



Natural Language Interpretability for ML-Based QoT Estimation via Large Language Models

Downloaded from: <https://research.chalmers.se>, 2026-01-12 08:04 UTC

Citation for the original published paper (version of record):

Ayoub, O., Natalino Da Silva, C., Troia, S. et al (2025). Natural Language Interpretability for ML-Based QoT Estimation via Large Language Models. International Conference on Transparent Optical Networks. <http://dx.doi.org/10.1109/ICTON67126.2025.11125132>

N.B. When citing this work, cite the original published paper.

© 2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, or reuse of any copyrighted component of this work in other works.

Natural Language Interpretability for ML-Based QoT Estimation via Large Language Models

Omran Ayoub¹, Carlos Natalino², Sebastian Troia³, Cristina Rottondi⁴,
Davide Andreoletti¹, Francesco Lelli¹, Silvia Giordano¹, Paolo Monti²

¹University of Applied Sciences of Southern Switzerland, Switzerland, ²Chalmers University of Technology, Sweden,

³Politecnico di Milano, Italy, ⁴Politecnico di Torino, Italy

Abstract—As Machine Learning (ML) systems become integral to network management, the need for transparent decision-making grows. While post-hoc explainability methods provide insights into model behavior, their technical nature often limits accessibility. We explore Large Language Models (LLMs) for translating complex ML model explanations, extracted using explainable artificial intelligence frameworks, into natural language to simplify user understanding and interpretability. Using direct prompting and self-reflection-based prompting, we generate explanations for a lightpath Quality of Transmission (QoT) estimation model. Empirical evaluations confirm the correctness and usefulness of LLM-generated interpretations in about 65% of the cases, highlighting the benefits of self-reflection in enhancing explanation quality. The study also remarks on the necessity of devising enhancements to improve the results achieved so far.

Index Terms—Explainable Artificial Intelligence; Shapley Additive Explanations; Empirical Evaluation.

I. INTRODUCTION

As Machine Learning (ML) systems become increasingly integrated into network management, ensuring the transparency of their decision-making processes is paramount [1]–[4]. Explaining ML-driven decisions can be particularly important in network operations where human intervention may be required to validate or override automated decisions [5], [6].

Post-hoc explainability techniques have emerged as standard tools for interpreting the behavior of trained ML models [7]. These methods, ranging from quantifying feature importance to generating counterfactual explanations, offer valuable insights into model behavior. However, such explanations are frequently presented in abstract or numerical forms requiring specialized expertise (see example explanation in Fig. 1). Even for domain experts, making sense of such explanations can be time-consuming and cognitively demanding [8], [9]. A potential solution to alleviate this challenge is the application of Large Language Models (LLMs) to translate ML model’s explanations, such as feature importance plots, into human interpretable language [10], [11].

The adoption of LLMs for automating and optimizing various networking tasks is increasing [12], [13], with a growing emphasis on optical networks [14], [15]. In [16], authors use LLMs to enable the use of natural language to perform the creation, search, and explanation of network slices.

In [17], authors use LLMs to help automate Intent-Based Networking (IBN) by translating high-level operator intents into optimization code and autoconfiguration policies, making network operations more efficient and interoperable. In [18], the authors propose a digital-twin-enhanced LLM framework to improve autonomous optical networks by integrating real-time network state updates and strategy pre-verification before deployment. In [19], authors explore how LLMs can enhance decision-making in real-time operations by automating key tasks like failure prediction and lightpath QoT estimation. Lastly, in [20], the authors leverage LLMs for log analysis to improve log parsing, anomaly detection, and report generation, demonstrating their potential to enhance operational efficiency and reliability.

