LLM-Guided Explanation Generation Optimization for Classification Tasks

Thesis Supervision

Louenas Bounia (Associate Professor) and **Hanane Azzag** (Full Professor)
Members of LIPN, UMR 7030, A3 team, Sorbonne Paris Nord University

Keywords:
XAI, Interpretability, LLM, Explanation Guidance,
Hybrid Neuro-Symbolic Approaches, Democratization of Explanation

1 PhD Research Project

Scientific context. The use of **LLMs** has seen rapid expansion due to their ability to generalize across diverse tasks without requiring fine-tuning and their adaptability in handling complex problems. These models are now employed to automatically generate explanations for decisions made by machine learning systems in classification tasks, combining their contextual understanding with methods of *post-hoc rationalization*. This need for LLM-generated explanations addresses the demand from non-expert human users who struggle to comprehend state-of-the-art methods that typically produce explanations based on feature importance [20], local rules [19], or even rigorous formal explanations [2].

However, in the absence of appropriate control or guidance mechanisms, LLMs exhibit significant limitations. They can produce unreliable explanations, or even completely hallucinated ones [11], which considerably diminishes their usefulness for end-users. As demonstrated by [8], current methods for automatic explanation generation using LLMs do not show clear superiority, and the majority of these explanations are deemed irrelevant by end-users in empirical evaluations. To address these challenges, we propose a research topic aimed at developing innovative strategies to guide LLMs in generating both factual and counterfactual explanations. Our approach targets three essential criteria: the efficiency of the generation process, the faithfulness of the explanations to the model's reasoning, and the interpretability of the produced justifications. This work seeks to address a dual objective: on one hand, to provide explanations that are naturally understandable by humans [15] by aligning the model's reasoning with human cognitive patterns, and on the other hand, to explore neuro-symbolic approaches that combine the flexibility of LLMs with the transparency of symbolic systems.

State of the art. We briefly recall the main methods in Explainable Artificial Intelligence (XAI). On the one hand, model-agnostic approaches such as LIME [20], SHAP [13], Anchors [21], or contrastive explanations [5] aim to explain model decisions without knowledge of their internal structure. Although widely used, they rely on local perturbations of inputs and present several limitations: the same explanation can justify opposite predictions [10], Shapley values lack formal grounding [12], and the produced explanations are often not robust [1]. These weaknesses make them poorly suited for critical domains (medical, financial, legal), where consistency and rigor are essential. Conversely, formal approaches tailored to the model's structure guarantee rigorous explanations consistent with predictions [2, 10]. Often based on prime implicants (PI-implicants) [4], these methods, however, struggle to scale for complex models such as neural networks or ensemble models.

Both model-agnostic and model-specific methods, while capable of providing situationally appropriate explanations, tend to be more specialized and understandable primarily by expert users. Indeed, they rely on technical concepts requiring a certain level of expertise, making them inaccessible to non-experts. This issue is highlighted by Mavrepis et al. [14], who emphasize that XAI primarily targets experts, limiting its broader adoption. Their work aims to democratize XAI through a customized large language model (LLM) developed with ChatGPT Builder, generating clear and tailored explanations. To make

XAI more accessible, recent studies, such as that of Mavrepis et al., propose using customized LLMs (e.g., ChatGPT Builder) to produce clear and non-technical explanations. Other works follow this direction, including [22], which explores LLMs for generating recommendation explanations.

However, recent state-of-the-art methods still lack a robust framework for using LLMs in explanation generation, as shown by studies from [9], [17], and [6], which underscore major risks of unguided LLM use in XAI, especially in critical settings. These include: (1) **inconsistent explanations**, with contradictory justifications for similar inputs, (2) **critical hallucinations** such as fictitious references , and (3) **statistical mimicry**, favoring surface patterns over grounded reasoning [17].

To address these limitations, we propose combining three elements: a semantic validation pipeline, symbolic anchors inspired by neuro-symbolic approaches [17], and an interactive guidance protocol aligned with verifiable facts. Unlike the methods discussed in [18], our approach integrates a traceability mechanism combining verification $\mathcal{V}(e_i)$ and constraints $\mathcal{C}(e_i)$ for each explanation e_i, thereby meeting systematicity requirements while leveraging LLMs. Our main contributions are: controlled LLM integration, safeguards against hallucinations, and a quantitative reliability assessment, forming an **optimized guidance framework** that overcomes current LLM-based explanation limits.

2 Research directions

This thesis aims to democratize explanation by developing strategies to guide LLMs in generating explanations. To this end, the following directions will be explored:

Integration of Classifier Internal Information. We aim to enhance LLMs through systematic integration of target classifier information. The core idea is to avoid considering LLMs as "blind" explanation generators, but rather to provide them with maximum usable context about the underlying classification model. This includes structural information (such as a neural network's topology, decision tree rules, or linear model coefficients), model complexity characteristics (VC-dimension or algorithmic stability), as well as elements regarding the intrinsic difficulty of producing certain types of explanations [7] (such as the computational cost of model-agnostic or model-specific explanations, explanation fidelity, or ambiguities in post-hoc methods). The objective is to design an explanatory interface where LLMs are guided by theoretical and empirical anchors, enabling them to produce more reliable explanations aligned with the model's logical structure, and adapted to the difficulty level required by each instance or explanatory method.

