Concept-RidgeAIME: LLM-Guided Automatic Concept-Based Explanations via Ridge-Regularized Inverse Operators for Trustworthy AI

Anonymous authors
Paper under double-blind review

Abstract

Concept-based explanations overcome the limitations of low-level feature importance and focus on high-level, human-understandable concepts to explain the decision-making behind machine learning models. However, achieving model independence and the simultaneous presentation of global and local information within a single framework has been difficult. This study extends the concept of approximate inverse model explanations (AIME) and proposes Concept-RidgeAIME, which simultaneously obtains global and local explanations via concepts by utilizing a regularized linear approximate inverse mapping as its core. The proposed method learns a two-stage structure—an inverse operator mapping from the model output to the input and an inverse operator mapping from the concept to the input—only once. Subsequently, it efficiently calculates the contribution and ratio of concepts for any individual using simple matrix-vector operations. Without requiring access to internal representations or gradients, it presents global (concept importance ranking) and local (individual concept contributions) information within the same framework, thereby achieving model independence with low overhead. Using the global feature importance as a foundation, this study demonstrates a workflow in which a large language model automatically synthesizes rule concepts composed of normalization thresholds and one-hot equations, then validates the syntax and excludes zero/positive cases to ensure robustness. Evaluations quantified the reconstructability (completeness) of black-box outputs and coverage (projection completeness) at the concept base level using tabular benchmarks (Adult, German Credit, and COMPAS). Stability and efficiency were verified using bootstrap confidence intervals and inference time (millisecond-level). Results showed that Concept-RidgeAIME demonstrated practical advantages over conventional concept-based methods (ConceptSHAP, CBM, and TCAV) and the application of generic SHAP to the concept space. These advantages are achieved by Concept-RidgeAIME through a model-independent implementation that requires no additional training and can handle global, local, and concept mappings in an integrated manner.

1 Introduction

With the deployment of high-performance machine learning models in society, the demand to explain *the reasoning behind a decision* using a human conceptual vocabulary has been increasing. Post-hoc explanations, represented by methods such as local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016) and Shapley additive explanations (SHAP) (Lundberg & Lee, 2017), visualize local feature contributions based on perturbations near the input-output neighborhood. However, because output units remain confined to low-level features such as pixels or one-hot encoding, even experts find it difficult to connect these explanations to a causal or counterfactual understanding of the decision-making process. By contrast, *concept-based explanations* such as TCAV (Kim et al., 2018), ConceptSHAP (Yeh et al., 2020), and concept bottleneck models (CBMs) (Koh et al., 2020) hold an advantage: they can use human vocabulary (e.g., "highly educated" or "managerial position") to represent feature importance. However, many methods require gradient access or additional training, making it difficult to satisfy both model independence and the simultaneous presentation of global and local information within a single computational framework.

This research reexamines this gap from the perspective of approximate inverse problems. Approximate inverse model explanations (AIME) (Nakanishi, 2023) constructs a single linear operator that approximately inverts the mapping from *output* → *input* in a least-squares manner, presenting a unified framework for reading its *sequence* (global) and *action* (local). AIME requires no gradient or internal parameters and provides global and local explanations with only one precomputation and matrix-vector multiplication. Consequently, it can be directly applied to gradient-discontinuous tree models and hidden APIs. This study proposes (i) RidgeAIME, which introduces Tikhonov regularization to enhance numerical stability and explanation consistency while preserving these advantages, and (ii) Concept-RidgeAIME, which elevates the explanation unit from features to human-readable concepts. The latter is novel because it connects the inverse operators of the output-to-input and concept-to-input mappings solely through linear algebra, thereby providing both *global concept formation* and *individual concept contribution (local)* under the same inference rule.

Furthermore, to maintain the design cost of the concepts at a practically acceptable level, this study uses the global feature importance (GFI) from AIME as a scaffold. It then employs a large language model (LLM) to automatically synthesize rule forms for minority literals (e.g., normalization thresholds or one-hot encoding). After generation, the program performs *syntax and data sanitization* by (a) matching feature names, (b) imposing range constraints (e.g., numerical values in $[0, 1]$), and (c) excluding zero-positive rules. Only the *concepts that pass this process* are adopted as the basis for the concept space, thereby minimizing manual trial-and-error while endowing a linear inverse mapping system—which is model-independent, gradient-free, and low-overhead—with concept-level readability.

This study makes three contributions. First, it introduces **RidgeAIME**, which enhances AIME's inverse mapping (Nakanishi, 2023) with *ridge* regularization, to stably ensure the coefficient of determination (BB completeness) and reproducibility of contribution rankings, even under conditions of high correlation, few samples, and many classes. Second, **Concept-RidgeAIME** combines two inverse operators—output-to-input and concept-to-input—to simultaneously provide, through *single linear algebra*, (a) global concept formation (column vectors) and (b) individual concept contribution vectors (actions). Third, it establishes two evaluation metrics: AIME-style emphreconstruction-based completeness (BB R^2) and the local contribution's emphconcept basis coverage (projection completeness). Finally, it presents a concept design workflow assisted by LLMs (GFI → rule generation → syntactic/zero-positive verification → adoption), enabling the automatic reproduction of global, local, and concept contributions for tabular data (Adult, German Credit, and COMPAS) without additional retraining.

Thus, Concept-RidgeAIME achieves (a) model independence and gradient-free operation; (b) unification of global, local, and conceptual contributions within the same linear framework; (c) completeness (BB projection) and stability (CI); and (d) sub-millisecond execution efficiency (post-precomputation). It satisfies these four requirements, surpassing existing concept-based methods (e.g., TCAV (Kim et al., 2018), ConceptSHAP (Yeh et al., 2020), and CBMs (Koh et al., 2020)) and SHAP-based methods, while providing *highly readable explanations suitable for real-world deployment.*

The rest of this paper is organized as follows: the related works are described in Section 2, the implementation of Concept-RidgeAIME is explained in Section 3, the experiments are discussed in Section 4, and the conclusions are provided in Section 5.

2 Related Works

Recent concept-based explanations (concept-based explainable artificial intelligence (XAI)) can be broadly categorized into *(A) post-hoc methods measuring the sensitivity and contribution of externally defined concepts*, *(B) methods embedding concepts into the model structure*, *(C) methods using examples or prototypes as concepts*, and *(D) operational approaches applying general feature attribution methods to conceptual representations.* This classification is useful for positioning these approaches relative to each other in terms of the timing of concept introduction (training/post-training), requirements for accessing gradients or intermediate representations, and units of explanation output (global, local, or interventional). A recent systematic survey (Khoozani et al., 2024) traversed this diversifying landscape, organizing key issues around the quality control of concept definitions, evaluation metrics for faithfulness, and connections to automated concept discovery and counterfactual operations.

(A) For post-hoc concept sensitivity/contribution, TCAV (Kim et al., 2018) uses directional derivatives with respect to the concept activation vector (CAV) learned from a few positive examples to quantify concept sensitivity for class predictions. Because it requires gradient access and internal representations, it is difficult to apply to nondifferentiable models, such as tree-based models or hidden APIs. However, it has been widely used as a standard method for measuring *global relevance per class*, primarily in the image domain. As an automation of TCAV, Ghorbani et al. (2019) proposed ACE, which first extracts candidate concepts by oversegmenting and clustering images and then assigns importance using TCAV. Methods such as **ICE** (Zhang et al., 2021), which extends CAV from linear to region-based, and **CAR** (Crabbé & van der Schaar, 2022), which generalizes feature regions occupying concepts, are positioned within the trend aimed at improving concept separability and fidelity. The game-theory-based ConceptSHAP defines the sufficiency of concept sets and ensures the axiomatic validity of global importance by allocating marginal contributions via Shapley values; however, it incurs a high computational load owing to combinatorial growth in subset evaluations. Furthermore, **CCE** (Abid et al., 2022), which enables counterfactual debugging at the concept level, constructs *meaningful concept counterfactuals* for each individual to perform causal attributions, thus demonstrating the feasibility of ex-post concept manipulation.