In this work, we explore the usage of LLMs to interpret explanations generated by ML models. As a case study, we focus on the Quality of Transmission (QoT) estimation problem in optical networks. We develop an XGBoost (XGB) model to estimate the QoT of a lightpath and employ Shapley Additive Explanations (SHAP) as eXplainable AI (XAI) framework [21]. SHAP provides insights into the model’s decisions by quantifying feature importance. More precisely, we extract local model explanations, i.e., instance-specific interpretations that highlight how different features influence individual QoT predictions (see [3], [22] for an overview on SHAP’s application to lightpath QoT estimation). To enhance the human interpretability of these explanations, we leverage ChatGPT as LLM to process and interpret the SHAP-based explanations, providing human-readable insights that facilitate decision-making in optical network management. Figure 2 illustrates the translation of complex explanations (extracted using SHAP) to natural language.

We employ ChatGPT via two distinct approaches: (1) a direct prompting method and (2) a self-reflection strategy designed to enhance the quality of the interpretations. To validate our method, we conduct an empirical study involving expert evaluation. Domain experts assess the interpretations generated by the LLM and provide scores based on predefined evaluation metrics. We evaluate the interpretations in terms of correctness, scope, and usefulness. Our empirical results indicate that LLMs can generate useful interpretations of ML model explanations. The direct prompting and the self-

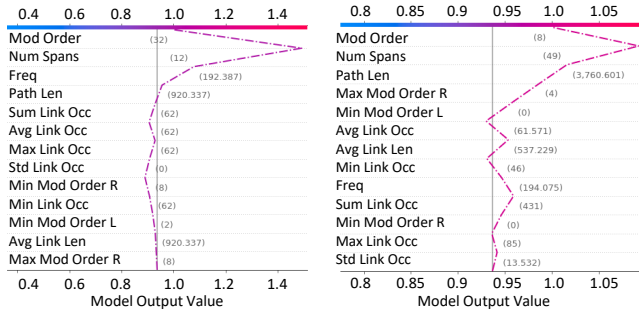


Fig. 1: Two examples of SHAP local explanations of a QoT estimation model.

reflection approaches achieve acceptable performance across evaluation metrics. However, the study highlights the necessity for enhancements to ensure the correctness and practicality of the generated interpretations.

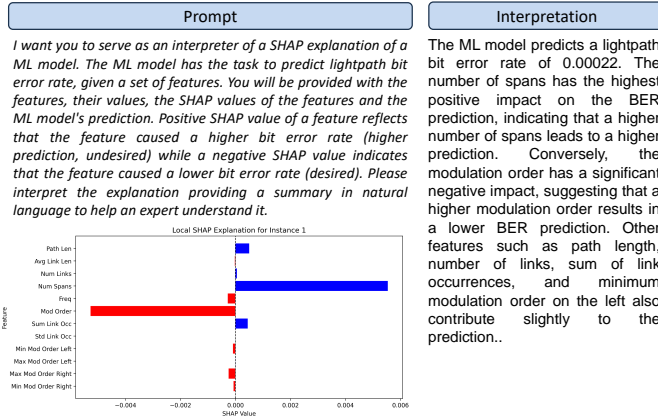


Fig. 2: Illustrative example of translating a SHAP local explanation into natural language. Note that we have transformed the SHAP local plot into a simpler bar plot for simplicity.

II. LLM-BASED FRAMEWORK FOR INTERPRETING MODEL'S EXPLANATION

Our framework consists of three main modules: (1) an ML model, (2) an XAI framework, and (3) an LLM, as illustrated in Fig. 3. The objective is to interpret the ML model's decisions by extracting explanations via SHAP and translating them into natural language by means of the LLM.

ML Model. Our framework has the potential to be applied to ML models that perform any task. Specifically, in this work, we develop an ML model to estimate the QoT of a lightpath [23]–[25]. The lightpath QoT estimation problem is formulated as a regression task, where the objective is to predict the expected bit error rate (BER) at the receiver side for each candidate lightpath. Each prospective lightpath is characterized by a set of descriptive features, capturing both its intrinsic properties (e.g., path length, number of links and spans, minimum/maximum link length, modulation format in use) and its spectral context (e.g., characteristics

of the spectrally adjacent left/right neighbors, and the minimum/maximum/average spectral occupation of the traversed links).

