Problem of Overtrust in XAI Tools for Educational Data: Analysis of LLM Interpretations

Semyon Bosonogov[1,*], Alena Suvorova[1]

[1]HSE Universiry, Saint-Petersburg, 16 Soyuza Pechatnikov Street, St Petersburg, 190121

Abstract

The integration of Large Language Models (LLMs) into explainability workflows has been proposed to democratize access to machine learning interpretability by converting complex SHAP visualizations into natural language explanations. However, prior research documents systematic rationalization biases in how humans interpret XAI tools. This study tests whether LLMs reproduce these biases by replicating Kaur et al. (2020) in an educational data context. We trained a gradient boosting model on the Open University Learning Analytics dataset and artificially introduced bias by replacing studied credits with 125 for 10% of high-scoring students. Five LLMs (DeepSeek-R1, Qwen3-max, Sonar, Gemini-3, ChatGPT-5.2) were prompted to explain the anomalous concentration of points at 125 credits. All models rationalized the anomaly through plausible institutional narratives rather than flagging it as a data quality issue. To examine whether rationalization is context-dependent, we conducted a comparative analysis of LLM responses to the legitimate 120-credit institutional value (standard full-time load). Notably, LLMs similarly rationalized the natural 120-credit pattern, but with greater convergence across models and stronger alignment with institutional frameworks. Although some models acknowledged data error as possible, none prioritized it as primary explanation in either condition. These findings suggest that LLMs inherit and amplify overreliance and rationalization biases, and that institutional legitimacy of data patterns actively suppresses critical evaluation. We conclude that LLM-generated explanations of XAI outputs require careful validation and cannot substitute for domain expertise, particularly in high-stakes educational contexts where institutional knowledge may reinforce uncritical acceptance of AI-generated narratives.

Keywords

Explainable AI, Large Language Models, Machine Learning, Interpretability, Learning Analytics, Educational Data Mining

CEUR-WS.org/Vol-4328/paper_3.pdf

1. Introduction

The rapid proliferation of machine learning (ML) models across diverse fields, including education, has created unprecedented opportunities for data-driven decision-making [1]. However, this expansion has simultaneously exposed a critical challenge: the opacity of complex ML models, particularly deep neural networks and ensemble methods, makes it difficult for stakeholders to understand, trust, or validate model predictions [2]. This "black box" problem is especially acute in educational contexts, where decisions informed by ML predictions can directly impact student opportunities, institutional resource allocation, and educational equity [3].

In response to this challenge, the field of Explainable Artificial Intelligence (XAI) has emerged as a crucial research area dedicated to making AI systems transparent, interpretable, and trustworthy [1]. Within XAI, post-hoc explanation techniques—methods that provide interpretability after model training—have gained particular prominence. SHAP (SHapley Additive exPlanations), a game-theoretic approach grounded in Shapley values from cooperative game theory, has become one of the most widely adopted post-hoc explanation methods [4]. SHAP generates both local explanations (for individual predictions) and global explanations (for overall model behavior), typically presented through various visualizations such as summary plots, dependence plots, and force plots [5].

XAI-Ed 2026: Demystifying AI in Education and Learning Analytics through Explainability, Agency, and Transparency Workshop (XAI-Ed@LAK26), 27 April, 2026, Bergen, Norway

sdbosonogov@edu.hse.ru (S. Bosonogov); asuvorova@hse.ru (A. Suvorova)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-13

1.1. The Current State of XAI and Interpretability in Machine Learning

The field of machine learning interpretability has evolved along two parallel trajectories: intrinsic interpretability (models designed to be inherently interpretable, such as generalized additive models or decision trees) and post-hoc explainability (techniques applied to complex "black box" models to explain their decisions) [6]. Advances in both areas have been substantial. SHAP, in particular, has been successfully applied across numerous domains—from healthcare and criminal justice to finance and scientific research—offering practitioners a principled, theoretically grounded method for model explanation [7].

However, the widespread adoption and deployment of XAI tools has created new challenges. Despite the availability of interpretability methods, research indicates that practitioners often struggle to understand, correctly apply, or critically evaluate the explanations generated by these tools [8]. A seminal study by Kaur and colleagues demonstrated that data scientists frequently over-trust interpretability tools, misinterpreting visualizations even when provided with detailed tutorials. Practitioners sometimes rationalize anomalies in data as genuine patterns rather than recognizing them as data quality issues or model artifacts. This phenomenon—termed "overreliance" on ML explanations—reflects a broader tension between automation and human judgment in data-driven decision-making [9].

1.2. XAI and Interpretability in Educational Contexts

Educational Machine Learning and Learning Analytics (LA) represent a particularly sensitive domain for the application of AI technologies. Educational data mining and ML models are increasingly deployed to predict student performance, identify at-risk learners, recommend personalized interventions, and inform institutional decision-making [10]. Decisions informed by educational ML models can affect student trajectories, resource allocation, and equity outcomes [11].

Within educational contexts, the role of explainability becomes especially critical. Educational stakeholders—including educators, administrators, students, and parents—span diverse levels of technical expertise [12]. Unlike specialized technical audiences, these stakeholders require explanations that are not merely technically accurate but also meaningful, actionable, and aligned with their educational goals and values [13]. Recent research has emphasized the importance of what has been termed "learner-centered" or "human-centric" XAI in education, where explanation design is grounded in learning science principles and the cognitive needs of end users [3].

Several studies have begun to examine the application of XAI techniques specifically in educational domains. Swamy and colleagues compared five state-of-the-art explanation methods (LIME, PermutationSHAP, KernelSHAP, DiCE, and CEM) for student performance prediction, finding that different explainers produce conflicting feature importance rankings—highlighting that the choice of explanation method substantially influences interpretation [14]. Other research has explored using SHAP specifically for educational data mining tasks, demonstrating applications in predicting academic performance, identifying factors associated with student success, and supporting recommendations for intervention [15]. A growing body of work has investigated how different presentation formats and explanation techniques affect student and educator trust, satisfaction, and comprehension [16].