(B) For concept internalization (during learning), **CBMs** (Koh et al., 2020) explicitly predict concepts at intermediate layers and infer final labels on top of them, thereby reconciling *interventional capability* (rewriting concepts to control output) and *global and local explanations*. *Post-hoc CBM* has also been proposed for pretrained black-box models (Yüksekgönül et al., 2022). **SENN** (Alvarez-Melis & Jaakkola, 2018) simultaneously learns “interpretable base concepts” and linear readout, ensuring separability and stability through regularization. **Concept whitening** (Chen et al., 2020) enhances the interpretability of internal representations by inserting a whitening layer into convolutional neural networks (CNNs) to align latent axes with known concepts. Although internalization methods require concept supervision or retraining, they offer the advantage of providing highly coherent explanations through *intervention/constrained learning*.

(C) As a method for treating examples or prototypes as concepts, **ProtoPNet** (Chen et al., 2019) performs classification based on similarity to *prototypes (typical patches)* learned per class, presenting *example-based local explanations* of the “this looks like that” type. **Net2Vec** (Fong & Vedaldi, 2018) quantifies the correspondence between filters and concepts, while **Concept Attribution** (Wu et al., 2020) constructs *global explanations* for CNNs. Furthermore, Kumar et al. (2021) and Kamakshi et al. (2021) proposed implementation approaches—MACE and PACE, respectively—that extract and summarize concepts from visual models as model- and architecture-independent *posterior concept extractors*.

(D) Regarding the application of generic attribution methods to concept representations, **Concept Space SHAP** is not a formally defined method but rather an operational approach that treats *existing concept representations* (such as rule scores or one-hot concepts) as features and applies generic SHAP (Lundberg & Lee, 2017). While it yields *local and global contributions* based on the Shapley axioms, the explanation quality depends on the *computational load of the sampling approximation* and the *fidelity and granularity of concept representations*. By contrast, applying SHAP (Lundberg & Lee, 2017) to *arbitrary concept representations* is also common; however, this is merely an *operational practice* of replacing SHAP features with concept features (user-designed or automatically extracted) and is not a distinct method name. Similarly, while local Shapley values per concept can be obtained, computational overhead from sampling or approximation remains unavoidable. Moreover, it does not provide the *global, local, or inter-concept mappings* within a unified linear framework.

The *Concept-RidgeAIME* method proposed in this study extends the approximate inverse operator of AIME (Nakanishi, 2023) to the concept space. It simultaneously presents *global (concept formation) and local (concept contribution) explanations through a single linear algebra operation* in a *closed form with regularization* while remaining *model-independent and gradient-free*. Post-processing systems (TCAV, ACE, ConceptSHAP, CCE) provide concept relevance, completeness, and counterfactuality but have computational and access constraints that limit their use. Internalization systems (CBM, SENN, concept whitening) offer intervenability and high fidelity but require retraining. Concept Space SHAP is versatile but depends on the computational cost of approximation and concept representation quality. By contrast, Concept-RidgeAIME enables the evaluation of *BB completeness (reconstruction R^2)* and *concept projection completeness (Projection)* within a single implementation via *one-time precomputation and matrix-vector multiplication* and adapts to *gradient-independent operational environments* such as tabular, tree-based, and confidential APIs. Furthermore, to address the challenges identified in the survey (Khoozani et al., 2024) (quality control and automatic discovery of concept definitions, connection with counterfactual explanations), this study combines *lightweight rule synthesis and validation using LLMs* with *reconstruction evaluation of linear inverse mappings*, making the *concept design, verification, and presentation* reproducible, which holds practical significance.

Yu et al. (2025) comprehensively reviewed the challenges and solutions for achieving "Trustworthiness" in LLM-based agents and multiagent systems, systematizing all aspects from attack and defense methods to evaluation approaches under the TrustAgent framework. Mumuni & Mumuni (2025) comprehensively organized the development of XAI, considering structurally explainable models, black-box models, and even automated explanation generation by using LLMs and vision-language models, presenting their strengths and challenges. Bilal et al. (2025) comprehensively discussed the latest technologies, applications, and evaluation methods for utilizing LLMs in XAI, presenting user-friendly explanation generation, evaluation, and prospects for real-world applications. Furthermore, Basheer et al. (2025) proposed a framework for predicting vulnerability patches using LLM-based BERT models in the cybersecurity domain and discussed practical applications integrated with reliability enhancements. Benk et al. (2025) investigated user expectations and values regarding LLM reliability standards, identified a perception gap between developers and users, and examined the challenges for the standardization of these models. These studies demonstrate the rapid expansion of XAI initiatives incorporating LLMs. However, most remain at the conceptual organization or application stage, and challenges persist in establishing an implementation foundation that uniformly provides global-, local-, and conceptual-level explanations within a model-independent framework.

By contrast, our approach is unique in that Concept-RidgeAIME, an extension of AIME, simultaneously achieves explanations across these three layers using only linear algebra operations, which are independent of gradients or internal representations. Furthermore, by combining LLM-based automatic rule generation with consistency verification, a reproducible framework that streamlines the design of the conceptual space is provided, thereby addressing the limitations of existing research.

Overall, Concept-RidgeAIME simultaneously achieves (a) model and gradient independence; (b) simultaneous presentation of global, local, and conceptual contributions; (c) completeness evaluation based on reconstruction metrics; (d) numerical stability and reproducibility through regularization; and (e) operational feasibility of concept design via LLM assistance. Thus, it complements the strengths of TCAV, Concept-SHAP, CBM, and Concept Space SHAP while filling practical gaps in black-box explanations for tabular data (gradient independence, low computational load, and local concept contribution).

3 Method

This section extends the framework of AIME (Nakanishi, 2023) to a notation where samples are arranged in rows. $X \in \mathbb{R}^{d\times n}$ (d-dimensional features, n samples), $Z \in \mathbb{R}^{k\times n}$ (k-dimensional downstream representation, including class one-hot, logit, intermediate activations). The core function of AIME is to approximate the unknown forward linearization $A \in \mathbb{R}^{k\times d}$ under the assumption that it yields $z \approx Ax$. This approximation is performed *from the output to the input side* by estimating the approximate inverse operator $A^\dagger \in \mathbb{R}^{d\times k}$ that maps from data (X, Z) to $z \approx Ax$ via least-squares estimation. This single linear operator simultaneously extracts global and local explanations.

In column-vector notation, the approximate inverse operator $W \in \mathbb{R}^{d\times k}$ is defined as

$$\min_{W\in\mathbb{R}^{d\times k}} \left\| X - WZ \right\|_F^2 \quad \Longrightarrow \quad W^\star = X Z^\dagger = A^\dagger, \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This formulation corresponds to the minimum-norm solution using the Moore–Penrose pseudoinverse $Z^\dagger$. Here, A denotes the conceptual forward operator representing the local linearization (mean Jacobian) of the model, and it is directly approximated from the data matrix as $A^\dagger \approx XZ^\dagger$.

