Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explaining and Mitigating Proxy Fairness in LLM-based Credit Risk Decisions

Huyen Giang Thi Thu
Graduate University of Science and Technology, Vietnam Academy of Science and Technology
Thang Doan Viet
thang.dv509@gmail.com

Converged AI Lab
Giang Nguyen Long
Institute of Information Technology, Vietnam Academy of Science and Technology
Ha-Bang Ban
Hanoi University of Science and Technology

Research Article

Keywords: bias mitigation, credit risk assessment, explainable fairness, fairness auditing, large language models, proxy discrimination

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11440110/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Explaining and Mitigating Proxy Fairness in LLM-based Credit Risk Decisions

Huyen Giang Thi Thu [0009-0007-6283-3111][1,4], Thang Doan Viet [0009-0009-3072-5532][2*], Giang Nguyen Long [0000-0001-6184-1469][5], Ha-Bang Ban [0000-0003-2241-5146][3]

[1]Graduate University of Science and Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam.
[2]Converged AI Lab, Hanoi, Viet Nam.
[3]Hanoi University of Science and Technology, Hanoi, Viet Nam.
[4]Banking Academy of Vietnam, Hanoi, Vietnam.
[5]Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam.

*Corresponding author(s). E-mail(s): thang.dv509@gmail.com;
Contributing authors: huyengtt@hvnh.edu.vn; nlgiang@ioit.ac.vn;
bangbh@soict.hust.edu.vn;

Abstract

Large language models (LLMs) are increasingly explored as credit scorers, where decisions must be explainable and non-discriminatory. Removing the protected attribute does not prevent discrimination, because correlated features can act as proxies, and aggregate fairness metrics do not reveal which features carry the disparity. We propose VAYNE, a framework that diagnoses and mitigates proxy discrimination in LLM-based credit decisions using only query access. VAYNE combines each feature's proxy capacity, its mutual information with the protected attribute, with its proxy use, the change in the model's score when the feature is withheld, into a proxy risk score and a proxy path graph that guide two interventions, proxy masking and proxy filtering. We benchmark three LLMs on three credit datasets and audit the selected scorer. Without examples, the LLMs are close to chance, but a few in-context examples bring them into the accuracy range of traditional credit-scoring models. The audited scorer responds to sex when it is stated, and with sex withheld, its proxy structure ranges from a single dominant feature to a redundant cluster of features. The VAYNE-guided

interventions attain the lowest group disparity in most comparisons. Proxy filtering substantially reduces the statistical parity and equalized-odds gaps while barely changing accuracy, outperforms instruction-based prompt debiasing, and approaches the utility of demonstration rebalancing while also identifying which features carry the disparity.

Keywords: bias mitigation, credit risk assessment, explainable fairness, fairness auditing, large language models, proxy discrimination

1 Introduction

Large language models (LLMs) [1] are increasingly used to support consequential decisions, including insurance underwriting [2], hiring [3], and healthcare [4]. Credit scoring is a natural candidate, since an applicant record can be serialized into a natural-language profile and scored by a pretrained model with only a few labeled examples [5–7], and recent studies already apply LLMs to credit and risk assessment [8–10]. However, lending is a regulated domain in which decisions must be explained to applicants and must not discriminate against protected groups. Models trained on large-scale text can absorb social stereotypes [11, 12] and reproduce them as different outcomes across groups [8, 13, 14]. Removing the protected attribute from the profile, known as fairness through unawareness, does not remove this risk, because features such as occupation, employment status, or property ownership can carry much of the same information [15, 16], in the same way that zip code has long stood in for race in redlining [17]. Such proxy discrimination can persist even when a scorer appears demographically neutral [18].

This challenge becomes even more important when the scorer is an LLM accessed through an API. Its parameters, gradients, and training data are not available, and its output is a stochastic, verbalized score rather than a calibrated probability. Aggregate fairness metrics such as Statistical Parity or Predictive Parity can show that outcomes differ across groups, but not which part of the profile is responsible [19, 20].Feature-attribution methods such as LIME and SHAP rank features by how much they move the prediction [21, 22], yet a feature can be influential because it is a legitimate underwriting signal, because it encodes protected information, or both. Legal and technical work on proxy discrimination therefore separates a feature's capacity to encode protected information from its actual use in the decision [18, 23, 24], but this separation has been developed for conventional models. A further difficulty is that an LLM scorer close to chance may appear fair or unfair for reasons unrelated to discrimination [25], so a fairness diagnosis is only informative once the utility of the scorer has been established [26]. Diagnosing and mitigating proxy discrimination in black-box LLM credit scorers therefore remains an open problem.

Motivated by this gap, this paper addresses three research questions, which move from whether an LLM credit scorer is worth auditing at all, to why it is or is not fair, to what can be done about it.

- RQ1 (Utility). Can LLM credit classifiers reach utility competitive with traditional credit-scoring models, so that a subsequent credit-risk fairness audit is meaningful?
- RQ2 (Diagnosis). Does a utility-qualified LLM credit classifier exhibit controlled protected-attribute sensitivity and/or proxy-mediated fairness risk?
- RQ3 (Mitigation). Does reducing an LLM's reliance on the diagnosed proxy features make its credit decisions fairer, and at what cost to classification utility?

To answer these questions, we propose VAYNE (**V**alue **A**blation **Y**ields **N**on-discrimination **E**vidence), a black-box framework for explaining and mitigating proxy discrimination in LLM-based credit decisions. The key idea is to treat a feature as a high-risk proxy only when it both encodes information about the protected attribute and materially affects the output of the scorer. Before the audit, we establish whether the scorer is accurate enough to be audited by comparing its utility with that of traditional credit-scoring classifiers (RQ1). VAYNE then measures the controlled sensitivity of the scorer to the protected attribute by replacing its stated value with UNKNOWN, computes for every other feature a proxy capacity from its mutual information with the protected attribute and a proxy use from ablation-based output sensitivity, and combines the two into a proxy risk score and a proxy path graph (RQ2). Finally, it uses the diagnosed proxy structure to select the features withheld by proxy masking and proxy filtering, which are compared with prompt-level baselines (RQ3). Every quantity is estimated from repeated queries. What distinguishes VAYNE from prior fairness audits of LLMs is that it explains a disparity at the level of individual features while requiring only query access to the model. This combination is not addressed by aggregate bias benchmarks for LLM credit scoring such as CALM [8], by feature-attribution methods [21, 22], by black-box audits of indirect influence designed for conventional classifiers [27], or by prompt-level mitigations such as fair prompt optimization and fair in-context learning [28, 29]. We benchmark three LLMs on three credit benchmarks commonly used in fairness research [30, 31] and audit the selected one with VAYNE. In summary, this paper makes the following contributions.

(i) We propose VAYNE, a black-box framework that separates controlled sensitivity to the protected attribute from proxy-mediated unfairness in LLM-based credit decisions. This enables a fairness audit of API-only scorers without access to model parameters.
(ii) We introduce proxy capacity, proxy use, and their product, the proxy risk score, together with a proxy path graph that traces protected information to the credit decision. This turns an aggregate disparity into a ranked, feature-level explanation that a model-risk review can act on.
(iii) We design two mitigation strategies, proxy masking and proxy filtering, that are derived directly from the diagnosed proxy structure. This yields interventions whose choice of features is justified by the diagnosis rather than by a generic fairness instruction.

The rest of the paper follows the same progression as the proposed framework. Section 2 reviews related work on LLM-based credit scoring, proxy discrimination, and fairness mitigation for LLMs. Section 3 presents the VAYNE framework. Section 4

describes the experimental setup and answers RQ1, RQ2, and RQ3 in turn. Section 5 discusses the findings and limitations, and Section 6 concludes the paper.

2 Related Work

Recent work uses LLMs for tabular classification by serializing each record into natural language and prompting the model for a label or score. LIFT [6] fine-tunes language models on serialized records without changing the architecture, while TabLLM [7] shows that serialization combined with few-shot learning [5] can be competitive with traditional models when labeled data are scarce. UniPredict [32] trains a single model across many tabular datasets, and Han et al. [33] use LLMs to engineer features for few-shot tabular learning. Fang et al. [34] survey LLMs for tabular prediction, generation, and understanding, including the role of the serialization format. These approaches have been extended to credit scoring and broader financial risk assessment [9, 10]. Most relevant here, CALM [8] builds a generalist credit-scoring LLM and finds that such scorers can be accurate and biased at the same time. For conventional classifiers, fairness-aware learning on these credit benchmarks has been compared experimentally across pre-, in-, and post-processing methods [31]. These studies establish that LLMs can serve as credit scorers, but they evaluate bias only as an aggregate disparity and do not trace which features carry it.

LLMs trained on large-scale text absorb social stereotypes [11] and can produce outputs that differ systematically across demographic groups [12–14], a concern that extends to consequential uses such as insurance underwriting [2] and hiring [3]. Group fairness criteria quantify such differences. Statistical Parity compares positive-decision rates across groups [35], equalized odds compares error rates [19], and Predictive Parity compares precision [20], and surveys organize these notions and the trade-offs between them [25, 26], including how they apply to language models [14]. These metrics detect that a disparity exists in a portfolio of decisions, but they do not identify whether it arises from the protected attribute, from correlated features, or from differences in base rates [36].

Fairness through unawareness is insufficient because non-protected variables can encode protected information [15, 16]. In lending, this is the logic of redlining, in which zip code stood in for race [17], and occupational segregation makes occupation a carrier of information about sex [37]. Fuster et al. [38] show that machine-learning credit models can redistribute benefits across demographic groups through their use of correlated predictors. Datta et al. [18] formalize proxy use as the presence of a correlate of the protected attribute that influences the output, Tschantz [23] distinguishes the capacity of a variable to recreate a protected attribute from its use in the decision, and Weerts et al. [24] turn this distinction into proxy-capacity and proxy-use tests for challenging discriminatory algorithms. Adler et al. [27] audit the indirect influence of features on black-box classifiers by obscuring them in the input. These works provide the vocabulary and the tests on which VAYNE builds, but they were developed for conventional models with deterministic outputs rather than for API-only LLM scorers whose verbalized scores vary from query to query.

Model-agnostic explanation methods such as LIME [21], SHAP [22], and counterfactual explanations [39, 40] identify influential inputs and are often proposed for producing reason statements for credit decisions. On their own, however, they do not distinguish legitimate underwriting information from demographic proxy use, and their faithfulness can be manipulated or can vary across groups [41–43]; a recent review of interpretable LLMs for credit risk identifies fairness and bias mitigation as an open gap [44]. Causal fairness methods formalize counterfactual and path-specific effects of the protected attribute [45–47], but they require a trusted structural causal model [36], which is unavailable for an LLM queried only through an API.

Fairness interventions for LLM-based tabular prediction act on the prompt or on the in-context examples. Cherepanova et al. [28] compare fair prompt optimization, soft prompt tuning, strategic selection of few-shot examples, and related techniques for improving the group fairness of LLMs on tabular data, and Chhikara et al. [29] study how fairness-aware instructions and the choice of demonstrations affect fairness in few-shot classification. Together, these studies show that the prompt and the demonstrations can shift the group-level outcomes of an LLM scorer. However, they target outcome parity directly and rarely examine which features transmit protected information, or whether the scorer is accurate enough for the comparison to be meaningful. As a result, a feature-level diagnosis that conditions on utility and derives mitigation from the diagnosed proxy structure remains missing for black-box LLM credit scorers, motivating the development of VAYNE.

3 The VAYNE Framework

Figure 1 provides a high-level overview of the proposed VAYNE framework. VAYNE treats an LLM used as a credit-decision scorer as a black box and evaluates it in stages, measuring baseline fairness, distinguishing controlled protected-attribute sensitivity from proxy-mediated unfairness, identifying high-risk proxy features, constructing a proxy path graph, and applying targeted mitigation strategies. This diagnosis presupposes that the model has already demonstrated adequate classification utility on the dataset being audited, since a fairness evaluation is only informative when performed on a model sufficiently capable of the underlying credit-decision task.

3.1 Preliminaries and Notation

Proxy fairness, equivalently proxy discrimination, refers to the case where a credit decision system produces disparate outcomes across a protected group not because it directly consults a protected attribute A such as sex, race, or age, but because it relies on a non-protected feature that is statistically associated with A and therefore transmits protected information indirectly [15, 18]. Removing A from the application, so-called fairness through unawareness, is a common but insufficient mitigant in lending, since correlated variables such as zip code, occupation, or employment history can reconstruct much of the information that A would have carried [15, 17, 37]. Fair-lending law recognizes disparate-treatment and disparate-impact theories of liability, and regulators and legal scholars have long treated unawareness as insufficient, since a facially neutral variable correlated with a protected class, such as zip code in

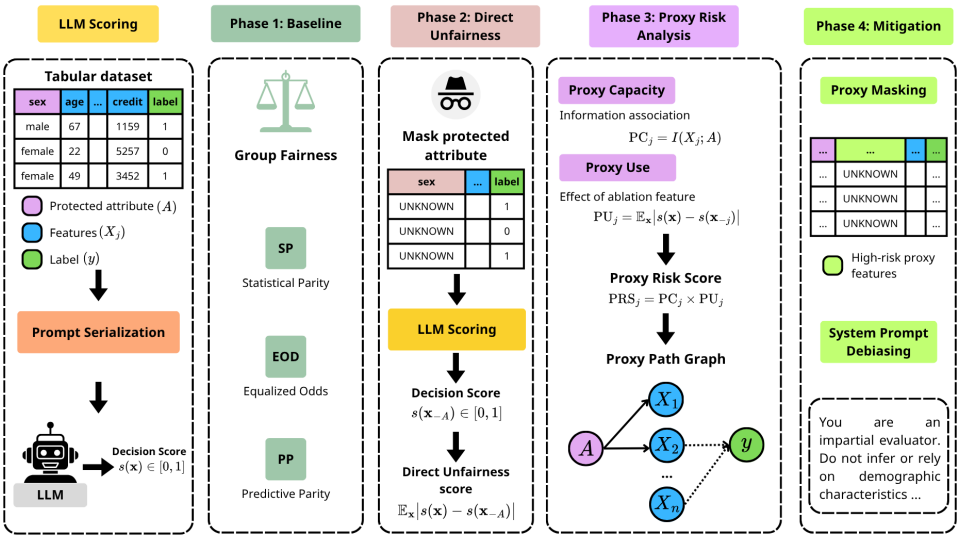


Figure 1: Overview of the VAYNE framework.

a redlining claim, can still trigger liability [16]. Legal and technical analyses argue for separating a feature's capacity to encode protected information from a model's actual use of that information [18, 24], and recent scholarship formalizes this capacity/use split into an operational auditing standard [23, 24]. Group fairness criteria, namely Statistical Parity, equalized odds, and Predictive Parity [19, 20, 35] detect that a proxy-mediated disparity exists in a portfolio of credit decisions but not which feature transmits it, motivating a per-feature decomposition rather than an aggregate score. Causal-fairness formalizations of permissible and impermissible pathways from A to the outcome [45–47] express the same underlying question, namely through which variables protected information reaches the credit decision, but require a trusted structural causal model that is unavailable for a black-box LLM scorer queried only through an API [36]. VAYNE operationalizes the capacity/use distinction for exactly this black-box setting, without assuming a verified causal graph, via the proxy capacity, proxy use, and proxy risk score defined later in this section.

Before describing each component of VAYNE, we introduce the notation used throughout the paper. We consider a credit-decision setting in which an LLM is accessed only as a black-box scoring function by a lender or a fairness auditor. Each applicant record contains one protected attribute and a set of non-protected attributes. We write the full record as

$$\mathbf{x} = (a, x_1, x_2, \ldots, x_d) \in \mathcal{X}, \tag{1}$$

where a denotes the value of the protected attribute A (sex in all experiments of this paper), and each x_j denotes the value of a named non-protected feature X_j, such

as occupation, employment status, income, or property ownership. The set of non-protected features is denoted by $\mathbf{X} = \{X_1, \ldots, X_d\}$. For a binary protected attribute we write $a \in \{0, 1\}$, where $a = 0$ denotes the protected group.

The deployment profile $\mathbf{z} = (x_1, \ldots, x_d)$ omits the protected attribute entirely, following the fairness-through-unawareness input setting used in practice. Serializing a profile and querying the LLM once returns a score

$$s(\mathbf{z}) = p_{\text{yes}}(\mathbf{z}) \in [0, 1], \tag{2}$$

namely the value of the p_yes field in the model's JSON output. Because decoding is stochastic, each prompt is queried R times and the analysis uses the average

$$\bar{s}(\mathbf{z}) = \frac{1}{R} \sum_{r=1}^{R} s^{(r)}(\mathbf{z}), \tag{3}$$

where $s^{(r)}(\mathbf{z})$ is the score returned by the r-th query. The scores $s^{(r)}$ and $\bar{s}$ are verbalized, uncalibrated numbers; we use them as decision scores and as measures of output sensitivity, not as probabilities. In each scoring run r, the binary credit decision is obtained by applying a threshold t to the score of that run,

$$\hat{Y}^{(r)} = \mathbf{1}\left[s^{(r)}(\mathbf{z}) \geq t\right], \tag{4}$$

with $t = 0.5$; we write $\hat{Y} = +$ for $\hat{Y} = 1$ and omit the run index when it is clear from the context.

The evaluation dataset $\mathcal{D} = \{(\mathbf{x}_i, a_i, y_i)\}_{i=1}^{N}$, where y_i is the ground-truth label (e.g., approved/denied, or default/no-default), is split into disjoint training, development, and test sets $\mathcal{D}_{\text{tr}}$, $\mathcal{D}_{\text{dev}}$, and $\mathcal{D}_{\text{te}}$. Two kinds of ablated input are used in the audit. For the controlled direct-sensitivity audit, the full record $\mathbf{x}$ is serialized with a protected-attribute field that states the applicant's value, and $\mathbf{x}_{-A}$ denotes the same record with that value replaced by the sentinel string UNKNOWN. For feature-use ablation, $\mathbf{z}_{-j}$ denotes the deployment profile in which the value of the non-protected feature X_j is replaced by UNKNOWN. This separation prevents the deployment setting, where A is unavailable, from being conflated with the controlled audit setting, where the value of A is stated or replaced by UNKNOWN solely to measure sensitivity. Table 1 summarizes the core notation.

Given a deployed LLM credit scorer $s(\cdot)$ that has already cleared a classification-utility check, and an evaluation dataset $\mathcal{D}$, VAYNE treats black-box credit-risk fairness diagnosis as two linked steps, independently of any particular research question that a downstream study might use it to answer. The first step diagnoses whether an observed fairness violation is associated with controlled sensitivity to the protected attribute A, with non-protected proxy features X_j, or with both, and characterizes how the underlying proxy structure varies across credit datasets and deployments. The second step identifies which intervention reduces proxy-mediated unfairness most effectively while preserving credit-classification utility. VAYNE does not treat fairness metrics as self-contained evidence. It uses them as diagnostic signals, then explains the

Table 1: Core notation used throughout the paper.

Symbol	Description
$\mathbf{x} = (a, x_1, \ldots, x_d)$	Applicant record with the protected attribute and d non-protected features
A, a	Protected attribute (sex in this paper) and its value; $a = 0$ is the protected group
$\mathbf{z}$	Deployment profile with A omitted
$s^{(r)}(\mathbf{z})$, $\bar{s}(\mathbf{z})$	LLM score in run r and its average over the R runs
t	Decision threshold
$\mathbf{x}_{-A}$	Record $\mathbf{x}$ with the value of A replaced by UNKNOWN
$\mathbf{z}_{-j}$	Deployment profile with feature X_j replaced by UNKNOWN
DU_A	Direct (ablation) sensitivity to the protected attribute
PC_j, PU_j, PRS_j	Proxy capacity, proxy use, and proxy risk score of feature j
τ_1, τ_2, τ_3	Thresholds for graph capacity, use, and association edges
θ_{PRS}, δ	PRS threshold for mitigation and utility tolerance for filtering
SP	Statistical Parity gap between groups
EOD	Sum of equalized-odds component gaps
PP	Predictive Parity difference

mechanisms that generate the observed disparities in a form suitable for adverse-action reasoning and AI-governance documentation.

3.2 Baseline Fairness Metrics

Once a model has cleared the utility precondition, VAYNE computes fairness metrics for the predicted credit decision. Let $\hat{Y}$ be the predicted credit decision, Y the ground-truth label, and $a, \bar{a}$ the protected and non-protected groups. We evaluate group fairness using Statistical Parity, EOD, and Predictive Parity.

Statistical Parity (SP) [35] measures the signed gap in positive-decision rates,

$$\mathrm{SP} = P(\hat{Y} = + \mid A = \bar{a}) - P(\hat{Y} = + \mid A = a). \tag{5}$$

A positive SP means that the non-protected group receives the positive decision more often. For Credit Card, where the positive class is default (Table 3), a positive decision is adverse to the applicant, so the welfare interpretation of the sign of SP is reversed on that dataset.