XAI Framework. To explain the model's decisions, we apply SHAP, a post-hoc explainability technique that assigns a Shapley value (i.e., a feature importance score) to each input feature [21]. These values quantify the contribution of each feature to the model's prediction. These values indicate how much a feature has influenced the prediction in comparison to a baseline (e.g., the average prediction across the dataset). A positive SHAP value means the feature has pushed the predicted bit error rate (BER) higher, while a negative SHAP value indicates it has contributed to lowering the predicted BER. This allows for a detailed, interpretable breakdown of how each feature affects the model's output. We extract local explanations, which provide insight into the predictions of the model for each sample/inference. The extracted SHAP explanations are then used as part of the input to the LLM.

Large Language Model. Once the explanations are extracted, we process them by prompting an LLM to translate the raw SHAP values into natural language explanations. To identify a prompt for generating high-quality explanations, we experimented with several prompt templates that varied both the explanation format and the accompanying text. We evaluated the quality of the resulting interpretations and iteratively refined our prompts based on these assessments and suggestions provided by the LLM itself.

We employ two different prompting strategies to guide the LLM in generating the interpretations:

a) *Direct Prompting Approach:* in this approach, we provide the LLM with a structured prompt that includes the data point, the model's decision, and the corresponding SHAP-extracted explanation. The LLM then generates a natural language interpretation based on the provided input. The prompt used in the *direct prompting* strategy is shown in Fig. 2.

b) *Self-Reflection Prompting Approach:* this method extends direct prompting by introducing a self-reflection step. Specifically, self-reflection is a foundational design pattern of agentic AI systems that provide an approach to advancing the capabilities of LLMs [26]–[28]. It focuses on enabling models to self-correct and iteratively improve their outputs through feedback mechanisms [27], [29], [30] over multiple iterations. After generating an initial explanation, we prompt the LLM to critically evaluate and refine its response, enhancing clarity and completeness. To optimize the prompt design, we experimented with different formats and phrasing, iteratively refining them based on the quality of interpretations and the LLM's own suggestions. Figure 2 shows the prompt used initially in our analysis while the prompt used for the *Self-reflection prompting* strategy is the following: *Please critically analyze the interpretation of the SHAP explanation provided below, and provide an improved interpretation.*

III. EXPERIMENTAL AND EMPIRICAL RESULTS

To evaluate the performance of the proposed framework, we train an XGB regressor model that performs QoT estimation.

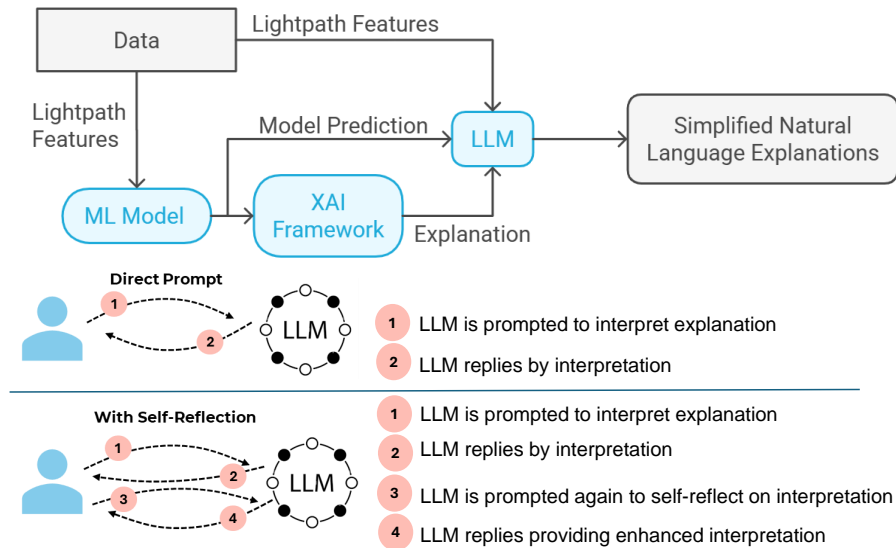


Fig. 3: Schematic representation of the framework.

