hillebrand, Armin Berger, Mohamed Khaled, Sarah Heiden, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. Uncovering inconsistencies and contradictions in financial reports using large language models. In *Proc. BigData*, pages 2814–2822. IEEE, 2023.

[4] Stefan Erben and Andreas Waldis. ScamSpot: Fighting financial fraud in Instagram comments. In *Proc. EACL*, pages 71–81, 2024.

[5] Jorge Galindo and Pablo Tamayo. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational economics*, 15:107–143, 2000.

[6] Noujoud Ahbali, Xinyuan Liu, Albert Nanda, Jamie Stark, Ashit Talukder, and Rupinder Paul Khandpur. Identifying corporate credit risk sentiments from financial news. In *Proc. NAACL*, pages 362–370, 2022.

[7] Ana Clara Teixeira, Vaishali Marar, Hamed Yazdanpanah, Aline Pezente, and Mohammad Ghassemi. Enhancing credit risk reports generation using llms: An integration of bayesian networks and labeled guide prompting. In *Proc. ICAIF*, page 340–348, 2023.

[8] Tobias Deußer, Max Hahnbück, Tobias Uelwer, Cong Zhao, Christian Bauckhage, and Rafet Sifa. Resource-efficient anonymization of textual data via knowledge distillation from large language models. In *Proc. COLING*, pages 243–250, 2025.

[9] Min-Yuh Day and Chia-Chou Lee. Deep learning for financial sentiment analysis on finance news providers. In *Proc. ASONAM*, pages 1127–1134, 2016.

[10] Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. Financial sentiment analysis: An investigation into common mistakes and silver bullets. In *Proc. COLING*, pages 978–987, 2020.

[11] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proc. ICAIF*, page 349–356, 2023.

[12] Mengzhen Fan, Dawei Cheng, Fangzhou Yang, Siqiang Luo, Yifeng Luo, Weining Qian, and Aoying Zhou. Fusing global domain information and local semantic information to classify financial documents. In *Proc. CIKM*, page 2413–2420, 2020.

[13] Samuel A. Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E. Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proc. ICAIF*, 2021.

[14] Nataliya Tkachenko. Opportunities for synthetic data in nature and climate finance. *Frontiers in Artificial Intelligence*, 6, 2024.

[15] Xiangwu Zuo, Anxiao Jiang, and Kaixiong Zhou. Reinforcement prompting for financial synthetic data generation. *The Journal of Finance and Data Science*, 10, 2024.

[16] Tobias Deußer, Lorenz Sparrenberg, Armin Berger, Max Hahnbück, Christian Bauckhage, and Rafet Sifa. A survey on current trends and recent advances in text anonymization. In *Proc. DSAA*, 2025.

[17] Akshar Prabhu Desai, Tejasvi Ravi, Mohammad Luqman, Ganesh Mallya, Nithya Kota, and Pranjul Yadav. Opportunities and challenges of generative-ai in finance. In *Proc. BigData*. IEEE, 2024.

[18] Rolandas Gricius and Igoris Belovas. On the generation of synthetic invoices for training machine learning models. *IEEE Access*, 2025.

[19] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the ACL: ACL 2024*, 2024.

[20] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. In *Proc. NeurIPS*, 2022.

[21] Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. Generative data augmentation for commonsense reasoning. In *Findings of the ACL: EMNLP 2020*, 2020.

[22] Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. A French corpus and annotation schema for named entity recognition and relation extraction of financial news. In *Proc. LREC*, 2020.

[23] Lars Hillebrand, Tobias Deußer, Tim Dilmaghani, Bernd Kliem, Rüdiger Loitz, Christian Bauckhage, and Rafet Sifa. Kpi-bert: A joint named entity recognition and relation extraction model for financial reports. In *Proc. ICPR*, 2022.

[24] Tobias Deußer, Syed Musharraf Ali, Lars Hillebrand, Desiana Nurchalifah, Basil Jacob, Christian Bauckhage, and Rafet Sifa. Kpi-edgar: A novel dataset and accompanying metric for relation extraction from financial documents. In *Proc. ICMLA*, 2022.

[25] Tobias Deußer, Lars Hillebrand, Christian Bauckhage, and Rafet Sifa. Informed named entity recognition decoding for generative language models. In *Proc. BigData*, 2024.

[26] Rafet Sifa, Anna Ladi, Maren Pielka, Rajkumar Ramamurthy, Lars Hillebrand, Birgit Kirsch, David Biesner, Robin Stenzel, Thiago Bell, Max Lübbering, et al. Towards automated auditing with machine learning. In *Proc. DocEng*, 2019.

[27] Armin Berger, Lars Hillebrand, David Leonhard, Tobias Deußer, Thiago Bell Felix De Oliveira, Tim Dilmaghani, Mohamed Khaled, Bernd Kliem, Rudiger Loitz, Christian Bauckhage, et al. Towards automated regulatory compliance verification in financial auditing with large language models. In *Proc. BigData*, 2023.

[28] Hanchi Gu, Marco Schreyer, Kevin Moffitt, and Miklos Vasarhelyi. Artificial intelligence co-piloted auditing. *International Journal of Accounting Information Systems*, 2024.

[29] Mengzhen Fan, Dawei Cheng, Fangzhou Yang, Siqiang Luo, Yifeng Luo, Weining Qian, and Aoying Zhou. Fusing global domain information and local semantic information to classify financial documents. In *Proc. CIKM*, 2020.

[30] Kumar Akuthota, A Ganesh, SivaKumar Depuru, et al. Machine learning models for classification of sensitive financial documents. In *Proc. ICCCMLA*, 2023.

[31] Masaru Isonuma, Toru Fujino, Junichiro Mori, Yutaka Matsuo, and Ichiro Sakata. Extractive summarization using multi-task learning with document classification. In *Proc. EMNLP*, 2017.

[32] Daniele Faraglia and Other Contributors. Faker. `https://github.com/joke2k/faker`, 2025.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL*, 2019.