International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 11 November-2026 | Impact Factor: 3.5



Bias-Aware and Hallucination-Resistant Financial AI for Credit Scoring

Dr. Neelotpal Dey[1], Ms. Apoorva Singh[2], Mr. Shani Kumar Jaiswal[3], Ms. Akarshika Pandey[4]

[1] HOD of CS Dept and TnP / Microtek Group of Institution, Varanasi, India
[1] PhD Computer Application / Vikrant University, Gwalior, Madhya Pradesh, India
[2] Management Department / Microtek Group of Institution, Varanasi, India
[3] Management Department / Microtek Group of Institution, Varanasi, India
[3] Management Department / Sunbeam College for Women, Varanasi, India

Corresponding Author Email: dr.neelotpaldey@gmail.com | ORCID: orcid.org/ 0009-0008-4759-7664

How to Cite this Article:

Dey, N., Singh, A., Jaiswal, S. K. & Pandey, A. (2026). Bias-Aware and Hallucination-Resistant Financial AI for Credit Scoring. International Journal of Creative and Open Research in Engineering and Management, 02(11), 1-9. https://doi.org/10.55041/ijcope.v2i11.056

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.







https://doi.org/10.55041/ijcope.v2i11.056

Abstract:

Financial institutions are increasingly exploring Large Language Models (LLMs) for financial document analysis and credit assessment. However, hallucinated financial information and demographic bias can undermine the reliability and fairness of AI-assisted lending decisions. This study proposes an integrated framework for trustworthy financial AI that combines hallucination detection, bias mitigation, and fair credit-risk assessment. The framework uses an LLM to extract structured financial information from source documents, followed by an evidence-verification module that identifies unsupported claims through source matching, semantic verification, numerical consistency checks, and confidence assessment. Verified features are then processed by a separate credit-risk model, while a fairness module evaluates demographic and intersectional disparities in extracted information and credit predictions. The framework will be evaluated using public financial question-answering and credit-risk datasets, supplemented by controlled synthetic financial documents containing known factual and numerical errors. Experiments will examine hallucination detection performance, counterfactual demographic bias, credit-risk prediction quality, and the propagation of incorrect financial information into downstream credit decisions. Evaluation will use claim-level precision and recall, false acceptance rate, AUROC, Brier score, calibration, demographic parity, equal opportunity, equalized odds, and credit-decision flip rate. The study aims to determine whether integrating evidence-based verification and fairness controls can reduce the influence of unsupported financial information while maintaining predictive performance and improving the consistency of AI-assisted credit assessment.

Keywords - Trustworthy Financial AI; Large Language Models; Hallucination Detection; Bias Mitigation; Fair Credit Scoring; Credit Risk Assessment; Evidence Verification; Algorithmic Fairness.

I. INTRODUCTION

Artificial Intelligence (AI) is increasingly being embraced in financial services for financial study, credit-risk assessment, and lending decisions, broadly speaking; more accurately, Large Language Models (LLMs) generalize these capabilities by pulling and taking as information from financial statements, credit reports, and other financial entries, in real-world scenarios, but their application introduces two significant roadblocks, namely hallucination and algorithmic bias, really, hallucination is emerging when an LLM inaccurately informs, such as incorrect income figures, outstanding debts, or repayment details, which may warp later credit-risk predictions, in actual application. Financial hallucination benchmarks, covering PHANTOM and FAITH, play up the need to gauge the factual reliability of LLM-generated financial information [1], [2]; long ago, credit-scoring systems may produce demographic imbalances due to biases in historical data, feature representation, or model behaviour [3][4][5][6][7]; these hurdles are proposing that predictive accuracy alone is insufficient to establish



the reliability and fairness of AI-assisted financial decisions, in real-world scenarios.

