IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 February 2026, accepted 6 March 2026, date of publication 10 March 2026, date of current version 24 March 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3672699

RESEARCH ARTICLE

AGORA: Agentic Green Orchestration Architecture for Beyond 5G Networks

RODRIGO MOREIRA[1], (Member, IEEE), LARISSA FERREIRA RODRIGUES MOREIRA[1], (Member, IEEE), MAYCON LEONE MACIEL PEIXOTO[2], AND FLÁVIO DE OLIVEIRA SILVA[3], (Senior Member, IEEE)

[1]Institute of Exact and Technological Sciences, Federal University of Viçosa (UFV), Viçosa 36570-900, Brazil
[2]Institute of Computing, Federal University of Bahia (UFBA), Salvador 40170-110, Brazil
[3]Department of Informatics, ALGORITMI Centre, University of Minho (UMinho), 4800-058 Guimarães, Portugal

Corresponding author: Rodrigo Moreira (rodrigo@ufv.br)

This work was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, under Finance Code 001; and in part by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) Ministry of Science, Technology, Innovations and Communications (MCTIC)/Brazilian Internet Steering Committee (CGI) Research Project SFI2—Slicing Future Internet Infrastructures and Fundação para a Ciência e Tecnologia (FCT) within the Research and Development (RD) Units Project of Centro ALGORITMI under Grant 2018/23097-3. The work of Rodrigo Moreira was supported by Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) under Grant APQ00923-24.

ABSTRACT Effective management and operational decision-making for complex mobile network systems present significant challenges, particularly when addressing conflicting requirements such as efficiency, user satisfaction, and energy-efficient traffic steering. The literature presents various approaches aimed at enhancing network management, including the Zero-Touch Network (ZTN) and Self-Organizing Network (SON); however, these approaches often lack a practical and scalable mechanism to consider human sustainability goals as input, translate them into energy-aware operational policies, and enforce them at runtime. In this study, we address this gap by proposing the AGORA: Agentic Green Orchestration Architecture for Beyond 5G Networks. AGORA embeds a local tool-augmented Large Language Model (LLM) agent in the mobile network control loop to translate natural-language sustainability goals into telemetry-grounded actions, actuating the User Plane Function (UPF) to perform energy-aware traffic steering. The findings indicate a strong latency-energy coupling in tool-driven control loops and demonstrate that compact models can achieve a low energy footprint while still facilitating correct policy execution, including non-zero migration behavior under stressed Multi-access Edge Computing (MEC) conditions. Our approach paves the way for sustainability-first, intent-driven network operations that align human objectives with executable orchestration in Beyond-5G infrastructures.

INDEX TERMS Agentic, B5G, intents, energy-aware, MEC, monitoring.

I. INTRODUCTION

Future mobile network architectures are required to deliver cognitive services to end users while simultaneously ensuring sustainable network management [1], [2]. The provision of cognitive services allows users to utilize Artificial Intelligence (AI) capabilities according to their needs at the network edge [3], [4], [5]. Although Beyond Fifth Generation (B5G) necessitates complex closed-loop systems for management and orchestration, it will drive innovation across industries and smart societies by offering not only unprecedented network Key Performance Indicators (KPIs) capabilities but also human-centered features and applications [6], [7]. This evolution presents challenges for management and operations across various network segments, particularly as modern mobile networks incorporate Space-Air-Terrestrial-Sea Integrated Networks (SATSI) and facilitate the Internet of Smart Things [8], [9]. Nevertheless, AI has enhanced the robustness of network management in addressing complex, multi-operator, and vendor-related challenges. Yet,

The associate editor coordinating the review of this manuscript and approving it for publication was Abbas Kiani.

turning high-level operational goals, especially sustainability goals, into AI-native, safe, measurable, and executable actions across the data plane remains a largely unresolved challenge [6], [7], [10].

In the literature, we found different approaches to manage complex mobile network architectures, as expected in 6th Generation Mobile Network (6G). 3rd Generation Partnership Project (3GPP) standardized Self-Organizing Network (SON) mechanisms, and the field has since progressed toward Zero-Touch Network (ZTN) and Intent-Based Networking (IBN), reshaping how network operations requirements are specified and automated [11], [12], [13], [14], [15], [16]. As autonomy increases, the control logic shifts from static policies toward distributed agent ecosystems. More recently, a new paradigm has been emerging: the Internet of Agents (IoA). The IoA establishes a distributed framework in which autonomous, AI-driven agents interconnect through standardized protocols to collaborate, exchange knowledge, and orchestrate workflows across decentralized environments [17], [18]. In this context, IoA refers to the interconnected multi-agent ecosystem, whereas agentic characterizes the autonomous reasoning and goal-driven behavior exhibited by individual agents within that ecosystem. A common thread across these approaches is the pursuit of higher autonomy through standardized closed loops and intent-based specifications [19]. However, sustainability objectives are often treated as secondary constraints, and they rarely translate into concrete, telemetry-grounded control actions at the 5th Generation Mobile Network (5G) core data plane.

Problem Statement: Current intent-based and zero-touch frameworks lack a practical mechanism to translate human sustainability goals into verifiable, tool-executed User Plane Function (UPF) actions that are continuously grounded in real-time telemetry.

This gap becomes more pronounced in Beyond 5G deployments, where decisions must span heterogeneous domains, including the data plane, Multi-access Edge Computing (MEC), and the 5G core, under time-varying loads. Leveraging an Large Language Model (LLM), the agentic framework can perform reasoning and planning, use tools, manage memory, and collaborate with other agents, thereby enhancing network management with unprecedented capabilities. To the best of our knowledge, existing approaches do not directly actuate the 5G UPF to achieve green network management objectives. Consequently, this study proposes AGORA: Agentic Green Orchestration Architecture for Beyond 5G Networks, which integrates the local deployment of LLMs into data plane management. This integration aims to translate aspirations for greener network management into complex network management tasks, thereby augmenting the established ZTN or SON management approaches. Building on this design, we implemented an end-to-end prototype. We evaluated it under controlled MEC stress to quantify the energy footprint, policy compliance, and User Equipment (UE) perceived Quality of Service (QoS).

The main contributions of this paper are as follows: (i) an agentic LLM driven closed-loop architecture that translates natural language intents into telemetry-grounded tool calls and UPF routing actions for sustainable B5G management; (ii) a policy compliance evaluation methodology for tool-augmented agents, including an intent suite with phrasing variations, tool and action compliance indicators, and energy and UE QoS metrics aligned to decision windows; (iii) an end-to-end prototype to create controlled MEC energy asymmetry; and (iv) an experimental comparison of multiple local LLMs, including a non-English capable model, quantifying energy footprint, latency, and migration behavior under an energy threshold policy. Accordingly, our evaluation focuses on whether AGORA closes this gap by (i) grounding decisions on telemetry via tool calls and (ii) enforcing the resulting green policy directly at the UPF under MEC stress.

