White Box Finance: Interpreting AI Decisions in Finance through Rules and Language Models

Oluwafemi Azeez
Pastel Africa
femi@pastel.africa

Samson Tontoye
Pastel Africa
samsont@pastel.africa

Olorunleke White
Pastel Africa
olorunleke@pastel.africa

Adenike Alli
Pastel Africa
adenike@pastel.africa

Abuzar Royesh
Pastel Africa
abuzar@pastel.africa

Abstract

Although loan defaults continue to cause substantial financial losses, this study focuses on improving how AI credit risk models are explained. Beyond developing a predictive model based on the demographics of the borrower, the attributes of the loan, and the credit history, the core contribution lies in introducing and comparing explanation methods. Specifically, we evaluated two ways to provide explanations. One method is a module that integrates SHAP values and GPT-4 to generate human-friendly narratives, a second is a rule-based logic explanation. This approach aims to enhance interpretability and trust, offering a clearer understanding of model predictions than traditional explanation techniques.

1 Introduction

1.1 Literature Review

Model explainability is central to promoting transparency in machine learning applications, especially within high-stakes domains like finance. Several techniques have been proposed in the literature to demystify black-box models. Lundberg and Lee [1] introduced SHAP, a unified framework based on game-theoretic Shapley values that attributes the contribution of each feature to a model's output. This technique has become a standard for post hoc interpretability across tree-based models, including XGBoost. Ribeiro et al. [2] proposed LIME, which builds locally linear interpretable models around each prediction. Although effective, its sensitivity to perturbations often limits its robustness. SHAP overcomes this by ensuring consistency and local accuracy. More recently, attention has shifted to language-based explanations. Tools such as LLMExplainer [4] and GPT-4-based methods demonstrate how large language models (LLMs) can augment feature-based explanations with human-readable justifications. In the financial domain, explainability tools have seen application in loan and credit risk modeling [3]. These efforts highlight the growing importance of visual and textual explanations in improving end-user trust, regulatory compliance, and auditability.

1.2 Background

Financial institutions face significant losses due to loan defaults, which occur when borrowers fail to meet repayment obligations. Traditional rule-based credit scoring systems struggle to adapt to nonlinear borrower behavior and may misclassify borrowers with atypical but reliable profiles. Machine learning (ML) models, such as gradient-boosting trees, have recently improved predictive performance for loan default detection. However, their black-box nature remains a major barrier

to adoption in highly regulated domains such as finance. Stakeholders, including loan officers, compliance teams, and regulators, require clear justifications for automated decisions. This has led to the rise of explainable artificial intelligence (XAI), which aims to provide transparency in ML predictions. Prominent XAI methods include SHapley Additive exPlanations (SHAP), which assign feature-level attributions to model outputs. Additionally, recent advances in large language models (LLMs) such as GPT-4 enable the generation of human-readable natural language explanations. In this study, we propose and compare two methods:

1. Rule-based logic to capture high-level decision heuristics

2. local SHAP visualizations to attribute feature contributions at the individual prediction level, and GPT-generated textual rationales to translate explanations into business-friendly language.

2 Methodology

2.1 Data Preprocessing

The dataset comprised anonymized records of loan applicants and their repayment behaviors. The key attributes included demographic data, employment details, financial history, and credit bureau characteristics. Initial data cleaning steps involved; Removing a negligible fraction of records with missing values, encoding categorical variables using frequency encoding to retain ordinal relationships, and conserving numerical column scales, as XGBoost inherently handles unscaled data effectively.

2.2 Feature Engineering

To capture non-linear signals, several derived features were introduced and features selected by removing highly correlated variables (Pearson > 0.9) and low-variance columns.

2.3 Model Training

An XGBoost classifier was employed for its robustness and ability to handle missing values and nonlinear interactions in tabular data. A key challenge was class imbalance: loan defaults were significantly less frequent than non-defaults. To address this, the $scale\_pos\_weight$ parameter was computed as follows:

$$weight = \frac{\sum_{y=0}}{\sum_{y=1}} \tag{1}$$

The model evaluation used 5-fold stratified cross-validation to preserve class proportions across splits. The final training was conducted on the complete training set. Performance was evaluated, and this yielded discrimination capability and calibration between predicted probabilities and actual default outcomes.

2.4 Model Validation

The model achieved an AUC of 0.74, with good balance between precision and recall. Confusion matrix analysis indicated that the classifier maintained conservative decision boundaries to minimize false positives, a crucial metric in risk-sensitive applications.

2.5 Post-Processing

The explanation module Predictions and probability scores from the XGBoost model were fed into the explanation module to derive rationale for each instance, enhancing transparency and trustworthiness of automated decisions.

3 Explanation Methods

The explanation module is designed to provide intelligible justifications for individual loan default predictions using a combination of SHAP (SHapley Additive exPlanations), rule-based logic and GPT-4o. This modular design enables explainability in both structured numerical formats and human-understandable narratives.

3.1 SHAP Explainer

We applied SHAP TreeExplainer to the trained XGBoost model to calculate the local feature attributions for each prediction. SHAP values provide additive feature importance scores for each input instance. These values are visualized using waterfall plots, where positive and negative contributors to the predicted probability are clearly distinguished. This graphical representation enables analysts to rapidly identify why a borrower was classified as high or low risk.

3.1.1 Prompt Generator

The prompt generator converts the top SHAP-ranked features into a domain-specific natural language prompt. It selects the top 3 to 5 contributing features, attaches their directionality. This structured prompt forms the basis for the next stage.

3.1.2 Language Generator (GPT-4o)

Each prompt is asynchronously sent to OpenAI's GPT-4o via the openai Python SDK. GPT-4o processes the inputs and returns a short, explainable narrative describing why the customer was flagged as a high or low default risk. The model runs with a low temperature (temperature=0.4) to ensure conciseness and factual consistency. Outputs are saved to the dataframe alongside SHAP plots for each record.