The regressor takes as input a set of features describing a candidate lightpath for deployment and provides an estimation of the Bit Error Rate (BER) as output, as in [31]. We use the dataset made publicly available in [32]. We split the data following a 90/10 training/testing split.

Based on the trained model, we extract 40 local explanations using SHAP. To maintain objectivity, interpretations are presented to the experts in a randomized order, preventing them from knowing which prompting strategy generated each one. Additionally, experts evaluate the interpretations independently, without access to each other’s assessments, thus minimizing potential bias.

The evaluation metrics we consider are: (i) *correctness*, assessing how accurately the LLM-generated interpretations reflect the underlying explanation; (ii) *scope*, assessing whether LLM’s interpretation, when correct, focused on the important aspects of the explanation¹; and (iii) *usefulness*, assessing, if correct, the utility in providing practical support for human understanding. For correctness and scope, evaluators assign a binary score (e.g., correct/incorrect, within/outside scope). For usefulness, they provide an assessment using a scale ranging from 0 to 5, reflecting the degree to which the interpretation contributes to their understanding of the model’s decision-making process.

Table I reports the empirical results in terms of the average *correctness*, *scope* and *usefulness*. We report results for *scope* and *usefulness* only when the expert evaluates the interpretation as correct. We also report the agreement among the experts across *correctness* and the standard deviation of *usefulness*.

Results indicate that the direct prompt strategy achieves an average correctness of 61.8% while the one based on

TABLE I: Results in terms average of the predefined metrics, *agreement* across experts for *correctness* and the standard deviation (std) of *usefulness*.

Metric	Direct Prompt		Self Reflection	
	Avg	Agreement/Std	Avg	Agreement/Std
Correctness	61.8%	79%	65.2%	82.1%
Scope	96.8%	NA	98.9%	NA
Usefulness	3.81	0.14	3.92	0.40

self-reflection yields a slightly higher average correctness of 65.2%. This marginal improvement suggests that a single iteration of self-reflection is insufficient to overcome the LLM’s limitations in accurately interpreting model explanations for the lightpath QoT estimation task. Although these scores are low, they are in line with other LLM benchmark scores for the same model. The evaluators’ agreement is around 80% for both cases. Despite relatively high agreement, results indicate some discrepancy among evaluators due to differences in scoring LLM’s interpretations. LLM’s interpretations can sometimes lack specificity, leading experts to interpret the same output differently. Both strategies perform exceptionally well in terms of scope. Direct prompt and self-reflection achieve high averages, 96.8% and 98.2%, respectively. This indicates that once the LLM correctly interprets an explanation, it effectively identifies and emphasizes the most relevant and influential factors. Results regarding usefulness indicate that the experts believe that the LLM explanations can improve understanding of the SHAP explanations. Direct prompt achieves a rate of 3.81 (std 0.14), while self-reflection achieves a rate of 3.92 (std 0.40). Despite the slight score advantage for the self-reflection strategy, it is worth noting that it received the two lowest ratings from the evaluators. This was because interpretations, in some cases, were more detailed than desired.

¹LLM’s interpretation can be correct but not capturing the most important aspects.

IV. CONCLUSION

Our study demonstrates that LLMs can effectively translate ML model explanations into natural language to improve their interpretability. We employ two prompting strategies, referred to as direct prompting and self-reflection prompting. Empirical evaluations confirm the effectiveness of the LLM-generated interpretations. Moreover, while the two strategies yield meaningful explanations, the self-reflection strategy shows an edge over direct prompting in terms of scope and usefulness.