The remainder of this paper is organized as follows: Section II contrasts our approach with those in the literature, while Section III presents our method. Section IV describes our evaluation testbed, and Section V presents the results and discussions of our approach. Finally, we present concluding remarks and possible research directions in Section VI.

II. RELATED WORK

The orchestration of B5G and 6G networks has recently transitioned toward intent-based and autonomous paradigms, driven by LLMs and agentic architectures. This section categorizes recent advancements into objective-driven orchestration, intent-to-action translation, and agentic control loops.

The selection of references in this study follows a strict inclusion protocol focused on the operational pillars of B5G management. Specifically, we prioritize works that contribute to: (i) intent-based and zero-touch orchestration frameworks; (ii) autonomous agentic control loops; (iii) UPF and MEC dynamic steering; and (iv) standardized sustainability metrics for energy-aware networking.

A. ZERO-TOUCH AND OBJECTIVE-DRIVEN ORCHESTRATION

Early efforts in green edge computing established the foundation for objective-driven frameworks. Guim et al. [20] introduced a multi-tier lifecycle management system using Kubernetes Custom Resource Definitions (CRDs) to translate performance targets into scaling actions. Similarly, Barrachina Muñoz et al. [21] operationalized Zero-Touch Management (ZSM) through the MonB5G architecture, employing a Monitor-Analyze-Decide-Execute (MAPE) loop for slice reconfiguration. While these works facilitate automated management, they primarily rely on deterministic policies or classical Machine Learning (ML) and lack the flexibility of LLM-based reasoning in complex, multi-domain environments. Decentralized approaches, such as the SCHE2MA framework [22], utilize Reinforcement Learning (RL) to balance latency and energy consumption, yet often

treat resource descriptors and reward weights as opaque parameters.

B. LLM-ENABLED INTENT TRANSLATION AND OSS INTEGRATION

A significant body of work explores LLMs as intermediaries between natural language intents and network configurations. Dandoush et al. [23] and Tzanakaki et al. [24] proposed hierarchical agent workflows and transformer-based translators to map user requirements into standards-based descriptors, such as Open Source MANO (OSM) and YANG templates. In the realm of Operations Support Systems (OSS), Mekrache et al. [15], [25] developed GPT-based interfaces that decompose intents into ordered API calls across heterogeneous domains. While effective for provisioning, these approaches prioritize the initial deployment and intent reconciliation as seen in the MAESTRO [26] and AIORA [27] architectures rather than continuous, telemetry-grounded closed-loop control.

C. AGENTIC CONTROL LOOPS AND SUSTAINABILITY

The latest frontier involves ''agentic'' orchestration, where LLMs autonomously invoke tools to manage live infrastructure. Recent frameworks like Brodimas et al. [28] and Elkael et al. [29] leverage tool-calling mechanisms and the Model Context Protocol (MCP) to execute Kubernetes and protocol-stack commands. Chatzistefanidis et al. [30], [31], [32] further extended this to Open RAN environments, using agent graphs to enforce blueprints via R1/E2 interfaces.

Although these works introduce sophisticated reasoning, sustainability metrics, such as real-time carbon intensity or power efficiency, are often treated as secondary constraints or omitted entirely. Notable exceptions include Habib et al. [33], [34] and Chergui et al. [35], who integrate RL, Digital Twins, and risk metrics to optimize power consumption; however, their focus remains mainly on Radio Access Network (RAN) scheduling rather than End-to-End (E2E) sustainability orchestration.

Dev et al. [36] outline a multi-agent architecture for next-generation networks, emphasizing Radio Access Network (RAN) cloudification, serverless orchestration, and constrained AI for energy efficiency. Their work proposes a broad 6G framework and validates Agentic AI through conceptual V2X scenarios. Our approach instead targets concrete enforcement at the UPF: AGORA closes the intent-to-action loop through native tool-calling tied to real-time testbed telemetry, yielding measured energy and QoS outcomes rather than high-level orchestration or simulation results.

Unlike prior intent-based and zero-touch frameworks that mainly stop at descriptor generation or control plane reconfiguration, AGORA closes the loop with verifiable UPF-level data-plane actions driven by live telemetry. It translates intents into native tool calls and enforces traffic steering at the UPF with intent/tool/action compliance gating to ensure safe execution.

D. RESEARCH GAP

Table 1 presents a comparative analysis of the proposed architecture with the current state-of-the-art architecture, evaluated across five critical technical pillars. We denote (●) when the approach achieves the feature and (○) when it does not. The ''LLM Agentic'' column distinguishes frameworks that employ autonomous reasoning agents from those restricted to passive intent translation or deterministic logic. The ''Native Tool-Calling'' criterion identifies systems capable of direct, programmatic interaction with live infrastructure Application Programming Interfaces (APIs) via function calls, rather than merely generating static deployment descriptors. The ''Telemetry Feedback'' criterion evaluates the presence of a dynamic closed-loop integration that utilizes real-time time-series data to inform agentic decision-making. Additionally, the ''E2E B5G Scope'' assesses whether the orchestration encompasses the entire Beyond-5G landscape, including Core and MEC domains, rather than being confined to isolated segments such as the RAN. Finally, the ''Sustainability Focus'' indicates whether energy, power, or carbon metrics are prioritized as primary orchestration objectives rather than secondary constraints.

In contrast to the cited literature, our approach unifies real-time observability with an agentic tool-calling framework specifically optimized for E2E sustainability. While prior efforts focus on intent-to-descriptor translation or narrow-domain closed loops, our work operationalizes an LLM agent that ingests high-fidelity telemetry and fault-injection signals to autonomously call functions across Kubernetes, 5G Core, and MEC interfaces. This positioning ensures that power, energy, and carbon efficiency are treated as first-class objectives guiding the automated re-orchestration process.

III. PROPOSED METHOD

Mobile networks increasingly require energy-aware closed-loop orchestration that can translate high-level operational goals into concrete, real-time actions across heterogeneous infrastructures. Figure 1 summarizes AGORA, the end-to-end workflow of our sustainable agentic orchestration in the 5G testbed. In Phase (0), the user expresses a high-level sustainability intent that defines a green policy goal, according to Table 2. In Phase (1), a controlled workload is injected into the MEC environment using a chaos tool to emulate realistic [37], time-varying operating conditions and create energy asymmetry across MEC sites, where MEC2 is Graphics Processing Unit (GPU)-enabled for cognitive services and MEC1 is Central Processing Unit (CPU)-only as the greener target. In Phase (2), the agentic controller queries the monitoring stack to estimate the current system state, including the power and performance signals. In Phase (3), the agent enforces the policy by actuating the 5G data plane at the UPF and updating the routing rules toward the greener MEC target based on the observed telemetry.

TABLE 1. Comparison of related approaches across key capabilities.