Global/Local Readout. Once the approximate inverse operator $W^\star = A^\dagger$ is obtained, the GFI (a k-dimensional weight vector for feature j) can be derived from $(W^\star)_{j:} \in \mathbb{R}^{1\times k}$ (row j), with the scalar importance metric $\|(W^\star)_{j:}\|_p$ (e.g., $p = 2$). The local feature contribution of sample i (column vector $x_i \in \mathbb{R}^d$, $z_i \in \mathbb{R}^k$) is

$$\ell_i \;:=\; \big(W^\star z_i\big) \;\circ\; x_i \;\in\; \mathbb{R}^d, \tag{2}$$

where ($\circ$ denotes the Hadamard product). The sign indicates the direction of change (boost or suppression).

3.1 RidgeAIME: Approximate Inverse with Tikhonov Regularization

The pseudoinverse $Z^\dagger$ may become numerically unstable under conditions such as high correlation, few samples, or many classes. Therefore, Tikhonov (ridge) regularization is applied as follows:

$$\min_W \big\|\, X - WZ \,\big\|_F^2 \;+\; \lambda \,\|W\|_F^2, \qquad \lambda \geq 0, \tag{3}$$

with the regularized solution

$$W_\lambda \;=\; X\,Z^\top\,(ZZ^\top + \lambda I_k)^{-1} \;=\; X\,Z_\lambda^\dagger \;=\; A_\lambda^\dagger, \tag{4}$$

where $Z_\lambda^\dagger := Z^\top(ZZ^\top + \lambda I_k)^{-1}$ denotes the Tikhonov pseudoinverse. As $\lambda \to 0$, this reduces to $W_\lambda \to W^\star$. The readout rule is identical to equation 2, with $W^\star$ replaced by W_λ:

$$\text{GFI (feature } j\text{): } (W_\lambda)_{j:}, \qquad \text{LFI (sample } i\text{): } \ell_i := (W_\lambda z_i) \circ x_i. \tag{5}$$

This regularization improves the condition number of W_λ, enhancing robustness against outliers and noise and increasing the reproducibility of the GFI rankings.

3.2 Concept-RidgeAIME: Two-Stage Extension to Concept Space

To use *human-readable concepts* rather than features as the units of explanation, a two-stage approximate inverse is constructed while retaining the column-wise notation.

Concept Score Matrix. The concept score matrix $C \in \mathbb{R}^{q\times n}$ (q concepts) is evaluated for each sample from LLM-generated or manually defined rules. Numerical features are handled using normalization thresholds, and categorical features are represented using one-hot encoding.

Approximate Inverse from Concept to Feature. To re-express the local feature contribution vector ℓ_i on this concept basis,

$$\min_{U\in\mathbb{R}^{d\times q}} \big\|\, X - UC \,\big\|_F^2 \;+\; \gamma \,\|U\|_F^2 \quad\Longrightarrow\quad U_\gamma \;=\; X\,C^\top\,(CC^\top + \gamma I_q)^{-1}. \tag{6}$$

Here, (U_γ) is the approximate inverse operator mapping "concepts to t features."

Local Concept Contribution. Mapping ℓ_i from equation 5 to the concept basis yields

$$v_i \;:=\; U_\gamma^\top\,\ell_i \;\in\; \mathbb{R}^q, \qquad \text{ratio}_{i,c} \;:=\; \frac{v_{i,c}}{\sum_{c'=1}^{q} |v_{i,c'}|}. \tag{7}$$

Here, $v_{i,c} > 0$ (< 0) indicates that concept c boosts (suppresses) the judgment, and i indexes the sample, with $i = 1, \ldots, n$. The ratio ratio represents the relative contribution of each concept and is highly interpretable.

3.3 Completeness Metric

The adequacy of the explanation is evaluated using two R^2 metrics based on AIME's reconstruction accuracy. Specifically, the Frobenius norm is used in BB completeness to capture the overall reconstruction error across all samples and features, whereas the Euclidean norm (2-norm) is used in projection completeness to measure the fidelity of each individual contribution vector when projected onto the concept basis.

BB Completeness (Output→Input).

$$R^2_{\mathrm{BB}} \;:=\; 1 - \frac{\left\| X - W_\lambda Z \right\|_F^2}{\left\| X - \bar{X} \right\|_F^2}, \qquad \bar{X} \;=\; \frac{1}{n} X \mathbf{1}\mathbf{1}^\top, \tag{8}$$

where $\mathbf{1}$ denotes a vector of all 1s with length n). For $\lambda = 0$, this coincides with AIME in R^2.

Projective Completeness (Feature Contribution to Conceptual Basis).

$$R^2_{\mathrm{Proj}} \;:=\; 1 - \frac{\sum_{i=1}^n \left\| \ell_i - U_\gamma v_i \right\|_2^2}{\sum_{i=1}^n \|\ell_i\|_2^2} \;=\; 1 - \frac{\sum_{i=1}^n \left\| \ell_i - \Pi_{\mathcal{C}}(\ell_i) \right\|_2^2}{\sum_{i=1}^n \|\ell_i\|_2^2}, \tag{9}$$

($\Pi_{\mathcal{C}}$ is the orthogonal projection of $\mathrm{span}(U_\gamma)$); Because U_γ is obtained via least squares, $U_\gamma v_i$ typically coincides with the optimal projection.

3.4 Numerical Implementation and Computational Complexity Key Points

The operator $W_\lambda = A_\lambda^\dagger = XZ^\top(ZZ^\top + \lambda I_k)^{-1}$ is obtained by solving a symmetric positive-definite system of size $k \times k$, which can be computed quickly and stably. Local description requires only matrix-vector multiplication $W_\lambda z_i$ and element-wise multiplication per column, taking $\mathcal{O}(dk)$. Conceptually, U_γ is a $q \times q$ system, which is particularly lightweight when $q \ll d$. In terms of implementation, (i) column centering and normalization of Z; (ii) scale normalization of λ (e.g., $\lambda = \alpha, \mathrm{tr}(ZZ^t op)/k$); and (iii) SPD solvers such as Cholesky (falling back to pinv upon failure) are effective. Because it does not rely on the model's internal gradients or parameters, the method is *completely model-independent*.

3.5 Setting the Regularization Coefficient λ in RidgeAIME

Background and Role. AIME/RidgeAIME learns a linear operator that performs an "approximate inverse mapping" from the model outputs (or downstream representations) to the inputs. The notation herein represents each sample as a column vector, such that $X \in \mathbb{R}^{d\times n}$ denotes the inputs and $Z \in \mathbb{R}^{k\times n}$ denotes the downstream representations, where n is the number of samples. The approximate inverse operator of "$Z \to X$" is defined as

$$A_\lambda \in \mathbb{R}^{d\times k}$$

. RidgeAIME solves the following *Tikhonov-regularized* least-squares problem:

$$\min_{A\in\mathbb{R}^{d\times k}} \;\Big(\|X - AZ\|_F^2 \;+\; \lambda \|A\|_F^2 \Big), \qquad \lambda \geq 0. \tag{10}$$

with the closed-form solution

$$A_\lambda \;=\; XZ^\top (ZZ^\top + \lambda I_k)^{-1} \;=\; X\, Z_\lambda^\dagger, \qquad Z_\lambda^\dagger := Z^\top (ZZ^\top + \lambda I_k)^{-1}, \tag{11}$$

where I_k is the $k \times k$ identity matrix. As $\lambda \to 0$, $A_\lambda \to XZ^\top(ZZ^\top)^{-1}$ (when Z is full rank), matching the unregularized AIME. The readout for RidgeAIME is identical to that in the previous section. GFI corresponds to the column vectors of A_λ, and the local feature importance (LFI) for sample i is

$$\ell_i \;=\; (A_\lambda z_i) \,\circ\, x_i \;\in\; \mathbb{R}^d \tag{12}$$

where z_i and x_i denote the ith columns of Z and X, respectively, and $\circ$ denotes the element-wise product. Regularization improves the condition number of $(ZZ^\top)$, enhancing numerical stability and mitigating the effects of outliers and multicollinearity.

