Human-Grounded Evaluation of Large Language Models for Optical Network Automation

Kiarash Rezaei*, Omran Ayoub†, Paolo Monti*, Carlos Natalino*

* Department of Electrical Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden
{kiarashr, mpaolo, carlos.natalino}@chalmers.se

†University of Applied Sciences and Arts of Southern Switzerland, 6928 Lugano, Switzerland
omran.ayoub@supsi.ch

***Abstract*—Large language models (LLMs) are increasingly adopted for network automation, yet their output quality and inference cost can vary substantially across LLMs families. We present HuGLEN, a stepwise evaluation pipeline that uses an LLM-as-a-judge together with a small set of expert ratings to enable scalable and reproducible comparison of candidate LLMs, and to rank them using a quality efficiency score (QES). We demonstrate HuGLEN on translating outputs from an explainable artificial intelligence (XAI) model for optical network quality of transmission (QoT) estimation task into operator-friendly explanations. Our results show that a medium-sized LLM (12B parameters) achieves the highest QES, indicating the best trade-off between explanation quality and efficiency. Overall, HuGLEN reduces the human-labeling burden while supporting consistent model selection for operator-facing automation tasks.**

***Index Terms*—large language models (LLMs), LLM evaluation, network automation, explainable AI (XAI), quality of transmission (QoT), energy efficiency, Quality-Efficiency Score (QES), human-grounded evaluation.**

I. Introduction

Large language models (LLMs) are transforming network automation by enabling unprecedented capabilities in configuration generation, fault diagnosis, and operational decision-making. From automating complex network configurations that traditionally required specialized expertise to providing natural language interfaces for network management, LLMs demonstrate remarkable potential across diverse networking domains.

Recent research has demonstrated LLM applications across networking domains, including optical network control and lifecycle management [1], [2] and wireless network design and optimization [3], [4]. Beyond domain-specific applications, the broader networking community has embraced LLM-based automation tools, including benchmark frameworks for network-configuration tasks such as NetConfEval [5]. In parallel, telecom-focused evaluation resources and benchmarks are emerging to better characterize LLMs performance under domain-specific requirements [6]. Related work has also integrated LLMs with XAI-enabled anomaly detection for zero-touch service management (ZSM) in 6G microservices, using the LLM to translate numerical explainable artificial intelligence (XAI) outputs into human-readable explanations and to autonomously execute corrective actions for service level agreement (SLA) violations [7].

However, this rapid adoption presents critical challenges. The performance of LLMs vary widely across model families and task types, their energy consumption can be substantial, and reliable evaluation methods that align with human judgment remain elusive. While deploying state-of-the-art LLMs may appear attractive, doing so for routine network tasks often leads to unnecessary computational overhead and energy expenditure, especially when smaller, more efficient ones could deliver comparable performance. Recent benchmarking results indicate that some LLMs may consume over 70 times more energy per query than streamlined alternatives [8], [9], underscoring the need for evaluation frameworks that balance output quality with computational efficiency. Existing evaluation methods offer only partial solutions: text-similarity metrics (e.g., BLEU/ROUGE) capture surface overlap but miss deeper dimensions like correctness and clarity [10], [11], while human evaluations are informative but costly and difficult to scale [12], [13]. Together, these limitations reveal a fundamental gap: the absence of a scalable and trustworthy way to evaluate LLMs that can both reflect human judgment and account for computational efficiency. Bridging this gap is essential to guide the resource-aware adoption of LLMs in network automation.

One approach to address this challenge is LLM-as-a-judge, where an LLM evaluates the outputs of other LLMs using explicit, task-specific rubrics [14]–[17]. This automation significantly reduces the reliance on extensive human annotation, thereby improving scalability and mitigating individual cognitive bias. However, prior work has shown that LLMs employed as evaluators may exhibit reduced agreement with subject-matter experts in tasks requiring specialized domain knowledge. This limitation underscores the importance of careful rubric design and human involvement, particularly in technical domains such as telecommunications [18].