3.1.3 Integration Pipeline

The entire explanation module runs in asynchronous batches (typically 100–200 records) to optimize throughput while respecting rate limits. The final output includes; SHAP plots saved as .png for visual explanation, Textual justifications appended to each record, Audit-ready logs for compliance and decision traceability

3.2 Rule-Based Logic Method

We create an histogram of categorical variables normalized across the target variable class and plotted together e.g figure 1b and KDE plot of numerical variables grouped by target class e.g figure 1a. The visualization is then used to determine boundary variables and business logical rules that could be used to provide explanations after a prediction is made. e.g

```
if row["Age"] < 40:
    explanations.append("Young age may indicate lack of financial experience.")
```

If any of these business rules are triggered, a tag is added to the explanation record, reinforcing the decision from both a statistical and deterministic perspective. This enhances model reliability and auditability by aligning predictions with institutional underwriting policies.
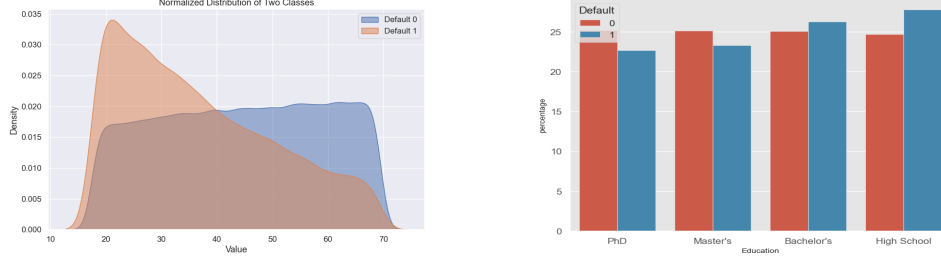
The various approach statistical(SHAP), rule-based, and language-based ensures robust, interpretable and human aligned explanations for credit risk predictions.

4 Results

1. **gpt_explanation**: Based on the provided information, the risk of this customer defaulting on their loan can be explained by examining the key factors and their respective SHAP impacts:

    (a) **Interest Rate**: The interest rate on the loan is quite high at 23.96%. This significantly increases the cost of borrowing, making it more challenging for the customer to manage

Table 1: A Test Dataset sample for prediction

Age	Income	LoanAmount	CreditScore	MonthsEmployed	InterestRate	DTIRatio	Education
36	80846	179949	347	20	23.96	0.9	PhD

(a) KDE Plot of Age (continuous) for default 0 and default 1 class.

(b) Histogram Plot of Education (categorical) for default 0 and default 1 class.

Figure 1: Plots of numerical and categorical boundaries for rule based decision

their monthly payments. The high SHAP impact of 0.95 indicates that this factor is a strong contributor to the default risk.

(b) **DTI Ratio**: The debt-to-income ratio of 0.9 suggests that the customer's monthly debt payments are very high relative to their income. However, the negative SHAP impact of -0.38 implies that, in this context, the DTI ratio is somewhat mitigating the default risk, possibly because the model expects even higher DTI ratios for high-risk cases.

(c) **Months Employed**: The customer has been employed at their current job for 20 months. While this is a moderate duration, the positive SHAP impact of 0.32 suggests that the model views this employment length as a slight risk factor, possibly due to the lack of longer-term job stability.

2. **rule_explanation**: From the KDE plot in Figure 1a for example, you would notice that you can visually create a business logic on numerical variable age based on the boundary of 40, The age variable in the table 1 is 36 and less than 40 so a good explanation about young age listed below would be reasonable

   (a) Short employment duration may indicate job instability.
   (b) High interest rate increases financial burden, raising risk.
   (c) Low credit score indicates high risk of default.
   (d) Young age may indicate lack of financial experience.
   (e) High debt-to-income ratio indicates financial strain.
   (f) High loan amount increases risk of default.

5   Conclusion

In financial applications of AI, especially those that involve risk-sensitive tasks such as credit scoring, the ability to generate understandable and trustworthy explanations is crucial. This paper introduced two explanation methods designed to improve the interpretation of the model at the individual prediction level.

The first method leverages local SHAP values to identify the most influential features in a prediction, and then utilizes GPT-based natural language generation to produce human-readable explanations. By pairing the importance of quantitative characteristics with qualitative descriptions, this approach allows contextual and user-friendly interpretations of the behavior of the model.

The second method focuses on the extraction of rules through descriptive statistics. By analyzing feature distributions (e.g., via histograms) across classes, simple yet effective logical rules can be derived. These rules are used to construct transparent, rule-based explanations that can be applied post-prediction to help clarify why a particular decision was made.

Together, these approaches strike a balance between statistical rigor and semantic clarity, providing a pathway toward more interpretable and actionable AI systems in finance. Future work may involve validating the effectiveness of these explanations through user studies and expanding the logic framework to support more complex feature interactions.

References

[1] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[3] Shreya and Harsh Pathak. Explainable artificial intelligence credit risk assessment using machine learning, 2025.

[4] Jiaxing Zhang et al. LLMExplainer: Large Language Model Based Bayesian Inference for Graph Explanation Generation. *arXiv preprint arXiv:2407.15351*, jul 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract captured the exact methodologies used

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations were briefly discussed.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper discusses everything needed to reproduce results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We are still working on finishing up open access to the code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper discussed all the procedure necessary to produce the most important parts of the experiments

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper doesn't go into much details about this.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The paper does not go into this detail.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The paper follows NeurIPS code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses briefly about societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper doesn't pose any risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Authors are properly cited for every dataset used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not create any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Usage of LLM was properly described in the paper given that it's a core methodology

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.