Policy for λ in this implementation (fixed value). In the submitted program, *reproducibility and computational efficiency were prioritized.* For all datasets (Adult, German Credit, COMPAS),

$$\lambda = 10^{-3}$$

The same order of magnitude is used even when combined with the conceptual linear mapping setting. All features are normalized to the scale $[0,1]$ (min-max for continuous variables and one-hot for categorical features). Under this scale, $\lambda = 10^{-3}$ (i) keeps the condition number of $ZZ^\top + \lambda I_k$ within a safe range, (ii) does not degrade BB completeness (R^2_{BB}) or projective completeness (R^2_{Proj}), and (iii) suppresses the bootstrap-induced variance. This practical trade-off confirms it as a stable choice *Preliminary verification* (Grid search is not performed; and a fixed value is used).

Notes on Numerical Implementation. equation 11 requires solving only a single $k \times k$ symmetric positive-definite system, which can be computed quickly and stably using Cholesky decomposition (falling back to pseudoinverse via SVD if it fails). Furthermore, column centering and scaling of Z, together with alignment of the one-hot basis, improve the estimation of A_λ and the stability of ell_i.

4 Experiments

The effectiveness of Concept-Ridge AIME was evaluated on tabular datasets (Adult/German Credit/COMPAS) using four criteria: (i) black-box reproducibility completeness (**BB Completeness**: R^2 when externally fitting the black-box logit model using only the concept matrix C); (ii) projection completeness of local contributions onto the conceptual space (**Projection Completeness**: norm ratio $\|\Pi_{\mathcal{C}}\ell\|_2/\|\ell\|_2$ when projecting the AIME-derived local feature contribution vector ℓ onto the concept basis); (iii) stability of the metric (**95% CI**: estimated from 200 bootstrap samples); and (iv) computational efficiency (**Latency**: average inference time per instance, including warm-up).

Experimental Setup

For each dataset, missing values were imputed using the median for numeric features and the mode for categorical features. Numeric values were normalized to $[0,1]$, and categorical features were one-hot encoded. Training and evaluation followed a fixed stratified 8:2 split. The black-box model was LightGBM (learning rate = 0.05, number of trees = 300, other parameters set to defaults). Concepts were generated by first extracting the top features and one-hot indicators from AIME's GFI, then feeding them into an LLM context to generate AND/OR rules (thresholds in $[0,1]$; one-hot indicators expressed as `==1`). Rules were adopted after syntactic validation (feature-name matching and exclusion of zero or positive cases). If fewer than six rules were generated, backup concepts based on quantile points were automatically added to ensure concept set coverage. In COMPAS, race and gender were excluded from the predictive concept set and treated separately as auxiliary concepts for auditing. Comparison methods include ConceptSHAP (concept Shapley), CBM (two-stage logistic), TCAV (finite difference approximation), Concept Space SHAP (Ridge approximation + KernelExplainer), Feature SHAP (Tree/Kernel), and LIME. All of these models can be deployed and managed via a common model-agnostic API. Finally, the inverse operator for AIME/ConceptAIME was computed only once, and subsequent outputs were obtained solely through matrix-vector operations, yielding inference speeds several orders of magnitude faster.

Results and Discussion

Table 1 reports *BB completeness (R^2) + projection completeness + 95% CI + inference time (ms/instance)* across the three datasets. For Adult, BB completeness was high at 0.725222 (95% CI: 0.723265–**0.727076**), indicating that even with only six concepts, the black-box logit could be well reproduced externally. Projection completeness was also high at **0.851456** (**0.834869–0.865143**), indicating that the local contribution ℓ adheres closely to the conceptual basis with high precision and minimal information loss. German Credit exhibited a BB completeness of **0.339209** (**0.324459–0.352034**), whereas projection completeness was stably high at **0.828336** (**0.822406–0.834705**). Therefore, even with only hard binary rules, local explanations can be sufficiently expressed in the conceptual space. However, external reproducibility of black-box logits could be improved by expanding the number of concepts or using softened (probabilistic) outputs. COMPAS showed a BB completeness of **0.221209** (**0.216630–0.225862**), but projection completeness was the highest at **0.901333** (**0.893052–0.908658**), demonstrating that local contributions can be compressed extremely efficiently into conceptual coordinates. Inference time for ConceptAIME remained around **0.013–0.014 ms/instance** for all datasets, which is significantly faster than that of ConceptSHAP (**60.6–109.7 ms**), Concept Space SHAP (**54.9–133.9 ms**), LIME (**177.5–323.7 ms**), and TCAV (**2.55–2.64 s** on Adult/German Credit). CBM was also fast (1–1.6 ms), but ConceptAIME is fundamentally different in that it can *present both local and global information simultaneously without modifying the trained black box.* Figures 1, 2, and 3 visually compare the rankings of ConceptAIME, ConceptSHAP, CBM, TCAV, Concept Space SHAP, and Feature SHAP across each dataset. They highlight domain-valid, high-level concepts, such as education level, marital status, and occupation (Adult); duration, credit amount, and age (German Credit); and criminal record, juvenile delinquency, and prosecution rate (COMPAS).

Table 1: Completeness (BB+Projection), 95% CI, and inference time (ms/item)

Dataset	#Feat	#Concepts	BB completeness R^2			Projection completeness			Latency (ms/instance)						
			Mean	CI_{low}	CI_{high}	Mean	CI_{low}	CI_{high}	ConceptAIME	Concept					
SHAP	CBM	TCAV	C-space												
SHAP	SHAP	LIME													
Adult	105	6	0.725222	0.723265	0.727076	0.851456	0.834869	0.865143	0.013388	109.711768	1.073232	2550.506928	133.936217	5.695561	323.665807
German Credit	61	7	0.339209	0.324459	0.352034	0.828336	0.822406	0.834705	0.013841	108.194204	1.247247	2635.219289	98.577847	4.521723	177.510727
COMPAS	17	7	0.221209	0.216630	0.225862	0.901333	0.893052	0.908658	0.013638	60.640338	1.577662	—	54.897113	3.372619	187.923336

Overall, ConceptAIME demonstrated superiority in three aspects: (1) high projection completeness with minimal information loss in local explanations; (2) BB completeness that can be monotonically improved through concept count, AND rules, and softening, achieving $R^2 \approx 0.73$ even with few concepts in Adult; and (3) *inference that is orders of magnitude faster*, making it suitable for interactive visualization and auditing. This demonstrates a comprehensive advantage in three key areas. In particular, for datasets such as German Credit and COMPAS where strong nonlinear interactions are suspected, the high projection metric confirms that "explainability in conceptual coordinates" is maintained. Meanwhile, the potential for improving BB R^2 will be further investigated through supplementary soft conceptualization and automatic concept expansion.