Approach	LLM Agentic	Native Tool-Calling	Telemetry Feedback	E2E B5G Scope	Sustainability Focus
Guim et al. [20]	○	○	●	○	●
Dalgkitsis et al. [22]	○	○	●	○	●
Dandoush et al. [23]	●	○	○	●	○
Maestro [26]	●	○	○	○	○
MonB5G [21]	○	○	●	●	○
Tzanakaki et al. [24]	○	○	○	●	○
DMO-GPT [15]	●	●	○	●	○
OSS-GPT [25]	●	●	○	○	○
Brodimas et al. [28]	●	●	○	○	○
Symbiotic [30]	●	○	●	○	○
MX-AI [31]	●	●	●	○	○
AGORAN [32]	●	○	○	○	○
AgentRAN [29]	●	●	●	○	○
AIORA [27]	○	○	○	●	○
Habib et al. [33], [34]	○	○	●	○	●
Chergui et al. [35]	●	○	○	○	○
Dev et al. [36]	●	○	○	●	●
Our Proposal	●	●	●	●	●

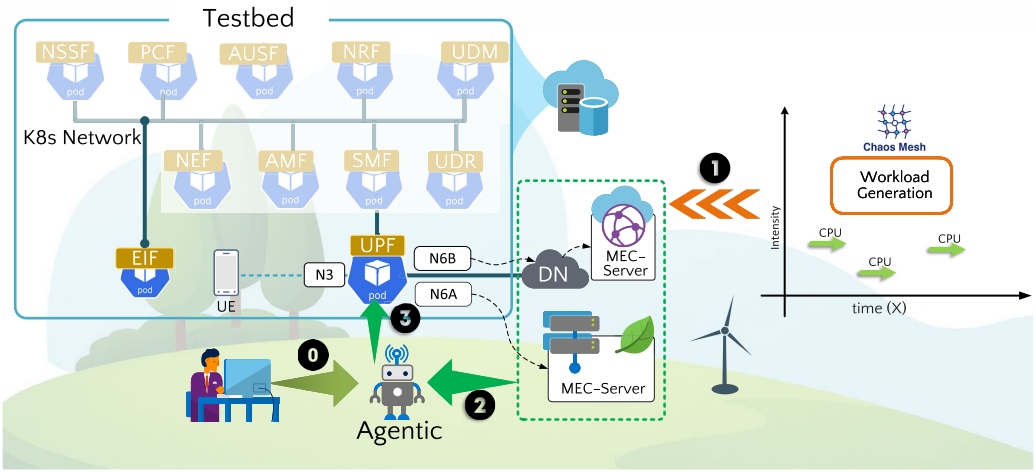


FIGURE 1. Agentic closed-loop sustainability orchestration in a 5G MEC testbed.

A. OVERVIEW AND RESEARCH GOAL

This work evaluates whether a tool-augmented local LLM can act as an agentic network manager in a B5G control loop. Here, "agentic" refers to reliably translating natural-language intents into executable actions via telemetry-grounded tool use and verifiable UPF actuation, with correctness defined by the ability to (i) interpret user intents, (ii) select and invoke the appropriate external tools in the correct sequence, and (iii) enforce an energy policy through UPF routing updates. The policy rule is intentionally simple to keep decision semantics fixed and isolate the non-trivial part of the problem, intent interpretation, tool use, and action execution, from policy design and testbed-specific factors, which are reported separately.

A deterministic threshold script can enforce the same rule once the telemetry source, threshold, and actuation API are set. However, it does not address intent variability, multilingual phrasing, or failures in tool selection and sequencing, which are central to our evaluation and are quantified using tool/action compliance and the oracle baseline.

Our method specifies the decision loop, the tool interfaces, the intent suite, the compliance criteria, and the derived metrics. Implementation and deployment details such as cluster configuration, UPF realization, telemetry stack versions, and hardware resources are described in the experimental testbed section to avoid conflating algorithmic design with platform-specific factors.

B. AGENTIC CONTROL LOOP

We model network management as a closed loop with four stages: intent ingestion, tool-based state estimation, policy evaluation, and actuation. Let t be a decision step triggered by a user request. The agent receives a natural language

intent I_t, queries measurements through tools to obtain an observed state $\hat{s}_t$, evaluates a policy $\pi(\hat{s}_t, I_t)$, and issues an action a_t that updates the UPF target.

The actuation space is discrete and corresponds to selecting the serving MEC site (according to Equation 1):

$$a_t \in \{\texttt{route\_to\_MEC1}, \texttt{route\_to\_MEC2}\}. \quad (1)$$

Although we evaluate a binary action for clarity, the same loop extends to K MECs by selecting $m^{\text{sel}} \in \{1, \ldots, K\}$ from telemetry and policy constraints, and can support partial steering by parameterizing actuation (e.g., split weights or per-flow rules) in the UPF tool interface.

Our method is implemented with a chat-oriented LLM bound to two executable tools:

- `energy_mean_last_time(mec, delta)` returns an energy proxy for the requested MEC, derived from infrastructure telemetry.
- `upf_set_target(mec)` updates the UPF egress selection so that traffic is routed toward the chosen MEC.

In standards-compliant deployments, the agentic tool can be backed by energy telemetry exposed by the 3GPP Energy Information Function (EIF) [38], ensuring that power/energy KPIs are obtained via the management plane rather than ad hoc instrumentation.

To enable consistent cross-log analysis, each intent execution is associated with a decision interval $\Delta_t = [t_t^{\text{start}}, t_t^{\text{end}}]$ covering the LLM inference and any tool invocations. All infrastructure telemetry and UE probing samples are aligned to Δ_t by timestamp, using the nearest-neighbor matching when sampling periods differ.

C. INTENT SUITE AND CLASSIFICATION

To probe the robustness of energy-aware tool selection via AGORA reasoning and policy enforcement, we used a compact suite of user intents that varied in phrasing, urgency, and constraint expression. Each intent encodes a threshold-based policy and requires the agent to consult measurements before making a decision. The four intents used in the evaluation are listed in Table 2.

TABLE 2. User prompt phrasing variations for energy-aware traffic migration.

Type	User Instruction Prompt.
Threshold	*Monitor MEC2 power. If $> \theta$, migrate to MEC1.*
Policy-based	*Is there any green policy violation on MEC2? If energy > 20 W, trigger migration.*
Contextual	*Compare MEC power and move traffic if MEC2 usage is too high.*
Urgent	*Execute migration to MEC1 now if MEC2 usage > 20 W.*

We denote by $\theta(I_t)$ the threshold implicitly specified by the intent text. In our suite, $\theta(I_t)$ equals θ for the baseline threshold intent and equals 20 W (empirically defined in AGORA as a power threshold) for the policy-based and urgent variants. This allows the same compliance rule to be evaluated across heterogeneous phrasings. The value 20 W was empirically calibrated for our testbed to separate nominal from stressed MEC2 operation under the induced workloads while still yielding near-threshold cases that require telemetry consultation. We emphasize that θ is a policy parameter expected to vary with hardware and load; changing θ mainly shifts the migration rate (lower θ triggers earlier/more frequent migrations, higher θ reduces them), without altering our main qualitative findings on latency-energy coupling and the need for correct tool/action compliance.

