

First Experimental Demonstration of Full Lifecycle Automation of Optical Network through Fine-Tuned LLM and Digital Twin

Chenyu Sun^{(1,2,3)*}, Xin Yang^(1,4), Nicola Di Cicco⁽⁴⁾, Reda Ayassi⁽¹⁾, Venkata Virajit Garbhapu⁽¹⁾, Photios A. Stavrou⁽²⁾, Massimo Tornatore⁽⁴⁾, Gabriel Charlet⁽¹⁾, Yvan Pointurier⁽¹⁾

(1) Huawei Paris Research Center, Boulogne-Billancourt, France, *chenyu.sun1@huawei.com

(2) Eurecom, Sophia Antipolis, France

(3) Sorbonne Université, Paris, France

(4) Politecnico di Milano, Milan, Italy

Abstract We experimentally demonstrate the full lifecycle of a multi-OMS network leveraging AI-Agents with LLM and a digital twin. We achieve 100% accuracy of management API-calling, 7x speed-up alarm-log analysis, and 83% reduction in computation resources through LoRA fine-tuning. ©2024 The Author(s)

Introduction

Generative AI, powered by transformer-based Large Language Models (LLMs) [1], can automate a broad spectrum of tasks by comprehending and generating natural language text. In autonomous optical networks, LLM-based AI-agents can significantly reshape network management (e.g., configure network parameters, monitor performance metrics, and diagnose issues in real-time...). This level of automation not only alleviates the operational burden on engineers but also significantly reduces the risk of human error, leading to more robust and resilient network performance.

Past research aimed at a) automating network configuration by mapping Application Programming Interfaces (APIs) calls based on user intent [2]-[6] and b) alarm log analysis automation [7]-[9]. LLM-based operation can be seen as advanced intent-based networking, but utilization of a public cloud for the largest models raises concerns about data privacy. Additionally, high-accuracy calling is yet to be achieved.

In this paper, we demonstrate the first fine-tuned LLM for automating an optical network during the 4 key steps of its lifecycle: design, deployment, maintenance (including power re-equalization and troubleshooting), and upgrade. By deploying and fine-tuning LLMs *locally*, we avoid

using a public cloud and automate a wide range of tasks on a testbed based on commercial optical network products, such that network operation is automated, secure, accurate, and fast.

Building AI-Agent for specific domain/task

Our framework leverages prompt engineering + retrieval-augmented generation (RAG) (Fig. 1a, top) and fine-tuning (Fig. 1a, bottom). Different tasks require different skills, so we implement a multi-agent framework where specific tasks are allocated to dedicated agents (Fig. 1b).

Prompt engineering and RAG

Prompt engineering involves crafting input prompts to maximize the effectiveness and accuracy of LLMs' responses. *RAG* leverages external knowledge bases (e.g., technical documentation) to combine retrieval-based methods with generation-based models to produce more accurate and contextually relevant responses. Our demonstration leverages both.

Fine-tuning the model

An LLM pre-trained on a generic knowledge base can be fine-tuned to learn a new, specific domain through model weight adjustment. Here, we employ Low-Rank Adaptation (LoRA) [10], which lowers hardware requirements by re-training only a fraction of the LLM parameters.

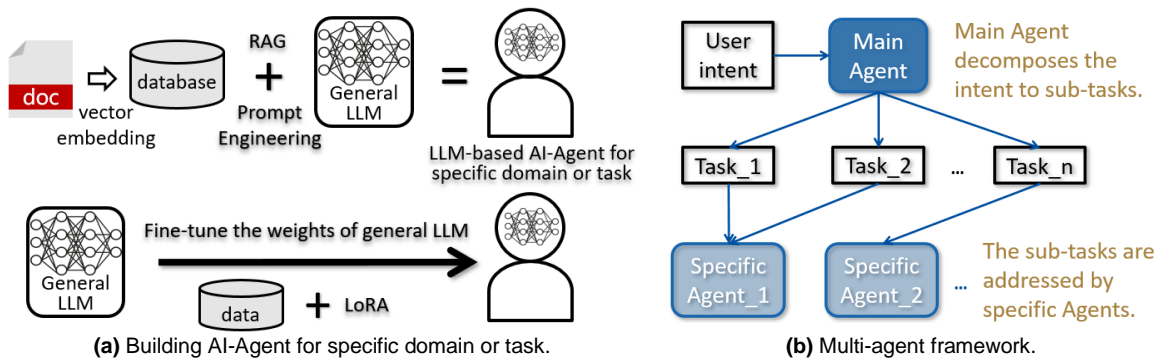


Fig. 1: (a) Per domain AI-Agent and (b) multi-agent framework.