To this end, this paper introduces **HuGLEN** (Human-Grounded Auto LLM Evaluation), a modular framework that bridges scalable LLM judge evaluation with human expert involvement in the assessment of LLMs, with a particular focus on network automation scenarios and use cases where

the quality of LLM outputs directly impacts operational decisions. In this work, we use standardized prompts (instead of task-specific fine-tuning) because task-specific training data are often unavailable in operational networks [1].

HuGLEN employs a small set of human-based evaluations to establish a baseline for LLM performance, which is then leveraged to calibrate an automated evaluation pipeline. This automation significantly reduces the reliance on extensive human annotation, making the evaluation process more scalable and less prone to individual cognitive bias. To guide resource-aware deployment and adoption, HuGLEN introduces the quality efficiency score (QES) concept, a novel metric that jointly considers the quality of LLM-generated outputs and computational efficiency. Jointly, the innovations proposed in HuGLEN enable more efficient and consistent selection of LLMs for network automation tasks, helping to prevent unnecessary resource consumption while ensuring that LLM choices remain grounded in human-relevant criteria. We evaluate the proposed framework in a representative use case, i.e., assessing LLMs used to generate natural language explanations for a quality of transmission (QoT) estimation model in optical networks [13]. Results show that a mid-sized LLM (12B parameters) achieves the highest QES, indicating the best trade-off between explanation quality and computational efficiency for this use case.

II. HuGLEN: Human-grounded Auto LLM Evaluation

This section describes the HuGLEN workflow (Fig. 1). HuGLEN takes as input task data and a task specification, which are converted into a set of standardized prompts via a prompt generation module. The workflow then proceeds through four stages: *(i) LLMs Inference*, where candidate LLMs generate outputs and the human-centered evaluation metrics are defined; *(ii) LLM Judge Selection*, where candidate LLM judges are compared against expert ratings using agreement scores to select a high-agreement LLM judge; and *(iii) Automatic Evaluation*, where the selected LLM judge scores all candidate outputs and the QES is computed to rank the LLMs.

While task-specific adaptation (e.g., instruction tuning or LoRA fine-tuning) may further improve performance [19], [20], it typically requires collecting and curating task data, which can be difficult in operational networks due to data access constraints (e.g., proprietary datasets, privacy), or limited labeling capacity. Therefore, in this work we deliberately focus on prompt engineering and standardized prompting (rather than fine-tuning).

A. Task Definition

The *Task Definition* step specifies the target task, inputs, and a standardized prompting template (via a prompt generation module) to ensure all candidate LLMs are evaluated under identical conditions. The framework itself is domain- and task-agnostic. It can be adapted to any application that leverages an LLM, with the prompt-generation module tailored to the specific objective.

As illustrated in the first (purple) block of Fig. 2, *Task Definition* maps available *data* to a concrete *task* specification and then uses the prompt generation module to convert the task inputs into a set of standardized prompts.

[1]To support reproducibility, we will release the code, prompts, and evaluation artifacts upon acceptance.

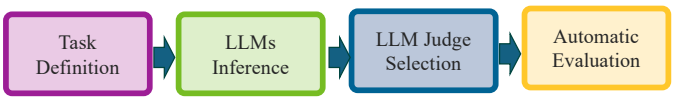


Fig. 1: High-level overview of HuGLEN Framework.