AGORA can be extended to non English energy aware prompts, for example Portuguese [39], by relying on multilingual LLMs or a lightweight translation and normalization step before intent parsing, which makes the approach scalable across languages and LLM backends. In this study, however, we evaluated the mapping from higher level user phrasing to correct tool usage and green compliant actions using English prompts only, leaving a systematic Portuguese evaluation as future work.

D. WORKLOAD INDUCTION AND ENERGY ASYMMETRY

The evaluation requires measurable, repeatable differences in energy consumption across MEC sites. Therefore, we induced asymmetric load episodes over MEC workloads, producing stress intervals with timestamps. The stress generator applies a randomized CPU load with different parameter ranges per MEC, yielding a sustained higher load on MEC2 and lighter or shorter stress on MEC1. Each stress event (e_i) was logged according to Equation 2.

$$e_i = \langle \texttt{MEC}_i, \texttt{begin}, \texttt{end}, \texttt{cpu\_load}, \texttt{workers} \rangle \quad (2)$$

The resulting stress log is later time-aligned with the agent decision traces to interpret behavior under stressed operating conditions.

This procedure does not assume that AGORA has direct access to the stress schedules. Instead, the agent must infer the operational condition from telemetry and reasoning insights obtained via tools and monitoring platforms, or even 3GPP EIF.

E. POLICY AND DECISION COMPLIANCE

An energy-aware policy defines the target behavior. Let $\hat{P}_2(t)$ be the observed power proxy for MEC2 obtained by the tool at decision step t, and θ be the threshold specified by the intent. The policy compliance rule is given by Equation 3.

$$a_t = \begin{cases} \texttt{route\_to\_MEC1} & \text{if } \hat{P}_2(t) > \theta, \\ \texttt{route\_to\_MEC2} & \text{otherwise.} \end{cases} \quad (3)$$

Because intent phrasing can be ambiguous, we evaluate compliance at two levels:

- Tool compliance, whether the agent invokes energy measurement tools before acting.

- Action compliance, whether the resulting UPF target matches the threshold-based rule implied by the intent.

An execution is considered valid only if the agent completes the loop in the order of state estimation, followed by actuation. Actuation without prior measurement is considered noncompliant, even if the final action coincidentally matches the threshold rule.

F. NON-LLM BASELINE FOR VERIFICATION

To verify AGORA's performance without the bias of model-specific stochasticity, we define a *Deterministic Oracle Baseline*. This baseline represents a theoretical non-generative controller that executes the policy in Eq. (3) with 100% tool and action compliance. It serves as a benchmark for the best-case execution of a single fixed policy, against which we compare the LLM backends' ability to handle intent variability and multilingual prompts while maintaining operational correctness.

This oracle is equivalent to a traditional automation script that polls the telemetry and applies a fixed UPF steering rule without any language reasoning. We used it as a non-LLM baseline to contextualize latency, energy, QoS, and compliance when compared to the LLM backends.

G. QUALITY OF SERVICE OBSERVATION AT THE USER EQUIPMENT

To quantify the user-perceived impact of AGORA actions, the UE continuously probes the active (MEC_i) target selected by the UPF. Each probing window logs the round-trip latency and User Datagram Protocol (UDP) quality indicators, producing records according to Equation 4.

$$u_j = \langle t_s, t_e, \texttt{target\_MEC}, \texttt{ping\_avg\_ms}, \texttt{udp\_jitter\_ms}, \texttt{udp\_loss\_pct} \rangle. \quad (4)$$

These measurements were later aligned with agent decisions to analyze how policy-driven migration affects perceived service quality.

Each UE probing record u_j is mapped to the most recent UPF target observed at the start of its window, enabling the attribution of latency and UDP quality to the active energy-aware routing choice during that time interval.

H. EXPERIMENTAL PROCEDURE FOR MODEL COMPARISON

We evaluate AGORA with multiple LLM backends as the energy-aware decision engine. For each model M, we repeat R independent runs and execute the same intent suite $\mathcal{I}$ to estimate the average behavior under stochastic system conditions. As summarized in Algorithm 1, each run initializes the agent and binds the external tools, then iterates over intents by (i) issuing the intent, (ii) querying telemetry via tool calls to estimate the current state, (iii) selecting the policy action, and (iv) actuating the UPF. Finally, we persist the decision traces and align them with MEC energy telemetry and UE probing logs for cross-model aggregation.

Algorithm 1 Agentic Policy Evaluation per Model

Input: Model M, intent set $\mathcal{I}$, repetitions R, threshold θ
for $r \leftarrow 1$ **to** R **do**
  Initialize the agent with model M and bind tools
  foreach $I_t \in \mathcal{I}$ **do**
    Provide I_t to the agent
    Query telemetry via tool calls to obtain $\hat{P}_1(t)$ and $\hat{P}_2(t)$
    Apply the policy $\pi(\hat{P}_2(t), \theta)$ and select a_t
    Execute a_t by calling `upf_set_target`
    Log tool calls, inference latency, token counts, and action outcome
  Collect MEC energy telemetry snapshots aligned to the run
  Collect UE probing records during the run window
Aggregate metrics across R runs for model comparison

I. MEASURED VARIABLES AND DERIVED METRICS

The method logs the decision time, token usage, tool call traces, and infrastructure energy proxies. For each AGORA intent execution, we recorded our local LLM inference time T_t and observed power proxy $\hat{P}_m(t)$ for the active MEC m. We derive the energy consumed by the serving MEC during decision execution using Equation 5.

$$E_t^{\text{MEC}} = \hat{P}_m(t) \cdot T_t, \quad (5)$$

where m equals MEC1 when migration is active; otherwise, it equals MEC2. We also compute the energy normalized by the output volume when applicable, according to Equation 6.

$$E_t^{\text{token}} = \frac{E_t^{\text{gpu}}}{N_t^{\text{gen}}}, \quad (6)$$

where E_t^{gpu} is the GPU energy proxy during the decision interval and N_t^{gen} is the number of generated tokens.

Finally, policy compliance is summarized using two binary indicators per intent execution, as defined in Equation 7.

$$C_t^{\text{tool}} \in \{0, 1\}, \quad C_t^{\text{act}} \in \{0, 1\}. \quad (7)$$

Here, $C_t^{\text{tool}} = 1$ if the AGORA invokes the telemetry measurement tools before acting, and $C_t^{\text{act}} = 1$ if the resulting UPF target matches the threshold-based decision rule implied by the intent at step t. The resulting dataset enables a unified comparison of decision quality, energy footprint, and user-perceived QoS impact across AGORA LLM backends.