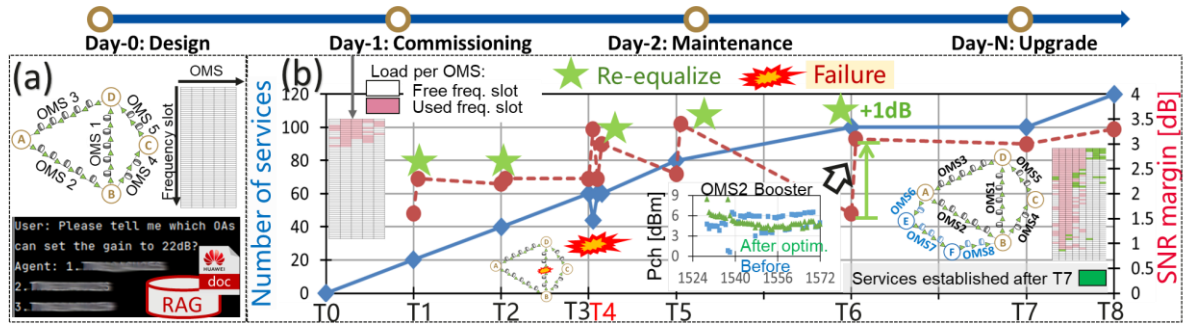


Fig. 2: Lifecycle management and automation.

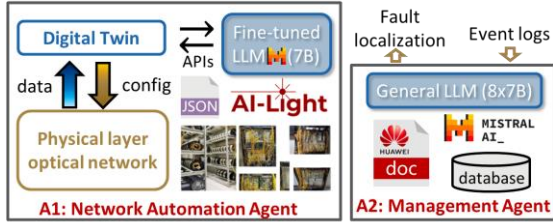


Fig. 3: Proposed AI agents.

Experimental Setup

Hardware: Fig. 2 depicts the workflow and processing of day-0 to day-N management and automation based on our testbed [11] with 8 OMS (6 THz C band, 25 spans, 1980 km of fiber) and 1 real-time transponder (used at 400Gb/s PCS-16QAM, 100 GHz spacing) for SNR measurement. Channel loading is emulated with an Amplified Spontaneous Emission (ASE) source.

Software: We locally deploy open-source Mistral AI LLMs (7B-Instruct [12], 8x7B-Instruct [13]) on a multi-GPU (256G) server. Then, we build two task-specific AI-Agents: A1 for network automation, fed with API descriptions for interaction with our AI-Light digital twin [14], and A2 for network management, which embeds product documentation for system design and log analysis, Fig. 3.

D-0: Network design

With RAG, we integrate the documentation of our optical products, AI-Agent A2 then becomes an expert in commercial products, who knows all their specifications. We use A2 to select the available devices based on digital twin-designed network configurations to build a 5-OMS network; see Fig. 2a, which illustrates AI-assisted optical amplifier selection.

D-1: Network deployment: service establishment and commissioning

We then perform service commissioning. The agent for network automation A1 translates service human language requests into valid predefined APIs in JSON format (compliance to standards is possible, but is out of the scope of this paper). The SDN controller then calls the APIs to implement the intent.

Specifically, we use our custom SDN API `add_och(source, destination, path, frequency, bandwidth, board_mode)` to add a (possibly multi-

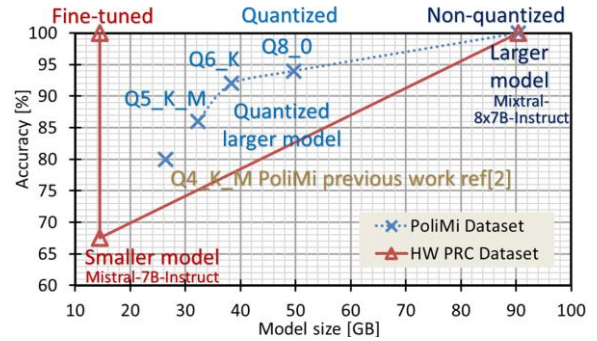


Fig. 4: API-calling accuracy of different models and datasets. The different quantization precisions are used for PoliMi's datasets (blue). Models without quantization are used for Huawei Paris Research Center (HW PRC) datasets (red).