An combined framework for bias-aware and hallucination-resistant financial AI for credit scoring is being suggested by the study, in actual application; indeed, its process begins with financial data acquisition, followed by LLM-based extraction of structured credits such as income, debts and repayment history, in many respects. The pulled information is passing through an evidence-verification module that's spotting unsupported claims using source matching, semantic verification, numerical consistency checks, and confidence assessment; uncertain or inconsistent information is flagged for exclusion or human review, in many respects; in fact, demographic and intersectional bias assessment is then undergone by checked traits, with fitting mitigation techniques applied where vital, on the whole. Default probability is being estimated by a separate credit-risk prediction model, but still, and what's more, meanwhile, its outputs are found for fairness and predictive reliability; indeed, what wow and, what the framework makes is an explainable risk assessment supported by inspected proof and related fairness indicators for human review. In plain words, the study will likely gauge the framework through the controlled experiments measuring hallucination detection, counterfactual bias, credit-risk prediction performance, and the extent to which incorrect financial information influences downstream credit decisions, to a noticeable degree.

II. LITERATURE REVIEW

Recent research has checked out the application of Large Language Models (LLMs) in financial investigation, credit scoring, and quantitative reasoning, in many respects. In the meantime, finQA [4] is offering a benchmark for numerical reasoning over financial data, so also, bozagiu et al. [3] compared standard credit-scoring models with generative AI techniques, to a important degree. The reliability of LLM-generated financial information has received increasing attention through PHANTOM [1], which finds hallucinations in long-context financial question answering, and FAITH [2], which centers on tabular hallucinations in financial reports. Janiak et al. [8] further stressed shortcomings in existing hallucination-detection check methods, showing the importance of trustworthy and semantically grounded assessment. These studies usefully found for financial information extraction and verification, but don't fully tackle the propagation of incorrect pulled information into the downstream credit-risk predictions.

Research on algorithmic fairness has studied demographic disparities in credit scoring and methods for their mitigation, in many respects; alloula et al. [7] showed subgroup definitions, particularly intersectional groupings, influence the effectiveness of bias-mitigation techniques, to a significant degree. Hurlin et al. [5] examined fairness in credit-scoring models, so mariscal et al. [6] studied fairness assessment and mitigation in banking credit scoring; indeed, chen et al. [9] examined fairness in credit ratings, and Bahlool et al. [10] reviewed the relationship between predictive performance, fairness and explainability in AI-based credit scoring. Even as these studies are tackling important aspects of reliable and fair financial decision-making, the combined evaluation of LLM hallucination propagation and demographic bias, within a unified credit-scoring framework is remaining an area requiring further investigation; in fact, the proposed study tackles this gap, through the integrated evidence verification, fairness assessment, and credit-risk prediction.

III. RESEARCH GAP AND OBJECTIVES

Existing research is tackling financial hallucination detection, LLM-assisted financial examination, and fairness in credit scoring largely as separate roadblocks, in real-world scenarios; restricted attention has been given to how incorrect LLM-extracted financial information propagates into the credit-risk predictions, in many respects. As, well, whether such deviations are creating diverse effects, across the demographic groups; this study is tackling this gap by blending evidence-based verification and fairness assessment within a combined credit-scoring framework.

The primary target is to formulate and gauge a bias-aware, hallucination-resistant financial AI framework. The precise aims are:

1. **Metrics, namely Precision, Recall, F1-score, and false acceptance rate; so, hallucination detection:** Identify unsupported financial claims employing evidence verification, *in actual applicat*ion.
2. **Bias assessment,** namely Analyse demographic and intersectional imbalances in financial feature extraction; *indeed, metrics, namely* Counterfactual consistency and feature discrepancy.



3. **Metrics, such** as Demographic Parity Difference, Equalized Odds Difference, AUROC and Brier score; in turn, fair credit-risk prediction: Use fairness-mitigation procedures while sustaining predictive performance, to a noticeable degree.

4. **Metrics, such as Default-probability shift, risk-category flip rate, fairness differences, and processing time; so, end-to-end check:** Measure the effect of hallucinated information and combined mitigation on credit-risk predictions, in many respects.

The study studies four research questions:

- **RQ1:** How basically can proof verification detect hallucinated financial information?
- **RQ2:** Do demographic credits influence LLM-based financial feature extraction?
- **RQ3:** Can fairness mitigation reduce demographic imbalances without greatly degrading predictive performance?
- **RQ4:** Does the combined framework reduce the propagation of incorrect financial information into the credit-risk predictions?