The above definitions specify the AGORA independently of the deployment. The experimental testbed section details the platform used to instantiate MEC sites, collect telemetry, execute workload induction, and run UE probing. In contrast, the present section defines how decisions, compliance, and energy and QoS metrics are operationalized and compared across models.

IV. EXPERIMENTAL TESTBED

Here, we describe the testbed used to instantiate and evaluate AGORA in a container-based 5G core. We summarize the compute platform, the 5G core with two MEC sites interconnected through a controllable UPF, the local LLM serving stack and evaluated models, and the monitoring instrumentation. We then describe the workload induction, UE QoS probing, and repeated run procedures used to ensure reproducible comparisons.

A. REPRESENTATIVENESS AND ASSUMPTIONS

Our testbed was designed to be operator-like while remaining completely reproducible. The deployment follows a cloud-native model with Kubernetes-orchestrated network functions, a 5G core with a programmable UPF anchoring the user plane, and a telemetry-driven closed loop, which reflects common directions in operator deployments of CNF-based cores and edge platforms. The evaluated workflow (telemetry → decision → UPF steering → UE probing) mirrors the practical control path required for intent-based management to generate observable data-plane effects.

Simultaneously, we acknowledge that the environment is controlled and does not capture all dimensions of a production operator network. In particular, our setup simplifies (i) scale (single cluster and limited number of MECs), (ii) workload diversity (synthetic stress episodes to induce repeatable energy asymmetry), and (iii) multisite heterogeneity (e.g., diverse hardware generations, transport variability, and background traffic mix). These assumptions align with our research goal of isolating the latency–energy cost and compliance behavior of tool-augmented LLMs in a closed loop. Therefore in, weterpret the results as evidence of the feasibility and trade-offs of UPF-level enforcement under telemetry-grounded control, and we outline larger-scale and more diverse validations as future work.

B. INFRASTRUCTURE AND COMPUTE PLATFORM

All experiments were conducted on the FABRIC testbed [40] using an OpenStack-based compute node equipped with an AMD EPYC 7543 CPU and 64 GiB of system memory. The system exposes an NVIDIA A30 GPU with 24GB of device memory for LLM inference. The platform runs a Kubernetes cluster with both client and server versions at v1.28.15 that orchestrates the 5G core functions, MECs, the monitoring stack, and workload components.

C. 5G CORE AND MEC SETUP

We deploy free5GC, a 5G core, to provide end-to-end connectivity between the radio access and data networks. The user plane is anchored at a UPF that exposes a simple control interface used by the agent to update the active routing target. Two MEC instances, denoted MEC1 and MEC2, are deployed as separate pods and are reachable from the UPF through its data network interfaces. In the considered scenario, MEC1 represents the preferred green execution site, whereas MEC2 represents the stressed and non-green site used to evaluate policy compliance under adverse conditions. In our setup, MEC2 is GPU-enabled to serve user cognitive workloads, whereas MEC1 is CPU-only and represents the greener but less capable execution site.

D. AGENT EXECUTION AND LLM SERVING

The AGORA decision engine is implemented as a tool-augmented chat LLM that runs locally and issues actions to the network through external tools. Model inference was performed using vLLM with an OpenAI-compatible endpoint, enabling a consistent invocation interface across models and facilitating tool calling during inference. The serving configuration enables automatic tool selection and structured tool-call parsing, allowing the model to request telemetry and perform actuation within the same decision loop.

The evaluated models are:

- **Qwen2.5 1.5B Instruct** [41]: Selected as the primary compact baseline due to its state-of-the-art performance in structured tool-calling at a small parameter scale. It is utilized to evaluate whether a lightweight model, optimized for function-calling protocols, can maintain high action compliance within the AGORA control loop with minimal energy overhead.
- **Mistral 7B Instruct v0.2** [42]: Represents a mid-size, general-purpose instruction baseline. This model is used to assess if higher parameter counts and broader reasoning capabilities translate into better intent-to-action mapping, or if they primarily contribute to increased inference latency without improving UPF actuation accuracy.
- **Phi 3.5 Mini Instruct** [43]: An efficient, small-footprint model optimized for long-context and reasoning. It serves to test the trade-off between rapid token generation and the precision required for binary decision compliance in time-sensitive MEC migration scenarios.
- **OLMoE 1B/7B Instruct** [44]: A Mixture-of-Experts (MoE) model chosen to evaluate the impact of sparse activation on energy proportionality. By activating only a fraction of its parameters (1B out of 7B) per token, it allows for a direct comparison of infrastructure energy savings versus dense model architectures during the decision loop.

E. ENERGY AND TELEMETRY INSTRUMENTATION

Infrastructure telemetry is collected through Kepler, which exposes container-level energy-related metrics to Prometheus. The monitoring stack continuously scrapes both the Kepler and vLLM metrics endpoints, enabling the synchronized collection of MEC power proxies and inference-side counters, such as generated tokens, prompt tokens, and request inference time. All measurements were timestamped and later aligned with the agent decision

windows to compute the derived energy metrics, such as active MEC and GPU energy proxies.

To ensure consistent access to telemetry during automated runs, Prometheus is accessed via a local port forward, and vLLM exposes its metrics and health endpoint on the local host. This configuration supports automated batch execution across repeated cold-start runs while maintaining a stable metric collection interface.

F. WORKLOAD INDUCTION AND STRESS LOGGING

To create a measurable energy asymmetry between MEC sites, we injected synthetic CPU stress into MEC pods using Chaos Mesh [45]. The stress generator selects a target MEC and applies randomized load parameters, producing prolonged higher load intervals on MEC2 and shorter, lighter intervals on MEC1. Each stress episode is recorded in a structured log with start and end timestamps, load level, worker count, and duration. This log is used for post hoc contextualization of agent decisions, whereas the agent itself has no direct access to the stress schedule and must rely on telemetry obtained through the tool calls.

G. USER EQUIPMENT QOS PROBING

The user-perceived impact is measured at the UE by periodic probing of the current UPF routing target. The probing process repeatedly queries the target selection interface, extracts the active target IP address, and measures the round-trip time using `ping`. In addition, it optionally runs UDP probing with `iperf3` to estimate the jitter and loss when available. All UE records are timestamped and stored in a Comma-Separated Values (CSV) log, allowing alignment with the agent decisions and infrastructure telemetry during analysis.

H. EXPERIMENTAL EXECUTION PROCEDURE

We induce controlled and repeatable load asymmetry across MEC sites using Chaos Mesh by injecting CPU stress episodes into the Kubernetes pods that host MEC services. Each episode is defined by a target pod, CPU load percentage, number of worker threads, and duration, and is applied as a `StressChaos` resource. The generator enforces a single active stressor at a time by cleaning up any existing `StressChaos` objects before scheduling the next episode. It logs every injected event, including begin and end timestamps, to a CSV file. This produces a sustained higher load on MEC2 and a lighter load on MEC1, creating a consistent operating regime to evaluate whether agent decisions respond to measurable stress without exposing the stress schedule to the agent.