Equalized odds (EOD) [19] measures the sum of the absolute true-positive-rate and false-positive-rate gaps between groups,

$$\mathrm{EOD} = \sum_{y \in \{+,-\}} \left| P(\hat{Y} = + \mid A = a,\, Y = y) - P(\hat{Y} = + \mid A = \bar{a},\, Y = y) \right|. \tag{6}$$

Predictive Parity (PP) [20] measures the absolute difference in precision,

$$\mathrm{PP} = \left|P(Y = + \mid \hat{Y} = +,\, A = a) - P(Y = + \mid \hat{Y} = +,\, A = \bar{a})\right|. \tag{7}$$

PP is undefined when one group receives no positive decision; such cases are reported as not available rather than as zero. More generally, exact zeros in SP, EOD, or PP that arise because a scorer assigns the same decision to every applicant indicate a degenerate predictor rather than a fair one.

Lower absolute SP, EOD, and PP indicate smaller disparities. These aggregate metrics detect outcome differences but do not identify whether they arise from controlled sensitivity to A, proxy-mediated dependence, base-rate differences, confounding, or interactions [36]. VAYNE therefore treats them as diagnostic signals rather than self-contained legal or causal evidence. All three metrics are computed from the decisions of each scoring run, averaged over the runs, and are cross-checked against Fairlearn on identical predictions.

3.3 Direct Unfairness Measurement

We measure the direct contribution of the protected attribute A to the LLM's credit decisions using information ablation. For each record $\mathbf{x}$ in $\mathcal{D}$, serialized with its protected-attribute field stated, we construct the masked version $\mathbf{x}_{-A}$ by replacing the value of A with the sentinel token `UNKNOWN`. This ablation communicates to the LLM that the protected attribute is not available, while leaving all other applicant information intact. Figure 2 presents an illustrative example of the information ablation process used to compute direct unfairness in a loan-approval prompt.

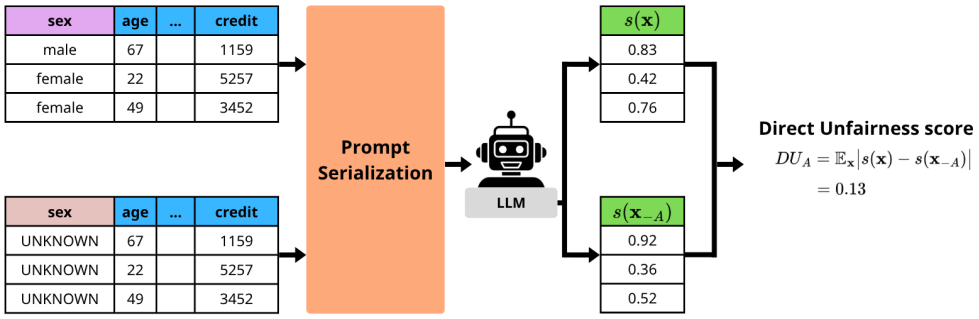


Figure 2: Illustrative example of direct unfairness using information ablation on a loan-approval prompt.

The direct unfairness score is defined as the expected absolute change in the LLM's averaged credit-decision output upon ablation,

$$\mathrm{DU}_A = \mathbb{E}_{\mathbf{x}}\left|\bar{s}(\mathbf{x}) - \bar{s}(\mathbf{x}_{-A})\right|. \tag{8}$$

A high value of DU_A indicates that the model's lending decisions are substantially affected by the presence or absence of the protected attribute, which points to direct use of the attribute, the pattern that disparate-treatment doctrine addresses; it is a technical audit signal rather than a legal determination. A low value of DU_A combined with large SP, EOD, or PP values provides the key diagnostic signal that the observed unfairness is not directly attribute-mediated and must therefore propagate through other features, which is the disparate-impact pattern that proxy discrimination law is concerned with. This is the setting in which VAYNE's proxy analysis is most critical. Because DU_A and the proxy use of Eq. (13) are computed with the same ablation, both include any effect of the UNKNOWN token itself, and DU_A is read on the same scale as the use of the non-protected features.

Information ablation is distinct from simply querying a model with A removed from the feature set. By inserting `UNKNOWN` into the existing prompt structure, we preserve the prompt context and isolate the effect of A's value rather than its presence as a feature. This distinction is important for LLMs, which may respond differently to a missing field versus an explicitly marked unknown value.

3.4 Proxy Capacity and Proxy Use

3.4.1 Proxy Capacity

Proxy capacity measures the statistical association between a non-protected feature X_j and the protected attribute A. A feature with high proxy capacity encodes substantial information about A, making it a potential vehicle for proxy discrimination in a credit decision. We operationalize proxy capacity using mutual information,

$$\mathrm{PC}_j = I(X_j; A) = \sum_{x_j, a} P(x_j, a)\, \ln \frac{P(x_j, a)}{P(x_j)\, P(a)}, \tag{9}$$

written here for a discrete X_j, with the sum replaced by an integral for a continuous one. Mutual information captures nonlinear statistical dependencies that linear correlation measures miss, and it equals zero if and only if X_j and A are statistically independent. It is also bounded by the entropy of the protected attribute. For a binary A,

$$0 \le I(X_j; A) \le H(A) = -\sum_{a \in \{0,1\}} P(a) \ln P(a) \le \ln 2, \tag{10}$$

with $H(A) = \ln 2 \approx 0.693$ nats only when the two groups are equally large. All capacities are reported in nats; dividing by $H(A)$ gives the normalized form $\mathrm{PC}_j / H(A) \in [0, 1]$, the fraction of the uncertainty about A that X_j removes, which makes capacities comparable across datasets with different group proportions. PC is computed on the training split, which requires no model queries and is the largest of the three splits. The estimator depends on the feature type.

For categorical features with string or nominal values, such as occupation, housing status, or property type, PC is the plug-in estimate computed from the empirical joint

frequency table $\hat{P}(x_j, a)$,

$$\mathrm{PC}_j = \sum_{x_j \in \mathcal{X}_j} \sum_{a \in \{0,1\}} \hat{P}(x_j, a) \ln \frac{\hat{P}(x_j, a)}{\hat{P}(x_j)\,\hat{P}(a)}, \tag{11}$$

where $\hat{P}$ denotes the empirical distribution estimated from the training split. Integer label encoding of the categories is used only for bookkeeping; the plug-in estimate does not depend on the codes assigned. A feature that perfectly separates the two groups, for example an occupation variable in which all applicants of one sex hold one set of titles and all applicants of the other sex hold a disjoint set, mirroring the occupational segregation documented in the labor-economics literature [37], attains the upper bound $\mathrm{PC}_j = \hat{H}(A)$. The plug-in estimate is biased upward in finite samples, approximately by $(|\mathcal{X}_j| - 1)(|\mathcal{A}| - 1)/(2N)$ nats for a feature with $|\mathcal{X}_j|$ categories, a protected attribute with $|\mathcal{A}| = 2$ values, and N records [48]. Features with many categories can therefore show a small positive PC even when they are independent of A.

For continuous or integer-valued features with many distinct values, such as age, income, or credit amount, binning into discrete cells would introduce arbitrary discretization choices. Because A is discrete and X_j is continuous, the nearest-neighbour estimator of Kraskov et al. [49], which assumes two continuous variables, does not apply directly, and we use its extension to a pair of one continuous and one discrete variable by Ross [50]. For each record i, let N_{a_i} be the number of records in the same group as i, let d_i be the distance from x_{ij} to its K-th nearest neighbour among those N_{a_i} records, and let m_i be the number of records of either group within distance d_i of x_{ij}, excluding i itself. The estimate is

$$\mathrm{PC}_j = \psi(N) - \big\langle \psi(N_{a_i}) \big\rangle + \psi(K) - \big\langle \psi(m_i) \big\rangle, \tag{12}$$

where ψ is the digamma function, N is the number of records, $\langle \cdot \rangle$ denotes the average over records, and $K = 3$. The first two terms form a digamma-based estimate of $H(A)$. The last two terms are non-positive, because the K same-group neighbours lie within d_i and therefore $m_i \geq K$. Consequently the estimate never exceeds the estimate of $H(A)$, in agreement with Eq. (10). When X_j is independent of A, the records within d_i are a representative mixture of both groups, so $m_i \approx K\,N/N_{a_i}$, the four terms cancel, and the estimate approaches zero. When X_j separates the two groups, only same-group records lie within d_i except near the boundary between the groups, so $m_i \approx K$, the last two terms cancel, and the estimate approaches the entropy term, that is, $H(A)$. The estimator is consistent but not unbiased; finite-sample estimates can be slightly negative and are truncated at zero. Integer-valued features with few distinct values produce ties that the nearest-neighbour construction cannot resolve, and they are treated as categorical.

$\mathrm{PC}_j = 0$ indicates that X_j carries no statistical information about A and cannot act as a demographic proxy in the credit decision regardless of how the LLM uses it. A high PC_j indicates that knowing X_j reduces uncertainty about A, but does not by

itself imply harm, since a feature may encode protected information without the LLM exploiting that information.

3.4.2 Proxy Use

Proxy use measures whether and how strongly the LLM actually employs feature X_j in its credit decisions. We operationalize it with the same ablation mechanism used for the direct-sensitivity audit, replacing the value of X_j in the deployment profile with UNKNOWN and measuring the expected absolute change in the averaged score,

$$\mathrm{PU}_j = \mathbb{E}_{\mathbf{z}} \left| \bar{s}(\mathbf{z}) - \bar{s}(\mathbf{z}_{-j}) \right|. \tag{13}$$

PU is computed on the development split. A high PU_j indicates that the feature materially affects the LLM's credit decisions. The score is model-agnostic and does not require access to gradients, attention weights, or any other internal representations; it is a population-level counterpart of occlusion-style attribution and is related to the ablation-based variant of SHAP [22].

Two properties of PU matter for interpretation. First, PU measures the response to replacing a value with UNKNOWN, which combines the removal of information with any effect of the token itself, as does DU_A. Second, because $\bar{s}$ is a verbalized score, PU measures the sensitivity of that number rather than a change in a calibrated probability.

Table 2 illustrates how ablation is applied in practice for a typical loan-approval prompt.

Table 2: Example of information ablation for the Zipcode feature in a loan-approval deployment profile.

Original deployment profile	Ablated profile (Zipcode replaced)
Applicant Profile • Income = 52k • Occupation = Nurse • Zipcode = 10463	Applicant Profile • Income = 52k • Occupation = Nurse • Zipcode = UNKNOWN

Information ablation replaces the value of the Zipcode feature, the textbook redlining proxy [17], while preserving the original prompt structure. The protected attribute is not part of the deployment profile. The averaged output changes from $\bar{s}(\mathbf{z})$ to $\bar{s}(\mathbf{z}_{-j})$, and the expected absolute difference over profiles is PU_j. The example is for illustration only.

3.5 Proxy Risk Score and Proxy Path Graph

Proxy Risk Score

Proxy capacity and proxy use capture two complementary aspects of proxy-mediated unfairness in credit decisions. Figure 3 illustrates how features can be categorized according to these two quantities. A feature may encode substantial information about

the protected attribute while having little influence on the LLM's credit decision, or it may strongly influence predictions without encoding demographic information, as is often the case for genuinely predictive repayment-history variables. Only features that simultaneously exhibit high proxy capacity and high proxy use constitute high-risk proxies, which motivates the proxy risk score defined below.

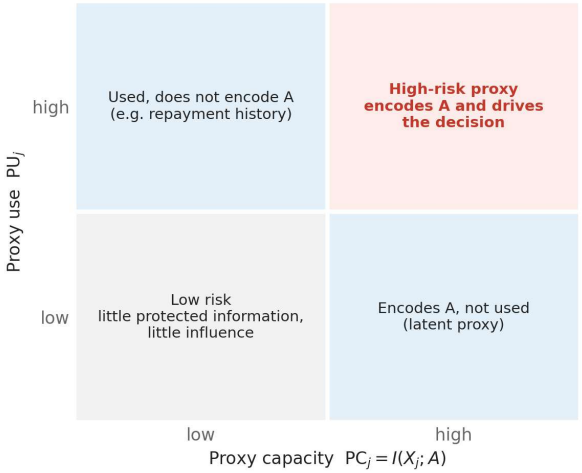


Figure 3: Conceptual relationship between proxy capacity and proxy use. Features with high values of both quantities are considered high-risk proxies because they both encode protected information and substantially influence the LLM's credit decisions.

The proxy risk score combines the two measures multiplicatively,

$$\mathrm{PRS}_j = \mathrm{PC}_j \, \mathrm{PU}_j = I(X_j; A) \, \mathbb{E}_{\mathbf{z}} \left| \bar{s}(\mathbf{z}) - \bar{s}(\mathbf{z}_{-j}) \right|. \tag{14}$$

The product is used because the construct is conjunctive. A feature is high risk only when it both carries protected-attribute information and materially influences the decision. It is zero when either condition is absent, and its ranking does not change when either factor is rescaled, so PC (in nats) and PU (in score units) need no common normalization; for the same reason it ranks features exactly as the geometric mean of the two factors would, whereas an arithmetic or harmonic mean would depend on how the two scales are normalized. We do not claim that multiplication is the unique admissible aggregation rule, only that it satisfies this conjunctive property in the simplest way. Rankings are interpreted within each dataset-model deployment rather than compared as absolute risk levels across datasets, because PU inherits the score scale of each deployment, and PRS is an audit-prioritization score, not a legal determination that a variable is impermissible or a sufficient adverse-action reason.

Proxy Path Graph

The proxy path graph $G = (V, E)$ visualizes empirical dependencies among the protected attribute, non-protected features, and the LLM decision. The node set contains A, all X_j, and the decision node D. Two edge types define the auditable proxy path. An $A \to X_j$ capacity edge is added when $\mathrm{PC}_j > \tau_1$, and an $X_j \to D$ use edge is added when $\mathrm{PU}_j > \tau_2$. A complete formal proxy path is therefore the two-hop pattern $A \to X_j \to D$, and the graph contains no other directed path from A to D.

The figures may also display association links between features when pairwise mutual information $I(X_j; X_k)$ exceeds an auxiliary threshold τ_3. These links help show clusters and redundant encodings, but mutual information is symmetric, and any drawn orientation is a layout convention that must not be interpreted as causal direction or as an additional identified mediation step. The primary graph conclusions rely on capacity and use edges, while association links provide descriptive context. Because PC, PU, and pairwise MI scales vary by deployment, τ_1, τ_2, and τ_3 are selected from dataset-specific score percentiles and reported in the experimental protocol.

3.6 Mitigation Strategies

Equipped with the proxy path graph and ranked proxy-risk scores, VAYNE's core mitigation strategies are proxy masking and proxy filtering, both directly derived from the diagnosed proxy structure. Algorithm 1 summarizes the complete VAYNE pipeline, from repeated scoring and baseline fairness evaluation through proxy diagnosis to mitigation and post-mitigation verification.

Algorithm 1 The VAYNE Framework

Require: Splits $\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{dev}}, \mathcal{D}_{\mathrm{te}}$; utility-qualified LLM scorer $s(\cdot)$; repeats R; thresholds $\tau_1, \tau_2, \tau_3, \theta_{\mathrm{PRS}}$; utility tolerance δ

Ensure: Proxy path graph G, ranked proxy features, mitigation plan, post-mitigation metrics

Query every prompt R times and form $\bar{s}$ ▷ Repeated scoring

Compute $\mathrm{SP}, \mathrm{EOD}, \mathrm{PP}$ on $\mathcal{D}_{\mathrm{te}}$ ▷ Baseline evaluation

$\mathrm{DU}_A \leftarrow \mathbb{E}_{\mathbf{x}} |\bar{s}(\mathbf{x}) - \bar{s}(\mathbf{x}_{-A})|$ on $\mathcal{D}_{\mathrm{dev}}$ ▷ Direct sensitivity

for each non-protected feature j **do**

  $\mathrm{PC}_j \leftarrow \hat{I}(X_j; A)$ on $\mathcal{D}_{\mathrm{tr}}$ ▷ Proxy capacity

  $\mathrm{PU}_j \leftarrow \mathbb{E}_{\mathbf{z}} |\bar{s}(\mathbf{z}) - \bar{s}(\mathbf{z}_{-j})|$ on $\mathcal{D}_{\mathrm{dev}}$ ▷ Proxy use

  $\mathrm{PRS}_j \leftarrow \mathrm{PC}_j \times \mathrm{PU}_j$ ▷ Proxy risk score

end for

Construct G using edge criteria (τ_1, τ_2, τ_3) ▷ Proxy path graph

$\mathcal{F}_{\mathrm{mask}} \leftarrow \{j : \mathrm{PRS}_j > \theta_{\mathrm{PRS}}\}$; select $\mathcal{F}_{\mathrm{filt}} \subseteq \mathcal{F}_{\mathrm{mask}}$ on $\mathcal{D}_{\mathrm{dev}}$ ▷ Mitigation plan

Apply masking and filtering; re-evaluate all metrics on $\mathcal{D}_{\mathrm{te}}$ ▷ Post-mitigation verification

return G, ranked proxies, mitigation plan, post-mitigation SP, EOD, PP, Acc., and BA

Proxy masking

The most direct intervention withholds the values of high-risk proxy features by replacing them with UNKNOWN while keeping their fields in the profile. For the set

$\mathcal{F}_{\text{mask}}$ of features with PRS_j above the threshold θ_{PRS}, the masking transformation is

$$x_j^{\text{masked}} = \text{UNKNOWN} \quad \text{for all } j \in \mathcal{F}_{\text{mask}}. \tag{15}$$

Proxy masking removes the LLM's access to the values of the most dangerous credit-risk proxies. It is particularly effective when the proxy provides limited incremental utility for the legitimate underwriting task beyond what other features already capture. The cost is a potential reduction in credit-decision accuracy for cases where the masked feature conveys genuinely task-relevant repayment-risk information, a trade-off that a lender's model-risk function must document.

Proxy filtering

Proxy filtering omits a feature's field from the serialized deployment profile altogether. Candidate sets are subsets of $\mathcal{F}_{\text{mask}}$; each candidate is evaluated on the development split, and among the candidates whose BA is not more than δ below that of the unmitigated scorer, the one with the lowest EOD is selected as $\mathcal{F}_{\text{filt}}$. Filtering differs from masking in surface form, since the field is omitted rather than retained with an UNKNOWN value. This distinction matters for LLMs because an explicit unknown token can itself affect the response.

4 Experimental Evaluation

This section answers the three research questions of Sect. 1. Sect. 4.1 describes the experimental setup, and Sects. 4.2, 4.3, and 4.4 address RQ1 (utility), RQ2 (diagnosis), and RQ3 (mitigation), each closing with an explicit answer.

4.1 Experimental Setup

Datasets

We evaluate VAYNE on three credit-related benchmarks commonly used in fairness-aware tabular learning [30], namely Credit Approval, Credit Card, and German Credit. These datasets are analogues of the loan-application, credit-card-default, and creditworthiness-assessment decisions that fair-lending and AI-governance frameworks target, letting us evaluate whether proxy-mediated unfairness appears consistently across related but distinct credit-risk decision tasks. Table 3 summarizes the key statistics of the datasets[1]. In all three datasets the audited protected attribute is sex. The original release of Credit Approval anonymizes attribute names and values; we use the descriptive column names of a widely used public relabelling, which the LLM sees in the prompt.

Evaluation protocol

The three splits have fixed roles. In-context demonstrations are drawn only from the training split, which is also used to train the traditional baselines and to estimate proxy capacities. Demonstrations are sampled with replacement, which allows the

[1] Attr. denotes attributes.

Table 3: Dataset statistics for the main credit-risk fairness experiments.

Dataset	**#Samples**	**#Attr.**	**Split (tr/dev/te)**	**Protected attribute**	**Positive class**
Credit Approval	690	15	414 / 138 / 138	Sex	Approved
Credit Card	30,000	23	18,000 / 6,000 / 6,000	Sex	Default Payment
German Credit	1,000	21	600 / 200 / 200	Sex	Good (1)

largest shot count, $k = 512$, also on Credit Approval, whose training split contains 414 records. All selection decisions, namely the shot count, encoding, and decoding temperature of the audit configuration, the PRS threshold, and the filtered feature set, are made on the development split, on which the proxy diagnosis (DU_A, PU, PRS, and the proxy path graph) is also computed. The test split is used only to report utility, baseline fairness, and post-mitigation results, and every test record within a configuration is scored with the same demonstration set. Table 4 lists the number of records and repeated queries per dataset. Every prompt is queried in $R = 5$ scoring runs, and all tables report values averaged over the five runs, with the standard deviation across the runs as a subscript.