ACKNOWLEDGMENTS

This work has been partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on “Telecommunications of the Future” (PE00000001 - program “RESTART”) and by the EUREKA cluster CELTIC-NEXT project SUSTAINET-Advance funded by the Swiss Innovation Agency.

REFERENCES

- [1] Y. Wu, G. Lin, and J. Ge, “Knowledge-powered explainable artificial intelligence for network automation toward 6g,” *IEEE network*, vol. 36, no. 3, pp. 16–23, 2022.
- [2] S. Wang, M. A. Qureshi, L. Miralles-Pechuan, T. Huynh-The, T. R. Gadekallu, and M. Liyanage, “Applications of explainable ai for 6g: Technical aspects, use cases, and research challenges,” *arXiv preprint arXiv:2112.04698*, 2021.
- [3] O. Ayoub, S. Troia, D. Andreoletti, A. Bianco, M. Tornatore, S. Giordano, and C. Rottondi, “Towards explainable artificial intelligence in optical networks: the use case of lightpath QoT estimation,” *Journal of Optical Communications and Networking*, vol. 15, no. 1, pp. A26–A38, 2022.
- [4] O. Ayoub, C. Natalino, and P. Monti, “Towards explainable reinforcement learning in optical networks: The RMSA use case,” in *Optical Fiber Communications Conference and Exhibition (OFC)*, 2024, p. W41.6.
- [5] B. Dutta, A. Krichel, and M.-P. Odini, “The challenge of zero touch and explainable AI,” *Journal of ICT Standardization*, vol. 9, no. 2, pp. 147–158, 2021.
- [6] S. Wang, M. A. Qureshi, L. Miralles-Pechuán, T. Huynh-The, T. R. Gadekallu, and M. Liyanage, “Explainable ai for 6g use cases: Technical aspects and research challenges,” *IEEE Open Journal of the Communications Society*, 2024.
- [7] D. Vale, A. El-Sharif, and M. Ali, “Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law,” *AI and Ethics*, vol. 2, no. 4, pp. 815–826, 2022.
- [8] A. Hudon, T. Demazure, A. Karran, P.-M. Léger, and S. Sénécal, “Explainable artificial intelligence (XAI): how the visualization of ai predictions affects user cognitive load and confidence,” in *Information Systems and Neuroscience: NeuroIS Retreat 2021*. Springer, 2021, pp. 237–246.
- [9] L. Weber, S. Lapuschkin, A. Binder, and W. Samek, “Beyond explaining: Opportunities and challenges of XAI-based model improvement,” *Information Fusion*, vol. 92, pp. 154–176, 2023.
- [10] P. Mavrepis, G. Makridis, G. Fatouros, V. Koukos, M. M. Separdani, and D. Kyriazis, “XAI for all: Can large language models simplify explainable ai?” *arXiv preprint arXiv:2401.13110*, 2024.
- [11] A. Zyteck, S. Pidò, and K. Veeramachaneni, “LLMs for XAI: Future directions for explaining explanations,” *arXiv preprint arXiv:2405.06064*, 2024.
- [12] Y. Huang, H. Du, X. Zhang, D. Niyato, J. Kang, Z. Xiong, S. Wang, and T. Huang, “Large language models for networking: Applications, enabling techniques, and challenges,” *IEEE Network*, 2024.
- [13] H. Zhou, C. Hu, Y. Yuan, Y. Cui, Y. Jin, C. Chen, H. Wu, D. Yuan, L. Jiang, D. Wu *et al.*, “Large language model (LLM) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities,” *IEEE Communications Surveys & Tutorials*, 2024.
- [14] D. Wang, Y. Wang, X. Jiang, Y. Zhang, Y. Pang, and M. Zhang, “When large language models meet optical networks: paving the way for automation,” *Electronics*, vol. 13, no. 13, p. 2529, 2024.
- [15] S. Cruzes, “Revolutionizing optical networks: The integration and impact of large language models,” *Authorea Preprints*, 2024.
- [16] D. Adanza, C. Natalino, L. Gifre, R. Muñoz, P. Alemany, P. Monti, and R. Vilalta, “IntentLLM: An AI chatbot to create, find, and explain slice intents in TeraFlowSDN,” in *IEEE 10th International Conference on Network Softwarization (NetSoft)*, 2024, pp. 307–309.