1.3. The Emerging Role of Large Language Models (LLMs) in Explanation

A recent and distinct development in the XAI landscape is the integration of Large Language Models (LLMs) into explanation pipelines. Several researchers have proposed using LLMs—such as GPT-4, Claude, and other generative models—to convert technical visualizations and numerical SHAP outputs into natural language explanations [17]. The motivation is intuitive: SHAP graphics, while informative, are complex visualizations that require statistical and technical literacy to interpret. Natural language explanations generated by LLMs could potentially bridge the gap between technical outputs and human understanding, making XAI tools more accessible to non-technical stakeholders [18].

Some proposed systems leverage LLMs to generate contextual, natural language narratives from SHAP graphics, offering explanations that adapt to different stakeholder knowledge levels and goals.

This approach holds promise for democratizing access to ML model interpretability and supporting more equitable educational decision-making [19].

1.4. The Problem: Uncritical Reliance on LLM-Generated Explanations

However, a critical gap exists in the research on LLM-mediated explanations for ML models. While prior work has documented the problems of human overreliance on XAI tools—rationalization of anomalies, insufficient critical evaluation, misinterpretation of visualizations [8]—there has been limited investigation into whether LLM-generated explanations of SHAP graphics reproduce, amplify, or mitigate these problems.

LLMs themselves have known limitations and failure modes: they can hallucinate plausible-sounding but false information, rationalize unexplained patterns, and present confident-sounding explanations even when the underlying facts are uncertain or incorrect [20, 21]. When an LLM is asked to interpret a SHAP graphic—a visualization representing a potentially flawed or biased ML model applied to data with quality issues—the possibility emerges that the LLM might generate explanations that sound reasonable but lack critical scrutiny of the underlying model and data[22]. In educational contexts, where decisions affect students' lives and where practitioners often lack deep technical expertise in ML, uncritical acceptance of LLM-generated explanations could propagate or amplify biases, misinterpretations, and poor decision-making [23].

1.5. Research Motivation and Objectives

The present research investigates whether LLM-generated natural language explanations of SHAP graphics represent a solution to the problem of user misinterpretation of XAI visualizations, or whether they introduce new risks of rationalization and overreliance. We examine this question by replicating the experimental design of Kaur and colleagues [8], who studied how human analysts interpret SHAP graphics in the presence of intentional data anomalies. Rather than using human participants, we test how multiple LLMs respond to the same task.

The study is motivated by several key questions:

1. **Do LLMs reproduce human cognitive biases when interpreting SHAP graphics?** Specifically, when presented with SHAP visualizations that contain obvious data quality issues (e.g., suspicious anomalies), do LLMs rationalize these anomalies as genuine patterns, or do they recognize and flag them as potential data problems?
2. **How do LLM-generated explanations compare across different models?** Given the variability in LLM outputs, do different models produce consistent explanations, and does providing additional context (e.g., explanations of SHAP methodology) improve explanation quality?

1.6. Research Gap

This research addresses an important gap at the intersection of three critical areas: (1) explainable AI for machine learning, (2) learning analytics and educational data mining, and (3) the emerging applications of LLMs for automating and democratizing access to technical knowledge.

From a **technical perspective**, the study provides empirical evidence about the reliability of LLM-generated explanations for complex visualizations. This is important for any domain where LLMs are being integrated into explanation pipelines, but it is especially critical for high-stakes applications like education.

From an **educational ßperspective**, the research highlights both the opportunities and risks associated with using LLMs to make ML-based educational systems more interpretable. As educational institutions increasingly adopt ML-driven decision support systems—for student success prediction, personalized learning, early warning systems, and more—understanding how explanations are generated, interpreted, and acted upon becomes essential to ensuring equity and supporting good decision-making.

From an **ethical perspective**, the study contributes to broader conversations about responsible AI. If LLM-based explanation systems can rationalize or obscure problematic patterns in data and models, this represents a failure of XAI's core promise: to make AI systems more trustworthy and accountable [24]. Conversely, if LLMs can be designed to support critical evaluation of ML outputs, they may offer genuine benefits for practitioners seeking to use AI responsibly.

2. Study Description

2.1. Description of the Motivational Study Design

Kaur and colleagues [8] conducted a user-centered evaluation of interpretability tools by deliberately introducing data anomalies to assess whether practitioners could identify data quality issues using SHAP explanations. They artificially replaced the age values with 38 (the dataset mean) for 10% of high-income individuals in the Adult Income dataset, then applied SHAP analysis to visualize the relationship between age and income predictions. This manipulation created a pronounced spike at age 38 in the SHAP dependence plot.

The critical finding was that most of human participants rationalized this anomaly as a genuine economic pattern rather than recognizing it as a data quality problem, despite having access to the visualizations and explanatory information. This behavior demonstrated systematic overreliance on interpretability tools and a tendency to rationalize suspicious patterns as meaningful findings.

The experiment design of our study was build to replicate one of common scenarios when web-based general purpose LLM-chatbot used as an assistant that helps to explain unclear data plot. So we are attempting to evaluate if this additional layer of explanation can help to overcome the rationalization bias.

2.2. Data

For the model training we used the Open University Learning Analytics dataset (OULAD) [25]. It is a large-scale, anonymized collection of educational data encompassing student interactions, demographic information, and academic outcomes across seven courses (modules) at the Open University. The dataset includes student demographic characteristics (age, prior education level, socioeconomic status measured by IMD band), course enrollment information, Virtual Learning Environment (VLE) interaction logs, assessment submissions, and final scores ranging from 0 to 100.

This dataset is well-suited for educational machine learning research because it captures the complexity of distance learning contexts, including diverse student populations with varying educational backgrounds and engagement patterns.