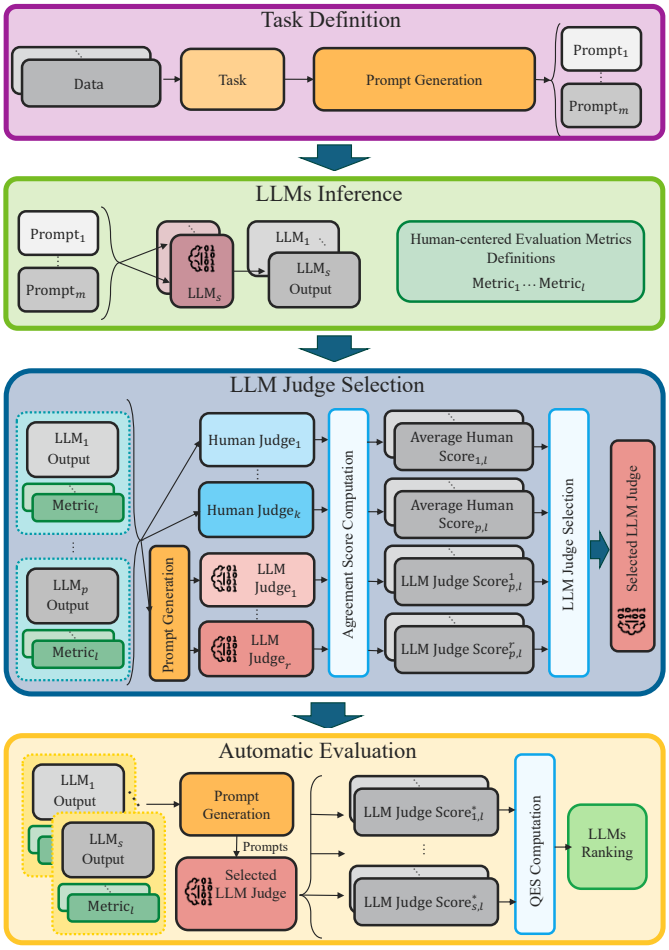


Fig. 2: Detailed workflow of the HuGLEN framework, expanding on the high-level overview in Fig. 1. The diagram illustrates each sequential stage: The task definition formalizes the transformation of input data into structured prompts via the prompt generation module, LLM inference including output generation and definition of human-centered evaluation metrics, LLM judge selection based on agreement with human ratings, and automated evaluation with QES-based ranking of candidate models.

B. LLMs Inference

In *LLMs Inference*, second (green) block in Fig. 1, all candidate LLMs generate outputs for the same set of prompts. We also define a small set of human-centered objective metrics, used consistently by both human evaluators and candidate LLM judges.

C. LLM Judge Selection

In the third stage, HuGLEN identifies the most reliable automated evaluator, referred to as the *LLM judge*, to score the outputs of the candidate LLMs. Here, "reliable" refers to the LLM judge's ability to produce scores that show high agreement with human judgments and remain consistent across different input samples. As illustrated in the third (blue) block of Fig. 2, the process begins with the set of outputs generated by the candidate LLMs in the previous stage. These outputs are first evaluated by multiple human judges using the human-centered metrics defined in Section II-B.

For each candidate LLM, the scores assigned by all human judges are averaged across each evaluation metric, yielding a human reference score per candidate LLMs and per metric. In parallel, the prompt generation module produces standardized prompts for a set of candidate LLM judges, which evaluate the same outputs using the same metrics. Similarly, for each LLM judge, the results are recorded separately for every candidate LLM and for each metric. This setup enables a direct comparison between human and automated evaluations, providing a basis for selecting the most reliable LLM judge. The LLM judge that demonstrates the highest overall alignment with human evaluations is selected to perform the scoring in the final automatic evaluation stage. Alignment can be quantified using different approaches, such as inter-rater agreement measures. This procedure have the potential to maximize consistency with human judgment while substantially reducing the need for manual evaluation. The achieved level of alignment is reported to indicate the reliability of the selected LLM judge.

D. Automatic Evaluation

The final stage of HuGLEN involves automated scoring of candidate LLMs using the selected LLM judge. As shown in the last (yellow) block of Fig. 2, the LLM judge evaluates each candidate LLM's outputs based on the human-centered metrics introduced in Section II-B. These individual metric scores are then aggregated into a single weighted average, yielding an overall quality score for each candidate LLM. By default, HuGLEN assigns equal weights to each metric, but practitioners can customize these weights to emphasize specific evaluation criteria according to their operational priorities.