For each model, we executed multiple independent runs to capture the average behavior under stochastic system conditions and cold-start effects. At the beginning of each run, the vLLM server was started for the selected model, and the monitoring endpoints were verified via health checks. The agent is then prompted with the intent suite, and it may issue tool calls to query the MEC power and update the UPF target. After each run, logs are persisted, including the AGORA decision traces, MEC energy CSV snapshots, and UE QoS measurements. The vLLM process is terminated between runs to enforce cold-start conditions and reduce cross-run cache effects, enabling a fair comparison across models.

TABLE 3. Summary of main KPIs per model.

Model	Energy (J)	Latency (s)	J/token	Tokens	Mig.
OLMoE 1B/7B Instruct	502.6	≤ 0.6	0.2654	2,450	0
Qwen2.5-1.5B-Instruct	910.3	0.7–1.0	0.3183	4,232	> 0
Mistral 7B Instruct v0.2	4252.3	> 1.0	1.1311	6,330	0
Phi-3.5-Mini-Instruct	6089.9	≥ 6.0	0.5307	14,350	0

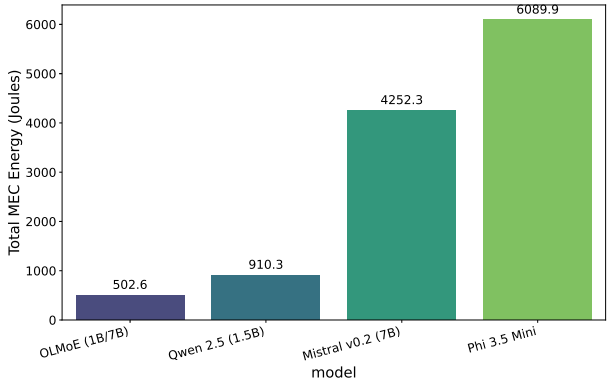


FIGURE 2. Total active MEC energy across all runs.

V. RESULTS AND DISCUSSION

Here, we discuss the AGORA test results for different local LLMs. We first examined how energy, delay, and UE QoS were affected by decision-making tools. Next, we examined how the workload context and GPU behaved under stress. We also report a deterministic automation (oracle) baseline that executes the same telemetry-to-UPF rule, providing a non-LLM reference for compliance and overhead measurements. Table 3 summarizes the main energy, latency, and compliance indicators per model.

A. ENERGY-QoS CHARACTERIZATION OF TOOL-AUGMENTED AGENTS

We quantified (i) infrastructure impact at the edge and (ii) perceived UE quality. Infrastructure energy is estimated by combining the Kepler instantaneous power at each MEC with the agent inference duration per prompt. Unless otherwise stated, active MEC energy denotes the energy spent at the MEC that effectively serves the UPF target during the agent decision loop, such as MEC1 when migration is active; otherwise, MEC2.

Aggregate infrastructure energy and latency. Figure 2 shows the total active MEC energy for all runs. The difference is clear: OLMoE 1B/7B Instruct uses the least energy (502.6 J), followed by Qwen2.5-1.5B-Instruct (910.3 J), while Mistral 7B Instruct v0.2 (4252.3 J) and Phi-3.5-Mini-Instruct (6089.9 J) consume significantly more. This ranking

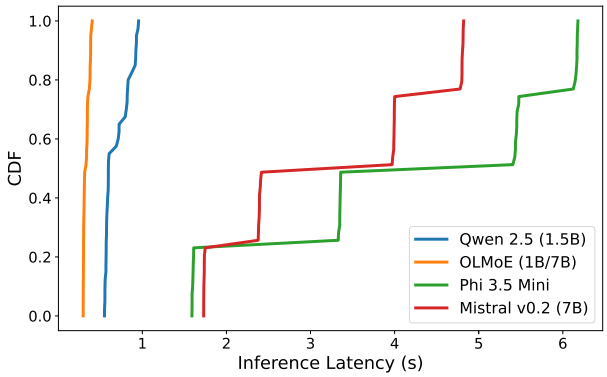


FIGURE 3. CDF of agent inference latency.

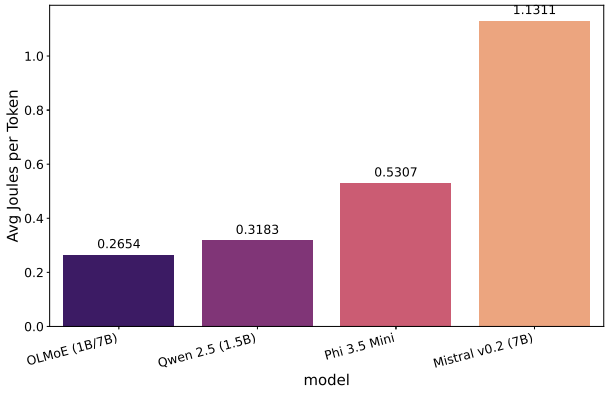


FIGURE 4. Average energy per generated token.

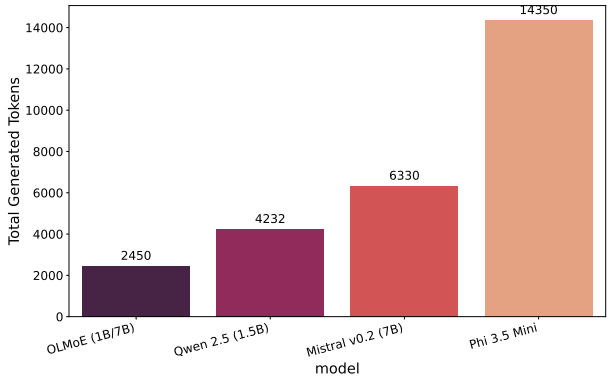


FIGURE 5. Total generated tokens across runs.

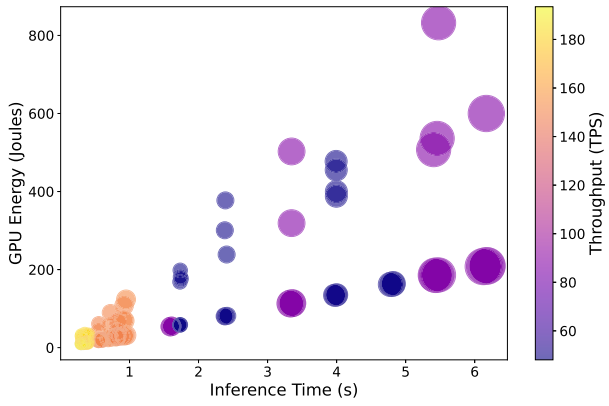


FIGURE 6. Inference time versus GPU energy with throughput and output volume.

demonstrates that, in this setting, the agent's runtime is the primary factor affecting the AGORA reasoning energy consumption, even when the MEC is under stress.