For the present study, we selected studied credits (the total number of course credits in which a student enrolled) and demographic features such as prior education level and deprivation index. The target variable was defined as the student's final score, ranging from 0 to 100:

1. imdband: specifies the Index of Multiple Depravation band of the place where the student lived during the module-presentation
2. studiedcredits: number of credits studied
3. ageband: age group
4. genbin: binary indication of gender (0 — female, 1 — male)
5. disbin: binary indication of disability (0 — no, 1 — yes)
6. numattempts: repeated test (0 – no, 1 – yes)
7. highest education: highest level of education

We replicated experimental design using Large Language Models as the "practitioners" interpreting SHAP graphics. We artificially introduced bias in the studied credits variable by replacing values with 125 (a mean-like value) for 10% of high-scoring students. By testing whether LLMs exhibit similar

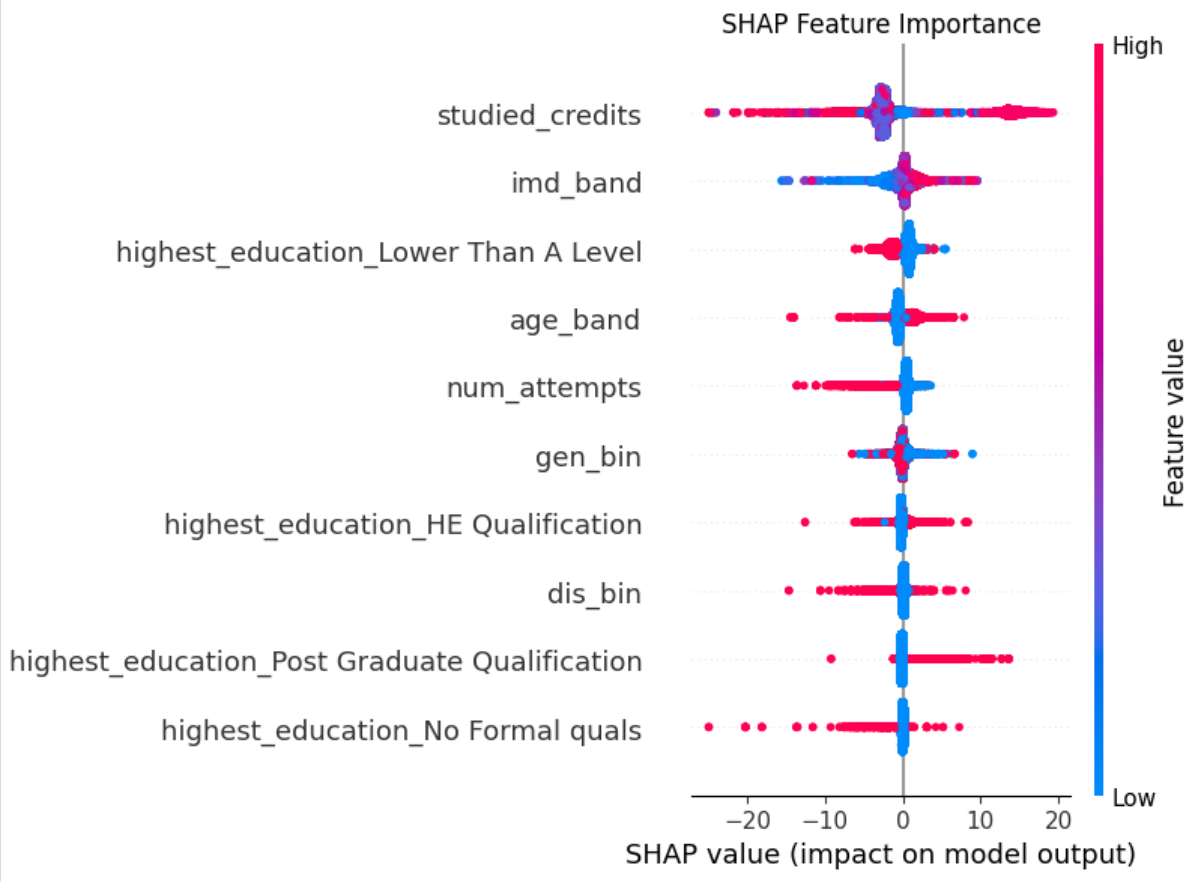


Figure 1: SHAP 125-Credit Summary Plot

rationalization tendencies when interpreting the resulting SHAP visualizations, we examine whether LLM-generated explanations inherit the overreliance biases documented in human practitioners [8].

2.3. Model

We trained several models, including Random Forest and LightGBM, and decided to settle on a gradient boosting regression model due to the better metrics and performance on the selected dataset. This model is recognized for its robustness and stability in prediction tasks. Model performance metrics on the test set included an R2 score of 0.1505, a mean absolute error (MAE) of 13.23 points, and a root mean squared error (RMSE) of 17.54 points. Cross-validation yielded an R2 score of 0.1470 ± 0.0085. These results indicate that the model explains approximately 15% of the variance in the data, with an average prediction error of 13–14 points.

2.4. SHAP analysis

For SHAP analysis, the Python shap library was employed. Several visualizations were made for interpretation, including feature importance (Figure 1) and dependence plots for studied credits, the variable subjected to manipulation (Figure 2).

Figure 1 (SHAP summary plot) summarizes the global behavior of the gradient boosting model trained on the Open University Learning Analytics dataset. The plot shows that studied credits is by far the most influential predictor of the final score, followed by contextual and demographic variables such as imd band, age band, and indicators of prior educational attainment. High values of studied credits (in pink) are generally associated with positive SHAP values and thus higher predicted scores, whereas low values (in blue) more often contribute negatively to the prediction. This pattern is consistent with