1) Composite Quality Index (CQI): We define the Composite Quality Index (CQI) as an aggregate quality metric for model i, combining correctness $C_i \in [0,1]$, scope $S_i \in [0,1]$, and usefulness $U_i \in [0,5]$. We first normalize usefulness as $\widehat{U}_i = U_i/5 \in [0,1]$, and then compute

$$\mathrm{CQI}_i = \frac{w_1\, C_i + w_2\, S_i + w_3\, \widehat{U}_i}{w_1 + w_2 + w_3}. \tag{1}$$

By default, $w_1 = w_2 = w_3 = 1$, hence $\mathrm{CQI}_i = (C_i + S_i + \widehat{U}_i)/3$.

2) Quality-Efficiency Score (QES): To jointly rank models by quality and efficiency, we normalize the CQI values across the evaluated model set:

$$\widehat{\mathrm{CQI}}_i = \frac{\mathrm{CQI}_i - \min_j(\mathrm{CQI}_j)}{\max_j(\mathrm{CQI}_j) - \min_j(\mathrm{CQI}_j)}. \tag{2}$$

We then define an efficiency factor based on parameter count P_i as a proxy for inference cost, relative to the largest model $P_{\max}$:

$$\eta_i = \frac{P_{\max}}{P_i}, \tag{3}$$

and normalize it to $[0,1]$ using the smallest model $P_{\min}$:

$$\widehat{\eta}_i = \frac{\eta_i - 1}{\frac{P_{\max}}{P_{\min}} - 1}. \tag{4}$$

Finally, we compute

$$\mathrm{QES}_i = 100\left(\alpha\, \widehat{\mathrm{CQI}}_i + (1-\alpha)\, \widehat{\eta}_i\right), \tag{5}$$

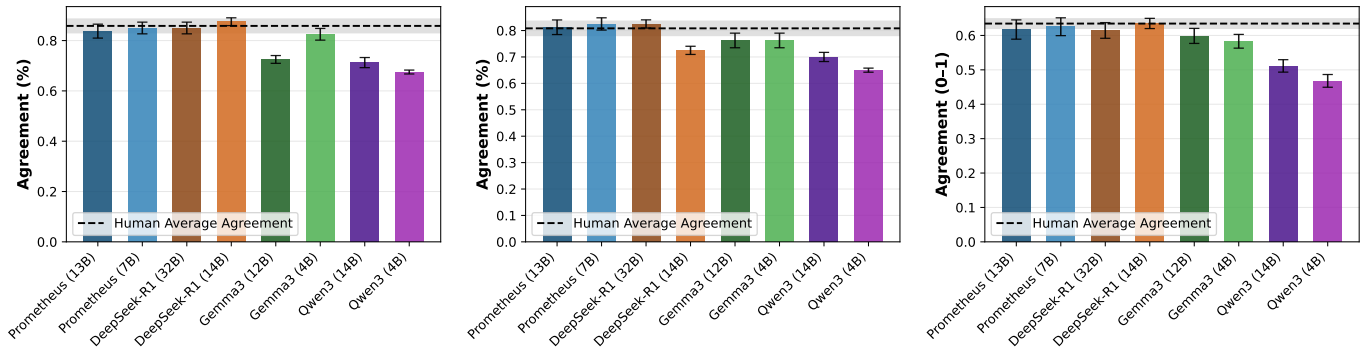
where $\alpha \in [0,1]$ controls the quality–efficiency trade-off (default $\alpha = 0.7$). In deployments, the proxy $\widehat{\eta}_i$ can be replaced by measured latency, energy, or monetary cost.

III. Use Case: AI-based Optical QoT Estimation

We illustrate HuGLEN on a representative network-automation scenario: artificial intelligence (AI)-based QoT estimation for unestablished lightpaths in optical networks. In this setting, the estimated QoT informs resource-allocation decisions that are supervised by an optical network engineer who is typically not an AI specialist.