IV. PROPOSED INTEGRATED FRAMEWORK

4.1 Framework Architecture and Problem Formulation

LLM-based financial information extraction, evidence-grounded hallucination detection, demographic bias assessment, and machine-learning-based credit-risk prediction are currently being blended by the suggested framework, broadly speaking; not like standard credit-scoring pipelines probably, it introduces a verification gate between information extraction and predictive modelling to prevent unsupported financial claims from directly changing credit-risk projections, on the whole.

The model consists of six interconnected stages:

1. Financial data acquisition and preprocessing.
2. LLM-based structured financial information extraction.
3. Claim-level hallucination detection and evidence verification.
4. Demographic and intersectional bias assessment and mitigation.
5. Fairness-aware credit-risk prediction.
6. Explainability, evidence provenance, and human review.

Let (D) represent the financial documents, (X) the available structured financial records, and (A) the demographic attributes reserved for fairness auditing. The LLM extracts financial claims from the available records, producing a set of attributes (C). Each claim contains a feature name, extracted value, unit, source reference, supporting evidence, and extraction confidence. The verification module evaluates these claims and constructs a verified feature set, $(X_{verified})$. Only accepted attributes are passed to the credit-risk model; rejected or uncertain claims are excluded or referred for human review.

4.2 LLM-Based Financial Information Extraction

The extraction module converts unstructured financial documents, including financial statements, income records, repayment histories, and credit reports, into the structured financial credits, in actual application, and relevant document segments are supplied to the LLM by a Retrieval-Augmented Generation (RAG) strategy, lowering dependence on information stored solely in model parameters, in real-world scenarios.

The extraction process entails:

1. **Document preprocessing:** Identify pertinent sections, tables, financial values, reporting periods, and units.
2. **Document segmentation, namely** Divide lengthy data into the meaningful sections while preserving table headers and contextual information, to a significant degree.
3. **Evidence retrieval, including** Retrieve relevant document segments employing semantic embeddings and similarity matching.
4. **Structured extraction:** Instruct the LLM to predefined field rather than unrestricted textual responses.
5. **Feature normalisation, like** Standardise currencies, quantitative formats, reporting periods, and missing-value representations.

Each pulled record preserves the following fields, namely feature name, pulled value, unit, reporting period, source document, page reference, evidence span, and extraction score, broadly speaking; the LLM confidence score is treated as an backup signal, not



proof of factual correctness, in real-world scenarios. Missing information is recorded as unavailable rather than inferred.

4.3 Claim-Level Hallucination Detection and Evidence Verification

This module acts as the primary reliability control; try this instead of finding an entire LLM response as a single unit, it decomposes the pulled information into the individual claims and checks each claim against its supporting source evidence, to an important degree.

The verification process combines four checks:

A. Source and semantic verification: Retrieved document segments are checked with pulled claims employing semantic similarity and Natural Language Inference (NLI). High semantic similarity alone isn't considered enough to establish factual correctness. The NLI model estimates entailment, contradiction, and neutral relationships between the claim and its supporting proof.

B. Numerical consistency verification: Financial values, currencies, reporting periods, and calculations are independently inspected wherever the needed source information is available. To take one instance, credit utilisation and debt-to-income ratios can be recalculated from their coring values. Quantitative mismatches are measured using:

$$[E_v = |V_e - V_r|]$$

What's more, where(V_e) is the pulled value and (V_r) is the independently checked value, which is why field-specific tolerances account for rounding and quantitative precision, in real-world scenarios.

C. Evidence-support scoring: The individual verification results are normally combined into a claim-level score:

$$\left[S_j = \frac{\sum_{k=1}^{m} w_k \, r_{jk}}{\sum_{k=1}^{m} w_k}\right]$$

Weights and acceptance thresholds are found out employing validation data. Where represents the verification score of claim , is the normalised finding of verification check , and is its allocated weight$(S_j)(j)$, (r_{jk})broadly speaking$(k)(w_k)$.