5 Conclusion

This study proposed Concept-RidgeAIME, which inherits the inverse problem perspective of AIME (Nakanishi, 2023), enhances numerical stability and reproducibility through Tikhonov (Ridge) regularization, and further extends the framework to the conceptual space. The method consists of two stages: an approximate inverse operator W_λ for output-to-input mapping and an approximate inverse operator U_γ for concept-to-input mapping. This design uniquely enables the consistent extraction of *global*, *local*, and *concept-level* contributions within a single linear algebra representation. While preserving the advantages of AIME (model independence, no gradient requirement, and millisecond-level inference after batch precomputation), it additionally enables decision-making to be described using higher-level conceptual units designed by a user or an LLM, thereby achieving interpretable explanations aligned with expert vocabulary.

Comparatively, ConceptSHAP excels in *global importance* based on game-theoretic axioms but faces challenges in terms of computational load from subset evaluation and the design of local explanations. TCAV presents sensitivity-based class-specific concept relevance but requires access to internal representations and gradients. CBM provides *interventional* internalized explanations but requires concept supervision during training, making it difficult to retrofit onto existing black-box models. Concept Space SHAP is an *operational* approach that applies SHAP to concept representations, inheriting the underlying theory but retaining the computational burden of sampling approximations. By contrast, Concept-RidgeAIME achieves *(i) model and gradient independence*; *(ii) simultaneous presentation of global, local, and conceptual contributions*; *(iii) quantification of fidelity using two completeness metrics (BB and projection)*; and *(iv) numerical stability and reproducibility via regularization.* This fills practical gaps in black-box explanations for tabular data, combining low computational load with local conceptual contributions.

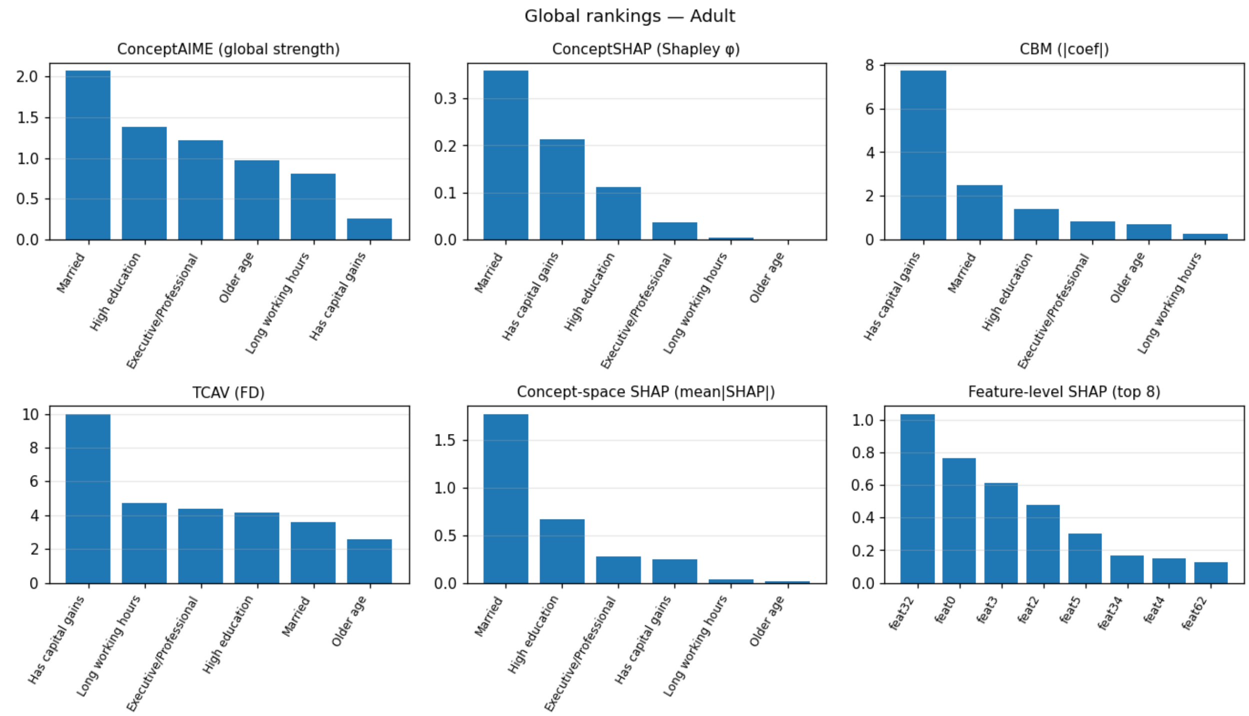

Figure 1: Global concept rankings (Adult; six panels)

Furthermore, this study prioritizes the operational feasibility of concept design by automatically synthesizing concept candidates with LLM assistance to reduce manual effort. By strictly constraining the rule format (only known feature names, thresholds in $[0, 1]$, one-hot equality, JSON-only) and performing syntax checks and zero/positive exclusion at the program level, this study enabled a workflow of *automatic generation → verification → adoption* to be reproduced within a single notebook, enhancing both the reproducibility and portability of the results.

However, this study has several limitations. First, *dependency on concept set quality*: inappropriate thresholds or extremely sparse rules reduce BB and projection completeness and introduce bias into local explanations. Second, *linear approximation limitations*: because the method relies on linear approximation, strong nonlinear interactions may be under- or over-estimated. Third, *scaling assumptions*: dependence on $[0, 1]$ normalization and one-hot encoding means that stability can be compromised by improper preprocessing. Fourth, *LLM generation validation*: verification and review are essential to prevent the introduction of unknown features or complex logic (e.g., OR or negation).

Reproducibility Statement

To ensure reproducibility, a comprehensive description of the proposed method is provided, including mathematical formulations (Sections 2–4), hyperparameter settings (Section 4), and implementation notes (Appendix A). All datasets used (Adult, German Credit, and COMPAS) are publicly available. All implementation details and source code, including a single executable notebook `Auto_Concept_AIME_ICLR2026.ipynb` that automatically downloads the datasets, trains the models, performs concept generation, and reproduces all figures and tables, are provided as anonymized supplementary material. Bootstrap confidence intervals are also reported for all key metrics to quantify robustness. These measures collectively ensure that independent researchers can reproduce both the methodology and results presented in this paper in a fully self-contained manner.

