Teaching “old” LLMs new tricks

Urs Baumann 30.05.2025

Urs Baumann

- Network Automation since 1503
- MSc Artificial Intelligence

chatGPT

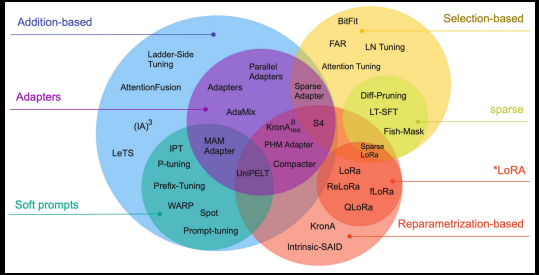
What is fine-tuning

- Fine-tuning image generation with art style

- https://stabilityai.notion.site/Stable-Diffusion-3-Medium-Fine-tuning-Tutorial-17f90df74bce4c62a295849f0dc8fb7e

Fine-tuning Methods Taxonomy

- arXiv:2303.15647
- Submitted on 28 Mar 2023 (v1), last revised 22 Nov 2024
- Addition-based
- Selection-based
- Reparametrization-based



https://arxiv.org/abs/2303.15647

LoRA

- **Lo**w-**R**ank **A**daptation of Large Language Models
- arXiv:2106.09685, 2021, Microsoft
- `https://github.com/microsoft/LoRA`
- Freezes the pretrained model weights
- Adding low-rank matrices into each transformer layer

Why LoRA?

- Minimizing memory consumption
- Enabling fine-tuning on hardware with limited resources
- Learning task-specific transformations without altering original model structure

Once upon a time ...

Preparing the stage

The network was run by CLI Cowboy's

chatGPT

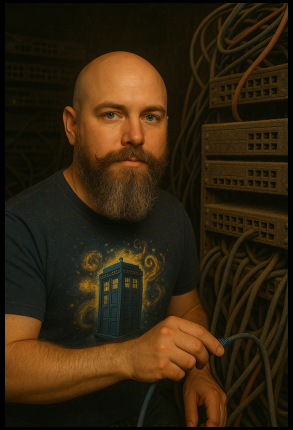
Preparing the stage

Shooting network configs
from the hips.

chatGPT

Preparing the stage

Network engineers working on the neglected network were dreaming about ...



chatGPT

Preparing the stage

...a modern declarative network.

chatGPT

Preparing the stage

And take overtime

chatGPT

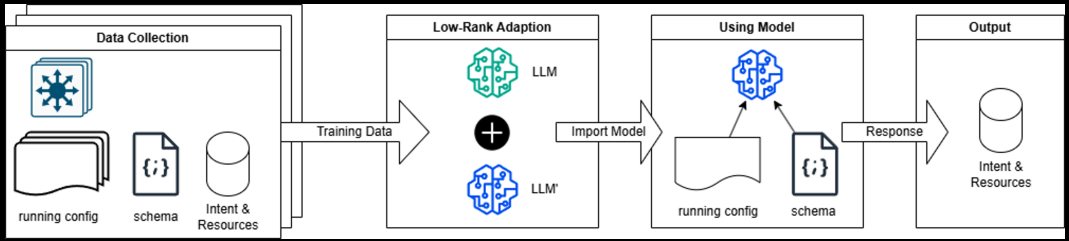
Preparing the stage

However, they needed to extract the intent and resources from the inconsitent network configuration.

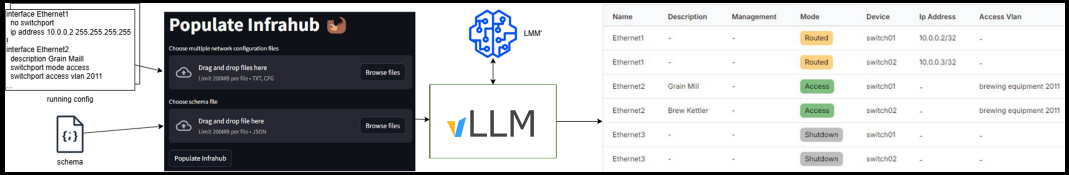
chatGPT

Intent and Resource Extraction from Network Device Configuration

Fine-tuning Pipeline



Using new model



Evaluation

`Qwen2.5-Coder-32B-Instruct`

- Data Extraction Match: 0.86
- JSON Validity: 1.0
- F1:

`OpenAI GPT-4.0`

- Data Extraction Match: 0.09
- JSON Validity: 0.98

Challanges

- Confusion between similar schemas
- IP Address + mask to CIDR
-

Lessons

- One model for each schema
- Generating parser template
-

Tool Stack Fine-tuning Pipeline

- unsloth
- vLLM
- MLflow

Thank you!