D. Claim classification and verification gate: Each claim receives one of three statuses:

Status	Condition	System response
Verified	Enough supporting proof and no material contradiction	Accept the attribute
Rejected	Contradictory indication or failed quantitative validation	Exclude the attribute
Uncertain	Insufficient or conflicting proof	Refer for human review

A high verification score can't override a critical contradiction or quantitative inconsistency; the resulting checked feature set includes accepted financial credits with their associateding evidence references and verification status, in real-world scenarios.

4.4 Demographic Bias Assessment and Mitigation

The fairness module is analyzing whether demographic information influences financial feature extraction or creates imbalances in credit-risk predictions; it operates at two levels, in actual application.

A. Counterfactual bias assessment

Matched applicant profiles are typically made with same financial traits, such as income, debt, credit limit, and repayment history, but controlled changes in demographic credits, to a noticeable degree; the same LLM, prompts, and configuration settings are, in effect, employed to each profile, to a important degree.

The difference in forecasted default probabilities is measured as:

$$[\Delta P_i = |P(default|x_i, a_i) - P(default|x_i, a_i')|]$$

where (x_i) represents unchanged financial credits and represent alternative demographic conditions; and thus, differences in pulled financial credits are also recorded to spot demographic sensitivity during the information extraction, broadly speaking$(a_i), (a_i')$.

B. Group-level and intersectional fairness assessment

Across demographic groups employing the following measures, predictive outcomes are checked:

1. **Demographic Parity Difference, like** Measures differences in projected outcome rates across the groups.



2. **Equal Opportunity Difference, such as** Measures differences in true-positive rates.

3. **Equalized Odds Difference, including** Measures differences in both true-positive and false-positive rates.

4. **Disparate Impact Ratio, namely** Compares favourable outcome rates between groups.

5. **Worst-Group Performance:** Identifies the group with the weakest predictive performance.

For datasets using default as the target variable, a positive prediction represents anticipated default, not loan approval, in actual application; and thus, clearly, fairness measures must be interpreted consistently (specifically in this setting) with the chosen target and decision policy, which is why intersectional groups are also assessed where enough sample representation is available, in real-world scenarios.

C. Bias mitigation protocols

Two mitigation approaches are generally considered:

- **Reweighing:** Designates diverse sample weights to lower the impact of imbalanced demographic-group and outcome combinations, in many respects.
- **Exponentiated Gradient Reduction, namely** Applies fairness constraints during the model training to spot predictors with lessened group disparities, in many respects.

The mitigation methods are usually checked against an unconstrained baseline to analyse fairness improvements and their effect on predictive performance, in many respects, and demographic credits are, in practice, mostly kept for auditing and controlled check rather than automatically employed as operational credit-scoring traits, in real-world scenarios.

4.5 Fair Credit-Risk Prediction Model

The LLM doesn't independently find out default probability or approve loan applications. But instead, confirmed financial credits are supplied to a separate machine-learning model that projections the applicant's probability of default:

$$[P(default|X_{verified})]$$

Three predictive algorithms are mainly considered:

1. **Logistic Regression:** Offers an interpretable baseline for default prediction.
2. **Random Forest -** Captures nonlinear relationships and feature interactions.
3. **XGBoost, like** Uses gradient-boosted decision trees to model complicated patterns in structured financial data.

All models use uniform training, validation and testing partitions; so, predictive performance is checked employing AUROC, AUPRC, F1-score, and Brier score, to a noticeable degree; decision thresholds are picked using validation data, keeping probability estimation separate from the final risk classification, in real-world scenarios.

4.6 Explainability, Evidence Provenance, and Human Review

SHAP (SHapley Additive exPlanations) is often employed to spot the insight of individual financial credits to projected default probability, in real-world scenarios, and model explanations with the evidence trail made during the extraction and verification are usually being combined by the final module, broadly speaking.

Each applicant's assessment record includes:

- Estimated default probability and risk grouping.
- Checked financial credits and supporting source references.
- Rejected and unresolved claims.
- Claim-level verification scores.
- SHAP-based feature explanations.
- Pertinent fairness-audit information.

High-impact unresolved claims, contradictory data, and low-confidence extraction cases are referred for human review; in turn, any LLM-generated explanation must remain uniform with the actual prediction and feature-attribution findings, to a important degree.