OMS) service in our testbed. The “path” and “frequency” may be explicit, or assigned to names as “shortest-path” and “first-fit” of allocation algorithms implemented in our digital twin. Channel bandwidth and modulation format can also be explicit (e.g., “PCS-16QAM” for 400Gb/s in 100GHz) or set to “auto” (i.e., chosen by the digital twin).

We load a new batch of 20 services each (see slot occupation in Fig. 2b) at T1, T2, T3, T5, T6; we measure SNR for 20% of the services.

AI-Agent A1 should call the correct APIs with 100% accuracy. However, there is a trade-off between performance and general LLM size, which relates to computation resource requirements. For instance, quantizing the LLM parameters is a strategy for reducing the model's size and inference costs. [2] achieved 80% accuracy by using Mistral-8x7B-Instruct with 4-bit quantization. Based on PoliMi's dataset (limited to 50 queries [2]), we test the model accuracy with different levels of quantization. Fig. 4 depicts the obtained model size/accuracy trade-off in blue.

To improve accuracy with smaller model size, we perform LoRA fine-tuning on Mistral-7B-Instruct for 5 APIs including add/delete service, estimate/measure quality of transmission (QoT), and equalization. We generate an augmented dataset with a general purpose LLM of 10k/API for fine-tuning; all 50000 queries are available in [15]. The loss curve is shown in Fig. 5.

We reduce the model size by 83% from 90.4 GB to 14.5 GB while maintaining 100% accuracy in API calls, as shown in Fig. 4 (red). Multi-

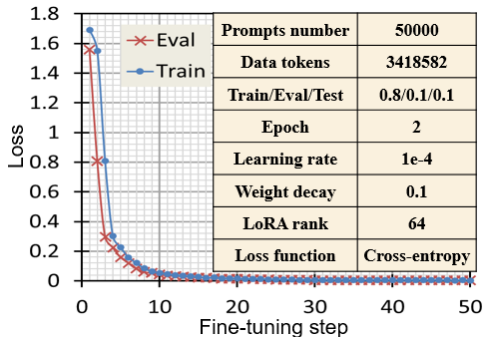


Fig. 5: Loss curve of fine-tuning Mistral-7B-Instruct.

language support is shown in Fig. 6. We test the model for syntax errors and mixing of multiple languages in the same query; Fig. 7 shows such robustness can be achieved with fine-tuning.

D-2: Network maintenance: re-equalization, troubleshooting and re-configuration

During the network lifecycle, the power profile varies due to the gain variation of amplifiers after loading new services. To avoid the resulting SNR margin degradation, we use periodic service launch power re-equalization as in [16].

By chatting with AI-Agent A1, we add services with “set & forget” fixed channel power (5 batches of 20 services). We periodically re-equalize to improve SNR (up to 1dB margin improvement is measured at T6). The *measured* network SNR margin before and after re-equalization for each of the 5 batches is shown in Fig. 2b (jumps in the red dashed line), and the digital twin *predicted* SNR gains for all services over network life in Fig. 8(inset). Some services may degrade but the network margin always improves.

The digital twin performs prediction and estimation of SNR before and after re-equalization, the pdf of the digital twin SNR prediction accuracy (prediction–measured) is shown in Fig. 8 (RMSE=0.3dB). Additionally, link failures may occur over time, requiring troubleshooting and

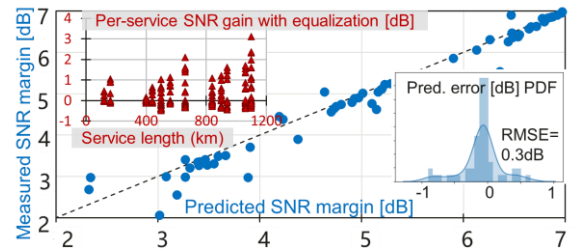


Fig. 8: SNR margin gain and digital twin prediction accuracy.

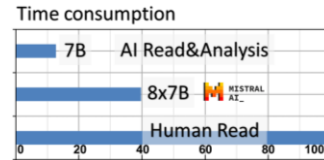


Fig. 9: Log analysis speed-up with AI agents.

network reconfiguration. As a natural language processing (NLP) technique, the LLM is highly effective in alarm log analysis. We unplug a connector to emulate a link failure on OMS1 (thereby 16 services dropping) at T4 and collect the logs, which are then analyzed by Agent A2. As shown in Fig. 9, by comparing with the reading speed of human beings [17], Agent A2 performs alarm log reading AND analysis 7x faster than a human would parse the logs (let alone analyze them).