When a lightpath decision is flagged as unusual (e.g., the predicted QoT contradicts operational expectations), the engineer must understand *why* the AI model produced that output. A common starting point is a SHAP-based local explanation, consisting of SHAP values (local feature importance) and SHAP feature-interaction plots (joint feature effects on the machine learning (ML) prediction), as illustrated in Fig. 5. However, these numerical attributions are difficult to interpret without AI expertise and may be prone to cognitive bias. We therefore use an LLM to translate the SHAP values into operator-facing, human-readable, contextualized explanations. Our expectation is that such contextualized explanations, when analyzed in conjunction with the SHAP values and interactions, will require a level of expertise from the engineer, and result in a better understanding of the predictions.

The translation task just described corresponds to the *Task Definition* stage of HuGLEN. As illustrated in Fig. 4, SHAP values, SHAP interactions, and predictions associated with each lightpath (LP) sample are integrated into a structured prompt through the prompt generation module. These structured prompts serve as inputs for the subsequent stages of HuGLEN.



(a) Agreement of correctness (b) Agreement of scope (c) Agreement of usefulness

Fig. 3: Agreement scores between candidate LLM judges and human evaluators across the three predefined metrics for QoT explanation assessment: (a) correctness, (b) scope, and (c) usefulness. Each plot shows the level of alignment for each LLM judge and average human reference across 40 explanation samples.

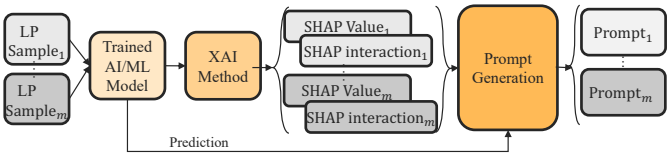
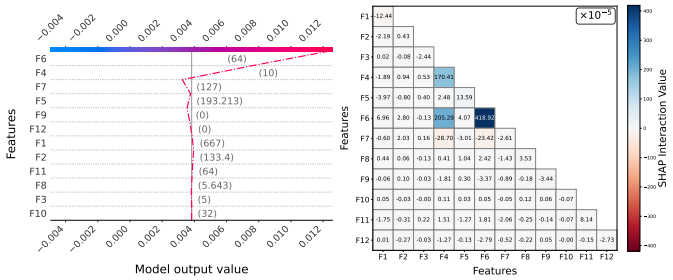


Fig. 4: Use case workflow (Task Definition stage).



(a) SHAP values (b) SHAP interaction

Fig. 5: An example of SHAP values and SHAP interaction plots, representing a local explanation generated by the XAI method for one of the sample lightpaths whose QoT was predicted by the artificial intelligence & machine learning (AI/ML) model.

IV. Experimental and Empirical Results

Building on the use case in Section III, we instantiate the *Task Definition* stage in Fig. 2 for QoT explanation generation. For this purpose, we selected a set of candidate LLMs, including DeepSeek-R1 (1.5B, 14B, 32B), Gemma3 (4B, 12B), and Qwen3 (4B, 14B).

The process begins with training an XGBoost (XGB) regressor model to predict the bit error rate (BER) for various lightpaths using a set of representative features from data samples in [21]. Once the model is trained, we apply SHapley Additive exPlanations (SHAP) as the XAI method [22], producing SHAP values (per-feature contributions) and SHAP interaction values (pairwise feature effects), as illustrated in Fig. 5. We focus on local explanations to support case-by-case analysis of individual lightpaths. The resulting SHAP outputs are embedded (by the prompt generation module) into structured prompts that are provided to candidate LLMs, enabling a direct comparison of their ability to produce operator-facing explanations.

In the *LLMs Inference* phase, all candidate LLMs were prompted with the structured prompts generated in the previous stage, aiming at translating the SHAP values and interaction values into natural language explanations. In terms of human-centered evaluation metrics, we rely on the following metrics [12]: *(i) correctness*, measuring how faithfully interpretations reflect the underlying model explanation; *(ii) scope*, determining if correct interpretations emphasize the most critical aspects; and *(iii) usefulness*, gauging how effectively explanations assist human understanding. Correctness and scope were assessed as binary outcomes, while usefulness was rated numerically from 0 (not helpful) to 5 (extremely helpful).