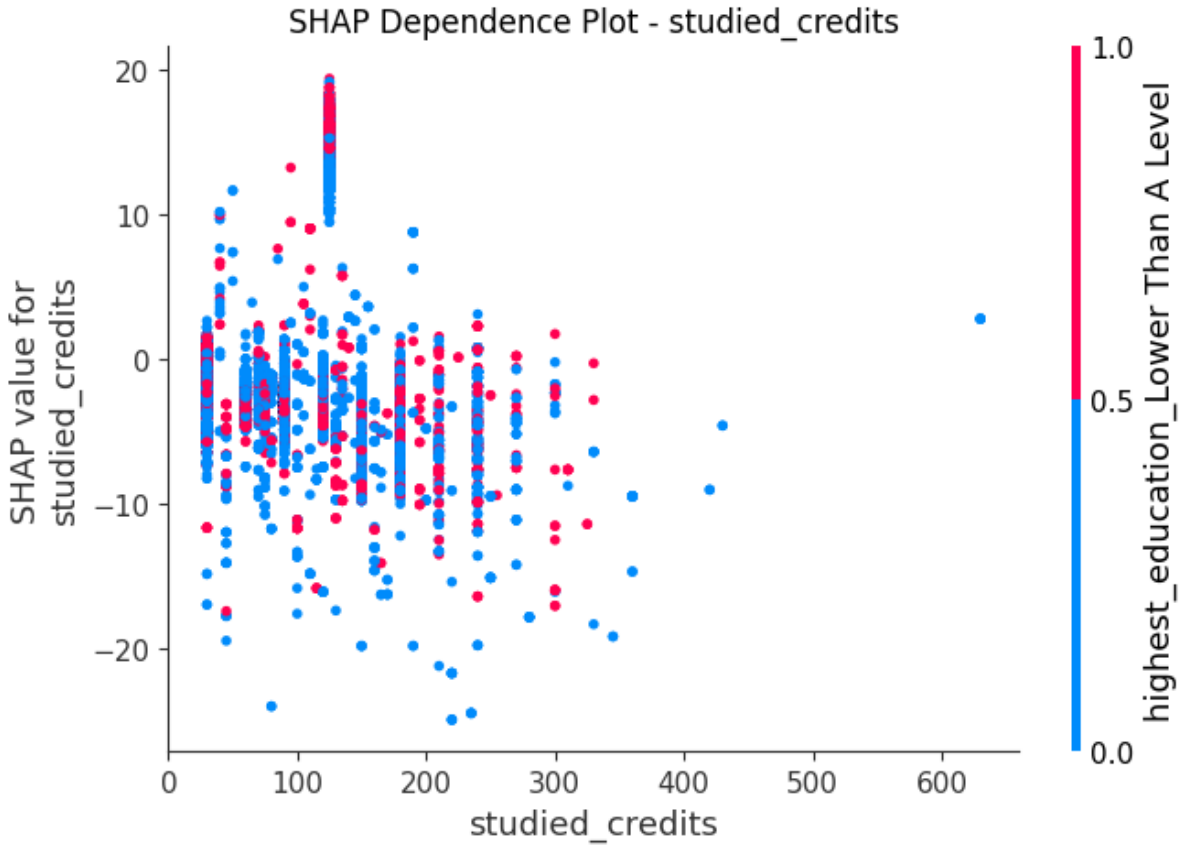


Figure 2: SHAP 125-Credit Dependence Plot

the intended semantic meaning of the feature, but it also suggests that any bias introduced into studied credits will strongly shape the model's explanations.

Figure 2 shows a more fine-grained view of how the manipulated feature affects individual predictions. Each point represents a student, with the x-axis showing their studied credits and the y-axis the corresponding SHAP value for this feature; color encodes whether the student's highest education level is lower than A level. Around 125 credits—the value used in the experimental manipulation—a vertical concentration of points with relatively high positive SHAP values is visible, indicating that the model systematically attributes a strong positive effect to this artificially inflated value. This pattern illustrates how the injected bias in studied credits is absorbed and rationalized by the model's explanations, reproducing the kind of overreliance and anomaly rationalization phenomena that the paper aims to study.

2.5. Asking LLMs

The following large language models were employed to generate explanations for the SHAP visualisations: DeepSeek-R1, Qwen3-max, Sonar (accessed via Perplexity), Gemini-3, and ChatGPT-5.2 (accessed via Perplexity). All models were queried in 3 separated dialogues and in their standard configuration without enabling any explicit reasoning or chain-of-thought features. The procedure was repeated with the addition of further information about SHAP and its visualizations. The models were provided with the SHAP plots (Figures 1, 2) and asked the initial question: "How does the studied credits attribute affect the final score?". The interaction protocol followed a two-step strategy: first, each model received a textual description of the dataset and the predictive model together with the initial question and SHAP visualisations; second, an additional prompt was issued requesting a possible explanation for the outlier in studied credits at the value of 125: "What is a possible explanation for the outlier values of

