Generative AI for Intent-Based Networking

draft-cgfabk-nmrg-ibn-generative-ai-01

Montreal, Nov 2025, IETF 124

Pietro Cassara'
Alberto Gotta
CNR-ISTI

Giuseppe Fioccola
Aldo Artigiani
Huawei Technologies

Riccardo Burrai
Emiljan Kolaj
SIRIUS Technology

Marco Martalo'
Virginia Pilloni
University of Cagliari

Background

This document describes how to specialize AI models in order to create highly targeted generative models for Intent-Based Networking (IBN)

The objective is to:

- explain how transfer learning techniques can be adopted to design generative AI specialized models for IBN.
- describe tools, such as Low-rank Adaptation (LoRA), for achieving efficient and scalable transfer learning in data networks for designing targeted generative models for IBN.
- propose a framework for the rapid adaptation and composition of specialized knowledge, addressing the challenges identified in draft-irtf-nmrg-ai-challenges and enabling the dynamic, multi-domain use cases of draft-irtf-nmrg-ibn-usecases.

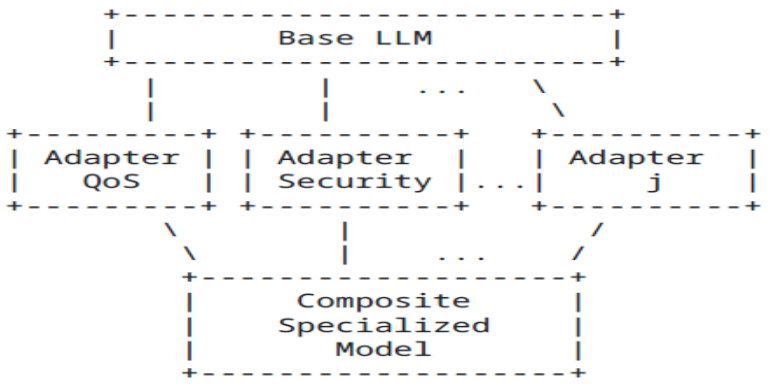
Specializing Generative AI

LLMs may take a long time and numerous resources to set up.

LoRA adds low-rank matrices into the layers of pre-trained models and adjusts the weights of the original models to get results specific to the context.

- It allows fine-tuning with significantly fewer parameters.
- It reduces the storage and computational footprint, enabling deployment of specialized models even on resource-constrained systems.

- ❑ A Hub is a structured repository where specialized adapters are stored, indexed, and shared. It enables efficient reuse and modularity. The architecture typically includes:
  - Adapter storage
  - Metadata index (domain, use case, version)
  - Dependency tracking for composite models
- ❑ Flow refers to the systematic combination of multiple adapters to form a new, composite model capable of handling complex, multi-domain intents.



Lifecycle of Specialized Models in IBN

Model Generation, Evaluation, and Deployment

- Generation: Fine-tuning on domain datasets
- Evaluation: Accuracy, latency, resource profiling
- Deployment: Containerized adapters for on-demand loading

Feedback Loops and Continuous Adaptation

In-network telemetry provides real-time feedback on policy effectiveness, enabling further fine-tuning or adapter updates, aligning with the adaptive model requirements.

Benchmarking Specialized Models: Accuracy, Latency, and Resource Consumption

Example of a Logical Architecture for Model Management

1. Intent Reception: User or application submits a high-level intent through the Intent Interface.

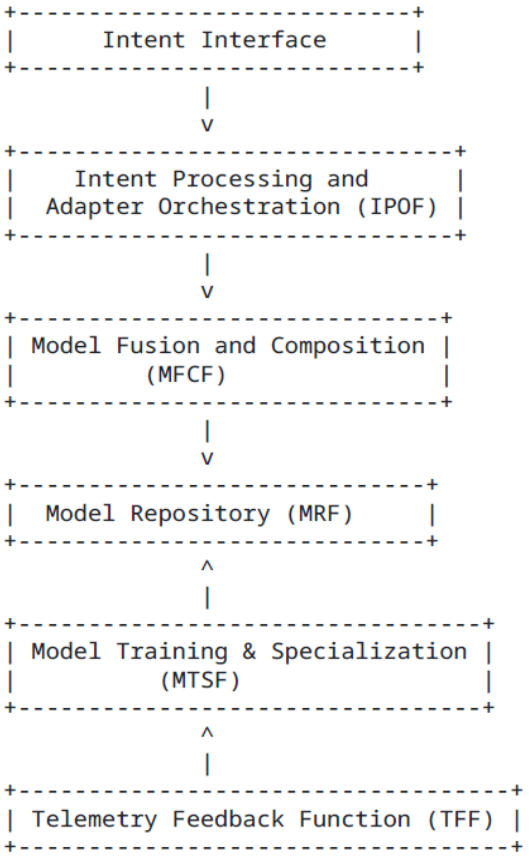
2. Adapter Orchestration: IPOF parses the intent, queries the MRF for relevant adapters, and orchestrates dynamic fusion via MFCF.

3. Model Fusion: MFCF composes adapters into a composite model tailored to the intent.

4. Deployment: The composite model is deployed into the IBN system to generate specific configurations.

5. Telemetry Feedback: The TFF monitors the applied configurations' impact on network KPIs and feeds insights back to IPOF and MTSF.

6. Re-Specialization: If performance deviates from expected thresholds, MTSF initiates LoRA re-training or fine-tuning using new data collected via TFF.



Changes from -00 to -01

New Section on "Network Digital Twin Ontology"

- The proposed generative AI–enabled IBN architecture can be empowered with the Network Digital Twin (NDT):
  - It provides a multi-layered abstraction that connects service requirements with traffic parameters and physical infrastructure, supporting both top-down intent translation and bottom-up feedback from telemetry data.
  - It operates as a continuously learning system, capable of dynamically optimizing network behavior, validating intents before deployment, and adapting to evolving conditions.

New coauthors: Marco and Virginia from University of Cagliari

Next Steps

Comments are welcome!

Thank You