During the *LLM Judge Selection* stage, we followed a structured process to identify the most reliable LLM judge for automated evaluation. First, we selected DeepSeek-R1 (32B) as the representative LLM from the previous phase and randomly extracted 40 natural language explanations. To minimize labeling effort and ensure a consistent and high-quality benchmark, we relied on a single LLM for this step.

Four human experts independently evaluated the explanations using the three previously defined human-centered metrics (correctness, scope, and usefulness), producing the baseline scores against which we later compared the LLM judges' assessments. Additionally, the experts conducted their evaluations individually and without access to each other's assessments, avoiding potential sources of bias. All participants provided informed consent for the use of their anonymized

evaluations in this study.

Next, we considered a set of candidate LLM judges, including Prometheus (7B-v2.0, 13B-v1.0), DeepSeek-R1 (14B, 32B), Gemma3 (4B, 12B), and Qwen3 (4B, 14B). Prometheus models were specifically designed for evaluation and benchmarking tasks [23].

Then, using the carefully designed prompt generation module, candidate LLM judges were given task descriptions and the extracted explanations and asked to rate them for correctness, scope, and usefulness. Their output evaluations were then compared against the human baseline, using agreement score as the primary selection criterion. For correctness and scope, agreement was computed as the percentage of matching ratings between the LLM judge and human evaluators. For usefulness, we measured the standard deviation of paired ratings, inverted and normalized it to a 0–1 scale so that higher values represent closer alignment [24].

Results showed that Prometheus-7B-v2.0 achieved the highest agreement scores on all three metrics, with Prometheus-13B-v1.0 close behind. These two models were the top performers across metrics, as illustrated in Fig. 3 (a)–(c). DeepSeek-32B fell slightly behind in usefulness agreement (Fig. 3 (c)), whereas DeepSeek-14B showed a noticeable drop in scope agreement (Fig. 3 (b)) despite competitive correctness (Fig. 3 (a)). Finally, based on its balanced and robust agreement with human ratings, Prometheus-7B-v2.0 was selected as the final LLM judge for the automated evaluation stage.

In the *Automatic Evaluation* phase, we leveraged the selected LLM judge (Prometheus-7B-v2.0) to automatically score 100 natural language explanation outputs from each candidate LLM. Fig. 6 summarizes the quality of the natural language explanations produced by LLMs, which are automatically evaluated by the LLM judge (Prometheus-7B-v2.0), across correctness, scope, and usefulness. Results show that most LLMs achieved near-optimal and very similar results for correctness and scope, reflecting the generally strong performance of all candidate LLMs on these metrics. In contrast, the usefulness scores showed a wider spread, providing a clearer basis for differentiating between LLMs.

Finally, to select the LLM that best balances explanation quality and computational efficiency, we computed QES. For this task, explanation quality is considered more important than efficiency, and thus we adopted the default weighting scheme mentioned in Section II-D. Fig. 7 shows the computed QES for all candidate LLMs alongside their number of parameters. Gemma3 (12B) achieved the highest QES, at approximately 78 units, representing the best trade-off between quality and efficiency. Interestingly, DeepSeek-R1 (32B) and Qwen3 (14B) achieved slightly lower scores despite their larger sizes, suggesting diminishing returns in quality relative to their computational cost. In contrast, smaller LLMs such as Gemma3 (4B) and Qwen3 (4B) scored roughly 35–34 units lower, highlighting that their computational advantages come at the cost of substantially reduced output quality. Overall, the results indicate that mid-sized LLMs (10–15B parameters) can offer the most attractive balance between explanation quality and efficiency for this task.