- [17] A. Tzanakaki, M. Anastasopoulos, and V.-M. Alevizaki, “Intent-based control and management framework for optical transport networks supporting B5G services empowered by large language models,” *Journal of Optical Communications and Networking*, vol. 17, no. 1, pp. A112–A123, 2024.
- [18] Y. Song, Y. Zhang, A. Zhou, Y. Shi, S. Shen, X. Tang, J. Li, M. Zhang, and D. Wang, “Synergistic interplay of large language model and digital twin for autonomous optical networks: Field demonstrations,” *IEEE Communications Magazine*, 2025.
- [19] S. Cruzes, “Enhancing optical networks with large language models: An era of automated efficiency,” *Authorea Preprints*, 2024.
- [20] Y. Pang, M. Zhang, Y. Liu, X. Li, Y. Wang, Y. Huan, Z. Liu, J. Li, and D. Wang, “Large language model-based optical network log analysis using LLaMA2 with instruction tuning,” *Journal of Optical Communications and Networking*, vol. 16, no. 11, pp. 1116–1132, 2024.
- [21] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.
- [22] O. Ayoub, A. Bianco, D. Andreoletti, S. Troia, S. Giordano, and C. Rottondi, “On the application of explainable artificial intelligence to lightpath QoT estimation,” in *Optical Fiber Communication Conference*. Optica Publishing Group, 2022, pp. M3F–5.
- [23] C. Rottondi, L. Barletta, A. Giusti, and M. Tornatore, “Machine-learning method for quality of transmission prediction of unestablished lightpaths,” *Journal of Optical Communications and Networking*, vol. 10, no. 2, pp. A286–A297, 2018.
- [24] S. Aladin, A. V. S. Tran, S. Allogba, and C. Tremblay, “Quality of transmission estimation and short-term performance forecast of lightpaths,” *Journal of Lightwave Technology*, vol. 38, no. 10, pp. 2807–2814, 2020.
- [25] G. Davoli, R. Di Tommaso, A. Giorgetti, and C. Raffaelli, “Impact of lightpath selection on end-to-end service orchestration in disaggregated optical networks,” in *International Conference on Optical Network Design and Modeling (ONDM)*, 2024.
- [26] J. Huang and K. C.-C. Chang, “Towards reasoning in large language models: A survey,” *arXiv preprint arXiv:2212.10403*, 2022.
- [27] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegrefe, U. Alon, N. Dziri, S. Prabhunoye, Y. Yang *et al.*, “Self-refine: Iterative refinement with self-feedback,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [28] A. Plaata, A. Wong, S. Verberne, J. Broekens, N. van Stein, and T. Back, “Reasoning with large language models, a survey,” *arXiv preprint arXiv:2407.11511*, 2024.
- [29] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: Language agents with verbal reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [30] Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, N. Duan, and W. Chen, “Critic: Large language models can self-correct with tool-interactive critiquing,” *arXiv preprint arXiv:2305.11738*, 2023.
- [31] O. Ayoub, D. Andreoletti, S. Troia, S. Giordano, A. Bianco, and C. Rottondi, “Quantifying features’ contribution for ML-based quality-of-transmission estimation using explainable AI,” in *2022 European Conference on Optical Communication (ECOC)*. IEEE, 2022, pp. 1–4.
- [32] G. Bergk, B. Shariati, P. Safari, and J. K. Fischer, “ML-assisted QoT estimation: a dataset collection and data visualization for dataset quality evaluation,” *Journal of Optical Communications and Networking*, vol. 14, no. 3, pp. 43–55, 2021.