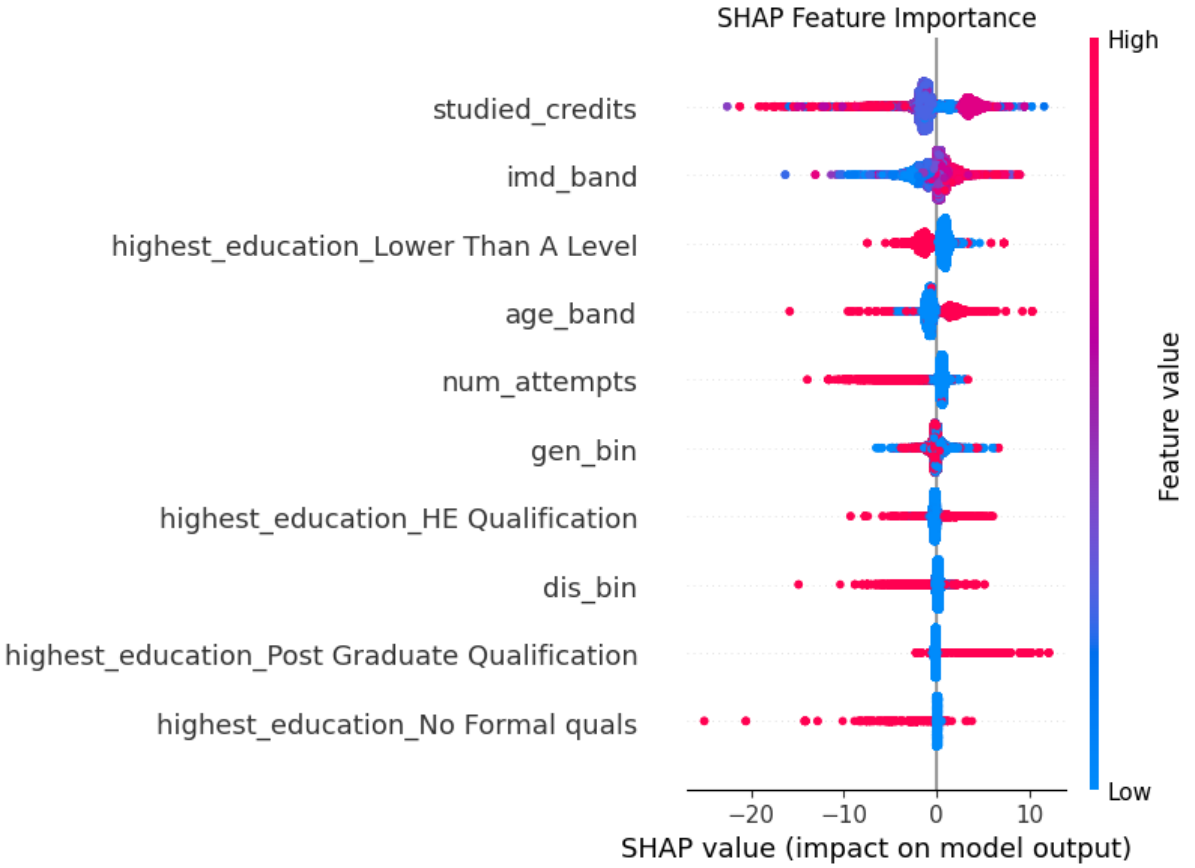


Figure 3: SHAP 120-credit Summary Plot

the studied credits, equal to 125?". For the similarity check of LLM responses, the embeddings were calculated.

2.6. Additional Comparative Experiment: 120-Credit Institutional Value

To examine whether LLM rationalization patterns are specific to artificial anomalies or generalize to legitimate data patterns, we conducted a second round of queries using the additional SHAP visualizations (Figures 3 and 4) with focus on the natural concentration of points at 120 studied credits. Unlike the artificial 125-credit value, which is not standardized for the educational context, 120-credit represents a legitimate, institutionally meaningful value in the Open University context—the standard full-time annual study load. This allowed us to observe how LLMs interpret and explain genuine data patterns that exhibit similar vertical dispersion in SHAP values but have clear institutional justification.

The experimental protocol paralleled the original 125-credit experiment: all five LLMs were presented with the same SHAP visualizations and asked the identical initial question: “How does the studied credits attribute affect the final score?” followed by a second prompt requesting explanation for the concentration and variability of SHAP values at 120 studied credits: “What is a possible explanation for the variability of SHAP values at the 120-credit value?”

This comparative design enabled us to assess whether LLM rationalization behavior persists when anomalies have legitimate institutional grounding, and whether the presence of institutional context modulates the models’ critical evaluation of data patterns and their propensity to identify potential data quality issues versus meaningful educational phenomena.

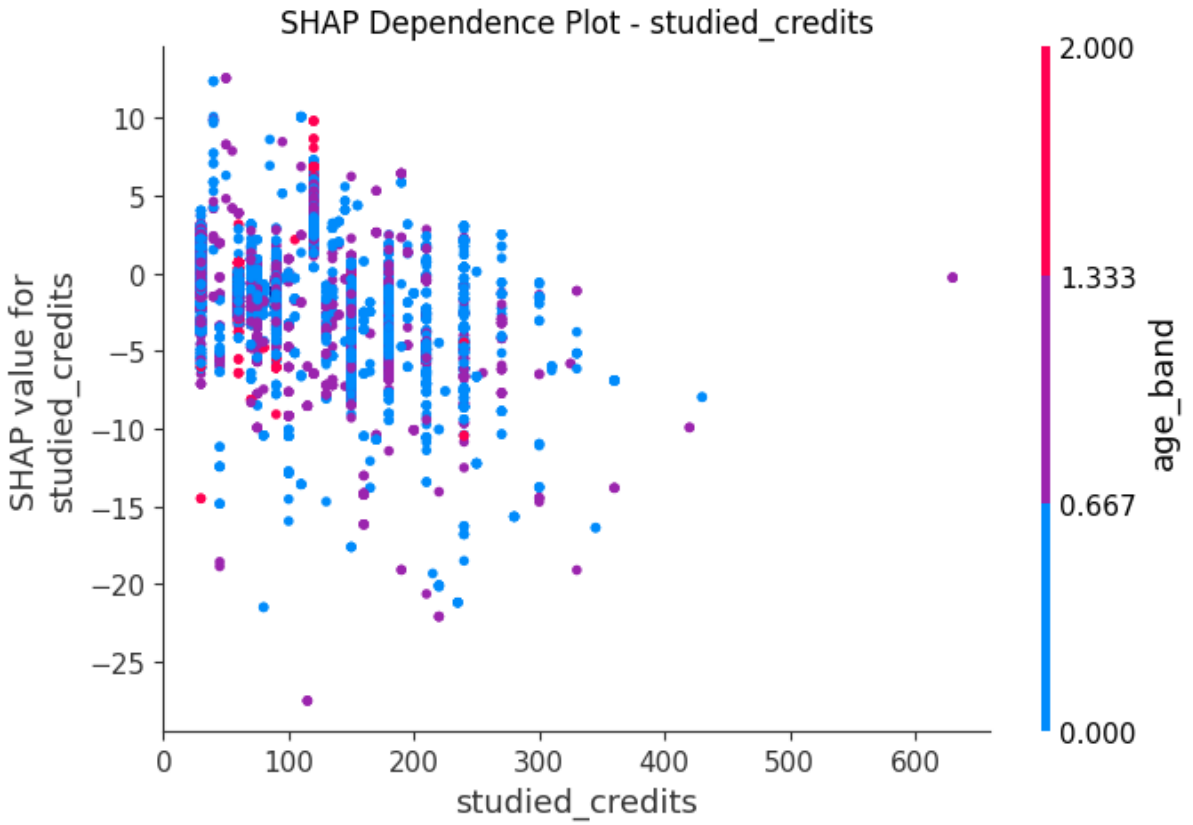


Figure 4: SHAP 120-credit Dependence Plot

3. Results

3.1. LLMs General Interpretations of SHAP Plots

The responses from the 5 language models (DeepSeek-R1, Qwen3-max, Sonar, Gemini-3, and ChatGPT-5.2) revealed consistent patterns in how LLMs interpreted the SHAP visualizations, while simultaneously exposing systematic tendencies toward rationalization of the introduced data anomaly. Within our sample, embedding similarities indicate a high degree of convergence across models: Gemini-3-3 shows the strongest overlap with the reference explanations (0.91), followed by Qwen3-max (0.87), DeepSeek-R1 (0.84), and GPT-5.2 (0.83), while Sonar remains only slightly lower at 0.79.