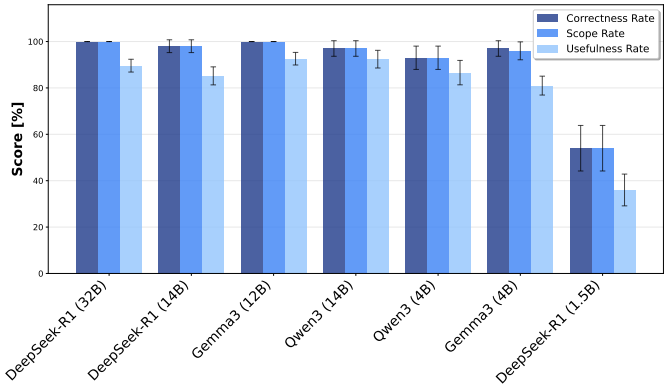


Fig. 6: Comparative performance analysis of candidate LLMs evaluated by Prometheus-7B-v2.0 across correctness, scope, and usefulness rates.

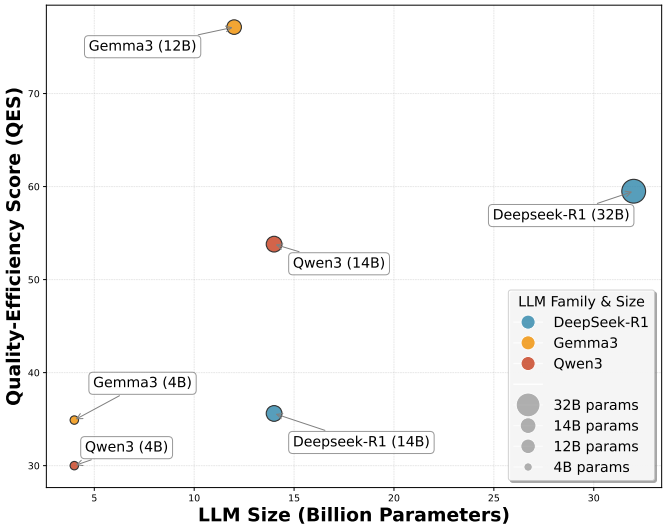


Fig. 7: Quality efficiency score (QES) comparison for candidate LLMs across different number of parameters.

V. Conclusion

In this paper, we introduced HuGLEN, a novel framework for evaluating large language models (LLMs) intended for network automation tasks. HuGLEN enables the automatic selection of the most trustworthy and resource-efficient LLM for a given task, leveraging human-centered evaluation metrics to ensure both quality and practical relevance. The proposed quality efficiency score (QES) allows to balance the LLM output quality with computational efficiency and promote sustainable, evidence-driven deployment decisions.

We demonstrated HuGLEN's effectiveness in a practical scenario, where the goal was to generate natural language explanations based on XAI interpretations of quality of transmission (QoT) predictions in optical networks. In this setting, LLMs translated the technical outputs of machine learning

(ML) models into clear, operator-friendly explanations that help network engineers better understand and trust automated decisions.

Experimental results showed that our approach can identify LLMs that provide strong explanatory performance while minimizing resource consumption. Notably, Gemma3 (12B) was found to deliver the best trade-off between quality and efficiency, as measured by the quality efficiency score (QES), while our automated LLM judge selection process reduced the need for extensive manual evaluation.

Acknowledgements

This work was supported by the Celtic-Next Flagship SUSTAINET-Advance project funded in Sweden by VINNOVA (2025-02987) and in Switzerland by Innosuisse (No. 119.588 INT-ICT). The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

[1] 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*, vol. 63, no. 6, pp. 90–96, 2025.

[2] X. Liu, Q. Qiu, Y. Zhang, Y. Cheng, L. Yi, W. Hu, and Q. Zhuge, "First field trial of LLM-powered AI agent for lifecycle management of autonomous driving optical networks," in *Optical Fiber Communication Conference (OFC)*. Optica Publishing Group, 2025, p. Th1A.2.

[3] K. Qiu, S. Bakirtzis, I. Wassell, H. Song, J. Zhang, and K. Wang, "Large language model-based wireless network design," *IEEE Wireless Communications Letters*, vol. 13, no. 12, pp. 3340–3344, 2024.