4.7 End-to-End Framework Algorithm

Algorithm 1, namely Bias-Aware and Hallucination-Resistant Financial Credit Scoring

Input: Financial records (D), structured credit records (X), demographic audit attributes (A), trained LLM, NLI model, credit-risk model, and verification thresholds, broadly speaking.



Output, like Checked financial credits, default probability, verification report, fairness report, and explanation record.

1. Preprocess financial data and structured credit records.
2. Segment documents while preserving tables, reporting periods, and source references.
3. Retrieve applicable proof for each financial credit.
4. Use the LLM to pull structured financial claims.
5. Confirm each claim using source matching, NLI, and quantitative validation.
6. Classify claims as Checked, Rejected, or Uncertain.
7. Build the verified trait set employing accepted claims.
8. Evaluate counterfactual consistency and group-level fairness.
9. Use the chosen fairness-mitigation method.
10. Estimate default probability employing the credit-risk model.
11. Explain and attach evidence provenance.
12. Refer unresolved or high-impact claims for human review.
13. Yield the final risk-assessment and evaluation reports.

This model therefore is separating financial information extraction, factual verification, fairness assessment, and credit-risk prediction into the distinct but connected stages, broadly speaking. What this separation makes possible is the insight of each component to be found independently and as part of the complete pipeline, in actual application.

V. RESEARCH METHODOLOGY

5.1 Research Design

Standard machine-learning models and component-level configurations will serve as baselines for comparison; in fact, a quantitative experimental strategy is being employed by this study to gauge the suggested framework through the four experiments covering hallucination detection, demographic bias, credit-risk prediction, and end-to-end performance, in real-world scenarios.

5.2 Datasets and Preprocessing

The study will use PHANTOM, FAITH, and FinQA for financial hallucination detection and quantitative reasoning, broadly speaking. The UCI Default of Credit Card Clients dataset will maybe support credit-risk prediction and demographic fairness assessment, in actual application. Home Credit Default Risk may be considered for extra validation. Synthetic financial entries will also be made from structured applicant entries, keeping the original values as ground truth for controlled hallucination experiments.

Preprocessing will include missing-value treatment, duplicate removal, quantitative standardisation, and categorical encoding, so demographic credits will be kept separately for fairness auditing, in actual application; in fact, data partitioning will be completed before model training and synthetic-document generation to prevent information leakage, in actual application. Structured credit records will (to some degree) be divided into the training, validation and testing sets employing a stratified 70:15:15 split.

5.3 Model Implementation

Python will be used by the implementation and established machine-learning libraries, and meanwhile, evidence retrieval will be supported by sentence Transformers and FAISS, on the whole; indeed, an instruction-tuned LLM will structuredly pull. Semantic consistency will be gauged by a Transformer-based NLI model, which is why as well, quantitative claims will be inspected by deterministic validation rules, in many respects; model versions, prompts, decoding settings, and verification thresholds will (in actual operation) be recorded for reproducibility. Logistic Regression, Random Forest, and XGBoost will be considered for credit-risk prediction, and fairness assessment and mitigation will be supported by fairlearn; while SHAP will explain.

5.4 Experimental Evaluation

This model will be found through the four experiments:

1. Hallucination Detection, including Compare source matching, NLI-based verification, and the suggested combined verification method employing Precision, Recall, F1-score, AUPRC and false acceptance rate, on the whole.
2. Measure counterfactual feature discrepancies, prediction differences, and group-level fairness indicators; indeed, Demographic Bias Assessment, like Evaluate matched applicant profiles with same financial credits but controlled demographic changes, on the whole.
3. Evaluate AUROC, AUPRC, F1-score, and Brier score. Credit-Risk Prediction, namely



Compare predictions based on original structured features, raw LLM-extracted attributes, and verified traits, in real-world scenarios.

4. Measure default-probability shifts, risk-category flip rates, fairness disparities, and processing overhead, and End-to-End Evaluation, such as Compare baseline, verification-only, fairness-only, and combined arrangements, in many respects.

5.5 Reproducibility and Experimental Controls

Uniform data partitions and recorded model settings will be employed by all setups. Findings from deliberately corrupted synthetic entries will be reported separately from naturally (or more precisely) starting LLM errors. Verification thresholds and hyperparameters will be chosen using validation data, without reference to held-out test results. And repeated experiments and statistical comparisons will be did where suitable.