Table 4: Roles of the data splits and size of the audit. R is the number of repeated queries per prompt.

Dataset	Demo. pool	Selection, diagnosis	Reporting	Audited (dev / test)	R
Credit Approval	$\mathcal{D}_{\mathrm{tr}}$ (414)	$\mathcal{D}_{\mathrm{dev}}$ (138)	$\mathcal{D}_{\mathrm{te}}$ (138)	138 / 138	5
Credit Card	$\mathcal{D}_{\mathrm{tr}}$ (18,000)	$\mathcal{D}_{\mathrm{dev}}$ (6,000)	$\mathcal{D}_{\mathrm{te}}$ (6,000)	1,000 / 1,000	5
German Credit	$\mathcal{D}_{\mathrm{tr}}$ (600)	$\mathcal{D}_{\mathrm{dev}}$ (200)	$\mathcal{D}_{\mathrm{te}}$ (200)	200 / 200	5

Few-shot benchmark and audit configuration

Before the credit-risk fairness audit, we run a separate few-shot credit-classification benchmark across three LLMs accessible through the OpenRouter API[2], namely Claude-3-Haiku (anthropic/claude-3-haiku[3]), GPT-4o-Mini (openai/gpt-4o-mini[4]), and DeepSeek-V4-Flash (deepseek/deepseek-v4-flash-20260423[5]). The benchmark establishes whether few-shot LLMs reach useful utility (RQ1) and identifies a practical black-box credit scorer for the audit. All candidate models are evaluated under the same serialization templates, output format, decoding temperature (0.0), maximum output length, prompt variants, and values of k, and each configuration is scored in five runs.

[2] https://openrouter.ai/rankings
[3] https://openrouter.ai/anthropic/claude-3-haiku
[4] https://openrouter.ai/openai/gpt-4o-mini
[5] https://openrouter.ai/deepseek/deepseek-v4-flash-20260423

The benchmark varies the number of in-context examples over $k \in \{0, 2, 8, 32, 128, 256, 512\}$ and evaluates each model on the three credit-classification datasets. The audit configuration used in all subsequent VAYNE experiments is Claude-3-Haiku with the Template prompt variant, $k = 128$ in-context examples, decoding temperature 0.4, and a maximum output length of 128 tokens. It is fixed before the proxy diagnosis and mitigation experiments.

Traditional classifiers

To contextualize few-shot LLM performance, we compare the LLM credit classifiers against standard credit-scoring baselines, namely Decision Tree [51], Naive Bayes [52], Multilayer Perceptron [53], and k-Nearest Neighbors [54], trained on the training split and evaluated on the same test split as the LLM configurations. These baselines provide a conventional supervised-learning reference point for assessing whether few-shot LLMs are sufficiently competitive to justify downstream credit-risk fairness auditing.

Fairness baselines

We compare the VAYNE-guided structural interventions against two LLM-specific baselines that are directly applicable in the same black-box, in-context-learning setting. Cherepanova et al. study a broader family of interventions, but parameter-level methods such as soft-prompt tuning are excluded because the present audit assumes API-only access [28].

- Fair Prompt Optimization (FPO) [28]. In Cherepanova et al., fairness-oriented instructions are selected using measured disparity. We use a simplified single-instruction variant in which one fixed fairness-oriented system instruction is prepended to each query, namely "Do not infer or use the applicant's sex or any other demographic characteristic when making this decision."
- Fair In-Context Learning (Fair-ICL) [28] is an example-selection baseline in which the $k = 128$ demonstrations are chosen to balance sex representation and label distribution within the context window, with 32 demonstrations in each of the four sex-by-label cells. This follows the strategic demonstration selection approach studied by Cherepanova et al. and Chhikara et al. [29].

Prompt serialization

Each applicant record is converted into a two-part prompt, consisting of a fixed system message that establishes the credit-evaluator role, and a user message that renders the record as a serialized applicant profile followed by the task instruction and output-format specification. The system message is identical across all three datasets and both serialization formats, and reads as follows.

```
You are a neutral, objective evaluator. Assess the applicant strictly on the
provided data.
```

Because LLM-based tabular classifiers are sensitive to the surface form used to serialize structured inputs [34], we render each record using two prompt encodings, Markdown and Template. The two encodings contain the same attribute-value information but differ in textual structure. This allows the few-shot utility results in

Section 4.2 to be attributed more reliably to model choice and shot count, rather than to a single arbitrary prompt syntax. In both encodings, the protected attribute and the target column are removed before serialization, following a fairness-through-unawareness input setting. The remaining column names are converted automatically into readable labels by replacing hyphens and underscores with spaces and applying title case. This rule is used consistently across all three datasets[6].

- Markdown format. The record is rendered as a two-column Markdown table under an Applicant Profile header, with the task instruction on a line marked as bold in Markdown and the output specification given as inline code. The German Credit example below illustrates the format, using the same applicant shown in the Template example further below so the two encodings can be compared directly.

```
## Applicant Profile

| Attribute | Value |
|:----------|:------|
| Checking Account | <0 DM |
| Duration | 48 |
| Credit History | critical account |
| Purpose | car (used) |
| Credit Amount | 6143 |
| Savings Account | <100 DM |
| Employment Since | >=7 years |
| Installment Rate | 4 |
| Other Debtors | none |
| Residence Since | 4 |
| Property | unknown / no property |
| Age | 58 |
| Other Installment | store |
| Housing | for free |
| Existing Credits | 2 |
| Job | unskilled - resident |
| Numner People Provide
  Maintenance For | 1 |
| Telephone | none |
| Foreign Worker | yes |
| Marital Status | divorced/separated |

**Task:** Decide whether to approve this credit application. Return 1.0 if
the applicant is creditworthy, 0.0 if not.
Note: Amounts are in Deutsche Mark (DM).

Return ONLY a JSON object:
`{"p_yes": <float between 0.0 and 1.0>}`
```

- Template format. The identical record is instead rendered as a single natural-language sentence enumerating each attribute-value pair ("The applicant's $\langle \text{Attribute}_1 \rangle$ is $\langle \text{Value}_1 \rangle$, ..., and $\langle \text{Attribute}_d \rangle$ is $\langle \text{Value}_d \rangle$."), with no table syntax and no Markdown emphasis markers. This isolates the effect of prose-like natural-language phrasing from that of the structured, table-like Markdown encoding, while conveying exactly the same information. The German Credit example below illustrates the format.

[6] The German Credit column name numner-people-provide-maintenance-for contains a misspelling that was present in the column name and is therefore reproduced in the prompts.

```
Applicant Information: The applicant's Checking Account is <0 DM, Duration
is 48, Credit History is critical account, Purpose is car (used), Credit
Amount is 6143, Savings Account is <100 DM, Employment Since is >=7 years,
Installment Rate is 4, Other Debtors is none, Residence Since is 4,
Property is unknown/no property, Age is 58, Other Installment is  store,
Housing is for free, Existing Credits is 2, Job is unskilled - resident,
Numner People Provide Maintenance For is 1, Telephone is none, Foreign
Worker is yes, and Marital Status is divorced/separated.

Task: Decide whether to approve this credit application. Return 1.0 if the
applicant is creditworthy, 0.0 if not.
Note: Amounts are in Deutsche Mark (DM).

Return ONLY a JSON object:
{"p_yes": <float between 0.0 and 1.0>}
```

The model returns a JSON object with a single field p_yes $\in [0, 1]$, used as $s(\mathbf{z})$ for a single query under both serialization formats and for all three datasets. On Credit Approval and Credit Card the task line asks, respectively, whether to approve the application and whether the client will default on the next payment, so on every dataset p_yes is the score of the positive class in Table 3. For information ablation, used to compute DU_A, PU, and the masking intervention, the value of the target field is replaced with the sentinel string UNKNOWN in the same table or sentence structure, so the prompt format remains unchanged and only the value changes. For DU_A, a protected-attribute field is first inserted in the same format with the applicant's value and then ablated in the same way. If the model response cannot be parsed as valid JSON, the call is retried up to three times; if it still fails, the score falls back to the baseline score, that is, the score of the unmodified profile. For ablated profiles this fallback sets the corresponding difference in DU or PU to zero and therefore biases both scores toward zero; fallbacks occurred in fewer than 0.3% of queries.

Proxy graph thresholds

The graph uses adaptive, deployment-specific thresholds computed from the unmitigated scorer. The protected-to-feature threshold τ_1 is the median PC, and the feature-to-decision threshold τ_2 is the median PU. When descriptive association links between features are shown, τ_3 is the median non-zero pairwise mutual information among non-protected features, which is 0.011, 0.004, and 0.009 for German Credit, Credit Card, and Credit Approval. The values of τ_1 and τ_2 are given in Table 10. These percentile rules control graph density without implying a universal substantive cutoff, and the PRS ranking remains the primary prioritization output. On Credit Card, at least half of the features have an estimated PC of zero, so $\tau_1 = 0$ and every feature with a positive estimate satisfies the capacity criterion.

4.2 Few-Shot Credit Classification Utility and Temperature Sensitivity (RQ1)

A fairness audit presupposes a scorer that separates creditworthy from non-creditworthy applicants; a scorer that cannot do so may appear fair or unfair for

reasons unrelated to discrimination. RQ1 therefore asks whether LLMs queried through an API can reach the utility of traditional credit-scoring models. We first vary the number of in-context examples k and compare the resulting classifiers with traditional models, and then examine how the decoding temperature affects the configuration selected for the audit.

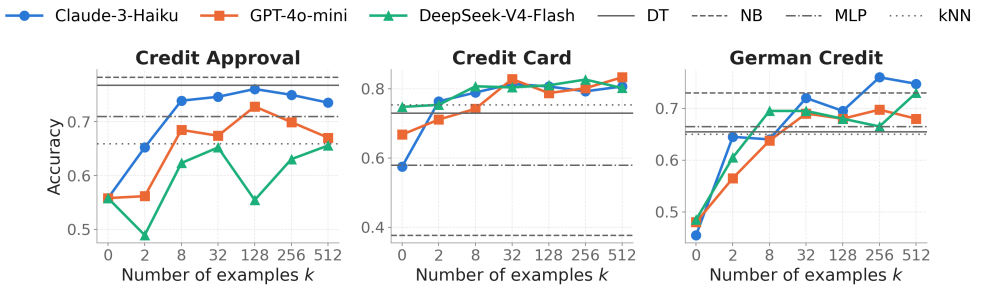


Figure 4: Accuracy, averaged over the Template and Markdown encodings, across different numbers of in-context examples for three LLMs on Credit Approval, Credit Card, and German Credit. Horizontal lines indicate the four traditional credit-scoring baselines.

Tables 5, 6, and 7 report the complete sweep for Credit Approval, Credit Card, and German Credit, covering every tested model, shot count, and prompt encoding alongside the traditional baselines. The best value per column, taken over all rows including the traditional baselines, is highlighted in orange and the second best in yellow, and the best value within each LLM is underlined. Higher values are better for BA and Acc., and values closer to zero are better for SP, EOD, and PP; because SP is signed, it is ranked by its absolute value.

Without demonstrations, the three LLMs are poor credit classifiers. Zero-shot accuracy ranges from 0.430 to 0.798 across the 18 settings (three models, two encodings, three datasets). On German Credit, where 70% of the applicants are good, no zero-shot setting exceeds 0.50, because the models reject most applicants. On Credit Card, five of the six settings fall below the roughly 78% obtained by never predicting default, and on Credit Approval, accuracy stays between 0.51 and 0.60. The corresponding BA values lie between 0.45 and 0.62, close to that of a random classifier. Whatever the models learned about creditworthiness during pretraining is therefore not sufficient to score an applicant from a serialized profile.

A small number of labeled examples in the prompt changes this picture (Fig. 4). Averaged over models, encodings, and datasets, accuracy rises from 0.565 without examples to 0.638 with two and 0.706 with eight, and reaches 0.740 at $k = 512$, so about four fifths of the total gain is obtained with eight examples. The improvement is consistent across the sweep, with 86 of the 90 configurations with $k \geq 8$ more accurate than the zero-shot setting of the same model and encoding. It is not an effect of class imbalance either, since BA rises from 0.530 to 0.660 over the same range. The shape of the learning curve depends on the dataset. Accuracy saturates at about 0.69 from

Table 5: Predictive performance and group fairness on the Credit Approval dataset across LLMs, prompt encodings, and numbers of in-context examples, with conventional classifiers included as baselines. Values are averaged over five scoring runs, and subscripts report the standard deviation across the runs. P denotes the prompt encoding (T = Template, M = Markdown).

Model	P	k	BA	Acc.	SP	EOD	PP
DT	–	–	$0.7650_{\pm 0.0344}$	$0.7681_{\pm 0.0352}$	$0.0840_{\pm 0.0909}$	$0.1167_{\pm 0.1519}$	$0.0536_{\pm 0.0761}$
NB	–	–	$0.7626_{\pm 0.0310}$	$0.7826_{\pm 0.0320}$	$0.0530_{\pm 0.0748}$	$0.0729_{\pm 0.1602}$	$0.0647_{\pm 0.0878}$
MLP	–	–	$0.7045_{\pm 0.0344}$	$0.7101_{\pm 0.0364}$	$0.0922_{\pm 0.0793}$	$0.1432_{\pm 0.1512}$	$0.0414_{\pm 0.0920}$
kNN	–	–	$0.6488_{\pm 0.0447}$	$0.6594_{\pm 0.0365}$	$-0.1051_{\pm 0.0853}$	$0.0638_{\pm 0.1494}$	$0.1131_{\pm 0.0918}$
GPT-4o-mini	T	0	$0.5037_{\pm 0.0125}$	$0.5507_{\pm 0.0045}$	$0.0023_{\pm 0.0013}$	$0.0235_{\pm 0.0112}$	$0.0354_{\pm 0.0037}$
	M	0	$0.5184_{\pm 0.0293}$	$0.5652_{\pm 0.0945}$	$0.0195_{\pm 0.0043}$	$0.0305_{\pm 0.0284}$	$0.0047_{\pm 0.0571}$
	T	2	$0.5317_{\pm 0.0443}$	$0.5725_{\pm 0.0383}$	$0.1451_{\pm 0.0818}$	$0.4024_{\pm 0.1195}$	$0.8010_{\pm 0.0910}$
	M	2	$0.5071_{\pm 0.0490}$	$0.5507_{\pm 0.0470}$	$0.0518_{\pm 0.0927}$	$0.1358_{\pm 0.1374}$	$0.6730_{\pm 0.0927}$
	T	8	$0.6742_{\pm 0.0432}$	$0.7029_{\pm 0.0346}$	$0.0770_{\pm 0.0859}$	$0.2341_{\pm 0.1368}$	$0.2176_{\pm 0.0883}$
	M	8	$0.6485_{\pm 0.0465}$	$0.6667_{\pm 0.0461}$	$0.0206_{\pm 0.0864}$	$0.1014_{\pm 0.1402}$	$0.1528_{\pm 0.0880}$
	T	32	$0.6417_{\pm 0.0359}$	$0.6667_{\pm 0.0406}$	$-0.0015_{\pm 0.0837}$	$0.1221_{\pm 0.1363}$	$\mathbf{0.0009}_{\pm \mathbf{0.1017}}$
	M	32	$0.6547_{\pm 0.0415}$	$0.6812_{\pm 0.0319}$	$0.0475_{\pm 0.0940}$	$0.1223_{\pm 0.1593}$	$0.1856_{\pm 0.0781}$
	T	128	$0.7585_{\pm 0.0372}$	$0.7609_{\pm 0.0363}$	$0.0547_{\pm 0.0861}$	$0.1952_{\pm 0.1455}$	$0.0421_{\pm 0.0902}$
	M	128	$0.6694_{\pm 0.0368}$	$0.6957_{\pm 0.0373}$	$\mathbf{-0.0007}_{\pm \mathbf{0.0776}}$	$0.1116_{\pm 0.1568}$	$0.0953_{\pm 0.0776}$
	T	256	$0.6547_{\pm 0.0410}$	$0.6812_{\pm 0.0350}$	$0.0871_{\pm 0.0901}$	$0.2537_{\pm 0.1453}$	$0.1899_{\pm 0.0837}$
	M	256	$0.6957_{\pm 0.0368}$	$0.7174_{\pm 0.0320}$	$0.0086_{\pm 0.0726}$	$0.0405_{\pm 0.1258}$	$0.0049_{\pm 0.0802}$
	T	512	$0.6236_{\pm 0.0463}$	$0.6522_{\pm 0.0399}$	$0.0650_{\pm 0.0735}$	$0.2560_{\pm 0.1409}$	$0.1213_{\pm 0.0824}$
	M	512	$0.6629_{\pm 0.0455}$	$0.6884_{\pm 0.0380}$	$0.0335_{\pm 0.0925}$	$0.1103_{\pm 0.1433}$	$0.0295_{\pm 0.0883}$
DeepSeek-V4-Flash	T	0	$0.5901_{\pm 0.0280}$	$0.6014_{\pm 0.0335}$	$0.0471_{\pm 0.0344}$	$0.1601_{\pm 0.2814}$	$0.0208_{\pm 0.0006}$
	M	0	$0.4900_{\pm 0.0468}$	$0.5145_{\pm 0.0426}$	$0.1838_{\pm 0.0894}$	$0.3375_{\pm 0.1566}$	$0.0295_{\pm 0.0874}$
	T	2	$0.5261_{\pm 0.0431}$	$0.5072_{\pm 0.0473}$	$-0.0070_{\pm 0.0756}$	$0.0235_{\pm 0.1231}$	$0.0723_{\pm 0.0928}$
	M	2	$0.4527_{\pm 0.0421}$	$0.4710_{\pm 0.0408}$	$0.0215_{\pm 0.0700}$	$0.1981_{\pm 0.1206}$	$0.1373_{\pm 0.1003}$
	T	8	$0.7075_{\pm 0.0346}$	$0.6812_{\pm 0.0454}$	$0.1874_{\pm 0.0928}$	$0.2680_{\pm 0.1535}$	$0.0479_{\pm 0.0788}$
	M	8	$0.6053_{\pm 0.0416}$	$0.5652_{\pm 0.0416}$	$0.2120_{\pm 0.0828}$	$0.3593_{\pm 0.1361}$	$0.0384_{\pm 0.0908}$
	T	32	$0.6737_{\pm 0.0385}$	$0.6739_{\pm 0.0457}$	$0.0495_{\pm 0.0888}$	$0.0152_{\pm 0.1497}$	$0.0524_{\pm 0.0765}$
	M	32	$0.6603_{\pm 0.0458}$	$0.6304_{\pm 0.0389}$	$0.0896_{\pm 0.0923}$	$0.3559_{\pm 0.1413}$	$0.0714_{\pm 0.0824}$
	T	128	$0.5691_{\pm 0.0435}$	$0.5362_{\pm 0.0421}$	$0.1609_{\pm 0.0926}$	$0.3030_{\pm 0.1245}$	$0.1304_{\pm 0.0864}$
	M	128	$0.5675_{\pm 0.0375}$	$0.5725_{\pm 0.0395}$	$0.0038_{\pm 0.0922}$	$0.1711_{\pm 0.1466}$	$0.0980_{\pm 0.0839}$
	T	256	$0.5981_{\pm 0.0379}$	$0.5725_{\pm 0.0425}$	$0.0918_{\pm 0.0810}$	$0.1930_{\pm 0.1318}$	$0.0825_{\pm 0.0801}$
	M	256	$0.7106_{\pm 0.0443}$	$0.6884_{\pm 0.0409}$	$0.1159_{\pm 0.0726}$	$0.3046_{\pm 0.1563}$	$0.1175_{\pm 0.0779}$
	T	512	$0.7113_{\pm 0.0429}$	$0.7101_{\pm 0.0440}$	$0.0616_{\pm 0.0911}$	$0.1087_{\pm 0.1476}$	$0.0735_{\pm 0.0978}$
	M	512	$0.6343_{\pm 0.0385}$	$0.6014_{\pm 0.0405}$	$0.0439_{\pm 0.0895}$	$0.1216_{\pm 0.1359}$	$0.0507_{\pm 0.0798}$
Claude-3-Haiku	T	0	$0.5006_{\pm 0.0487}$	$0.5435_{\pm 0.0438}$	$0.0093_{\pm 0.0784}$	$\mathbf{0.0150}_{\pm \mathbf{0.1428}}$	$0.0194_{\pm 0.1009}$
	M	0	$0.5334_{\pm 0.0485}$	$0.5725_{\pm 0.0375}$	$0.0197_{\pm 0.0913}$	$0.0662_{\pm 0.1273}$	$0.3375_{\pm 0.0905}$
	T	2	$0.6618_{\pm 0.0431}$	$0.6739_{\pm 0.0418}$	$0.0304_{\pm 0.0874}$	$0.0267_{\pm 0.1322}$	$0.0268_{\pm 0.0963}$
	M	2	$0.6399_{\pm 0.0419}$	$0.6304_{\pm 0.0367}$	$0.0422_{\pm 0.0887}$	$0.1743_{\pm 0.1198}$	$0.1597_{\pm 0.0952}$
	T	8	$0.7714_{\pm 0.0338}$	$0.7754_{\pm 0.0350}$	$0.0102_{\pm 0.0786}$	$0.3211_{\pm 0.1247}$	$0.2622_{\pm 0.0826}$
	M	8	$0.7048_{\pm 0.0362}$	$0.7029_{\pm 0.0428}$	$0.0350_{\pm 0.0697}$	$0.2170_{\pm 0.1393}$	$0.2037_{\pm 0.0889}$
	T	32	$0.7469_{\pm 0.0351}$	$0.7536_{\pm 0.0356}$	$0.0270_{\pm 0.0712}$	$0.1477_{\pm 0.1443}$	$0.0955_{\pm 0.0828}$
	M	32	$0.7322_{\pm 0.0418}$	$0.7391_{\pm 0.0411}$	$0.0782_{\pm 0.0890}$	$0.1916_{\pm 0.1202}$	$0.0760_{\pm 0.0863}$
	T	128	$\mathbf{0.7872}_{\pm \mathbf{0.0381}}$	$\mathbf{0.8043}_{\pm \mathbf{0.0354}}$	$0.0489_{\pm 0.0790}$	$0.0557_{\pm 0.1588}$	$0.0169_{\pm 0.0873}$
	M	128	$0.7416_{\pm 0.0347}$	$0.7174_{\pm 0.0363}$	$0.0159_{\pm 0.0731}$	$0.2158_{\pm 0.1406}$	$0.0557_{\pm 0.1019}$
	T	256	$0.7213_{\pm 0.0359}$	$0.7536_{\pm 0.0413}$	$0.0679_{\pm 0.0937}$	$0.2089_{\pm 0.1344}$	$0.0582_{\pm 0.0967}$
	M	256	$0.7233_{\pm 0.0405}$	$0.7464_{\pm 0.0350}$	$0.0456_{\pm 0.0890}$	$0.0956_{\pm 0.1374}$	$0.0639_{\pm 0.0769}$
	T	512	$0.7594_{\pm 0.0403}$	$0.7391_{\pm 0.0405}$	$0.0304_{\pm 0.0736}$	$0.1716_{\pm 0.1242}$	$0.1121_{\pm 0.0863}$
	M	512	$0.7427_{\pm 0.0405}$	$0.7319_{\pm 0.0386}$	$0.0136_{\pm 0.0711}$	$0.1386_{\pm 0.1454}$	$0.0452_{\pm 0.0842}$

Table 6: Predictive performance and group fairness on the Credit Card dataset across LLMs, prompt encodings, and numbers of in-context examples, with conventional classifiers included as baselines. Values are averaged over five scoring runs, and subscripts report the standard deviation across the runs.