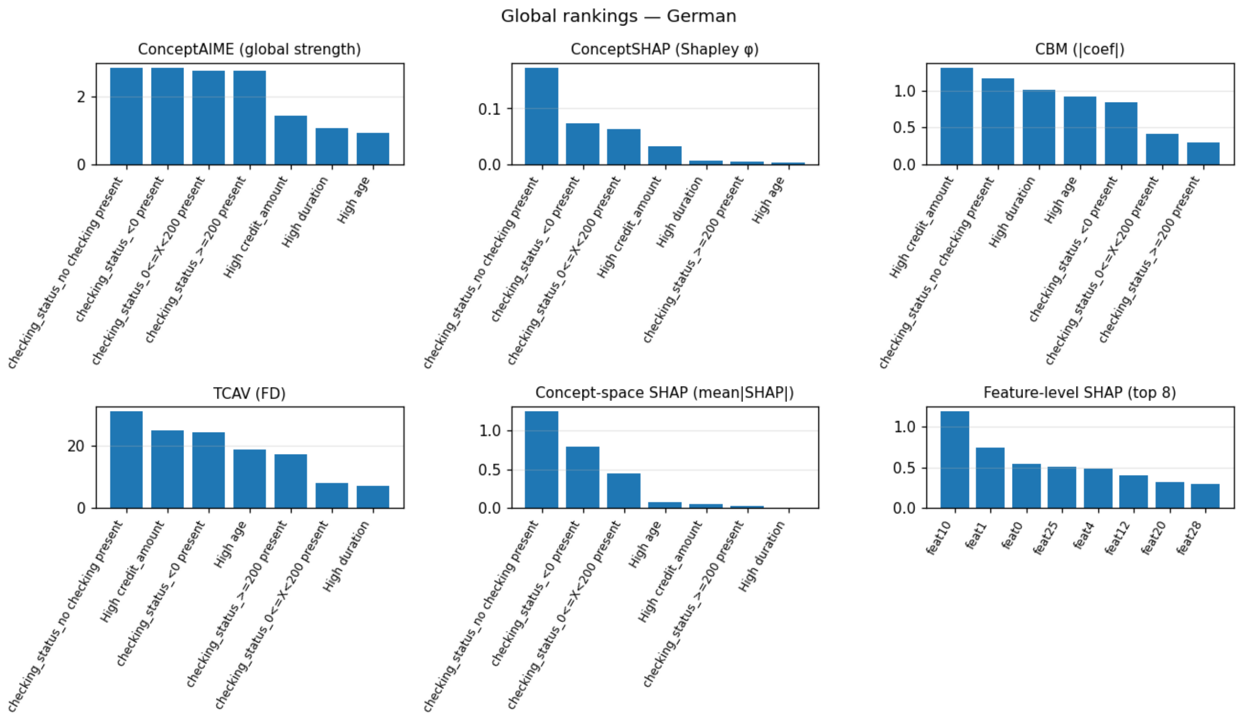

Figure 2: Global concept rankings (German; six panels)

Broader Impact Statement

This work aims to enhance the transparency, fairness, and trustworthiness of machine learning systems by providing human-understandable, concept-based explanations at both the global and local levels. Because Concept-RidgeAIME operates without accessing model internals or gradients, it can be applied to a wide range of real-world black-box systems in domains such as healthcare, finance, and public administration, where interpretability and accountability are essential.

Potential negative impacts include the risk of misunderstanding or overtrusting explanations that appear intuitive but may not fully reflect model behavior, and the possibility of biased or misleading concept definitions when LLMs are used for concept generation. To mitigate these risks, the method enforces strict syntactic validation, excludes zero/positive trivial rules, and supports deterministic, auditable concept generation. This study does not automate decision-making itself but instead provides tools to support human auditing and understanding of AI systems.

Promoting model transparency and concept-level interpretability while maintaining awareness of potential biases and misuse can contribute to the responsible and ethical deployment of AI technologies.

Author Contributions

The sole author was responsible for the conception of the study, development of the methodology, implementation of the algorithms, the experimental design and analysis, and for the writing and revision of the manuscript in its entirety.

Acknowledgments

The author thanks Editage [www.editage.com] for English language editing.

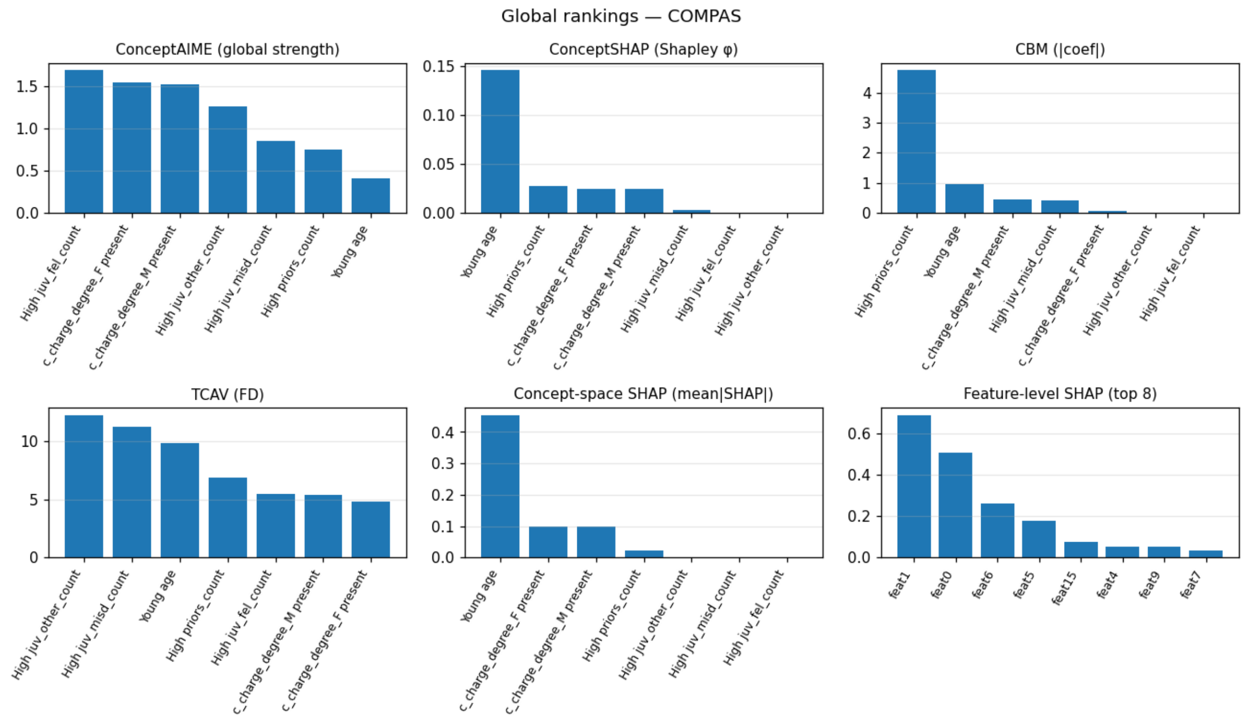

Figure 3: Global concept rankings (COMPAS; six panels)

Use of Generative AI

The proposed method incorporates an LLM, utilizing GPT4-o mini as its API. Additionally, GPT-5 Pro, DeepL, and Paperpal were employed to assist with implementing the method and refining the text of this paper.

The author takes full responsibility for the content of this paper.

References

Abubakar Abid, Mert Yüksekgönül, and James Zou. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research (PMLR)*, pp. 66–88, 2022. URL `https://proceedings.mlr.press/v162/abid22a.html`.

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 7786–7795, 2018.

Nihala Basheer, Shareeful Islam, Spyridon Papastergiou, Haralambos Mouratidis, and Nikolaos Papagiannopoulos. Vulnerability patch prediction using llm based bert model with trustworthy ai practice for cyber security enhancement. In Ilias Maglogiannis, Lazaros Iliadis, Andreas Andreou, and Antonios Papaleonidas (eds.), *Artificial Intelligence Applications and Innovations*, pp. 431–445, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-96235-6.

Michaela Benk, Léane Wettstein, Nadine Schlicker, Florian von Wangenheim, and Nicolas Scharowski. Bridging the knowledge gap: Understanding user expectations for trustworthy llm standards. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27197–27205, Apr. 2025. doi: 10.1609/aaai.v39i26.34928. URL `https://ojs.aaai.org/index.php/AAAI/article/view/34928`.