VII. DISCUSSION

The suggested framework is tackling two interconnected roadblocks in AI-assisted credit scoring, including the factual reliability of LLM-extracted financial information and the fairness of later credit-risk predictions, broadly speaking; and thus, its verification mechanism aims to prevent unsupported financial claims from changing predictive models; in turn, in the meantime, counterfactual testing and group-level fairness assessment analyse demographic imbalances, in real-world scenarios. Findings will be construed alongside subgroup sample sizes, model calibration, and computational running costs to assess whether claim-level verification decreases prediction instability without compromising predictive performance, and the planned experiments will compare raw and verified financial credits, as well as baseline and fairness-aware models, to gauge their individual and combined insights.

This model certainly is limiting. Controlled synthetic financial records may not represent the full range of departures in real-world financial entries, in actual application. Verification accuracy also rests on source-document quality, retrieval effectiveness, and NLI model capabilities; in turn, fairness outcomes may differ according to demographic group definitions and the picked credit-risk target, on the whole. Future validation should, of course, therefore encompass mixed financial datasets, independently annotated records, broader subgroup analysis, and evaluation under the real-world institutional bounds.

VIII. CONCLUSION AND FUTURE WORK

What this study puts forward is an blended framework for bias-aware and hallucination-resistant financial AI in credit scoring, in many respects; indeed, LLM-based financial information extraction, claim-level indication verification, demographic bias assessment, and machine-learning-based credit-risk prediction are being combined by this model, broadly speaking. By introducing a verification mechanism before financial traits enter the predictive model, the put forward approach aims to lower the impact of unsupported or inconsistent information on credit-risk assessment; it also includes counterfactual and group-level fairness evaluation to analyse demographic imbalances while upholding predictive reliability, really, the experimental design is offering a basis for finding he individual and combined advancements of hallucination detection and fairness mitigation, in many respects.

Future work will stress on employing and inspecting the framework, across the mixed financial datasets and LLM architectures, to a noticeable degree, in actual application; what's more study will study the potency of advanced hallucination detection methods, fairness-aware optimisation, and evidence retrieval, under the incomplete or conflicting financial entries, on the whole. Evaluation using independently marked up real-world financial entries will, for all practical purposes, also be vital to establish concrete applicability beyond controlled synthetic experiments, in many respects; extra work may try out continuous model monitoring, regulatory compliance, and human-in-the-loop decision support for responsible rollout in financial institutions.

XI. References

[1] L. Ji et al., “PHANTOM: A Benchmark for Hallucination Detection in Financial Long-Context QA,” *NeurIPS*, 2025.

[2] M. Zhang et al., “FAITH: A Framework for Assessing Intrinsic Tabular Hallucinations in Finance,” *ACM ICAIF*, 2025.

[3] A. M. Bozagiu et al., “A Comparative Analysis of Credit Scoring Models and Generative AI Techniques,” *International Conference on Business Excellence*, 2025.



[4] Z. Chen et al., "FinQA: A Dataset of Numerical Reasoning over Financial Data," *EMNLP*, 2021.

[5] C. Hurlin et al., "The Fairness of Credit Scoring Models," *Management Science*, 2026.

[6] C. Mariscal et al., "Implementing and Analyzing Fairness in Banking Credit Scoring," *Procedia Computer Science*, 2024.

[7] A. Alloula et al., "Subgroups Matter for Robust Bias Mitigation," *ICML*, 2025.

[8] D. Janiak et al., "The Illusion of Progress: Re-evaluating Hallucination Detection in LLMs," *EMNLP*, 2025.

[9] Y. Chen et al., "Measuring Fairness in Credit Ratings," *Expert Systems with Applications*, 2024.

[10] R. Bahlool et al., "Performance, Fairness, and Explainability in AI-Based Credit Scoring: A Systematic Literature Review," *Journal of Risk and Financial Management*, 2026.

[11] Y. Zou and C. Gao, "Extreme Learning Machine Enhanced Gradient Boosting for Credit Scoring," *Algorithms*, 2022.