The latency distribution in Fig. 3 matches the energy ranking of the workloads. OLMoE 1B/7B Instruct was mostly under 0.6s. Qwen2.5-1.5B-Instruct was between 0.7s and 1.0s. Mistral 7B Instruct v0.2 and Mistral 7B Instruct v0.2 have longer times, with Mistral 7B Instruct v0.2 exceeding 6s. These times are essential for understanding energy use. Longer decision times imply that the system uses more energy because the energy-greedy MEC remains active for a longer duration.

Energy efficiency normalized by output. The total energy alone does not indicate whether a higher output volume compensates for longer runtimes. Figure 4 shows the normalized energy by the number of generated tokens. OLMoE 1B/7B Instruct remained the most efficient (0.2654 J/token), followed by Qwen2.5-1.5B-Instruct (0.3183 J/token). Mistral 7B Instruct v0.2 increases the energy per token to 0.5307 J/token, whereas Mistral 7B Instruct v0.2 is the least efficient (1.1311 J/token). This suggests that a proportionally greater token output does not offset the longer execution time of Mistral 7B Instruct v0.2 and that verbosity is an unreliable proxy for energy efficiency.

Figure 5 contextualizes these findings by showing the total number of generated tokens. Mistral 7B Instruct v0.2 produces the most tokens (14,350), followed by Mistral 7B Instruct v0.2 (6,330), Qwen2.5-1.5B-Instruct (4,232), and OLMoE 1B/7B Instruct (2,450). When contrasted with Fig. 4, the results indicate an apparent decoupling between verbosity and efficiency; a model can generate many tokens while still being energy-inefficient per token.

System operating regimes. Figure 6 combines the inference time (x-axis) with the GPU energy (y-axis), encoding the output volume via the marker size and the throughput via the color. The plot shows strong runtime-energy coupling, with the points moving upward as inference time increases. OLMoE 1B/7B Instruct and Qwen2.5-1.5B-Instruct concentrate in high-throughput regions at shorter runtimes and lower GPU energy, whereas Mistral 7B Instruct v0.2 and Mistral 7B Instruct v0.2 populate lower-throughput regimes, where longer runtimes amplify energy. This evidence supports the interpretation that, for tool-driven decision loops, faster models improve energy proportionality not only by reducing the time-to-decision but also by operating in higher-throughput regions of the accelerator.

The relationship between the throughput and GPU power is shown in Fig. 7. OLMoE 1B/7B Instruct and Qwen2.5-1.5B-Instruct achieved higher throughput

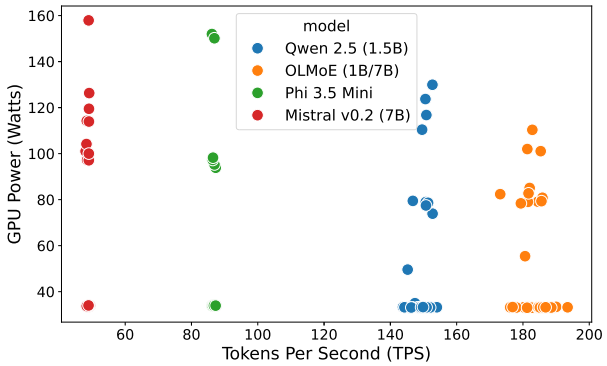


FIGURE 7. Throughput versus GPU power draw.

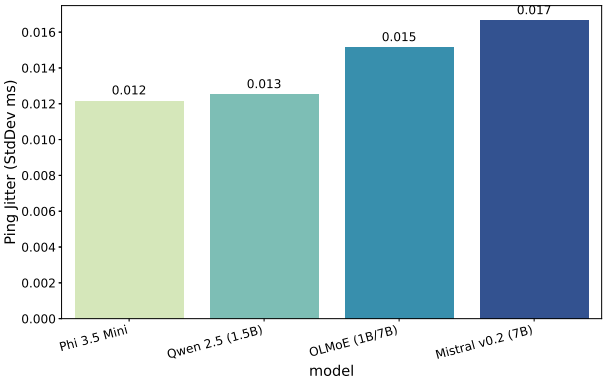


FIGURE 9. UE latency standard deviation per model.

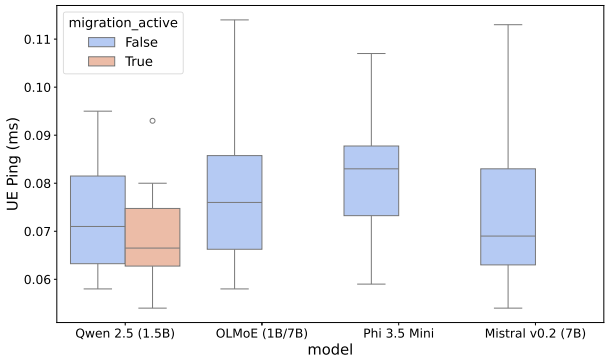


FIGURE 8. UE latency grouped by model and migration state.

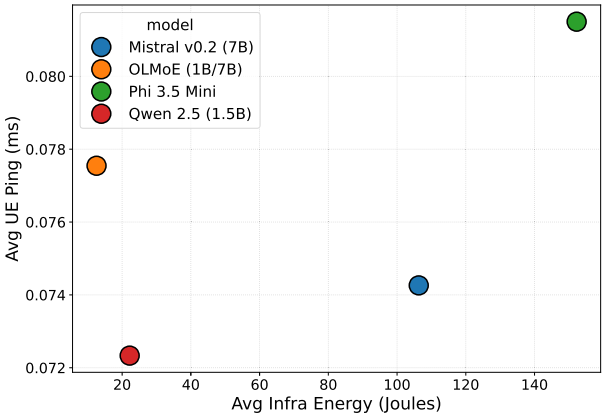


FIGURE 10. Average infrastructure energy versus average UE latency per model.

(approximately 140–195 tokens/s) under moderate power conditions. In contrast, Mistral 7B Instruct v0.2 and Mistral 7B Instruct v0.2 typically functioned below 90 tokens/s, yet they reached high power levels in several instances. Collectively, Figs 6 and 7 suggest that variations in throughput lead to distinct power-performance operating regimes, which in turn directly influence the energy consumed per decision.

QoS impact at the UE: latency and stability. Figure 8 shows that UE round-trip latency of the UE remains within a narrow band (tens of milliseconds) across the models, indicating that the radio and data plane baseline dominates the mean UE latency. However, the variability and tail behavior still revealed meaningful differences under stressed conditions. Qwen2.5-1.5B-Instruct is the only model that yields both migration states in the dataset, enabling a direct within-model comparison. When migration occurs, the distribution shifts slightly downward. It tightens, suggesting that routing away from the stressed MEC can reduce tail latency and improve stability, even if the average differences are modest.