Ahsan Bilal, David Ebert, and Beiyu Lin. Llms for explainable ai: A comprehensive survey, 2025. URL `https://arxiv.org/abs/2504.00125`.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K. Su. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019. URL `https://proceedings.neurips.cc/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf`.

Zhi Chen, Yijun Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2:772–782, 2020. doi: 10.1038/s42256-020-00265-z.

Jonathan Crabbé and Mihaela van der Schaar. Concept activation regions: A generalized framework for concept-based explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/11a7f429d75f9f8c6e9c630aeb6524b5-Paper-Conference.pdf`.

Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8730–8738, 2018. doi: 10.1109/CVPR.2018.00910.

Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL `https://papers.nips.cc/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html`.

Vidhya Kamakshi, Uday Gupta, and Narayanan C. Krishnan. Pace: Posthoc architecture-agnostic concept extractor for explaining cnns. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2021. URL `https://arxiv.org/abs/2108.13828`.

Zahra Shams Khoozani, Aznul Qalid Md Sabri, Woo Chaw Seng, Manjeevan Seera, and Kah Yee Eg. Navigating the landscape of concept-supported xai: Challenges, innovations, and future directions. *Multimedia Tools and Applications*, 83:67147–67197, 2024. doi: 10.1007/s11042-023-17666-y. URL `https://link.springer.com/article/10.1007/s11042-023-17666-y`.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/kim18d.html`.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/koh20a.html`.

Ashish Kumar, Karan Sehgal, Prerna Garg, Vidhya Kamakshi, and Narayanan C. Krishnan. Mace: Model agnostic concept extractor for explaining image classification networks. *IEEE Transactions on Artificial Intelligence*, 2(6):574–583, 2021. doi: 10.1109/TAI.2021.3111138.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Fuseini Mumuni and Alhassan Mumuni. Explainable artificial intelligence (xai): from inherent explainability to large language models, 2025. URL `https://arxiv.org/abs/2501.09967`.

Takafumi Nakanishi. Approximate inverse model explanations (aime): Unveiling local and global insights in machine learning models. *IEEE Access*, 11:101020–101044, 2023. doi: 10.1109/ACCESS.2023.3314336.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL `https://doi.org/10.1145/2939672.2939778`.

Wenbo Wu, Yaowei Su, Xizhou Chen, Shan Zhao, Irwin King, Michael R. Lyu, and Yu-Wing Tai. Towards global explanations of convolutional neural networks with concept attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8652–8661, 2020.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20554–20565. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/ecb287ff763c169694f682af52c1f309-Paper.pdf`.

Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pan, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, Bo An, and Qingsong Wen. A survey on trustworthy llm agents: Threats and countermeasures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 6216–6226, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3736561. URL `https://doi.org/10.1145/3711896.3736561`.

Mert Yüksekgönül, Yao Zhang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022. URL `https://arxiv.org/abs/2205.15480`.

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 11682–11690, 2021. URL `https://ojs.aaai.org/index.php/AAAI/article/view/17389`.

A Reproduction Steps and Environment.

Implemented and tested on Google Colab Pro+ (Python 3.12.11). The experiments used LightGBM 4.5.0, scikit-learn 1.5.2, SHAP 0.45.1, LIME 0.2.0.1, and the AIME package ('aime-xai 0.1' `https://github.com/ntakafumi/aime`). Concept candidate generation using the LLM was performed when an OpenAI API key was available.

B Concept Rules and Prompts (Generation, Sanitization, Adoption Set)

This study first generated concept candidates from LLMs in a *rule format* (numeric features expressed as normalized scale thresholds, categorical features represented as one-hot encoded equality) based on AIME's GFI. These candidates underwent syntactic validation (feature name matching, domain verification $[0, 1]$), exclusion of zero-positive rules, and were finally adopted. In this run (Patch–H), candidates were adopted directly across all datasets without requiring backup concepts (quantile scanning). Table 2 summarizes the candidate, adopted, and disqualification counts.

The final concept sets (concept names and corresponding rules) after selection are shown in Tables 3–5 for each dataset. Numeric attribute thresholds correspond to quantiles after normalization to the range ([0,1]), and one-hot encoding is represented by `==1` to indicate presence.

C Local Explanation (Additional Examples and Summary)

For each dataset, five examples of the concept contribution vector $v = F^{\top}\ell$ and the ratio ratio were calculated using ConceptAIME. The most boosting (Top+) concept and the most suppressing (Top−) concept are

Table 2: Concept candidate sanitation statistics (Candidates, Adopted, Rejected = Syntactic Inconsistency/Zero-Positive, etc.)

Dataset	Candidates	Adopted	Rejected
Adult	6	6	0
German Credit	7	7	0
COMPAS	7	7	0

Table 3: Final adopted concept set (Adult)

No.	Concept Name	Rule (Normalized Scale/one-hot)
1	High education	education-num ≥ 0.75
2	Long working hours	hours-per-week *ge* 0.65
3	Executive/Professional	occupation_Exec-managerial == 1
4	Married	marital-status_Married-civ-spouse == 1
5	Older age	age ≥ 0.60
6	Has capital gains	capital-gain ≥ 0.05

summarized in Tables 6 to 8. For Adult, *Married* consistently appeared as the Top+ concept consistently (+2.08 to +2.73; ratio 0.33 to 0.37), strongly supporting the model's high-income classification. Top− showed few significant negative contributions in this population, with *Has capital gains* appearing as the smallest positive contribution. In German Credit, the presence concept *checking_status* often appeared in Top+, while *High credit_amount* and *High duration* frequently showed negative contributions (Top−).In COMPAS, the degree of guilt (*c_charge_degree_M/F*) appeared in Top+, while a high number of juvenile prior offenses and young age appeared on the suppressing side (Top−) (Audit concepts are excluded from prediction and presented separately).

D Hard (Binary) vs. Soft (Continuous) Concept Ablation

This run used only the hard (binary) concepts. Table 9 restates the completeness (BB R^2 and Projection) for each dataset. Adult showed BB $R^2 \neq 0.725$ (95% CI: 0.723–0.727), and projection was also high at 0.851 (0.835–0.865). German Credit showed moderate BB R^2 at 0.339 but high projection at 0.828 (0.822–0.835), indicating sufficient conceptual space coverage for local contributions. COMPAS had a modest BB R^2 of 0.221, but its projection was the highest at 0.901 (0.893–0.909). The soft concept version is an extension where a learner (for example, logistic) assigns concept probabilities $\in (0, 1)$ from positive and negative examples of hard rules, replacing the binary vector with continuous scores. This is expected to increase BB R^2, while projection metrics often do not decrease significantly (implementation requires only linear algebra substitutions, with unchanged computational complexity). The soft version is planned for implementation and inclusion as an additional experiment in the Supplement to this paper.

E Quantifying Robustness Against Random Number Generators and Partitions (CI Width) and Inter-Method Rank Correlation

In this run, the 95% CI for each metric was calculated using 200 test-side bootstrap samples. While a more rigorous double bootstrap involving random number seed × partition combinations is planned for future supplementary experiments, the metrics were compared here using CI width as a proxy (Table 10). For Adult and COMPAS, both BB and projection exhibited narrow CI widths (≈ 0.004 and ≈ 0.009, respectively). For German Credit, BB showed a slightly wider width (≈ 0.028), whereas projection had a narrower width (≈ 0.012). Furthermore, the consistency in global concept rankings was evaluated using Spearman's rank correlations between ConceptAIME and other methods (ConceptSHAP/CBM/TCAV/Concept Space SHAP) (Table 11). German Credit showed high correlations with ConceptSHAP and Concept Space SHAP ($\rho \approx 0.893, 0.750$) and also exhibited high consistency with Concept Space SHAP in the Adult dataset ($\rho \approx 0.771$).