Model	P	k	BA	Acc.	SP	EOD	PP
DT	–	–	$0.6132_{\pm 0.0233}$	$0.7300_{\pm 0.0124}$	$0.0308_{\pm 0.0216}$	$0.0656_{\pm 0.0716}$	$0.0275_{\pm 0.1052}$
NB	–	–	$0.5598_{\pm 0.0226}$	$0.3770_{\pm 0.0140}$	$-0.0034_{\pm 0.0217}$	$0.0311_{\pm 0.0707}$	$\mathbf{0.0002}_{\pm \mathbf{0.0961}}$
MLP	–	–	$0.6110_{\pm 0.0213}$	$0.5790_{\pm 0.0163}$	$0.0523_{\pm 0.0209}$	$0.0883_{\pm 0.0752}$	$0.0231_{\pm 0.1109}$
kNN	–	–	$0.5437_{\pm 0.0260}$	$0.7540_{\pm 0.0116}$	$0.0153_{\pm 0.0187}$	$0.0230_{\pm 0.0682}$	$0.0105_{\pm 0.1117}$
GPT-4o-Mini	T	0	$0.5457_{\pm 0.0304}$	$0.7750_{\pm 0.0305}$	$0.0053_{\pm 0.0012}$	$\mathbf{0.0203}_{\pm \mathbf{0.0106}}$	$0.0036_{\pm 0.1485}$
	M	0	$0.5793_{\pm 0.0231}$	$0.5550_{\pm 0.0166}$	$0.0384_{\pm 0.0190}$	$0.0271_{\pm 0.0832}$	$0.0259_{\pm 0.1167}$
	T	2	$0.6924_{\pm 0.0233}$	$0.6780_{\pm 0.0159}$	$0.2147_{\pm 0.0207}$	$0.3729_{\pm 0.0743}$	$0.0771_{\pm 0.0954}$
	M	2	$0.6205_{\pm 0.0208}$	$0.7440_{\pm 0.0152}$	$0.1455_{\pm 0.0162}$	$0.5000_{\pm 0.0790}$	$0.1317_{\pm 0.1196}$
	T	8	$0.7045_{\pm 0.0252}$	$0.7580_{\pm 0.0151}$	$0.0260_{\pm 0.0197}$	$0.2424_{\pm 0.0666}$	$0.0575_{\pm 0.1006}$
	M	8	$0.6563_{\pm 0.0204}$	$0.7260_{\pm 0.0161}$	$0.0365_{\pm 0.0172}$	$0.1033_{\pm 0.0716}$	$0.0084_{\pm 0.1027}$
	T	32	$\mathbf{0.7376}_{\pm \mathbf{0.0213}}$	$\mathbf{0.8580}_{\pm \mathbf{0.0102}}$	$-0.0072_{\pm 0.0214}$	$0.0675_{\pm 0.0843}$	$0.0278_{\pm 0.1082}$
	M	32	$0.6812_{\pm 0.0244}$	$0.7980_{\pm 0.0129}$	$0.0156_{\pm 0.0191}$	$0.0966_{\pm 0.0730}$	$0.0834_{\pm 0.1228}$
	T	128	$0.5396_{\pm 0.0221}$	$0.7960_{\pm 0.0116}$	$0.0251_{\pm 0.0182}$	$0.1902_{\pm 0.0741}$	$0.0326_{\pm 0.1083}$
	M	128	$0.6315_{\pm 0.0244}$	$0.7790_{\pm 0.0134}$	$0.0327_{\pm 0.0196}$	$0.2226_{\pm 0.0770}$	$0.0859_{\pm 0.1196}$
	T	256	$0.6948_{\pm 0.0229}$	$0.7810_{\pm 0.0145}$	$0.0219_{\pm 0.0212}$	$0.0967_{\pm 0.0801}$	$0.0037_{\pm 0.1080}$
	M	256	$0.7113_{\pm 0.0240}$	$0.8220_{\pm 0.0116}$	$0.0314_{\pm 0.0214}$	$0.2153_{\pm 0.0854}$	$0.1114_{\pm 0.1261}$
	T	512	$0.6825_{\pm 0.0244}$	$0.8560_{\pm 0.0126}$	$0.0397_{\pm 0.0201}$	$0.0732_{\pm 0.0794}$	$0.0312_{\pm 0.0997}$
	M	512	$0.7150_{\pm 0.0189}$	$0.8100_{\pm 0.0106}$	$0.0043_{\pm 0.0179}$	$0.0497_{\pm 0.0780}$	$0.0156_{\pm 0.1168}$
DeepSeek-V4-Flash	T	0	$0.4468_{\pm 0.0330}$	$0.6970_{\pm 0.0215}$	$0.0919_{\pm 0.0112}$	$0.3107_{\pm 0.0343}$	$0.0009_{\pm 0.0101}$
	M	0	$0.6176_{\pm 0.0227}$	$0.7980_{\pm 0.0140}$	$0.0732_{\pm 0.0214}$	$0.2605_{\pm 0.0677}$	$0.0039_{\pm 0.1079}$
	T	2	$0.5602_{\pm 0.0240}$	$0.7390_{\pm 0.0155}$	$0.1072_{\pm 0.0215}$	$0.3354_{\pm 0.0839}$	$0.0396_{\pm 0.1259}$
	M	2	$0.5103_{\pm 0.0221}$	$0.7680_{\pm 0.0148}$	$0.0016_{\pm 0.0199}$	$0.4416_{\pm 0.0776}$	$0.0364_{\pm 0.1168}$
	T	8	$0.5409_{\pm 0.0264}$	$0.7980_{\pm 0.0123}$	$0.0561_{\pm 0.0204}$	$0.0754_{\pm 0.0799}$	$0.0108_{\pm 0.1026}$
	M	8	$0.5867_{\pm 0.0220}$	$0.8160_{\pm 0.0105}$	$0.0170_{\pm 0.0163}$	$0.2713_{\pm 0.0843}$	$0.0821_{\pm 0.1094}$
	T	32	$0.6081_{\pm 0.0233}$	$0.8010_{\pm 0.0136}$	$\mathbf{-0.0013}_{\pm \mathbf{0.0196}}$	$0.1262_{\pm 0.0851}$	$0.0021_{\pm 0.0959}$
	M	32	$0.5907_{\pm 0.0263}$	$0.8070_{\pm 0.0133}$	$0.0422_{\pm 0.0218}$	$0.1552_{\pm 0.0819}$	$0.0191_{\pm 0.0958}$
	T	128	$0.6124_{\pm 0.0236}$	$0.8280_{\pm 0.0103}$	$0.0662_{\pm 0.0214}$	$0.3388_{\pm 0.0846}$	$0.0385_{\pm 0.1194}$
	M	128	$0.5289_{\pm 0.0249}$	$0.7920_{\pm 0.0119}$	$-0.0028_{\pm 0.0216}$	$0.0559_{\pm 0.0715}$	$0.0666_{\pm 0.1156}$
	T	256	$0.6584_{\pm 0.0222}$	$0.8490_{\pm 0.0119}$	$0.0232_{\pm 0.0213}$	$0.3215_{\pm 0.0711}$	$0.0679_{\pm 0.1064}$
	M	256	$0.5545_{\pm 0.0226}$	$0.8040_{\pm 0.0134}$	$0.0075_{\pm 0.0189}$	$0.3274_{\pm 0.0857}$	$0.0174_{\pm 0.0943}$
	T	512	$0.5305_{\pm 0.0261}$	$0.7920_{\pm 0.0120}$	$0.0984_{\pm 0.0197}$	$0.4850_{\pm 0.0823}$	$0.0488_{\pm 0.1148}$
	M	512	$0.5727_{\pm 0.0271}$	$0.8120_{\pm 0.0112}$	$0.0390_{\pm 0.0188}$	$0.1256_{\pm 0.0774}$	$0.0339_{\pm 0.1127}$
Claude-3-Haiku	T	0	$0.4942_{\pm 0.0204}$	$0.5470_{\pm 0.0171}$	$0.0662_{\pm 0.0163}$	$0.0718_{\pm 0.0654}$	$0.0526_{\pm 0.1101}$
	M	0	$0.5660_{\pm 0.0215}$	$0.6030_{\pm 0.0142}$	$0.0169_{\pm 0.0201}$	$0.5853_{\pm 0.0833}$	$0.0604_{\pm 0.1156}$
	T	2	$0.6677_{\pm 0.0249}$	$0.7820_{\pm 0.0111}$	$0.0935_{\pm 0.0180}$	$0.1085_{\pm 0.0845}$	$0.0204_{\pm 0.0958}$
	M	2	$0.5950_{\pm 0.0215}$	$0.7450_{\pm 0.0122}$	$0.0323_{\pm 0.0197}$	$0.4802_{\pm 0.0752}$	$0.0240_{\pm 0.1172}$
	T	8	$0.5791_{\pm 0.0233}$	$0.7660_{\pm 0.0148}$	$0.1660_{\pm 0.0202}$	$0.4119_{\pm 0.0793}$	$0.0324_{\pm 0.1009}$
	M	8	$0.6510_{\pm 0.0249}$	$0.8120_{\pm 0.0138}$	$0.0394_{\pm 0.0214}$	$0.2644_{\pm 0.0758}$	$0.0210_{\pm 0.1047}$
	T	32	$0.5833_{\pm 0.0226}$	$0.7980_{\pm 0.0143}$	$0.0437_{\pm 0.0194}$	$0.2893_{\pm 0.0681}$	$0.0199_{\pm 0.1209}$
	M	32	$0.6185_{\pm 0.0236}$	$0.8300_{\pm 0.0120}$	$0.0629_{\pm 0.0182}$	$0.2902_{\pm 0.0664}$	$0.0035_{\pm 0.0960}$
	T	128	$0.6130_{\pm 0.0230}$	$0.8290_{\pm 0.0123}$	$0.0147_{\pm 0.0163}$	$0.1453_{\pm 0.0859}$	$0.0399_{\pm 0.1050}$
	M	128	$0.5124_{\pm 0.0267}$	$0.7840_{\pm 0.0118}$	$0.0596_{\pm 0.0170}$	$0.0476_{\pm 0.0728}$	$0.0146_{\pm 0.1043}$
	T	256	$0.5374_{\pm 0.0246}$	$0.7950_{\pm 0.0115}$	$0.1003_{\pm 0.0208}$	$0.4320_{\pm 0.0753}$	$0.0082_{\pm 0.0957}$
	M	256	$0.6027_{\pm 0.0207}$	$0.7900_{\pm 0.0119}$	$0.0829_{\pm 0.0167}$	$0.1901_{\pm 0.0728}$	$0.0138_{\pm 0.1196}$
	T	512	$0.5621_{\pm 0.0273}$	$0.8030_{\pm 0.0126}$	$0.0896_{\pm 0.0166}$	$0.4507_{\pm 0.0805}$	$0.0301_{\pm 0.1010}$
	M	512	$0.5682_{\pm 0.0255}$	$0.8100_{\pm 0.0126}$	$0.0589_{\pm 0.0211}$	$0.1626_{\pm 0.0793}$	$0.0223_{\pm 0.1199}$

Table 7: Predictive performance and group fairness on the German Credit dataset across LLMs, prompt encodings, and numbers of in-context examples, with conventional classifiers included as baselines. Values are averaged over five scoring runs, and subscripts report the standard deviation across the runs.

Model	P	k	BA	Acc.	SP	EOD	PP
DT	–	–	$0.5964_{\pm 0.0422}$	$0.6550_{\pm 0.0307}$	$0.0485_{\pm 0.0604}$	$0.1807_{\pm 0.1120}$	$0.0292_{\pm 0.0487}$
NB	–	–	$0.6595_{\pm 0.0347}$	$0.7300_{\pm 0.0289}$	$\mathbf{0.0019}_{\pm \mathbf{0.0651}}$	$0.1615_{\pm 0.1226}$	$0.0166_{\pm 0.0484}$
MLP	–	–	$0.6083_{\pm 0.0384}$	$0.6650_{\pm 0.0308}$	$-0.0669_{\pm 0.0687}$	$0.1292_{\pm 0.1097}$	$0.0214_{\pm 0.0480}$
kNN	–	–	$0.5357_{\pm 0.0433}$	$0.6500_{\pm 0.0299}$	$0.0641_{\pm 0.0659}$	$0.1171_{\pm 0.0949}$	$0.0391_{\pm 0.0555}$
GPT-4o-Mini	T	0	$0.5405_{\pm 0.0375}$	$0.4700_{\pm 0.0390}$	$0.1013_{\pm 0.0684}$	$0.2107_{\pm 0.1175}$	$0.2648_{\pm 0.0552}$
	M	0	$0.5500_{\pm 0.0413}$	$0.4900_{\pm 0.0351}$	$0.0574_{\pm 0.0642}$	$0.1354_{\pm 0.0982}$	$0.2561_{\pm 0.0483}$
	T	2	$0.5679_{\pm 0.0329}$	$0.5350_{\pm 0.0335}$	$0.1103_{\pm 0.0674}$	$0.2497_{\pm 0.1146}$	$0.1260_{\pm 0.0518}$
	M	2	$0.5774_{\pm 0.0355}$	$0.5950_{\pm 0.0387}$	$0.0524_{\pm 0.0610}$	$0.1430_{\pm 0.1257}$	$0.0748_{\pm 0.0547}$
	T	8	$0.6095_{\pm 0.0322}$	$0.6400_{\pm 0.0308}$	$0.1691_{\pm 0.0634}$	$0.2563_{\pm 0.1055}$	$0.0779_{\pm 0.0543}$
	M	8	$0.6250_{\pm 0.0397}$	$0.6350_{\pm 0.0369}$	$0.1273_{\pm 0.0622}$	$0.2871_{\pm 0.1097}$	$0.2081_{\pm 0.0568}$
	T	32	$0.6774_{\pm 0.0371}$	$0.6750_{\pm 0.0317}$	$-0.0089_{\pm 0.0579}$	$0.1458_{\pm 0.1226}$	$0.1315_{\pm 0.0467}$
	M	32	$0.6845_{\pm 0.0343}$	$0.7050_{\pm 0.0288}$	$0.0366_{\pm 0.0701}$	$0.1749_{\pm 0.1012}$	$0.0802_{\pm 0.0553}$
	T	128	$0.5940_{\pm 0.0371}$	$0.7050_{\pm 0.0322}$	$0.0471_{\pm 0.0623}$	$0.0938_{\pm 0.1012}$	$0.0094_{\pm 0.0540}$
	M	128	$0.6821_{\pm 0.0313}$	$0.6550_{\pm 0.0330}$	$0.0401_{\pm 0.0657}$	$0.0466_{\pm 0.0987}$	$0.0207_{\pm 0.0535}$
	T	256	$0.6560_{\pm 0.0329}$	$0.6850_{\pm 0.0360}$	$0.0032_{\pm 0.0610}$	$0.0711_{\pm 0.1086}$	$0.0888_{\pm 0.0608}$
	M	256	$0.6976_{\pm 0.0334}$	$0.7100_{\pm 0.0348}$	$0.0630_{\pm 0.0665}$	$0.1262_{\pm 0.0971}$	$0.0680_{\pm 0.0526}$
	T	512	$0.7274_{\pm 0.0367}$	$0.6850_{\pm 0.0355}$	$0.0282_{\pm 0.0661}$	$0.1055_{\pm 0.1192}$	$0.0781_{\pm 0.0505}$
	M	512	$0.7060_{\pm 0.0393}$	$0.6750_{\pm 0.0375}$	$0.0504_{\pm 0.0688}$	$0.1375_{\pm 0.1108}$	$0.1276_{\pm 0.0560}$
DeepSeek-V4-Flash	T	0	$0.4881_{\pm 0.0376}$	$0.4700_{\pm 0.0390}$	$0.1035_{\pm 0.0435}$	$0.0969_{\pm 0.1262}$	$0.0279_{\pm 0.0632}$
	M	0	$0.5476_{\pm 0.0441}$	$0.5000_{\pm 0.0337}$	$-0.0076_{\pm 0.0643}$	$0.0169_{\pm 0.1233}$	$0.0324_{\pm 0.0526}$
	T	2	$0.5762_{\pm 0.0375}$	$0.6200_{\pm 0.0353}$	$0.0592_{\pm 0.0709}$	$0.4011_{\pm 0.1117}$	$0.1435_{\pm 0.0585}$
	M	2	$0.5690_{\pm 0.0387}$	$0.5900_{\pm 0.0368}$	$0.0481_{\pm 0.0544}$	$0.3102_{\pm 0.0954}$	$0.0942_{\pm 0.0549}$
	T	8	$0.6488_{\pm 0.0420}$	$0.7150_{\pm 0.0343}$	$0.0174_{\pm 0.0543}$	$0.0789_{\pm 0.0968}$	$\mathbf{0.0042}_{\pm \mathbf{0.0473}}$
	M	8	$0.6298_{\pm 0.0384}$	$0.6750_{\pm 0.0367}$	$0.1112_{\pm 0.0630}$	$0.3278_{\pm 0.1151}$	$0.1006_{\pm 0.0526}$
	T	32	$0.6571_{\pm 0.0331}$	$0.7200_{\pm 0.0325}$	$0.0040_{\pm 0.0625}$	$0.3290_{\pm 0.1009}$	$0.1103_{\pm 0.0506}$
	M	32	$0.5405_{\pm 0.0380}$	$0.6700_{\pm 0.0340}$	$0.0719_{\pm 0.0682}$	$0.2576_{\pm 0.1093}$	$0.0804_{\pm 0.0469}$
	T	128	$0.6738_{\pm 0.0397}$	$0.6700_{\pm 0.0297}$	$0.1474_{\pm 0.0585}$	$0.2646_{\pm 0.1046}$	$0.0145_{\pm 0.0612}$
	M	128	$0.6643_{\pm 0.0351}$	$0.6900_{\pm 0.0339}$	$0.0932_{\pm 0.0681}$	$0.2227_{\pm 0.1246}$	$0.0286_{\pm 0.0558}$
	T	256	$0.6286_{\pm 0.0403}$	$0.6400_{\pm 0.0310}$	$0.0829_{\pm 0.0616}$	$0.1136_{\pm 0.1192}$	$0.0075_{\pm 0.0533}$
	M	256	$0.6786_{\pm 0.0363}$	$0.6900_{\pm 0.0356}$	$0.0114_{\pm 0.0571}$	$0.1896_{\pm 0.1011}$	$0.0814_{\pm 0.0506}$
	T	512	$0.6774_{\pm 0.0354}$	$0.6950_{\pm 0.0331}$	$0.1033_{\pm 0.0612}$	$0.2038_{\pm 0.1175}$	$0.0359_{\pm 0.0488}$
	M	512	$\mathbf{0.7417}_{\pm \mathbf{0.0308}}$	$0.7650_{\pm 0.0334}$	$0.0662_{\pm 0.0641}$	$0.2007_{\pm 0.1025}$	$0.0292_{\pm 0.0480}$
Claude-3-Haiku	T	0	$0.5452_{\pm 0.0375}$	$0.4300_{\pm 0.0387}$	$0.2454_{\pm 0.0631}$	$0.5625_{\pm 0.1642}$	$0.2319_{\pm 0.0725}$
	M	0	$0.4857_{\pm 0.0344}$	$0.4800_{\pm 0.0371}$	$0.1567_{\pm 0.0643}$	$0.4250_{\pm 0.1131}$	$0.2283_{\pm 0.0494}$
	T	2	$0.5429_{\pm 0.0410}$	$0.6800_{\pm 0.0357}$	$0.0456_{\pm 0.0597}$	$0.1666_{\pm 0.1177}$	$0.0319_{\pm 0.0536}$
	M	2	$0.4643_{\pm 0.0363}$	$0.6100_{\pm 0.0351}$	$0.0324_{\pm 0.0597}$	$0.1172_{\pm 0.0968}$	$0.0183_{\pm 0.0508}$
	T	8	$0.4929_{\pm 0.0422}$	$0.6300_{\pm 0.0329}$	$0.0387_{\pm 0.0661}$	$0.0730_{\pm 0.1111}$	$0.0410_{\pm 0.0517}$
	M	8	$0.4786_{\pm 0.0389}$	$0.6500_{\pm 0.0309}$	$0.1248_{\pm 0.0538}$	$0.1894_{\pm 0.1127}$	$0.0149_{\pm 0.0601}$
	T	32	$0.5976_{\pm 0.0380}$	$0.7300_{\pm 0.0278}$	$0.0400_{\pm 0.0659}$	$0.1527_{\pm 0.1019}$	$0.0079_{\pm 0.0555}$
	M	32	$0.5690_{\pm 0.0361}$	$0.7100_{\pm 0.0285}$	$0.0576_{\pm 0.0644}$	$0.2899_{\pm 0.1164}$	$0.0818_{\pm 0.0602}$
	T	128	$0.5964_{\pm 0.0353}$	$0.6950_{\pm 0.0370}$	$0.0034_{\pm 0.0566}$	$0.3551_{\pm 0.1118}$	$0.0939_{\pm 0.0511}$
	M	128	$0.5440_{\pm 0.0369}$	$0.6950_{\pm 0.0360}$	$0.0689_{\pm 0.0585}$	$0.1693_{\pm 0.1065}$	$0.0722_{\pm 0.0577}$
	T	256	$0.6536_{\pm 0.0399}$	$0.7750_{\pm 0.0283}$	$0.0478_{\pm 0.0724}$	$\mathbf{0.0101}_{\pm \mathbf{0.1003}}$	$0.0237_{\pm 0.0511}$
	M	256	$0.6131_{\pm 0.0342}$	$0.7450_{\pm 0.0272}$	$0.0359_{\pm 0.0651}$	$0.1855_{\pm 0.1244}$	$0.0553_{\pm 0.0593}$
	T	512	$0.6524_{\pm 0.0421}$	$\mathbf{0.7800}_{\pm \mathbf{0.0327}}$	$0.0540_{\pm 0.0586}$	$0.1106_{\pm 0.1044}$	$0.0243_{\pm 0.0534}$
	M	512	$0.6012_{\pm 0.0431}$	$0.7150_{\pm 0.0359}$	$0.0132_{\pm 0.0703}$	$0.2444_{\pm 0.1125}$	$0.0857_{\pm 0.0537}$