D-N: Network upgrade

To avoid blocking at end-of-life (T7), we add 3 OMSes and use Agent A1 to automate the addition of 20 services in the 8-OMS network, using the same method as in D-2, reaching a total of 120 services established, see Fig. 2b (right).

Conclusion

This is the first experimental demo of a fine-tuned LLM for digital twin optical network automation to enable non-specialists to operate a network. We present use cases for lifecycle management and automation that are 7x faster, 100% accurate, and decrease hardware memory usage by 83%.

```

User:Can you add a 400G service from A to D 🇬🇧
Agent:{"function": "add_och", "arguments": {"source": "A", "destination": "D", "path": "shortest-path", "frequency": "first-fit", "bandwidth": 100000, "board_mode": "PCS-16QAM-H3.0"}}
User:Je voudrais un service de A à D en utilisant la fréquence centrale de 196,5 THz 🇫🇷 Add a service from A to D with central frequency 196.5 (FR: 196,5) THz
Agent:{"function": "add_och", "arguments": {"source": "A", "destination": "D", "path": "shortest-path", "frequency": 196500000, "bandwidth": "auto", "board_mode": "auto"}}
User:Ich möchte einen optischen Kanal von A nach D über 'OMS2-OMS1' mit 200Gbit/s hinzufügen. 🇩🇪 Add a 200G service from A to D through 'OMS2-OMS1'
Agent:{"function": "add_och", "arguments": {"source": "A", "destination": "D", "path": "OMS2-OMS1", "frequency": "first-fit", "bandwidth": 75000, "board_mode": "QPSK"}}
User:Quiero añadir un servicio de A a D a través de 'OMS2-OMS4-OMS5' con 400Gbps 🇪🇸 Add a 400G service from A to D with through 'OMS2-OMS4-OMS5'
Agent:{"function": "add_och", "arguments": {"source": "A", "destination": "D", "path": "OMS2-OMS4-OMS5", "frequency": "first-fit", "bandwidth": 100000, "board_mode": "PCS-16QAM-H3.0"}}
User:Voglio aggiungere un servizio da C ad A con 200Gbps 🇮🇹 Add a 200G service from C to A
Agent:{"function": "add_och", "arguments": {"source": "C", "destination": "A", "path": "shortest-path", "frequency": "first-fit", "bandwidth": 75000, "board_mode": "QPSK"}}
User:Θέλω να προσθέσω μια υπηρεσία από το Α στο Δ με 400Gbps 🇬🇷 Add a 400G service from A to D
Agent:{"function": "add_och", "arguments": {"source": "A", "destination": "D", "path": "shortest-path", "frequency": "first-fit", "bandwidth": 100000, "board_mode": "PCS-16QAM-H3.0"}}

```

Fig. 6. Multi-language requests for adding a service. Based on the fine-tuned Mistral-7B-Instruct model, the Agent responds to the requirements in different European languages and calls the correct API of our network controller.

```

User:Plz ajouter a svc from A to D de 400G
User doesn't indicate the slot, so that "first-fit" should be used to avoid blocking
Agent:{"function": "add_och", "arguments": {"source": "A", "destination": "D", "path": "shortest-path", "frequency": "196625000", "bandwidth": 100000, "board_mode": "PCS-16QAM-H3.0"}}
User:Plz ajouter a svc from A to D de 400G
Agent:{"function": "add_och", "arguments": {"source": "A", "destination": "D", "path": "shortest-path", "frequency": "first-fit", "bandwidth": 100000, "board_mode": "PCS-16QAM-H3.0"}}

```

Fig. 7 “Please add a service from A to D at 400G”: Accurate answer even asking questions with typos (mixing of languages). Top: no fine-tuning; bottom: with fine-tuning.