Active Classifier Guidance in Explanations. We aim to strengthen the classifier's active role in the explanatory process by introducing a neuro-symbolic guidance mechanism [23] and systematic filtering of LLM-generated explanations. The core idea is to formalize this guidance through a hybrid framework combining symbolic representations (logical rules, semantic constraints, dependency graphs) as described in [3] and learned components (embeddings, neural models) to frame the explanatory generation with verifiable guarantees. These representations will encode expected properties of explanations - model fidelity, minimality, exhaustiveness, consistency - and translate them into selection or rejection criteria. Concretely, we will develop filtering modules guided by the classifier's structure and behavior, capable of evaluating LLM responses' relevance through explicit measures (prediction variability, local robustness, structural simplicity). The objective is to transform the classifier from a passive object to be explained into an active agent, enforcing neuro-symbolic constraints to ensure the quality and reliability of generated explanations.

Adaptive Interaction and Explanation Guidance Personalization. We will explore the interactive nature of explanation guidance by introducing an adaptive protocol that enables users to dynamically influence the explanation process. The goal is to generate personalized, contextualized, and understandable explanations that reflect user profiles (expertise, goals, constraints) and decision contexts. We propose an iterative architecture where users can adjust generation preferences [16] (e.g., detail level, argument types, language) while preserving core fidelity and consistency guarantees. We also study how human feedback impacts explanation robustness, especially under distribution shifts or uncertainty. An evaluation protocol will assess explanations along several axes: fidelity, alignment with classifier predictions, robustness to perturbations, and perceived utility by both expert and non-expert users.

References

[1] David Alvarez-Melis and Tommi S. Jaakkola. "On the Robustness of Interpretability Methods". In: *ICML Workshop on Human Interpretability in Machine Learning*. Stockholm, Sweden, 2018.

[2] G. Audemard et al. "On Preferred Abductive Explanations for Decision Trees and Random Forests". In: *Proc. of IJCAI'22*. 2022.

[3] Diego Calanzone, Stefano Teso, and Antonio Vergari. "Logically Consistent Language Models via Neuro-Symbolic Integration". In: *ICLR*. 2025.

[4] A. Darwiche and A. Hirth. "On the Reasons Behind Decisions". In: *Proc. of ECAI'20*. 2020.

[5] Amit Dhurandhar et al. "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives". In: *NeurIPS*. 2018.

[6] Nouha Dziri et al. "Faith and Fate: Limits of Transformers on Compositionality". In: *arXiv preprint arXiv:2305.18654* (2023). URL: https://doi.org/10.48550/arXiv.2305.18654.

[7] Yair Ori Gat et al. "Faithful Explanations of Black-box NLP Models Using LLM-generated Counterfactuals". In: *ArXiv* abs/2310.00603 (2023).

[8] Yair Ori Gat et al. "Faithful Explanations of Black-box NLP Models Using LLM-generated Counterfactuals". In: *The Twelfth International Conference on Learning Representations*. 2024.

[9] Gaël Gendron et al. "Large Language Models Are Not Strong Abstract Reasoners". In: *arXiv preprint arXiv:2305.19555* (2023). URL: https://arxiv.org/abs/2305.19555.

[10] A. Ignatiev, N. Narodytska, and J. Marques-Silva. "Abduction-Based Explanations for Machine Learning Models". In: *Proc. of AAAI'19*. 2019, pp. 1511–1519.

[11] Ziwei Ji et al. "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* (2023). URL: https://dl.acm.org/doi/10.1145/3571730.

[12] Olivier Letoffe, Xuanxiang Huang, and Joao Marques-Silva. "SHAP scores fail pervasively even when Lipschitz succeeds". In: *arXiv preprint arXiv:2412.13866* (2024). arXiv: 2412.13866 [cs.LG].

[13] S. Lundberg and S-I. Lee. "A Unified Approach to Interpreting Model Predictions". In: *Proc. of NIPS'17*. 2017, pp. 4765–4774.

[14] Philip Mavrepis et al. "XAI for All: Can Large Language Models Simplify Explainable AI?" In: *arXiv preprint arXiv:2401.13110* (2024). URL: https://doi.org/10.48550/arXiv.2401.13110.

[15] Tim Miller. "Explanation in Artificial Intelligence: Insights from the Social Sciences". In: *Artificial Intelligence* 267 (2019), pp. 1–38.

[16] Dimitry Mindlin et al. "Measuring User Understanding in Dialogue-based XAI Systems". In: *European Conference on Artificial Intelligence*. 2024.

[17] Iman Mirzadeh et al. "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models". In: *arXiv preprint arXiv:2410.05229* (2024).

[18] Fuseini Mumuni and Alhassan Mumuni. "Explainable artificial intelligence (XAI): from inherent explainability to large language models". In: *arXiv preprint arXiv:2310.09414* (2023).

[19] M-T. Ribeiro, S. Singh, and C. Guestrin. "Anchors: High-Precision Model-Agnostic Explanations". In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. 2018, pp. 1527–1535.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proc. of SIGKDD'16*. 2016, pp. 1135–1144.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin. "Anchors: High-Precision Model-Agnostic Explanations". In: *Proc. of AAAI'18*. 2018, pp. 1527–1535.

[22] Alan Said. "On Explaining Recommendations with Large Language Models: A Review". In: *arXiv preprint arXiv:2411.19576* (2024).

[23] Xin Zhang and Victor S. Sheng. "Neuro-Symbolic AI: Explainability, Challenges, and Future Trends". In: *ArXiv* abs/2411.04383 (2024).