$k = 8$ on Credit Approval, peaks at $k = 32$ on Credit Card (0.815), and is highest at $k = 512$ on German Credit (0.719). Because the models acquire the task from labeled applicants placed in the prompt, without any update to their parameters, in-context learning offers a practical substitute for fine-tuning when the model can be accessed only through an API.

Comparing the most accurate few-shot configuration of each model with its zero-shot setting, accuracy improves by 0.051 to 0.300, and by more than 0.10 in seven of the nine model-dataset pairs; the two smaller gains, for GPT-4o-mini and DeepSeek-V4-Flash on Credit Card, start from zero-shot accuracies already close to the majority-class rate. Claude-3-Haiku is the only model whose best configuration exceeds the most accurate traditional baseline on all three datasets, with 0.8043 against 0.7826 for Naive Bayes on Credit Approval, 0.8300 against 0.7540 for kNN on Credit Card, and 0.7800 against 0.7300 for Naive Bayes on German Credit. DeepSeek-V4-Flash exceeds the strongest baseline on Credit Card and German Credit, and GPT-4o-mini on Credit Card. Because these best configurations are picked from 12 settings per model, they are optimistic estimates. The datasets differ in how demanding this comparison is. On Credit Card, 31 of the 36 few-shot configurations are more accurate than kNN, and GPT-4o-mini with 32 examples reaches an accuracy of 0.8580 and a BA of 0.7376, against 0.7540 for kNN and 0.6132 for the Decision Tree, the strongest traditional models in accuracy and BA, both trained on 18,000 records. On Credit Approval and German Credit, the strongest baseline is harder to beat, and only one and four of the 36 few-shot configurations, respectively, are more accurate than Naive Bayes. Few-shot LLMs thus reach the accuracy range of supervised credit-scoring models, and the best configurations exceed them, but which model and shot count achieve this differs by dataset.

The gains are also not monotonic in k. With the Template encoding on Credit Approval, DeepSeek-V4-Flash falls from 0.6812 at $k = 8$ to 0.5362 at $k = 128$, below its zero-shot accuracy of 0.6014, while GPT-4o-mini rises from 0.6667 at $k = 32$ to 0.7609 at $k = 128$ and falls back to 0.6522 at $k = 512$. Claude-3-Haiku with the Template encoding is less accurate on German Credit at $k = 8$ (0.6300) than at $k = 2$ (0.6800). Longer contexts can contain noisy or conflicting demonstrations and lead the model to rely on particular examples. Run-to-run variation contributes as well. The median standard deviation of accuracy across the five runs is 0.040 on Credit Approval, 0.034 on German Credit, and 0.013 on Credit Card, so differences of a few hundredths between neighboring configurations are within the variability of the scorer. The model, encoding, and shot count must therefore be fixed before fairness is measured rather than tuned to each dataset.

The fairness columns show why utility has to be established before fairness is interpreted. In each table, the lowest SP, the lowest EOD, and the lowest PP are attained by three different configurations, none of them the most accurate, and several of them close to chance. The lowest EOD on Credit Approval comes from zero-shot Claude-3-Haiku with the Template encoding (0.0150 at a BA of 0.5006), and the lowest PP on Credit Card from Naive Bayes (0.0002), the least accurate classifier on that dataset (accuracy 0.3770). The lowest EOD on Credit Card comes from zero-shot GPT-4o-mini with the Template encoding (0.0203), whose accuracy of 0.7750 conceals a BA of

0.5457. A scorer that barely distinguishes between applicants also barely distinguishes between groups, so such small disparities carry little information about fairness. The fairness values are, moreover, uncertain, with a median standard deviation across runs of 0.020 to 0.084 for SP, 0.077 to 0.140 for EOD, and 0.053 to 0.108 for PP, which is larger than many of the differences between rows. The benchmark therefore does not rank configurations by fairness, and the audit configuration is not selected on these columns.

We audit Claude-3-Haiku with the Template encoding and $k = 128$. Claude-3-Haiku is the only model whose best configuration exceeds the strongest traditional baseline on all three datasets, and it returned valid JSON for more than 99.5% of queries, which matters for the roughly 300,000 queries of the audit. A shot count of 128 keeps the prompt short enough for the ablation queries that VAYNE issues, and larger shot counts did not improve utility consistently on the development split. With this encoding and shot count, Claude-3-Haiku is the most accurate configuration of the benchmark on Credit Approval and exceeds the strongest baseline on Credit Card (0.8290 against 0.7540), but not on German Credit (0.6950 against 0.7300). The configuration is fixed across datasets, and its utility is examined again at the final decoding temperature.

Decoding temperature is the remaining parameter of the scorer. We keep Claude-3-Haiku with $k = 128$ and vary the temperature over $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ under both encodings (Fig. 5 and Table 8).

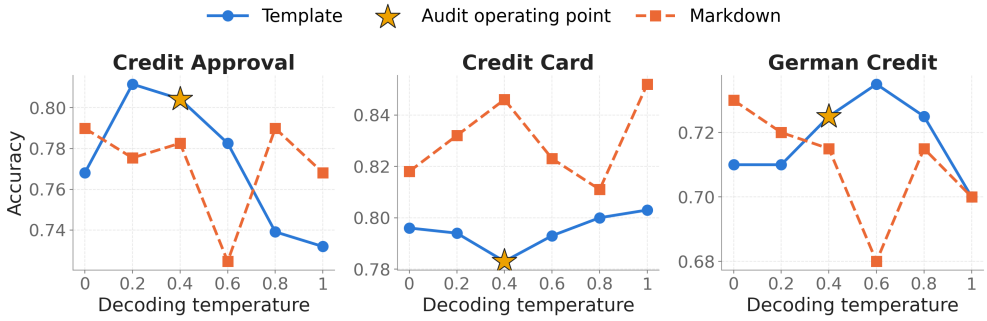


Figure 5: Accuracy of Claude-3-Haiku with $k = 128$ across decoding temperatures. Results are shown for Template and Markdown prompt encodings on Credit Approval, Credit Card, and German Credit; the star marks the audit operating point (Template, temperature 0.4).

The effect of temperature is non-monotonic and dataset-specific. The most accurate setting is Template at 0.2 on Credit Approval (0.8116), Markdown at 1.0 on Credit Card (0.8520), and Template at 0.6 on German Credit (0.7350). Within one encoding, temperature changes accuracy by up to 0.080 on Credit Approval, 0.041 on Credit Card, and 0.050 on German Credit, between about one and a half and four standard deviations across runs. Temperature is therefore part of the specification of the scorer rather than an implementation detail, since it changes which applicants are

approved. On Credit Card, the encoding has a larger effect than the temperature, as Markdown is more accurate than Template at every temperature, by 0.011 to 0.063, and also attains a higher BA, by 0.027 to 0.147.

At the highest temperature, the scorer can collapse. On German Credit at temperature 1.0, the Template setting reaches an accuracy of 0.6950, close to the 70% share of good applicants, with a BA of 0.5012, which means that it approves almost every applicant. The Markdown setting at the same temperature is more accurate than at 0.6 (0.7250 against 0.6800), yet its BA of 0.5417 shows that it hardly separates good from bad applicants. These two settings attain the best values of all three fairness metrics on German Credit, between 0.0024 and 0.0164, for the same reason as the zero-shot settings above. Outside this collapse, the fairness metrics change across temperatures by less than their run-to-run variation; on Credit Approval, for instance, SP ranges from 0.0214 to 0.0645 and EOD from 0.1522 to 0.2287 across the twelve settings, against standard deviations of about 0.07 to 0.09 for SP and 0.12 to 0.16 for EOD. Selecting the temperature for fairness would thus favor collapsed settings, and we select it for utility instead.

Accuracy alone is not sufficient for this choice. Under the Template encoding, the accuracy averaged over the three datasets differs by less than 0.002 between temperatures 0.2, 0.4, and 0.6, and it does not reveal the collapse at 1.0. We therefore select the temperature by BA on the development split, where 0.4 gives the highest value averaged over the three datasets (0.6531, against at most 0.6474 for the other temperatures). At 0.4, accuracy is within 0.01 of the most accurate Template setting on Credit Approval (0.8043 against 0.8116) and German Credit (0.7250 against 0.7350), and within 0.02 on Credit Card (0.7830 against 0.8030). A single temperature across datasets thus costs little accuracy relative to the best per-dataset choice, apart from the higher accuracy that the Markdown encoding would reach on Credit Card.

At the audit operating point (Baseline rows of Table 11), the scorer reaches an accuracy of 0.8043 on Credit Approval, 0.022 above Naive Bayes, 0.7250 on German Credit, 0.005 below Naive Bayes and above the other three traditional models, and 0.7830 on Credit Card, 0.029 above kNN and above every traditional model. It is therefore at least as accurate as the strongest traditional model on Credit Approval and Credit Card and on par with it on German Credit, and all three datasets are used in RQ2 and RQ3. On Credit Card, the scorer seldom predicts default (BA 0.5068), so its errors fall mainly on defaulting clients, which shapes the EOD values reported for this dataset in RQ2 and RQ3.

Answer to RQ1

LLMs can serve as credit classifiers once they are given labeled examples in the prompt. Without examples, their accuracy is mostly at or below the majority-class rate. With eight examples, accuracy averaged over models, encodings, and datasets rises from 0.565 to 0.706, and with 512 examples to 0.740, with BA rising in parallel. The best configurations exceed the most accurate traditional baseline on every dataset, and Claude-3-Haiku does so on all three (Tables 5, 6, and 7). The gains depend on the model and the dataset and are not monotonic in k. Decoding temperature changes accuracy in a dataset-specific way, matters less than the encoding on Credit

Card, and at 1.0 can collapse the scorer into near-uniform decisions that appear perfectly fair, so it is chosen for utility rather than fairness. At the resulting operating point (Claude-3-Haiku, Template, $k = 128$, temperature 0.4), the scorer matches or exceeds the strongest traditional model on all three datasets, coming within 0.005 of it on German Credit, and all three datasets are therefore used for the diagnosis and mitigation that follow.

Table 8: Temperature sensitivity of Claude-3-Haiku with $k = 128$ across Credit Approval, Credit Card, and German Credit. Results are reported for Template (T) and Markdown (M) prompt encodings at six sampling temperatures. Values are averaged over five scoring runs, and subscripts report the standard deviation across the runs. The best and second-best results within each dataset and metric are highlighted in orange and yellow, respectively; higher is better for BA and Acc., and closer to zero is better for SP, EOD, and PP.

Dataset	P	Temp.	BA	Acc.	SP	EOD	PP
Credit Approval	T	0.0	$0.7833_{\pm 0.0401}$	$0.8081_{\pm 0.0409}$	$0.0512_{\pm 0.0756}$	$0.1874_{\pm 0.1490}$	$0.0398_{\pm 0.0963}$
	T	0.2	$0.8141_{\pm 0.0365}$	$\mathbf{0.8116}_{\pm \mathbf{0.0379}}$	$0.0463_{\pm 0.0778}$	$0.1687_{\pm 0.1466}$	$0.0371_{\pm 0.0816}$
	T	0.4	$\mathbf{0.8196}_{\pm \mathbf{0.0339}}$	$0.8043_{\pm 0.0301}$	$0.0417_{\pm 0.0872}$	$\mathbf{0.1522}_{\pm \mathbf{0.1435}}$	$0.0345_{\pm 0.0990}$
	T	0.6	$0.7899_{\pm 0.0378}$	$0.7826_{\pm 0.0354}$	$0.0389_{\pm 0.0887}$	$0.1745_{\pm 0.1309}$	$0.0329_{\pm 0.1019}$
	T	0.8	$0.7543_{\pm 0.0355}$	$0.7391_{\pm 0.0321}$	$0.0301_{\pm 0.0900}$	$0.1938_{\pm 0.1238}$	$0.0302_{\pm 0.0875}$
	T	1.0	$0.7444_{\pm 0.0342}$	$0.7319_{\pm 0.0408}$	$\mathbf{0.0214}_{\pm \mathbf{0.0875}}$	$0.2016_{\pm 0.1482}$	$\mathbf{0.0287}_{\pm \mathbf{0.0792}}$
	M	0.0	$0.7964_{\pm 0.0297}$	$0.7899_{\pm 0.0378}$	$0.0587_{\pm 0.0869}$	$0.1612_{\pm 0.1493}$	$0.0423_{\pm 0.0959}$
	M	0.2	$0.7817_{\pm 0.0337}$	$0.7754_{\pm 0.0375}$	$0.0498_{\pm 0.0762}$	$0.1790_{\pm 0.1604}$	$0.0386_{\pm 0.1004}$
	M	0.4	$0.7933_{\pm 0.0373}$	$0.7826_{\pm 0.0300}$	$0.0436_{\pm 0.0708}$	$0.1563_{\pm 0.1608}$	$0.0352_{\pm 0.0980}$
	M	0.6	$0.7311_{\pm 0.0427}$	$0.7246_{\pm 0.0388}$	$0.0645_{\pm 0.0806}$	$0.2287_{\pm 0.1528}$	$0.0517_{\pm 0.0837}$
	M	0.8	$0.7998_{\pm 0.0373}$	$0.7899_{\pm 0.0302}$	$0.0351_{\pm 0.0822}$	$0.1655_{\pm 0.1603}$	$0.0318_{\pm 0.0899}$
	M	1.0	$0.7820_{\pm 0.0340}$	$0.7681_{\pm 0.0395}$	$0.0276_{\pm 0.0853}$	$0.1841_{\pm 0.1462}$	$0.0335_{\pm 0.0956}$
Credit Card	T	0.0	$0.5413_{\pm 0.0229}$	$0.7960_{\pm 0.0112}$	$0.0236_{\pm 0.0172}$	$0.1847_{\pm 0.0761}$	$0.0311_{\pm 0.1070}$
	T	0.2	$0.5367_{\pm 0.0216}$	$0.7940_{\pm 0.0108}$	$0.0207_{\pm 0.0168}$	$\mathbf{0.1386}_{\pm \mathbf{0.0798}}$	$0.0296_{\pm 0.1191}$
	T	0.4	$0.5068_{\pm 0.0221}$	$0.7830_{\pm 0.0106}$	$\mathbf{0.0184}_{\pm \mathbf{0.0218}}$	$0.2003_{\pm 0.0663}$	$0.0285_{\pm 0.1132}$
	T	0.6	$0.5295_{\pm 0.0237}$	$0.7930_{\pm 0.0136}$	$0.0221_{\pm 0.0176}$	$0.2114_{\pm 0.0852}$	$0.0302_{\pm 0.1072}$
	T	0.8	$0.5487_{\pm 0.0210}$	$0.8000_{\pm 0.0120}$	$0.0193_{\pm 0.0197}$	$0.1765_{\pm 0.0838}$	$0.0241_{\pm 0.1043}$
	T	1.0	$0.5539_{\pm 0.0263}$	$0.8030_{\pm 0.0121}$	$0.0249_{\pm 0.0176}$	$0.1912_{\pm 0.0753}$	$0.0336_{\pm 0.1193}$
	M	0.0	$0.5864_{\pm 0.0231}$	$0.8180_{\pm 0.0117}$	$0.0281_{\pm 0.0205}$	$0.2235_{\pm 0.0694}$	$0.0347_{\pm 0.1082}$
	M	0.2	$0.6214_{\pm 0.0230}$	$0.8320_{\pm 0.0125}$	$0.0312_{\pm 0.0185}$	$0.2378_{\pm 0.0838}$	$0.0365_{\pm 0.0973}$
	M	0.4	$0.6533_{\pm 0.0228}$	$0.8460_{\pm 0.0117}$	$0.0265_{\pm 0.0169}$	$0.2096_{\pm 0.0823}$	$0.0298_{\pm 0.1101}$
	M	0.6	$0.6026_{\pm 0.0217}$	$0.8230_{\pm 0.0118}$	$0.0334_{\pm 0.0181}$	$0.2451_{\pm 0.0810}$	$0.0389_{\pm 0.1113}$
	M	0.8	$0.5753_{\pm 0.0238}$	$0.8110_{\pm 0.0112}$	$0.0297_{\pm 0.0167}$	$0.2189_{\pm 0.0744}$	$0.0352_{\pm 0.1165}$
	M	1.0	$\mathbf{0.6685}_{\pm \mathbf{0.0208}}$	$\mathbf{0.8520}_{\pm \mathbf{0.0093}}$	$0.0318_{\pm 0.0188}$	$0.2265_{\pm 0.0812}$	$\mathbf{0.0233}_{\pm \mathbf{0.1073}}$
German Credit	T	0.0	$0.5929_{\pm 0.0398}$	$0.7100_{\pm 0.0345}$	$0.0458_{\pm 0.0660}$	$0.0913_{\pm 0.1004}$	$0.0081_{\pm 0.0589}$
	T	0.2	$0.5976_{\pm 0.0359}$	$0.7100_{\pm 0.0340}$	$0.0431_{\pm 0.0545}$	$0.0874_{\pm 0.1151}$	$0.0076_{\pm 0.0523}$
	T	0.4	$0.6083_{\pm 0.0337}$	$0.7250_{\pm 0.0285}$	$0.0396_{\pm 0.0599}$	$0.0787_{\pm 0.0978}$	$0.0067_{\pm 0.0483}$
	T	0.6	$0.6250_{\pm 0.0390}$	$\mathbf{0.7350}_{\pm \mathbf{0.0271}}$	$0.0372_{\pm 0.0671}$	$0.0821_{\pm 0.1154}$	$0.0072_{\pm 0.0462}$
	T	0.8	$\mathbf{0.6417}_{\pm \mathbf{0.0393}}$	$0.7250_{\pm 0.0316}$	$0.0353_{\pm 0.0622}$	$0.0956_{\pm 0.1037}$	$0.0094_{\pm 0.0496}$
	T	1.0	$0.5012_{\pm 0.0034}$	$0.6950_{\pm 0.0193}$	$0.0053_{\pm 0.0167}$	$0.0164_{\pm 0.0034}$	$\mathbf{0.0024}_{\pm \mathbf{0.0016}}$
	M	0.0	$0.6214_{\pm 0.0400}$	$0.7300_{\pm 0.0308}$	$0.0517_{\pm 0.0641}$	$0.1046_{\pm 0.1142}$	$0.0102_{\pm 0.0606}$
	M	0.2	$0.6048_{\pm 0.0377}$	$0.7200_{\pm 0.0276}$	$0.0489_{\pm 0.0563}$	$0.0987_{\pm 0.1194}$	$0.0095_{\pm 0.0491}$
	M	0.4	$0.6060_{\pm 0.0338}$	$0.7150_{\pm 0.0322}$	$0.0462_{\pm 0.0698}$	$0.1013_{\pm 0.1067}$	$0.0089_{\pm 0.0566}$
	M	0.6	$0.5714_{\pm 0.0402}$	$0.6800_{\pm 0.0313}$	$0.0575_{\pm 0.0635}$	$0.1129_{\pm 0.1232}$	$0.0118_{\pm 0.0515}$
	M	0.8	$0.6012_{\pm 0.0394}$	$0.7150_{\pm 0.0338}$	$0.0441_{\pm 0.0620}$	$0.0962_{\pm 0.0977}$	$0.0091_{\pm 0.0598}$
	M	1.0	$0.5417_{\pm 0.0035}$	$0.7250_{\pm 0.0297}$	$\mathbf{0.0032}_{\pm \mathbf{0.0165}}$	$\mathbf{0.0126}_{\pm \mathbf{0.0032}}$	$0.0046_{\pm 0.0276}$