[4] P. Wu, T. Wang, Y. Zhong, H. Zhang, Z. Wang, and F. Wang, "Deepform: Reasoning large language model for communication system formulation," 2025. [Online]. Available: https://arxiv.org/abs/2506.08551

[5] C. Wang, M. Scazzariello, A. Farshin, S. Ferlin, D. Kostić, and M. Chiesa, "Netconfeval: Can llms facilitate network configuration?" *Proc. ACM Netw.*, vol. 2, no. CoNEXT2, Jun. 2024. [Online]. Available: https://doi.org/10.1145/3656296

[6] S. Lee, D. Arya, S.-M. Cho, G.-e. Han, S. Hong, W. Jang, S. Lee, S. Park, S. Sek, I. Song, S. Yoon, and E. Davis, "Telbench: A benchmark for evaluating telco-specific large language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Miami, Florida, US: Association for Computational Linguistics, Nov. 2024, pp. 609–626. [Online]. Available: https://aclanthology.org/2024.emnlp-industry.45/

[7] A. Mekrache, M. Mekki, A. Ksentini, B. Brik, and C. Verikoukis, "On combining XAI and LLMs for trustworthy zero-touch network and service management in 6G," *IEEE Communications Magazine*, vol. 63, no. 4, pp. 154–160, 2025.

[8] N. Jegham, M. Abdelatti, L. Elmoubarki, and A. Hendawi, "How hungry is ai? benchmarking energy, water, and carbon footprint of llm inference," 2025. [Online]. Available: https://arxiv.org/abs/2505.09598

[9] J. Fernandez, C. Na, V. Tiwari, Y. Bisk, S. Luccioni, and E. Strubell, "Energy considerations of large language model inference and efficiency optimizations," 2025.

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[11] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[12] O. Ayoub, S. Troia, C. Natalino, C. Rottondi, D. Andreoletti, F. Lelli, S. Giordano, and P. Monti, "Natural language interpretability for ML-based QoT estimation via large language models," *International Conference on Transparent Optical Networks (ICTON)*, p. Tu.C2.4, 2025.

[13] K. Rezaei, O. Ayoub, S. Troia, F. Lelli, P. Monti, and C. Natalino, "Generative explainability for next-generation networks: LLM-augmented XAI with mutual feature interactions," in *2025 21th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2025, pp. 1–6.

[14] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-judge with MT-bench and chatbot arena," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[15] A. Maatouk, K. C. Ampudia, R. Ying, and L. Tassiulas, "Tele-LLMs: A series of specialized large language models for telecommunications," 2025. [Online]. Available: https://arxiv.org/abs/2409.05314

[16] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2511–2522. [Online]. Available: https://aclanthology.org/2023.emnlp-main.153/

[17] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, and I. Stoica, "Chatbot arena: An open platform for evaluating LLMs by human preference," in *Proceedings of the 41st International Conference on Machine Learning*, vol. 235. PMLR, 2024, pp. 8359–8388. [Online]. Available: https://proceedings.mlr.press/v235/chiang24b.html

[18] A. Szymanski, N. Ziems, H. A. Eicher-Miller, T. J.-J. Li, M. Jiang, and R. A. Metoyer, "Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks," in *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. New York, NY, USA: Association for Computing Machinery, 2025, pp. 952–966.

[19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.

[20] C. Xin, Y. Lu, H. Lin, S. Zhou, H. Zhu, W. Wang, Z. Liu, X. Han, and L. Sun, "Beyond full fine-tuning: Harnessing the power of LoRA for multi-task instruction tuning," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, May 2024, pp. 2307–2317. [Online]. Available: https://aclanthology.org/2024.lrec-main.206/

[21] 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.

[22] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[23] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo, "Prometheus 2: An open source language model specialized in evaluating other language models," 2024.

[24] A. S. Thakur, K. Choudhary, V. S. Ramayapally, S. Vaidyanathan, and D. Hupkes, "Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges," in *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, K. Dhole and M. Clinciu, Eds. Vienna, Austria and virtual meeting: Association for Computational Linguistics, Jul. 2025, pp. 404–430. [Online]. Available: https://aclanthology.org/2025.gem-1.33/