All 5 models correctly identified studied credits as the most influential feature in the predictive model, demonstrating accurate reading of the feature importance plot. The models uniformly described the relationship between studied credits and predicted final scores as non-linear, with SHAP values ranging from approximately -20 to +20. Each model noted that the effect varied substantially across individual students, acknowledging the wide vertical dispersion of points in the dependence plot. It was also found that responses of the models did not differ in content when descriptions of what SHAP is and its visualizations were added to the queries.

The models converged on several key observations: (1) moderate credit loads (approximately 100-200 credits) generally associated with positive SHAP values and higher predicted scores; (2) the relationship exhibited diminishing returns, with effects plateauing or reversing at higher credit values; and (3) interactions with other features, particularly prior education level, modulated the impact of studied credits on predictions. DeepSeek-R1 explicitly noted that the feature explained only approximately 15% of score variance, appropriately contextualizing the model's limited explanatory power.

Importantly, all models engaged with the color-coded interaction variable (highest education - Lower Than A Level) displayed in the dependence plot. DeepSeek-R1, Qwen3-max, and Gemini-3 observed

that students with lower prior education levels (represented by redxw/pink points) tended to show stronger positive effects from moderate credit loads, while Sonar emphasized that the credit load effect was "strongly moderated by prior education."

3.2. Rationalization of the 125-Credit Anomaly

When prompted to explain the vertical concentration of data points at 125 studied credits—the value artificially introduced during data manipulation—all 5 models engaged in rationalization rather than identifying it as a potential data quality issue or model artifact. This behavior closely mirrors the overreliance and rationalization patterns documented by Kaur et al. [8] in human data scientists.

DeepSeek-R1 characterized the 125-credit cluster as representing "a heterogeneous subgroup" with "widely scattered SHAP values," attributing the pattern to legitimate educational scenarios: atypical part-time enrollment, carryover credits from previous terms, module retakes, or administrative factors. The model framed 125 credits as slightly unusual (noting that "the standard full-time is often around 120 credits") but proposed substantive explanations for why such a load might occur and why it would produce variable performance outcomes.

Qwen3-max offered multiple interpretations, treating 125 credits predominantly as normative rather than anomalous, describing it as "a dense cluster (not rare)" representing "a common/typical load near the standard 120-credit full-time year." Importantly, Qwen3-max also acknowledged that one possible explanation involved data error, noting that 125 credits "may be atypical (potential data error, non-modular)." However, the model's overall consensus interpretation "leans toward institutional normality over anomaly," suggesting that even when LLMs recognize data quality concerns, they tend to privilege alternative explanations.

Sonar explicitly categorized 125 credits as an outlier but attributed it to special educational circumstances or data issues rather than experimental manipulation. The model suggested that the value "may arise from combinations of standard full-time loads plus extra smaller or legacy modules, or they may result from data entry, coding, or merging errors in the dataset." Sonar acknowledged the atypical behavior of these data points but rationalized their presence through plausible administrative explanations, cautioning that "the model predicts with more uncertainty and assigns extreme SHAP values" for this group.

ChatGPT-5.2 gave one variant of interpretation and similarly rationalized the 125-credit anomaly, describing it as corresponding to "special or atypical study configurations such as extra small modules, credit transfers, overlapping academic periods, or legacy program rules at the Open University." ChatGPT-5.2 framed these cases as reflecting "rare administrative or program structures" that produce "more variable or unusual prediction patterns that don't fit the typical credit-to-score relationship." The model did not question whether the concentration of points at this specific value might indicate a systematic data quality problem.

Gemini-3 uniquely characterized 125 credits as forming "a prominent vertical cluster" that acts as a "decision point" differentiating student outcomes based on prior education. Gemini-3 proposed that for low-education students, 125 credits signals "strong success (e.g., optimal full-time load like 120 + extra module)," while for higher-education students it predicts "failure (e.g., underachievement by not completing more)." This interpretation transforms the anomaly into a meaningful educational phenomenon reflecting differential effects of credit load across student subgroups.

3.3. LLMs Interpretations of 120-Credit Concentration

For the second round with 120 credits, which represents institutionally meaningful value in the Open University context, all five LLMs recognized and described the vertical dispersion of SHAP values without identifying it as a problematic anomaly. DeepSeek-R1 described the pattern as reflecting feature interactions with age band and deprivation, attributing the variability to behavioral differences among students selecting identical credit loads. ChatGPT-5.2 characterized 120 credits as a "data distribution artifact" resulting from administrative structure (maximum annual capacity), discrete

module packaging (30- and 60-credit blocks naturally summing to 120), and "hidden heterogeneity"—students with identical credit loads differing in other dimensions. Qwen3-max was notably more critical, explicitly acknowledging that while 120 credits is common, individual negative SHAP points around this value could stem from "data entry error" and recommending contextual investigation through local outlier detection and data lineage auditing. Sonar and Gemini-3 similarly interpreted 120 as the institutional norm for full-time engagement, framing the outlier-like vertical spread as meaningful evidence that heavy credit loads correlate with lower predicted performance rather than as a data quality concern.*f*

4. Discussion

4.1. RQ1: Do LLMs Reproduce Human Cognitive Biases When Interpreting SHAP Graphics?

The central finding of this study is that all 5 tested Large Language Models (DeepSeek-R1, Qwen3-max, Sonar, Gemini-3, and ChatGPT-5.2) reproduce the systematic rationalization behavior documented in human practitioners [8]. Rather than flagging the artificial 125-credit concentration as suspicious or requesting information about data preprocessing, all models generated post-hoc explanations that naturalized the anomaly within plausible institutional contexts. Notably, although Qwen3-max and Sonar acknowledged data error as a possible explanation, none prioritized it as the primary interpretation, instead promoting substantive educational narratives. This pattern reveals that LLMs inherit the cognitive vulnerabilities toward overreliance identified in human-centered XAI research [9, 21]. The absence of spontaneous critical questioning—no model requested data quality checks or flagged the pattern as suspicious—demonstrates that LLMs default to generating coherent narratives rather than maintaining skepticism about anomalies, directly mirroring human rationalization tendencies.