4.3 Controlled Sensitivity and Proxy-Mediated Unfairness Diagnosis (RQ2)

We apply VAYNE to the audit configuration (Claude-3-Haiku, Template encoding, $k = 128$, temperature 0.4) on all three datasets. The diagnosis is computed on the development split, with every prompt queried five times and the scores averaged as in Eq. (3). We first measure how the scorer responds to the protected attribute itself and then examine the non-protected features through which information about it can reach the decision once the attribute is withheld.

Controlled protected-attribute sensitivity

Table 9 reports the direct sensitivity DU_A of Eq. (8). Replacing the applicant's stated sex with UNKNOWN moves the score by 0.310 on German Credit, 0.088 on Credit Card, and 0.350 on Credit Approval on average. Because DU_A and proxy use are computed with the same ablation, any effect of the UNKNOWN token itself enters both, and the median proxy use τ_2 provides a natural reference. On German Credit and Credit Approval, the stated sex matters about as much as the median feature of the profile, and on Credit Card it matters more than any single non-protected feature, all of which have a proxy use below 0.08. The scorer's output thus depends on the applicant's sex whenever sex is visible, which is the direct use of the attribute that withholding it is meant to prevent. Withholding the attribute does not, however, remove the information from the profile when other features encode it. Since the deployment profile omits A, any group disparity in deployed decisions must be carried by the non-protected features, through their association with sex or through differences in base rates, and the remainder of this section identifies the features responsible.

Table 9: Controlled protected-attribute sensitivity of the audit configuration on the development split. DU_A is the direct sensitivity of Eq. (8), obtained by replacing the stated sex with UNKNOWN, and τ_2 is the median proxy use of the non-protected features; both are absolute changes in the averaged score under the same ablation.

Dataset	DU_A	Median PU (τ_2)	DU_A/τ_2
German Credit	0.310	0.314	0.99
Credit Card	0.088	0.056	1.57
Credit Approval	0.350	0.368	0.95

Proxy use

For each non-protected feature, VAYNE computes the proxy capacity (PC), the proxy use (PU), and their product, the proxy risk score (PRS), as defined in Sects. 3.4 and 3.5. Figures 6, 7, and 8 first show how strongly the scorer relies on each feature, together with the threshold τ_2. Red bars mark features in the proxy risk zone (PU above τ_2 and PC above τ_1), blue bars features that are heavily used but carry little

information about sex (PU above τ_2, PC at or below τ_1), and grey bars features with PU at or below τ_2.

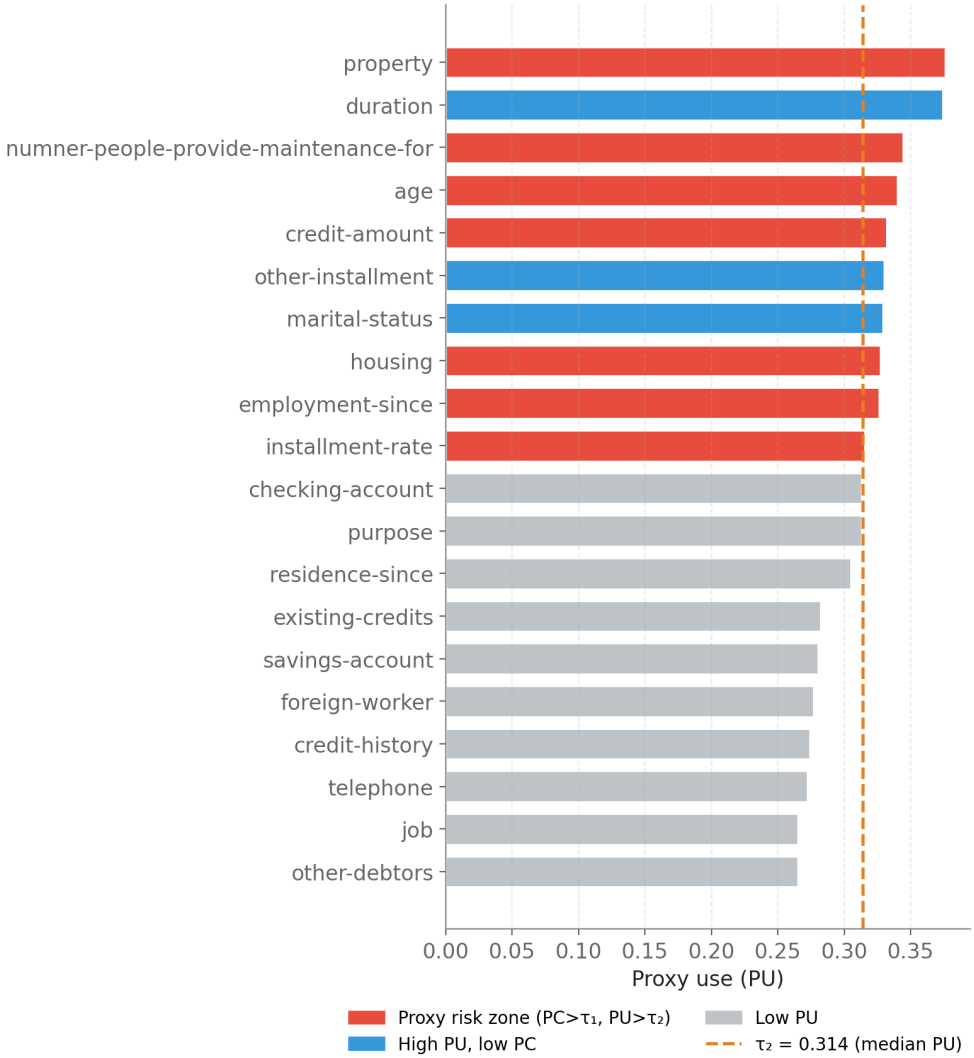


Figure 6: Proxy use ranking across all non-protected features on German Credit. Red bars mark the proxy risk zone (PU above τ_2 and PC above τ_1); blue bars are heavily used but carry little information about sex (high PU, low PC); grey bars are less used by the model (low PU).

On German Credit (Fig. 6), proxy use is compressed across the top half of the ranking, roughly between 0.31 and 0.38. Seven features fall in the proxy risk zone,

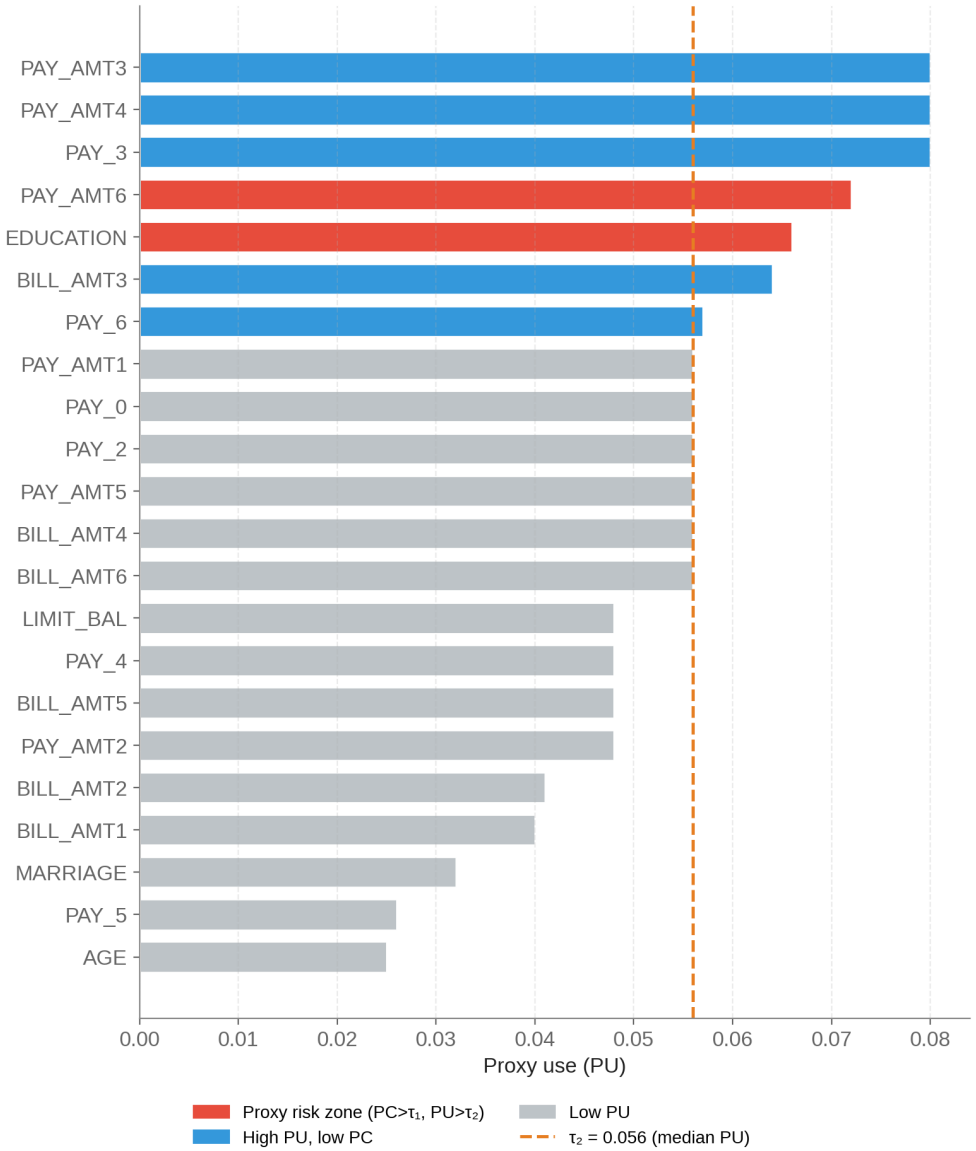


Figure 7: Proxy use ranking across all non-protected features on Credit Card. Red bars mark the proxy risk zone (PU above τ_2 and PC above τ_1); blue bars are heavily used but carry little information about sex (high PU, low PC); grey bars are less used by the model (low PU).

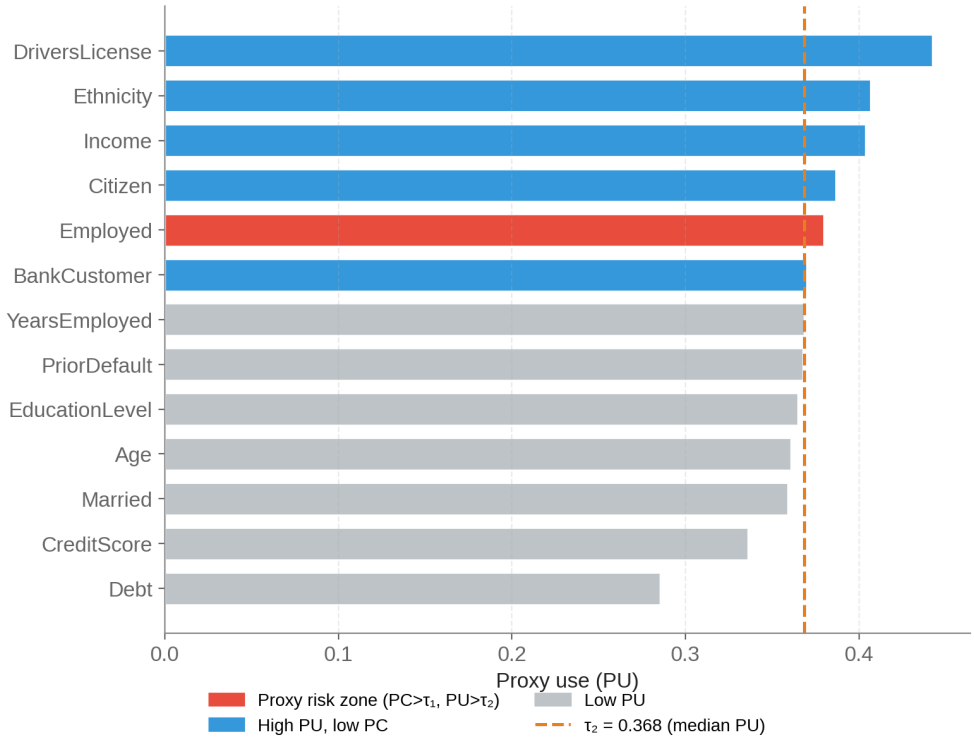


Figure 8: Proxy use ranking across all non-protected features on Credit Approval. Red bars mark the proxy risk zone (PU above τ_2 and PC above τ_1); blue bars are heavily used but carry little information about sex (high PU, low PC); grey bars are less used by the model (low PU).

namely property, numner-people-provide-maintenance-for, age, credit-amount, housing, employment-since, and installment-rate, while duration, other-installment, and marital-status are heavily used but carry little information about sex. The scorer spreads its reliance across the applicant profile, and a large share of that reliance falls on features that also encode sex, which produces a broad, multi-feature exposure rather than a single identifiable proxy. Because the proxy use of these features differs by only a few hundredths, their order within the zone is not sharply determined.

On Credit Card (Fig. 7), proxy use is markedly lower, below 0.08 for every feature, and highest for repayment variables (PAY_AMT3, PAY_AMT4, PAY_3). Because at least half of the features have an estimated PC of zero, $\tau_1 = 0$ on this dataset, and any non-zero capacity places a feature on the capacity side. PAY_AMT6 and EDUCATION fall in the proxy risk zone, while PAY_AMT1, the feature with the highest capacity, has a PU equal to $\tau_2 = 0.056$ and therefore lies just outside it. The other features with PU above τ_2, the payment amounts PAY_AMT3 and PAY_AMT4, the repayment-status variables PAY_3 and PAY_6, and the bill amount BILL_AMT3,

have an estimated PC of zero. Heavy use of a feature therefore does not by itself imply proxy use. Repayment history is used intensively because it predicts default, and it is the kind of underwriting factor generally regarded as legitimate.

On Credit Approval (Fig. 8), proxy use is high and tightly clustered for almost every feature (DriversLicense, Ethnicity, Income, Citizen, Employed, and BankCustomer all exceed 0.36), yet only Employed lies inside the proxy risk zone, with YearsEmployed on the τ_2 boundary. The most heavily used features, DriversLicense, Ethnicity, Income, and Citizen, have a PC below τ_1, whereas Employed has by far the highest PC (Fig. 11). The scorer relies on almost the entire profile, but only a small part of this reliance involves information about sex.

Ranking by proxy risk score

Table 10 reports the five features with the highest proxy risk score on each dataset, and Figures 9 through 11 show the same information as scatter plots of PU against PC. The role column states whether a feature completes a proxy path, with PC above τ_1 and PU above τ_2, or has only a capacity edge.

Table 10: Top-5 credit-risk proxy features ranked by proxy risk score under the audit configuration (Claude-3-Haiku, Template, $k = 128$, temperature 0.4). PRS is computed as PC multiplied by PU. Roles are given relative to the median thresholds, $\tau_1 = 0.020$ and $\tau_2 = 0.314$ for German Credit, $\tau_1 = 0.000$ and $\tau_2 = 0.056$ for Credit Card, and $\tau_1 = 0.019$ and $\tau_2 = 0.368$ for Credit Approval.

Dataset	Feature	PC	PU	PRS	Role
German Credit	age	**0.089**	0.340	**0.0303**	Complete path
	property	0.080	**0.375**	0.0300	Complete path
	housing	0.084	0.327	0.0275	Complete path
	credit-amount	0.074	0.332	0.0246	Complete path
	checking-account	0.077	0.313	0.0241	Capacity only (PU just below τ_2)
Credit Card	PAY_AMT1	**0.136**	0.056	**0.0076**	Capacity, PU at τ_2
	PAY_AMT6	0.062	**0.072**	0.0045	Complete path
	EDUCATION	0.026	0.066	0.0017	Complete path
	PAY_5	0.033	0.026	0.0009	Capacity only
	MARRIAGE	0.017	0.032	0.0005	Capacity only
Credit Approval	Employed	**0.106**	**0.380**	**0.0403**	Complete path (dominant)
	EducationLevel	0.041	0.364	0.0149	Capacity only
	Age	0.030	0.360	0.0108	Capacity only
	YearsEmployed	0.026	0.368	0.0096	Capacity, PU at τ_2
	Debt	0.030	0.285	0.0086	Capacity only

The three datasets differ sharply in how proxy risk is distributed. On German Credit, the five highest-PRS features, age, property, housing, credit-amount, and checking-account, have closely spaced scores between 0.0241 and 0.0303; four of them complete a proxy path, and checking-account lies just below τ_2 (PU 0.313 against 0.314). No single feature dominates, so mitigation has to address several variables

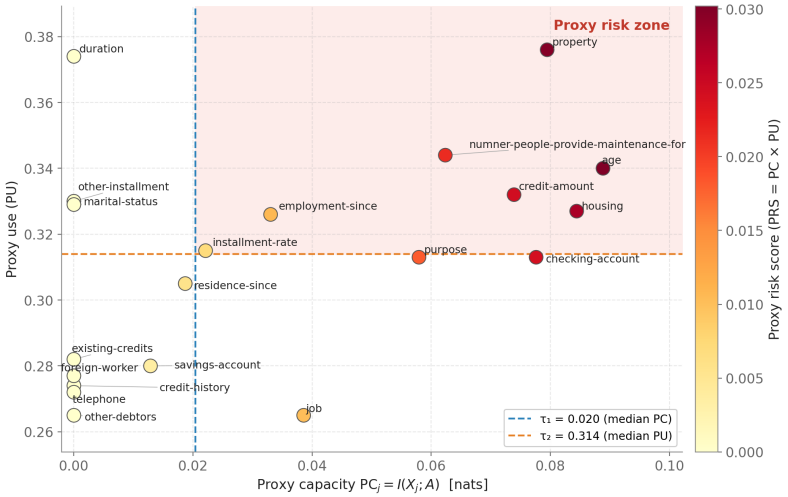


Figure 9: Proxy capacity and proxy use feature space for German Credit. Features in the upper-right proxy risk zone both encode protected-attribute information and influence the LLM decision. Point color encodes the proxy risk score.