References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *arXiv*, 2017. DOI: <https://doi.org/10.48550/arXiv.1706.03762>
- [2] Nicola Di Cicco, Memedhe Ibrahim, Sebastian Troia, Francesco Musumeci, and Massimo Tornatore, "Open Implementation of a Large Language Model Pipeline for Automated Configuration of Software-Defined Optical Networks," *50th European Conference on Optical Communication (ECOC)*, Frankfurt, Germany, 2024.
- [3] Anni Zhou, Yuchen Song, Yao Zhang, Min Zhang, and Danshi Wang, "Large Language Model-Driven AI Agent in SDN Controller Towards Intent-Based Management of Optical Networks," *50th European Conference on Optical Communication (ECOC)*, Frankfurt, Germany, 2024.
- [4] Daniel Adanza, Carlos Natalino, Lluís Gifre, Raul Muñoz, Pol Alemany, Paolo Monti and Ricard Vilalta, "IntentLLM: An AI Chatbot to Create, Find, and Explain Slice Intents in TeraFlowSDN," *2024 IEEE 10th International Conference on Network Softwareization (NetSoft)*, Saint Louis, USA, 2024. DOI: <https://doi.org/10.1109/NetSoft60951.2024.10588917>
- [5] Vignesh Karunakaran, Carlos Natalino, Behnam Shariati, Piotr Lechowicz, Johannes Karl Fischer, Achim Autenrieth, Paolo Monti, and Thomas Bauschert, "TAPI-based Telemetry Streaming in Multi-domain Optical Transport Network," *Optical Fiber Communications Conference and Exhibition (OFC)*, San Diego, USA, 2024. DOI: <https://doi.org/10.1364/OFC.2024.M3Z.9>
- [6] Ehsan Etezadi, Carlos Natalino, Vignesh Karunakaran, Renzo Diaz, Anders Lindgren, Stefan Melin, Achim Autenrieth, Lena Wosinska, Paolo Monti, Marija Furdek, "Demonstration of DRL-based intelligent spectrum management over a T-API-enabled optical network digital twin," *49th European Conference on Optical Communications (ECOC)*, Glasgow, UK, 2023. DOI: <https://doi.org/10.1049/icp.2024.1801>
- [7] Xiaotian Jiang, Min Zhang, Yuchen Song, Yao Zhang, Yidi Wang, Cheng Ju, and Danshi Wang, "OptiComm-GPT: a GPT-based versatile research assistant for optical fiber communication systems," *Optical Express (OE)*, vol. 32, no.12, p.p. 20776-20796, 2024. DOI: <https://doi.org/10.1364/OE.522026>
- [8] Yidi Wang, Chunyu Zhang, Jin Li, Yue Pang, Lifang Zhang, Min Zhang, and Danshi Wang, "AlarmGPT: an intelligent alarm analyzer for optical networks using a generative pre-trained transformer," *Journal of Optical Communications and Networking (JOCN)*, vol. 16, no. 6, p.p. 681-694, 2024. DOI: <https://doi.org/10.1364/JOCN.521913>
- [9] Danshi Wang, Yidi Wang, Xiaotian Jiang, Yao Zhang, Yue Pang, and Min Zhang, "When Large Language Models Meet Optical Networks: Paving the Way for Automation," *arXiv*, 2024. DOI: <https://doi.org/10.48550/arXiv.2405.17441>
- [10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv*, 2021. DOI: <https://doi.org/10.48550/arXiv.2106.09685>
- [11] Chenyu Sun, Reda Ayassi, Xin Yang, Gabriel Charlet, Photios A. Stavrou, and Yvan Pointurier, "Demonstration of LLM-based AI-Agent for Optical Network Management and Automation," *50th European Conference on Optical Communication (ECOC)*, Frankfurt, Germany, 2024.
- [12] Mistral AI, Mistral-7B-Instruct-v0.3, <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>
- [13] Mistral AI, Mixtral-8x7B-Instruct-v0.1, <https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1>
- [14] Alessio Ferrari, Venkata Virajit Garbhapu, Dylan Le Gac, Ivan F. de Jauregui Ruiz, Gabriel Charlet, and Yvan Pointurier, "Demonstration of AI-Light: an Automation Framework to Optimize the Channel Powers Leveraging a Digital Twin", *Optical Fiber Communications Conference and Exhibition (OFC)*, San Diego, USA, 2022. DOI: <https://doi.org/10.1364/OFC.2022.M3Z.14>
- [15] ECOC2024_LLM_API_Demo, GitHub, https://github.com/csun19/ECOC2024_LLM_API_Demo
- [16] Xin Yang, Alessio Ferrari, Dylan Le Gac, Gabriel Charlet, Massimo Tornatore and Yvan Pointurier, "Experimental Impact of Power Re-Optimization in a Mesh Network", *Journal of Optical Communications and Networking (JOCN)*, vol. 15, no. 7, p.p. C20-C28, 2023. DOI: <https://doi.org/10.1364/JOCN.482298>
- [17] Marc Brysbaert, "How many words do we read per minute? A review and meta-analysis of reading rate," *Journal of Memory and Language*, vol.109, 2019. DOI: <https://doi.org/10.1016/j.jml.2019.104047>