4.2. RQ2: How Do LLM-Generated Explanations Compare Across Different Models?

Across all five models, we observed consistent accuracy in identifying studied credits as the most influential feature and correctly describing the non-linear relationship to predicted scores. However, meaningful variability emerged in explaining the 125-credit anomaly: DeepSeek-R1, Sonar, and ChatGPT-5.2 attributed it to legitimate educational phenomena, Qwen3-max treated it as normative, while Gemini-3 framed it as a meaningful "decision point" differentiating outcomes by education level. This variability despite identical visualizations and prompts suggests different interpretive predispositions shaped by training and fine-tuning strategies [18]. Additionally, providing additional context about SHAP methodology did not substantially improve critical evaluation, which can be linked to the presence of information about what SHAP in the data on which the models were trained.

4.3. Natural Anomalies vs. Artificial Anomalies

Analysis of LLM responses to the legitimate 120-credit institutional value reveals that rationalization tendencies are context-dependent. While all five models rationalized the artificial 125-credit anomaly, they similarly rationalized the natural 120-credit concentration through educationally meaningful frameworks (administrative structure, module bundling, feature interactions). This refines RQ1: LLMs reproduce human rationalization biases, but institutional legitimacy actively suppresses critical evaluation. When data patterns align with known institutional structures, LLMs are more disposed toward substantive rationalization and less likely to flag data quality concerns—even when they possess the conceptual capability to do so (as demonstrated by Qwen3-max's suggested audits that remained deprioritized).

Regarding RQ2, all five models converged on similar explanatory frames for the legitimate 120-credit value, contrasting with the variability observed for the artificial 125-credit anomaly. This suggests

institutional context is a stronger moderator of critical evaluation than methodological context: providing SHAP methodology explanations did not improve skepticism about 125 credits, yet institutional legitimacy of 120 credits suppressed critical evaluation without prompting. These findings indicate that LLM-generated explanations of XAI outputs are particularly vulnerable to rationalization when data patterns align with practitioners' institutional knowledge and legitimate educational frameworks [18].

4.4. Risk Amplification Rather Than Mitigation

The deployment of LLM-based explanation systems in high-stakes educational contexts—where ML models inform decisions about student support, resource allocation, and academic pathways—carries a specific risk of amplifying rather than mitigating the rationalization and overreliance biases documented in prior research [17, 18]. Unlike human experts, LLMs do not tire, do not express uncertainty proportional to epistemic limits, and do not spontaneously acknowledge when information is insufficient to draw conclusions. An educator receiving an LLM-generated natural language explanation of a SHAP plot might reasonably perceive greater confidence and reliability than would be warranted by the underlying evidence. This concern is especially acute in educational contexts, where practitioners span diverse levels of ML expertise [12, 13].

4.5. Toward Responsible Integration of LLMs in XAI Workflows

These findings suggest several design considerations for the responsible integration of LLMs into XAI workflows. First, LLM-generated explanations should be accompanied by explicit warnings about the possibility of rationalization biases and should include prompts encouraging critical evaluation of data quality and preprocessing procedures. Second, LLM systems might be enhanced with reasoning components or logical checks that flag anomalies in visualizations and explicitly raise data quality concerns alongside substantive interpretations, as proposed by Abba Omar et al. [26]. Third, practitioners should be educated about the fallibility of both human and LLM-based interpretations of XAI outputs, with emphasis on the limitations of post-hoc rationalizations.

Most fundamentally, these results underscore that LLMs should be understood as amplifiers of existing interpretive patterns—including biases—rather than as correctives to them [27]. The promise of LLM-mediated explainability will only be realized if these systems are designed with explicit awareness of the cognitive biases they may reinforce, and if institutional and pedagogical structures are established to foster critical engagement with AI-generated explanations rather than passive acceptance.

5. Future Work

Several promising directions warrant investigation. First, hybrid validation systems combining LLM-generated explanations with automated anomaly detection algorithms could flag suspicious data patterns while leveraging LLMs' strengths in natural language generation [28]. Second, alternative prompting strategies—such as asking LLMs to generate competing hypotheses or explicitly articulate reasons why patterns might reflect data quality issues—could reduce rationalization tendencies [29]. Third, future work should investigate whether LLMs can be fine-tuned to adopt more skeptical epistemic stances when interpreting XAI visualizations, and whether users exhibit even higher levels of overreliance when engaging with LLM-generated explanations compared to raw SHAP visualizations.

A critical open question remains: what specific mechanisms—both technical and procedural—might effectively reduce rationalization risks in educational ML systems? Addressing this question requires investigation of data governance practices, multi-stakeholder review protocols, and institutional structures that foster critical engagement with AI-generated explanations. Furthermore, the comparison for human expert interpretations of the same SHAP visualizations with LLM responses in a controlled setting could be done. It would allow to quantify differences in interpretation quality Additionally, research should examine how different educational stakeholders (educators, administrators, students,

parents) interact with and act upon LLM-mediated explanations of model predictions, and whether stakeholder expertise modulates susceptibility to rationalization biases.