Table 4: Final adopted concept set (German Credit)

No.	Concept Name	Rule (Normalized Scale/one-hot)
1	High duration	duration $\geq$ 0.7
2	High credit_amount	credit_amount $\geq$ 0.7
3	High age	age $\geq$ 0.7
4	Checking_status_0<=X<200 present	Checking_status_0<=X<200 == 1
5	Checking_status_<0 present	Checking_status_<0 == 1
6	Checking_status_>=200 present	Checking_status_>=200 == 1
7	checking_status_no checking present	checking_status_no checking == 1

Table 5: Final adopted concept set (COMPAS)

No.	Concept Name	Rule (Normalized Scale/one-hot)
1	High priors_count	priors_count $\geq$ 0.6
2	High juv_fel_count	juv_fel_count $\geq$ 0.6
3	High juv_misd_count	juv_misd_count $\geq$ 0.6
4	High juv_other_count	juv_other_count $\geq$ 0.6
5	Young age	age $\leq$ 0.30
6	c_charge_degree_F present	c_charge_degree_F == 1
7	c_charge_degree_M present	c_charge_degree_M == 1

By contrast, COMPAS showed low correlations, likely due to differences in audit concept separation and rule sets (e.g., strong contribution from one-hot-encoded prosecution degrees).

F LLM Prompts

A standardized prompt was used to explicitly specify GFI top feature names and one-hot names; to apply thresholding using normalized numerical values; to reference one-hot values with "==1"; to allow up to two literal ANDs; to prohibit unknown feature names; and to reject candidates with zero positive counts (see the Methods section in the main text). Zero positive or unknown features did not occur in this run, and all the candidates were selected. Using AIME's top GFI features as clues, concept candidates were generated for the LLM using a *rule grammar* consisting solely of *normalized numerical features with* $[0, 1]$ *thresholds* and *categories represented by one-hot equality.* Each rule is represented as a *Conjunctive Normal Form approximation* with *AND at most three literals* or *OR at most two clauses*, and syntax other than `feature >= t`, `feature <= t`, or `feature == 1` (one-hot) is not permitted. Output is restricted to a *JSON array* (each element being `{"name": "...", ‘rule’: "..."}`), with no additional keys or explanatory text allowed. The program performs *syntax validation (feature name matching, domain, logical form)* and *exclusion of zero-positive rules*, and evaluates completeness (BB/projection), stability (CI), and efficiency (ms) only for valid concepts.

F.1 Minimal Template

```
System:
You are a careful scientist. Produce concept rules over preprocessed tabular features.
Use literals ‘feature >= t‘, ‘feature <= t‘, ‘feature == 1‘ (for one-hot).
Max 3 literals per AND clause, max 2 OR clauses.
Return EXACTLY {K} items in JSON: [{"name": "...", ‘rule’: "..."}].

User:
Dataset: {DATASET_NAME}
Top features (AIME GFI): {TOP_20_FEATURES}
Features (first 50): {FIRST_50_FEATURES} ... (+{REMAINING_COUNT} more)
Constraints:
- Use only provided feature names (exact match; case-sensitive).
- Thresholds in [0,1] for numeric features (min-max normalized).
- Do NOT invent features (e.g., "pos"/"neg"/class names).
- JSON only (no comments or extra keys).
```

Table 6: Additional local explanation examples (Adult; top positive/negative concept contributions and ratios)

Idx	Top+ Concept	Contribution	Ratio	Top− Concept	Contribution	Ratio
2088	Married	2.0823	0.330	Has capital gains	0.0763	0.012
8889	Married	1.9320	0.331	Has capital gains	0.0679	0.012
6607	Married	2.3992	0.349	Has capital gains	0.1356	0.020
7482	Married	2.6254	0.370	Has capital gains	0.0988	0.014
8034	Married	2.7336	0.365	Has capital gains	0.1038	0.014

Table 7: Additional local explanation example (German Credit; contribution and ratio of top positive/negative concepts)

Idx	Top+Concept	Contribution	Ratio	Top−Concept	Contribution	Ratio
1	checking_status_0<=X<200 present	3.6670	0.256	High credit_amount	-0.3427	0.024
5	checking_status_no checking present	4.5063	0.241	High credit_amount	-0.9859	0.053
56	checking_status_no checking present	5.6306	0.260	High credit_amount	-0.3100	0.014
32	checking_status_no checking present	5.3388	0.279	High duration	-0.1058	0.006
125	checking_status_0<=X<200 present	4.0486	0.242	High credit_amount	-0.6866	0.041

F.2 Rule examples (Grammar specification)

```
{"name": "High education",
"rule": "education-num >= 0.75"}

{"name": "Executive/Professional",
'rule': "occupation_Exec-managerial == 1"}

{"name": "Long hours OR Married",
'rule': "(hours-per-week >= 0.65 AND capital-gain >= 0.05)
OR (marital-status_Married-civ-spouse == 1)"}
```

F.3 Notes.

The model utilized the OpenAI API (default: gpt-4o-mini). The number of generated concepts was always `{K}`, but *the number of concepts selected after syntax verification and zero-positive exclusion* represents the final concept count, which matches the `#concepts=...` in the experiment logs.

Table 8: Additional local explanation examples (COMPAS; top positive/negative concept contributions and ratios)

Idx	Top+Concept	Contribution	Ratio	Top−Concept	Contribution	Ratio
470	c_charge_degree_M present	1.7945	0.480	High juv_other_count	-0.3227	0.086
1427	c_charge_degree_M present	1.2415	0.694	High juv_misd_count	-0.1035	0.058
1067	c_charge_degree_F present	1.9860	0.412	High juv_fel_count	-0.5772	0.120
87	c_charge_degree_M present	0.9445	0.276	High juv_fel_count	-0.9628	0.281
114	c_charge_degree_M present	1.1859	0.460	Young age	-0.0121	0.005

Table 9: Completeness based on hard (binary) concepts (this run)

Dataset	BB R^2 Mean	95% CI	Projection Mean	95% CI
Adult	0.725222	[0.723265, 0.727076]	0.851456	[0.834869, 0.865143]
German	0.339209	[0.324459, 0.352034]	0.828336	[0.822406, 0.834705]
COMPAS	0.221209	[0.216630, 0.225862]	0.901333	[0.893052, 0.908658]

Table 10: 95% Confidence interval width for metrics (Single Bootstrap 200 Times)

Dataset	BB R^2 CI Width	Projection CI Width
Adult	0.003811	0.030274
German	0.027575	0.012299
COMPAS	0.009232	0.015605

Table 11: Global concept rank correlation (Spearman's ρ; Comparison of ConceptAIME with other methods)

Dataset	Comparison Method	ρ
Adult	ConceptSHAP	0.371
Adult	CBM	0.143
Adult	TCAV	-0.657
Adult	C-space SHAP	0.771
German	ConceptSHAP	0.893
German	CBM	-0.107
German	TCAV	0.464
German	C-space SHAP	0.750
COMPAS	ConceptSHAP	-0.571
COMPAS	CBM	-0.786
COMPAS	TCAV	-0.571
COMPAS	C-space SHAP	-0.296