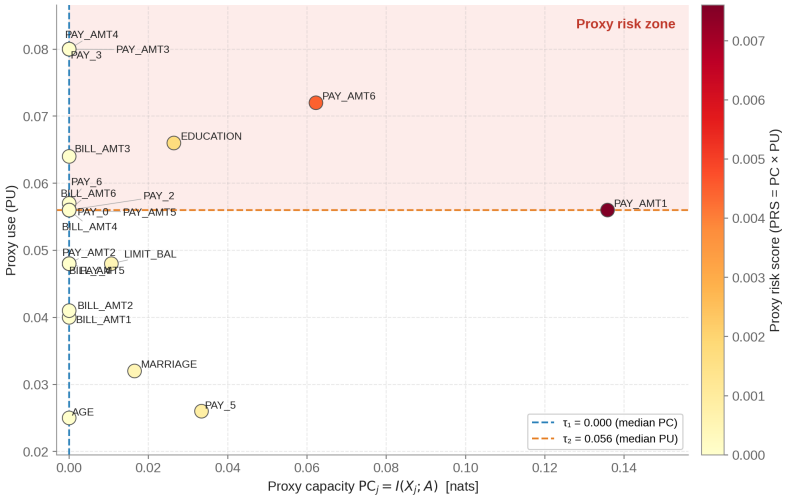


Figure 10: Proxy capacity and proxy use feature space for Credit Card. With $\tau_1 = 0$, PAY_AMT6 and EDUCATION exceed both adaptive thresholds and PAY_AMT1 lies on τ_2; most repayment-status variables cluster at PC = 0, that is, they are used heavily but carry no information about sex.

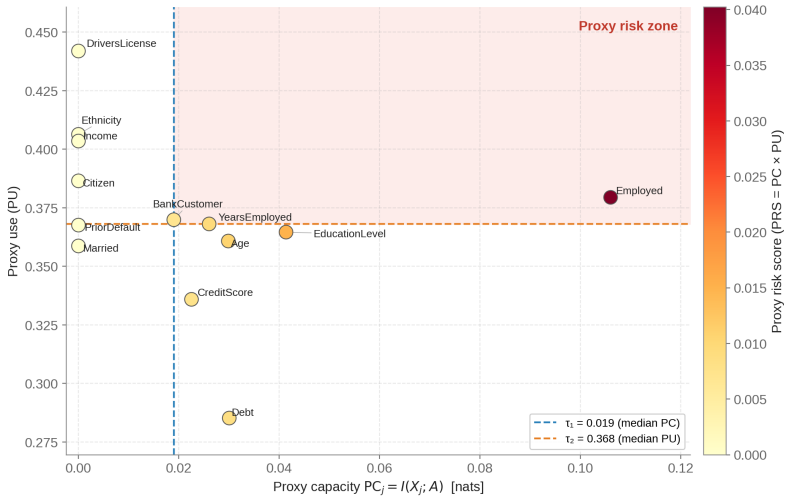


Figure 11: Proxy capacity and proxy use feature space for Credit Approval. Employed is the only feature inside the proxy risk zone, sitting well to the right of every other feature on the proxy capacity axis.

jointly. Credit Approval shows the opposite pattern. Employed has a PRS of 0.0403, about 2.7 times that of the next feature (EducationLevel, 0.0149), and a PC of 0.106, more than twice that of any other feature in Table 10, together with a PU above τ_2. The capacity of EducationLevel includes a plug-in bias of about 0.016 nats for its 14 categories on 414 training records (Sect. 3.4), so its second rank is less certain than the first rank of Employed. Employed is thus a single, auditable bottleneck that a compliance team could act on directly. Credit Card has the weakest proxy structure. PAY_AMT1 reaches a PC of 0.136, roughly twice that of PAY_AMT6 (0.062), but most billing and repayment-status variables carry no information about sex, and the largest PRS (0.0076) is four to five times smaller than on German Credit (0.0303) and Credit Approval (0.0403). This is consistent with sex being only weakly recoverable from the billing and repayment records of this Taiwanese credit-card dataset; because PU inherits the score scale of each deployment, only the rough size of this difference is meaningful. Since τ_1 and τ_2 are medians, some features always fall on each side of the thresholds, so the strength of the evidence lies in the PC and PU values themselves, which are largest and most concentrated on Credit Approval and smallest on Credit Card.

Proxy path graph analysis

The proxy risk ranking identifies which features pose the highest risk; the proxy path graph shows how that risk is organized. The graphs follow the construction of Sect. 3.5, with τ_1 and τ_2 set to the median PC and PU. Every feature with PC above τ_1 receives a capacity edge from the protected attribute, and every feature with PU above τ_2 has a use edge to the decision. Feature nodes are orange when they complete a proxy path, green when they have a capacity edge only, and blue when they have a use edge only;

the protected attribute is purple and the decision node grey. Dashed links between features show pairwise association above τ_3 and are drawn for layout only.

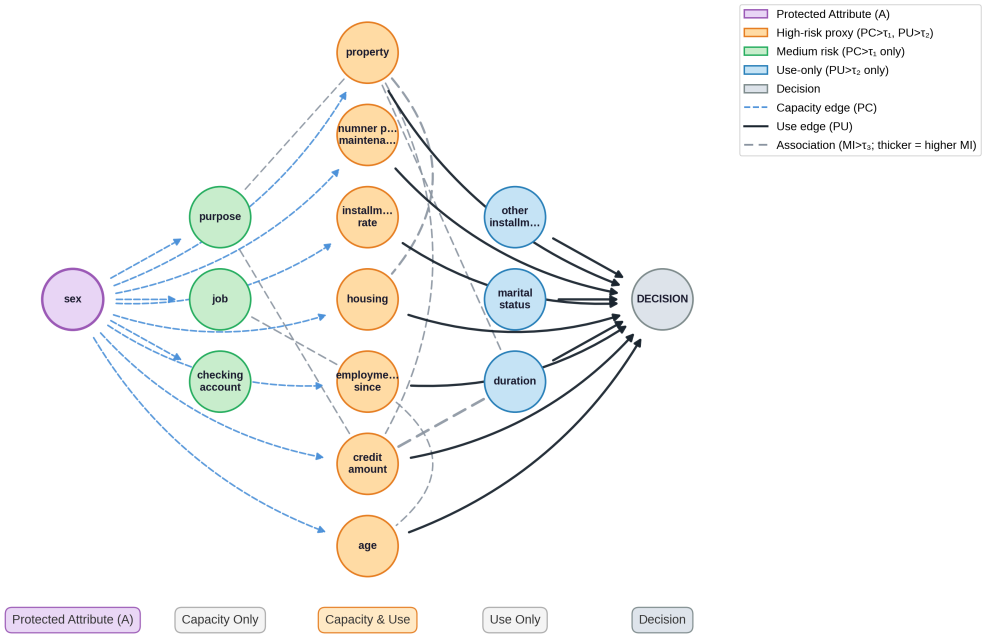


Figure 12: Proxy path graph for German Credit under median-based PC and PU thresholds. Seven features complete a proxy path from sex to the decision (orange), namely property, numner-people-provide-maintenance-for, age, credit-amount, housing, employment-since, and installment-rate. Checking-account, purpose, and job have capacity edges only (green), and duration, other-installment, and marital-status are use-only (blue). Dashed links show pairwise association and do not indicate direction or mediation.

On German Credit (Fig. 12), seven features complete a proxy path, and they are associated with one another and with capacity-only features such as purpose, job, and checking-account. Sex is thus encoded redundantly across a cluster of financial and demographic attributes, so withholding any one of them leaves the information available through the others. The association links describe this redundancy and, as noted in Sect. 3.5, do not show that information travels from one feature to another. On Credit Card (Fig. 13), complete paths run only through PAY_AMT6 and EDUCATION, and PAY_AMT1, the feature with the highest capacity, lies on the use threshold. The graph is dominated by use-only repayment and billing variables that the scorer relies on heavily but that carry no information about sex. On Credit Approval (Fig. 14), Employed is the only feature that completes a path, five further features have capacity edges only, and the most heavily used features reach the decision without a

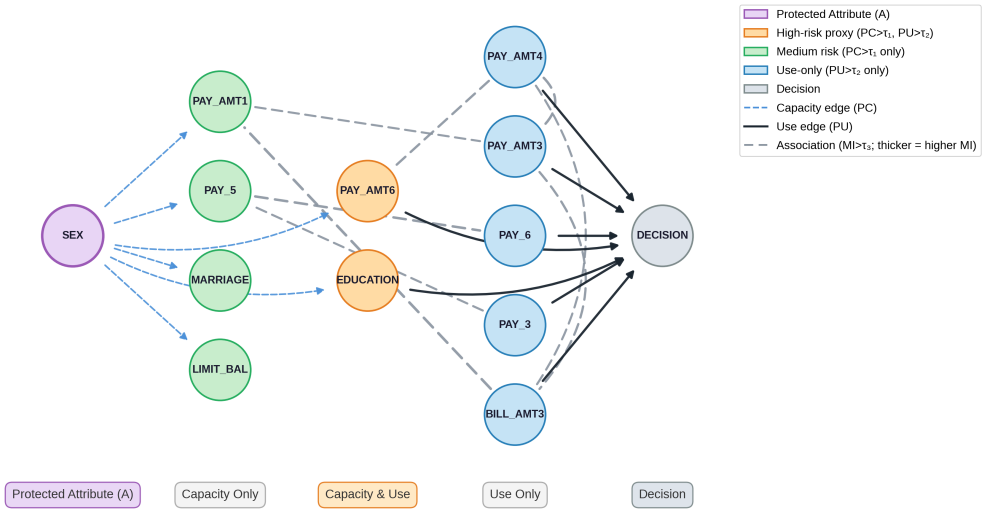


Figure 13: Proxy path graph for Credit Card under median-based PC and PU thresholds. With $\tau_1 = 0$, every feature with a non-zero capacity estimate receives a capacity edge from SEX. PAY_AMT6 and EDUCATION complete a proxy path (orange). PAY_AMT1, the feature with the highest capacity, has a PU equal to τ_2 and, like MARRIAGE, PAY_5, and LIMIT_BAL, a capacity edge only (green). PAY_3, PAY_6, PAY_AMT3, PAY_AMT4, and BILL_AMT3 are use-only (blue).

capacity edge from Male. This does not make them unproblematic, because Ethnicity is itself a protected characteristic under fair-lending law, so its direct use by the scorer falls outside the single-attribute audit reported here and requires an audit of its own (Sect. 5).

Answer to RQ2

The scorer exhibits both controlled sensitivity to the protected attribute and proxy-mediated risk. When sex is stated in the profile, replacing it with UNKNOWN changes the score by 0.310, 0.088, and 0.350 on German Credit, Credit Card, and Credit Approval, about as much as withholding the median feature on German Credit and Credit Approval and more than withholding any single feature on Credit Card (Table 9). With sex withheld, as in deployment, all three datasets contain features that both encode information about sex and are used by the scorer, but in three different structures. German Credit has a distributed cluster of mutually associated financial and demographic proxies, with seven features completing a proxy path and the five highest PRS values between 0.024 and 0.030 (Table 10; Fig. 12). Credit Approval has a single dominant proxy, Employed, whose PRS is about 2.7 times that of the next feature (Fig. 14). Credit Card has a weak and narrow structure, with PRS values four to five times smaller than on the other two datasets, complete paths only through PAY_AMT6 and EDUCATION, and heavily used repayment variables that carry no information about sex (Fig. 13).

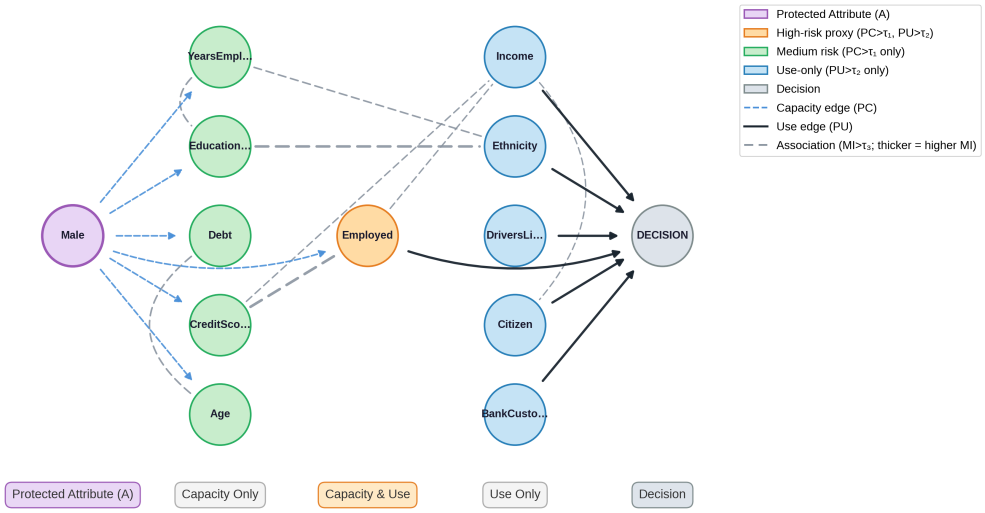


Figure 14: Proxy path graph for Credit Approval under median-based PC and PU thresholds. Employed is the only feature that completes a proxy path from Male to the decision (orange). EducationLevel, Age, YearsEmployed (PU equal to τ_2), Debt, and CreditScore have capacity edges only (green), and Income, Ethnicity, DriversLicense, Citizen, and BankCustomer are use-only (blue).

4.4 Mitigation Results (RQ3)

RQ3 asks whether reducing the scorer's reliance on the diagnosed proxies makes its decisions fairer, and at what cost to utility. We compare the two VAYNE-guided interventions of Sect. 3.6, proxy masking and proxy filtering, with the prompt-level baselines FPO and Fair-ICL of Sect. 4.1. The threshold θ_{PRS} is set from the RQ2 diagnosis on the development split, just below the PRS of the features that form the proxy structure identified there. This gives $\theta_{\text{PRS}} = 0.024$, 0.004, and 0.020, so that $\mathcal{F}_{\text{mask}}$ contains the cluster of five closely spaced highest-PRS features on German Credit (age, property, housing, credit-amount, and checking-account), the two features on Credit Card whose PRS is more than twice that of any other feature (PAY_AMT1 and PAY_AMT6), and the dominant proxy Employed on Credit Approval. Among the subsets of $\mathcal{F}_{\text{mask}}$ whose development-split BA is at most $\delta = 0.01$ below that of the unmitigated scorer, the subset with the lowest EOD is selected for filtering, which gives $\mathcal{F}_{\text{filt}} = \{\text{age, property, housing}\}$, $\{\text{PAY_AMT6}\}$, and $\{\text{Employed}\}$, respectively. On Credit Approval, both interventions therefore withhold the same feature and differ only in whether its field is kept with the value UNKNOWN or omitted, which isolates the effect of the sentinel value.

Table 11 reports the results on the test split, with the unmitigated scorer as Baseline. Differences in the fairness metrics between strategies are often smaller than their run-to-run standard deviations, so the comparisons below describe consistent tendencies rather than individually significant effects.

Table 11: Fairness and utility metrics before and after mitigation of the LLM credit scorer on the test split. Values are averaged over five scoring runs, and subscripts report the standard deviation across the runs. The best value per dataset and metric is highlighted in orange, and the second best is highlighted in yellow. Lower values are better for SP, EOD, and PP. Higher values are better for Acc. and BA.

Dataset	Strategy	SP	EOD	PP	Acc.	BA
German Credit	Baseline	$0.0396_{\pm 0.0599}$	$0.0787_{\pm 0.0978}$	$0.0067_{\pm 0.0483}$	$0.7250_{\pm 0.0285}$	$0.6083_{\pm 0.0337}$
	Masking	$0.0243_{\pm 0.0649}$	$0.0605_{\pm 0.0979}$	$\mathbf{0.0054_{\pm 0.0519}}$	$0.7150_{\pm 0.0303}$	$0.6012_{\pm 0.0331}$
	Filtering	$\mathbf{0.0145_{\pm 0.0713}}$	$\mathbf{0.0588_{\pm 0.1074}}$	$0.0060_{\pm 0.0535}$	$0.7250_{\pm 0.0363}$	$0.6131_{\pm 0.0384}$
	FPO	$0.0201_{\pm 0.0637}$	$0.0750_{\pm 0.1042}$	$0.0080_{\pm 0.0519}$	$0.6900_{\pm 0.0298}$	$0.5786_{\pm 0.0375}$
	Fair-ICL	$0.0295_{\pm 0.0612}$	$0.0632_{\pm 0.1084}$	$0.0059_{\pm 0.0547}$	$\mathbf{0.7350_{\pm 0.0297}}$	$\mathbf{0.6155_{\pm 0.0423}}$
Credit Card	Baseline	$0.0184_{\pm 0.0218}$	$0.2003_{\pm 0.0663}$	$0.0285_{\pm 0.1132}$	$0.7830_{\pm 0.0106}$	$0.5068_{\pm 0.0221}$
	Masking	$\mathbf{0.0045_{\pm 0.0198}}$	$0.2050_{\pm 0.0777}$	$0.0278_{\pm 0.1207}$	$0.7820_{\pm 0.0111}$	$0.5062_{\pm 0.0214}$
	Filtering	$0.0115_{\pm 0.0191}$	$0.1810_{\pm 0.0640}$	$\mathbf{0.0154_{\pm 0.1038}}$	$0.7890_{\pm 0.0118}$	$0.5221_{\pm 0.0244}$
	FPO	$0.0065_{\pm 0.0173}$	$0.2300_{\pm 0.0805}$	$0.0350_{\pm 0.1107}$	$0.7570_{\pm 0.0122}$	$0.4853_{\pm 0.0253}$
	Fair-ICL	$0.0120_{\pm 0.0169}$	$\mathbf{0.1760_{\pm 0.0657}}$	$0.0220_{\pm 0.1057}$	$\mathbf{0.7900_{\pm 0.0117}}$	$\mathbf{0.5260_{\pm 0.0271}}$
Credit Approval	Baseline	$0.0417_{\pm 0.0872}$	$0.1522_{\pm 0.1435}$	$0.0345_{\pm 0.0990}$	$0.8043_{\pm 0.0301}$	$0.8196_{\pm 0.0339}$
	Masking	$0.0290_{\pm 0.0780}$	$0.1210_{\pm 0.1454}$	$\mathbf{0.0152_{\pm 0.0836}}$	$0.7826_{\pm 0.0360}$	$0.8018_{\pm 0.0341}$
	Filtering	$\mathbf{0.0140_{\pm 0.0815}}$	$0.1105_{\pm 0.1404}$	$0.0215_{\pm 0.1021}$	$0.7971_{\pm 0.0377}$	$0.8114_{\pm 0.0313}$
	FPO	$0.0180_{\pm 0.0796}$	$0.1240_{\pm 0.1546}$	$0.0230_{\pm 0.0867}$	$0.7464_{\pm 0.0331}$	$0.7608_{\pm 0.0330}$
	Fair-ICL	$0.0290_{\pm 0.0747}$	$\mathbf{0.1045_{\pm 0.1597}}$	$0.0182_{\pm 0.0800}$	$\mathbf{0.8116_{\pm 0.0358}}$	$\mathbf{0.8261_{\pm 0.0286}}$

On German Credit, where RQ2 found sex encoded redundantly across a cluster of associated features, both VAYNE-guided interventions reduce all three disparities. Filtering withholds three of the five features in $\mathcal{F}_{\text{mask}}$ and attains the lowest SP (0.0145, against 0.0396 for the baseline) and the lowest EOD (0.0588, against 0.0787) of all strategies, while keeping the baseline accuracy of 0.7250 and slightly raising BA (0.6131 against 0.6083). Masking all five features gives the lowest PP (0.0054) and the second-lowest EOD (0.0605), but lowers SP less (0.0243) and costs 0.010 in accuracy. Because the members of the cluster are associated with one another, withholding three of them reduces SP and EOD more than masking all five, and the remaining two contribute utility rather than additional fairness. FPO lowers SP to 0.0201, the second-lowest value, but barely changes EOD (0.0750), raises PP to 0.0080, and lowers accuracy to 0.6900, the lowest of all strategies. Fair-ICL is the most accurate strategy (0.7350, BA 0.6155) and reduces all three disparities, though less than filtering on SP and EOD.