6. Limitations

This study has several important limitations that constrain the generalizability of findings. First, our sample of five LLMs represents a limited slice of the landscape of contemporary language models. The models tested (DeepSeek-R1, Qwen3-max, Sonar, Gemini-3, and ChatGPT-5.2) were chosen based on availability and accessibility, but they may not be representative of all LLMs or future developments in the field. More critically, all models were queried in their **standard configuration without enabling explicit reasoning or chain-of-thought features**, which are now available in several state-of-the-art systems (e.g., o1, Claude 3.7). Our findings therefore cannot speak to whether advanced reasoning capabilities might mitigate rationalization tendencies. Additionally, we conducted only a single experimental iteration with one educational dataset and one type of data anomaly (concentration of studied credits at 125). The generalizability to other domains, datasets, and types of anomalies (e.g., spurious correlations, missing data patterns, or subtle distributional shifts) remains unknown. Our prompting strategy was also relatively simple and uniform; more sophisticated prompt engineering techniques or few-shot examples might substantially alter LLM responses.

Second, our study did not directly measure the impact of LLM-generated explanations on end-user decision-making or behavior. While we documented that LLMs rationalize anomalies similarly to humans, we lack empirical evidence about whether educators or students who receive LLM-mediated explanations exhibit even higher levels of overreliance compared to those viewing raw SHAP visualizations, or whether certain user populations (e.g., those with domain expertise) are more or less susceptible to rationalization biases. Finally, the artificial manipulation of a single educational feature (studied credits) may not capture the complexity of real-world data quality issues or the interplay between multiple sources of bias in actual educational datasets.

Declaration on Generative AI

During the preparation of this work, the authors used Sonar model (accessed via Perplexity) in order to grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115.

[2] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.

[3] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, D. Gašević, Explainable artificial intelligence in education, Computers and Education: Artificial Intelligence 3 (2022) 100074.

[4] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems 30 (2017) 4765–4774.

[5] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022. URL: https://christophm.github.io/interpretable-ml-book/.

[6] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).

[7] I. Covert, S. M. Lundberg, S. Lee, Understanding global feature contributions through additive importance measures, CoRR abs/2004.00668 (2020). URL: https://arxiv.org/abs/2004.00668.

[8] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, J. Wortman Vaughan, Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–14.

[9] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, D. Weld, Does the whole exceed its parts? The effect of AI explanations on complementary team performance, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–16.

[10] G. Siemens, R. S. J. Baker, Learning analytics and educational data mining: towards communication and collaboration, in: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 2012, pp. 252–254.

[11] K. Holstein, S. Doroudi, Equity and artificial intelligence in education, in: The Ethics of Artificial Intelligence in Education, Routledge, 2022, pp. 151–173.

[12] U. Ehsan, E. A. Watkins, P. Wintersberger, C. Manger, S. S. Y. Kim, N. Van Berkel, A. Riener, M. O. Riedl, Human-centered explainable AI (HCXAI): Reloading explainability in the era of large language models (LLMs), in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–6.

[13] S. Maity, A. Deroy, Human-centric eXplainable AI in education, arXiv preprint arXiv:2410.19822, 2024.

[14] V. Swamy, B. Radmehr, N. Krco, M. Marras, T. Käser, Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs, in: Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022), 2022, pp. 1–12.

[15] I. E. Livieris, N. Karacapilidis, G. Domalis, D. Tsakalidis, An advanced explainable and interpretable ML-based framework for educational data mining, in: International Conference in Methodologies and Intelligent Systems for Technology Enhanced Learning (MIS4TEL 2023), Springer, 2023, pp. 87–96.

[16] S. Brdnik, V. Podgorelec, B. Šumak, Assessing perceived trust and satisfaction with multiple explanation techniques in XAI-enhanced learning analytics, Electronics 12 (2023) 2901. doi:10.3390/electronics12152901.

[17] C.-C. Hsu, I.-Z. Wu, S.-M. Liu, Decoding AI complexity: SHAP textual explanations via LLM for improved model transparency, in: 2024 International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan), IEEE, 2024, pp. 197–198.

[18] C. Singh, J. P. Inala, M. Galley, R. Caruana, J. Gao, Rethinking interpretability in the era of large language models, arXiv preprint arXiv:2402.01761 (2024).

[19] D. Slack, S. Krishna, H. Lakkaraju, S. Singh, Explaining machine learning models with interactive natural language conversations using TalkToModel, Nature Machine Intelligence 5 (2023) 873–883.

[20] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38. doi:10.1145/3571730.

[21] M. Turpin, J. Michael, E. Perez, S. R. Bowman, Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, in: Advances in Neural Information Processing Systems (NeurIPS), 2023. arXiv:2305.04388.

[22] A. Mishra, S. Rahman, H. Kim, K. Mitra, E. Hruschka, Characterizing large language models as rationalizers of knowledge-intensive tasks, in: Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, 2024, p. 8117–8139. doi:10.18653/v1/2024.findings-acl.484.

[23] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, et al., Ethical and social risks of harm from language models, arXiv preprint arXiv:2112.04359, 2021.

[24] European Commission, Ethics Guidelines for Trustworthy AI, Technical Report, European Commission, High-Level Expert Group on AI, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[25] J. Kuzilek, M. Hlosta, Z. Zdrahal, Open university learning analytics dataset, Scientific Data 4 (2017) 170171. doi:10.1038/sdata.2017.171.

[26] Z. Abba Omar, N. Nahar, J. Tjaden, I. M. Gilles, F. Mekonnen, J. Hsieh, C. Kästner, A. Menon, Beyond accuracy, SHAP, and anchors — on the difficulty of designing effective end-user explanations, arXiv preprint arXiv:2503.15512, 2025.

[27] H. Kaur, M. R. Conrad, D. Rule, C. Lampe, E. Gilbert, Interpretability gone bad: The role of bounded rationality in how practitioners understand machine learning, Proceedings of the ACM on Human-Computer Interaction 8 (2024) 1–34.

[28] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, R. Krishna, Explanations can reduce overreliance on AI systems during decision-making, Proceedings of the ACM on Human-Computer Interaction 7 (2023) 1–38. doi:10.1145/3579605.

[29] Z. Buçinca, M. B. Malaya, K. Z. Gajos, To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–21. doi:10.1145/3449287.