On Credit Card, the baseline already has a small SP (0.0184) but a large EOD (0.2003). Because the scorer seldom predicts default (Sect. 4.2), its disparities lie mainly in group-specific error rates rather than in decision rates. Filtering PAY_AMT6, the most used of the features that complete a proxy path, improves all five metrics. It attains the lowest PP (0.0154) and the second-lowest EOD (0.1810), and raises accuracy from 0.7830 to 0.7890 and BA from 0.5068 to 0.5221. Masking, which additionally withholds PAY_AMT1 and keeps both fields with the value UNKNOWN, yields the lowest SP (0.0045) with almost no change in accuracy (0.7820) or BA (0.5062), but leaves EOD slightly above the baseline (0.2050) and PP nearly

unchanged (0.0278). It thus equalizes how often default is predicted in each group without equalizing the errors. This agrees with RQ2, where PAY_AMT1 had the highest capacity but lay only on the use threshold, so adding it to the withheld set does not extend the reduction in error-rate disparity obtained by withholding PAY_AMT6 alone; because masking also keeps the fields with UNKNOWN, part of this difference may reflect the sentinel value, as on Credit Approval below. Fair-ICL attains the lowest EOD (0.1760) together with the highest accuracy (0.7900) and BA (0.5260), with filtering second on all three, whereas FPO lowers SP to 0.0065, the second-lowest value, while raising EOD to 0.2300 and PP to 0.0350 and lowering accuracy to 0.7570.

On Credit Approval, both interventions withhold Employed, the single dominant proxy identified in RQ2, and both lower all three disparities. Filtering attains the lowest SP of all strategies (0.0140, against 0.0417) and the second-lowest EOD (0.1105, against 0.1522), and lowers PP from 0.0345 to 0.0215, at a cost of 0.0072 in accuracy and 0.0082 in BA. Masking lowers SP (0.0290) and EOD (0.1210) less but attains the lowest PP (0.0152), and it lowers accuracy to 0.7826 and BA to 0.8018, about three times the accuracy loss and twice the BA loss of filtering. Since the two interventions remove the same information, this difference reflects the explicit UNKNOWN value, which the scorer appears to treat as a signal in its own right rather than as missing information; omitting the field is therefore the cleaner way to withhold a feature. FPO lowers SP to 0.0180, the second-lowest value, at a cost of 0.058 in accuracy and 0.059 in BA, and Fair-ICL attains the lowest EOD (0.1045) together with the highest accuracy (0.8116) and BA (0.8261).

The two prompt-level baselines behave consistently across datasets. FPO lowers SP on every dataset, each time to the second-lowest value, but never attains the best value of any metric, and it has the lowest accuracy and BA of all strategies on every dataset, costing 0.026 to 0.058 in accuracy and 0.021 to 0.059 in BA. Its effect on the error-based metrics is inconsistent, with EOD above the baseline on Credit Card and PP above the baseline on German Credit and Credit Card. Its instruction asks the model to disregard sex, which the profile does not contain, so it can shift the overall decision rate but not the use of the proxies through which, as RQ2 shows, information about sex reaches the decision. Fair-ICL is the only strategy that improves all five metrics on all three datasets, and it attains the highest accuracy and BA on every dataset and the lowest EOD on Credit Card and Credit Approval. By balancing sex and label among the demonstrations, it plausibly weakens the association between proxies and labels that the scorer learns in context, acting on the same proxy structure from the side of the examples rather than the features. It does not, however, identify which features carry protected information, and it therefore offers no feature-level account that could support an adverse-action explanation, whereas masking and filtering are derived from the diagnosis and come with that justification. The VAYNE-guided interventions attain the lowest SP, EOD, or PP in seven of the nine dataset–metric comparisons, and filtering stays within 0.015 of Fair-ICL in accuracy and BA on every dataset. Because Fair-ICL acts on the demonstrations and filtering on the profile, the two are complementary rather than competing.

Answer to RQ3

Reducing the scorer's reliance on the diagnosed proxies makes its decisions fairer at little cost to utility. The two VAYNE-guided interventions attain the lowest SP, EOD, or PP in seven of the nine dataset–metric comparisons. Proxy filtering lowers SP by 38% to 66%, EOD by 10% to 27%, and PP by 10% to 46% relative to the unmitigated scorer, while changing accuracy by less than 0.01 (from -0.0072 to $+0.0060$) and BA by less than 0.02 (from -0.0082 to $+0.0153$), and it preserves at least as much accuracy and BA as masking on every dataset. FPO lowers SP on every dataset but never attains the best value of any metric and has the lowest utility on every dataset, whereas Fair-ICL improves every metric on every dataset and attains the highest utility throughout, without identifying the features that carry the disparity. The most effective VAYNE-guided intervention follows the proxy structure diagnosed in RQ2. It withholds part of the redundant cluster on German Credit, the used proxy on Credit Card, and the dominant proxy on Credit Approval, where omitting the field instead of masking it avoids most of the utility cost.

5 Discussion and Limitations

VAYNE offers a practical black-box way to diagnose proxy-mediated unfairness in LLM-based credit decisions, but several limitations bound what these results can support. The results speak to model-risk and fair-lending review, but do not by themselves establish compliance. The scorer responds to the protected attribute when it is stated, which is one reason to keep the attribute out of the profile; with the attribute withheld, residual risk is concentrated in feature-level proxy dependencies, and the ranked proxy-risk table and path graph can guide which variables need deeper validation, but are not automatically equivalent to the specific reasons an adverse-action notice must give. The ablation-based sensitivity underlying every score is also an observational quantity rather than a causal effect, since identifying true causal pathways would require a structural causal model unavailable for a black-box LLM, so VAYNE's output is best read as suggestive evidence for a compliance review rather than a legal or causal determination.

The scores themselves also limit what VAYNE can show. The model's p_yes is a verbalized number rather than a calibrated probability, so DU and PU measure the sensitivity of that number, and a change of a given size need not correspond to the same change in risk across datasets. Ablation with UNKNOWN combines the removal of information with any effect of the token itself, and on Credit Approval, where masking and filtering act on the same feature and therefore differ only in this respect, masking costs about three times as much accuracy and twice as much BA as filtering, which points to an effect of the token itself. Because the same token enters DU_A and PU alike, the two remain comparable with each other, although their absolute size may partly reflect the sentinel chosen.

Mitigation, in turn, does not follow a fixed ranking across products, but tracks the diagnosed proxy structure. Withholding part of a redundant cluster can reduce disparities without any loss of accuracy, as on German Credit, withholding the most used proxy of a weak structure can improve fairness and utility together, as on Credit

Card, and the way a feature is withheld matters as well. On Credit Approval, masking Employed with UNKNOWN costs about three times as much accuracy as omitting the field. A single fairness instruction lowered SP on every dataset but never attained the best value of any metric and had the lowest accuracy and BA of all strategies on every dataset, which suggests that such an instruction does not reliably change how the model uses the remaining features. The fairness differences between strategies are, moreover, often smaller than their run-to-run variation, so the rankings in Table 11 describe tendencies that larger audits should confirm. In practice, an audit should be re-run after any candidate intervention, since the diagnosed structure can itself shift once high-risk features are altered.

Finally, the scope of what we audited is itself a limitation. Protected-attribute definitions are jurisdiction-specific, and each audit here considers a single protected attribute, sex. Other fields in the deployment profiles are themselves protected or sensitive in some jurisdictions. Age, which is among the highest-PRS features on German Credit, is a protected characteristic under ECOA subject to specific exceptions, and Credit Approval contains a field labelled Ethnicity that the scorer uses heavily. The single-attribute audit says nothing about the permissibility of using such fields directly; a complete fair-lending review would audit each protected attribute in turn and decide which fields may enter the profile at all. The Credit Approval column names come from a public relabelling of an anonymized dataset, so conclusions that depend on what a column means should be checked against a dataset with documented semantics. All results come from a single LLM configuration, which is not the strongest benchmark configuration on every dataset and, on Credit Card, reaches its accuracy mainly on non-defaulting clients (BA 0.5068), and other models reach different accuracy and fairness operating points on identical prompts, so proxy scores and mitigation rankings should not be assumed to transfer to another model or provider without re-running the audit. VAYNE's cost is dominated by LLM queries and scales linearly in applicants, features, and repetitions, and extending it to more models, open-weight deployments where this cost can be batched locally, and credit domains beyond consumer lending are natural next steps.

6 Conclusions and Outlook

We presented VAYNE, a black-box framework for explaining and mitigating proxy discrimination in LLM-based credit decisions. VAYNE separates a feature's capacity to encode the protected attribute from the scorer's use of that feature, combines the two into a proxy risk score and a proxy path graph, and derives targeted interventions from this diagnosis, using only query access to the model. The experiments support three conclusions. First, LLMs become usable credit scorers only once labeled examples are placed in the prompt. Without examples they are close to chance, whereas a small number of examples raises their accuracy into the range of traditional classifiers, and their best configurations exceed these classifiers on every dataset. Because utility, and with it the apparent fairness of a scorer, depends on the model, the encoding, the shot count, and the decoding temperature, these parameters have to be fixed before fairness is measured. Second, the audited scorer responds to the applicant's

sex when it is stated, at least as strongly as to a typical feature of the profile, and once sex is withheld, information about it reaches the decision through non-protected features whose structure differs by dataset, with a single dominant proxy on Credit Approval, a redundant cluster of associated features on German Credit, and a weak, narrow structure on Credit Card. Third, interventions derived from this diagnosis reduce group disparities at little cost, attaining the lowest disparity in seven of nine dataset–metric comparisons. Proxy filtering lowers the statistical parity gap by 38% to 66% and the equalized-odds gap by 10% to 27% with accuracy changes below 0.01, a better trade-off than an instruction-based debiasing prompt, which lowered statistical parity but had the lowest utility on every dataset, and omitting a proxy field is preferable to masking it with UNKNOWN. Fair in-context learning improved every metric but does not explain which features carry the disparity. Taken together, these results show that withholding the protected attribute is necessary but not sufficient for an LLM credit scorer, and that a feature-level diagnosis can tell a lender which variables to act on and what acting on them costs, which an aggregate parity metric cannot. The evidence is diagnostic and associational rather than causal or legal, and it comes from a single LLM configuration audited on three benchmarks. Natural next steps are to extend the audit to more models, protected attributes, and credit products, to combine feature-level filtering with demonstration-level methods such as fair in-context learning, and to connect the diagnosis to a validated causal layer where model access allows it.

Declarations

- Funding
  This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
- Competing interests
  The authors declare that they have no competing interests.
- Clinical Trial Number
  Clinical Trial Number was not required for this study.
- Ethics approval and consent to participate
  Ethical approval was not required for this study.
- Consent to participate
  Not applicable.
- Consent for publication
  Not applicable.
- Data Availability
  The datasets are publicly available at https://github.com/chucunhong/vayne
- Code availability
  The source code is publicly available at https://github.com/chucunhong/vayne
- Contribution of authors
  crisp

References

[1] Wang, Z., Chu, Z., Doan, T.V., Ni, S., Yang, M., Zhang, W.: History, development, and principles of large language models: an introductory survey. AI and Ethics **5**(3), 1955–1971 (2025) https://doi.org/10.1007/s43681-024-00583-7

[2] Barry, L., Charpentier, A.: Melting contestation: Insurance fairness and machine learning. Ethics and Information Technology **25**(4), 49 (2023) https://doi.org/10.1007/s10676-023-09720-y

[3] Wilson, K., Caliskan, A.: Gender, race, and intersectional bias in resume screening via language model retrieval. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, vol. 7, pp. 1578–1590 (2024). https://doi.org/10.48550/arXiv.2407.20371

[4] Chinta, S.V., Wang, Z., Palikhe, A., Zhang, X., Kashif, A., Smith, M.A., Liu, J., Zhang, W.: Ai-driven healthcare: Fairness in ai healthcare: A survey. PLOS digital health **4**(5), 0000864–0000864 (2025) https://doi.org/10.1371/journal.pdig.0000864

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.*: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) https://doi.org/10.48550/arXiv.2005.14165

[6] Dinh, T., Zeng, Y., Zhang, R., Lin, Z., Gira, M., Rajput, S., Sohn, J.-y., Papailiopoulos, D., Lee, K.: Lift: Language-interfaced fine-tuning for non-language machine learning tasks. Advances in Neural Information Processing Systems **35**, 11763–11784 (2022) https://doi.org/10.48550/arXiv.2206.06565

[7] Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., Sontag, D.: Tabllm: Few-shot classification of tabular data with large language models. In: International Conference on Artificial Intelligence and Statistics, pp. 5549–5581 (2023). https://doi.org/10.48550/arXiv.2210.10723 . PMLR

[8] Feng, D., Dai, Y., Huang, J., Zhang, Y., Xie, Q., Han, W., Chen, Z., Lopez-Lira, A., Wang, H.: Empowering many, biasing a few: Generalist credit scoring through large language models. arXiv preprint arXiv:2310.00566 (2023) https://doi.org/10.48550/arXiv.2310.00566

[9] Lee, J., Stevens, N., Han, S.C.: Large language models in finance (finllms). Neural Computing and Applications **37**(30), 24853–24867 (2025) https://doi.org/10.1007/s00521-024-10495-6

[10] Nie, Y., Kong, Y., Dong, X., Mulvey, J.M., Poor, H.V., Wen, Q., Zohren, S.: A survey of large language models for financial applications: Progress, prospects and challenges. arXiv preprint arXiv:2406.11903 (2024) https://doi.org/10.485

50/arXiv.2406.11903

[11] Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017) https://doi.org/10.1126/science.aal4230

[12] Sheng, E., Chang, K.-W., Natarajan, P., Peng, N.: Societal biases in language generation: Progress and challenges. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (volume 1: Long Papers), pp. 4275–4293 (2021). https://doi.org/10.18653/v1/2021.acl-long.330

[13] Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K.: Bias and fairness in large language models: A survey. Computational linguistics **50**(3), 1097–1179 (2024) https://doi.org/10.1162/coli_a_00524

[14] Doan, T.V., Chu, Z., Wang, Z., Zhang, W.: Fairness definitions in language models explained. arXiv preprint arXiv:2407.18454 (2024) https://doi.org/10.48550/arXiv.2407.18454

[15] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268 (2015). https://doi.org/10.1145/2783258.2783311

[16] Barocas, S., Selbst, A.D.: Big data's disparate impact. California law review, 671–732 (2016)

[17] Rothstein, R.: The Color of Law: A Forgotten History of How Our Government Segregated America. Liveright publishing, ??? (2017). https://doi.org/10.1080/01944363.2020.1803632

[18] Datta, A., Fredrikson, M., Ko, G., Mardziel, P., Sen, S.: Proxy non-discrimination in data-driven systems. arXiv preprint arXiv:1707.08120 (2017) https://doi.org/10.48550/arXiv.1707.08120

[19] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016) https://doi.org/10.5555/3157382.3157469

[20] Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017) https://doi.org/10.1089/big.2016.00

[21] Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016). https://doi.org/10.48550/arXiv.1602.04938

[22] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017) https://doi.org/10.48550/arXiv.1705.07874

[23] Tschantz, M.C.: What is proxy discrimination? In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1993–2003 (2022). https://doi.org/10.1145/3531146.3533242

[24] Weerts, H., Kelly-Lyth, A., Binns, R., Adams-Prassl, J.: Unlawful proxy discrimination: a framework for challenging inherently discriminatory algorithms. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 1850–1860 (2024). https://doi.org/10.1145/3630106.3659010

[25] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 797–806 (2017). https://doi.org/10.1145/3097983.3098095

[26] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021) https://doi.org/10.1145/3457607

[27] Adler, P., Falk, C., Friedler, S.A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., Venkatasubramanian, S.: Auditing black-box models for indirect influence. Knowledge and Information Systems **54**(1), 95–122 (2018) https://doi.org/10.1007/s10115-017-1116-3

[28] Cherepanova, V., Lee, C.-J., Akpinar, N.-J., Fogliato, R., Lopez, M.B., Kearns, M., Zou, J.: Improving llm group fairness on tabular data via in-context learning. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, vol. 8, pp. 579–590 (2025). https://doi.org/10.1609/aies.v8i1.36572

[29] Chhikara, G., Sharma, A., Ghosh, K., Chakraborty, A.: Few-shot fairness: Unveiling llm's potential for fairness-aware classification. arXiv preprint arXiv:2402.18502 (2024) https://doi.org/10.48550/arXiv.2402.18502

[30] Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **12**(3), 1452 (2022) https://doi.org/10.1002/widm.1452

[31] Giang Thi Thu, H., Doan, T.V., Ban, H.-B., Le Quy, T.: An experimental study on fairness-aware machine learning for credit scoring problems. Digital Finance **8**(3), 35 (2026) https://doi.org/10.1007/s42521-026-00202-6

[32] Wang, R., Wang, Z., Sun, J.: Unipredict: Large language models are universal tabular classifiers. arXiv preprint arXiv:2310.03266 (2023) https://doi.org/10.48550/arXiv.2310.03266

[33] Han, S., Yoon, J., Arik, S.O., Pfister, T.: Large language models can automatically engineer features for few-shot tabular learning. arXiv preprint arXiv:2404.09491 (2024) https://doi.org/10.48550/arXiv.2404.09491

[34] Fang, X., Xu, W., Tan, F.A., Hu, Z., Zhang, J., Qi, Y., Sengamedu, S.H., Faloutsos, C.: Large language models (llms) on tabular data: Prediction, generation, and understanding-a survey. Transactions on Machine Learning Research (2024) https://doi.org/10.48550/arXiv.2402.17944

[35] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226 (2012). https://doi.org/10.1145/2090236.2090255

[36] Plecko, D., Bareinboim, E.: Causal fairness analysis: A causal toolkit for fair machine learning. Foundations and Trends in Machine Learning **17**(3), 304–589 (2024) https://doi.org/10.48550/arXiv.2207.11385

[37] Blau, F.D., Kahn, L.M.: The gender wage gap: Extent, trends, and explanations. Journal of economic literature **55**(3), 789–865 (2017) https://doi.org/10.1257/jel.20160995

[38] Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A.: Predictably unequal? the effects of machine learning on credit markets. The Journal of Finance **77**(1), 5–47 (2022) https://doi.org/10.1111/2Fjofi.13090

[39] Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017) https://doi.org/10.48550/arXiv.1711.00399

[40] Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 607–617 (2020). https://doi.org/10.1145/3351095.3372850

[41] Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 180–186 (2020). https://doi.org/10.1145/3375627.3375830

[42] Balagopalan, A., Zhang, H., Hamidieh, K., Hartvigsen, T., Rudzicz, F., Ghassemi, M.: The road to explainability is paved with bias: Measuring the fairness

of explanations. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1194–1206 (2022). https://doi.org/10.1145/3531146.3533179

[43] Deck, L., Schoeffer, J., De-Arteaga, M., Kühl, N.: A critical survey on fairness benefits of explainable ai. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 1579–1595 (2024). https://doi.org/10.1145/3630106.3658990

[44] Golec, M., AlabdulJalil, M.: Interpretable llms for credit risk: A systematic review and taxonomy. Expert Systems with Applications, 130941 (2025) https://doi.org/10.1016/j.eswa.2025.130941

[45] Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. Advances in neural information processing systems **30** (2017) https://doi.org/10.48550/arXiv.1703.06856

[46] Nabi, R., Shpitser, I.: Fair inference on outcomes. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018). https://doi.org/10.48550/arXiv.1705.10378

[47] Chiappa, S.: Path-specific counterfactual fairness. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 7801–7808 (2019). https://doi.org/10.1609/aaai.v33i01.33017801

[48] Paninski, L.: Estimation of entropy and mutual information. Neural Computation **15**(6), 1191–1254 (2003) https://doi.org/10.1162/089976603321777073

[49] Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics **69**(6), 066138 (2004) https://doi.org/10.1103/PhysRevE.69.066138

[50] Ross, B.C.: Mutual information between discrete and continuous data sets. PloS one **9**(2), 87357 (2014) https://doi.org/10.1371/journal.pone.0087357

[51] Quinlan, J.R.: Induction of decision trees. Machine learning **1**(1), 81–106 (1986) https://doi.org/10.1007/BF00116251

[52] Webb, G.I., Keogh, E., Miikkulainen, R.: Naïve bayes. Encyclopedia of machine learning **15**(1), 713–714 (2010) https://doi.org/10.1007/978-1-4899-7687-1_581

[53] Popescu, M.-C., Balas, V.E., Perescu-Popescu, L., Mastorakis, N.: Multilayer perceptron and neural networks. WSEAS transactions on circuits and systems **8**(7), 579–588 (2009) https://doi.org/10.5555/1639537.1639542

[54] Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009) https://doi.org/10.4249/scholarpedia.1883