The jitter proxy shown in Fig. 9 reinforces this stability. Mistral 7B Instruct v0.2 and Qwen2.5-1.5B-Instruct showed the lowest dispersion (approximately 0.012–0.013), OLMoE 1B/7B Instruct was slightly higher (0.015), and Mistral 7B Instruct v0.2 exhibited the highest jitter (0.017). Notably, this ordering does not strictly follow the inference speed, indicating that UE stability is influenced more by edge stress conditions and routing outcomes than by the agent runtime alone.

Joint energy-QoS trade-off. Figure 10 illustrates the operating points of each model concerning the average infrastructure energy and average UE latency. Qwen2.5-1.5B-Instruct is positioned in the advantageous lower-left region, offering the lowest average infrastructure energy and the lowest average UE latency, thereby achieving the best joint operating point in this experiment. Although OLMoE 1B/7B Instruct further reduces energy consumption, it does so at a slightly higher average UE latency than Qwen2.5-1.5B-Instruct. Mistral 7B Instruct v0.2 is located in the upper-right region, indicating a less favorable operating point, whereas Mistral 7B Instruct v0.2 occupies an intermediate position but is not Pareto-optimal, given Qwen2.5-1.5B-Instruct's superior standing.

Policy execution under stress: migration behavior. Performance metrics alone do not reveal whether an agent follows the intended policy. Figure 11 shows the probability of triggering migration as a function of stressed MEC 2 power, discretized into bins. Only Qwen2.5-1.5B-Instruct triggered migrations with non-zero probability across bins (peaking at approximately 0.43), whereas all other models

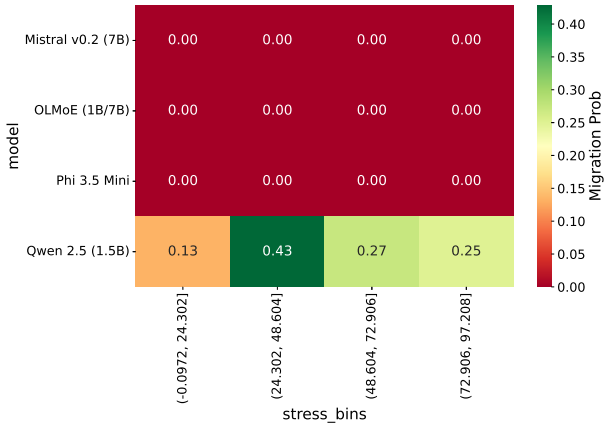


FIGURE 11. Migration probability by stressed MEC 2 power bin.

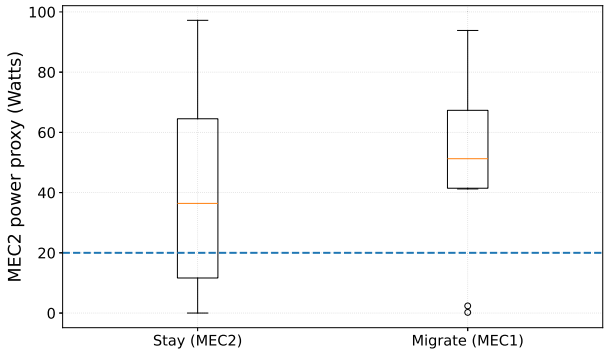


FIGURE 12. Qwen2.5-1.5B-Instruct migration selectivity.

remained at zero. This outcome indicates that, under the same tool-calling interface and prompts, Qwen2.5-1.5B-Instruct is the only model that reliably translates the energy threshold policy into an actionable UPF update. Consequently, part of the energy and QoS differences observed in earlier figures can be attributed not only to runtime but also to decision compliance.

Triggering migration in AGORA requires a strict sequence: (i) invoke the telemetry tool to obtain $\hat{P}_2(t)$, (ii) compare it with the threshold implied by the intent, and (iii) issue a correctly formatted `upf_set_target` actuation. Because the tool interface and intent prompts were identical across models, the differences in Fig. 11 primarily reflects model-to-tool alignment and instruction-following reliability rather than changes in the environment. We also validated tool availability and logging via the deterministic oracle baseline, indicating that zero-migration cases are mainly due to missed tool calls or incorrect actuation outputs, rather than missing telemetry.

Together, these results show that telemetry-grounded tool use can be translated into verifiable UPF actions, but decision compliance is model-dependent, making lightweight models with reliable tool/action alignment critical for sustainability-first closed-loop operation.

Our results expose a clear latency energy overhead of LLM-based control loops: longer inference times translate into higher active MEC and GPU energy, as shown by the energy/latency ranking and the strong coupling in the corresponding plots. However, this overhead is not always justified. When the policy and inputs are already fixed, a lightweight deterministic script can enforce the same rule at a negligible computational cost. The benefit of AGORA emerges when the control interface is a natural-language intent, where the system must handle phrasing variability (including multilingual prompts), select and sequence tool calls correctly, and produce verifiable UPF actuation with auditable compliance traces. Therefore, AGORA trades additional inference cost for policy expressiveness and operational flexibility, and our measurements quantify this trade-off across compact and large models.

To complement the Fig. 11, Fig. 12 provides a policy-grounded view of when AGORA uses Qwen2.5-1.5B-Instruct to evaluate migration. The boxplots compare the MEC2 power proxy (P_2) observed during decisions that *kept* traffic on MEC2 (Stay) versus decisions that *migrated* the UPF target to MEC1 (Migrate), with the dashed line indicating the threshold $\theta = 20\,\text{W}$ in Eq. (3). The dashed line shows a power limit of 20 W. Two main points are clear from this study. First, Qwen2.5-1.5B-Instruct tends to divert traffic when MEC2 uses more power. The power used during *Migrate* was higher than that during *Stay*. Second, the system is selective in nature. Although it correctly moved traffic only 28.57% of the time when needed (true positive rate), it was accurate 80.00% of the time when it did move traffic (positive predictive value). It also keeps the number of incorrect moves low at 15.38% (false positive rate). This shows that the system avoids unnecessary changes but still acts when required.

The fundamental mechanism for KPI improvement in AGORA is the UPF-level traffic steering triggered by energy thresholds. As shown in Fig. 8, routing traffic away from the stressed MEC2 reduces tail latency and improves stability. While the Deterministic Oracle achieves perfect compliance by design, it lacks the flexibility to interpret natural language or handle multilingual intents. Thus, AGORA's incremental value lies in its role as a robust intent-to-action translator that maintains QoS comparable to a deterministic script while offering human-centered, auditable orchestration traces.

B. WORKLOAD CONTEXT AND GPU OPERATIONAL BEHAVIOR

Stressor timeline and load asymmetry across MECs. Figure 13 summarizes the injected CPU workload over time. The experiment enforces an asymmetric regime in which MEC 2 sustains long intervals of elevated load, whereas MEC 1 remains predominantly near idle with only brief activity. This asymmetry is essential for interpreting policy behavior because it ensures that failure to migrate keeps the UPF target on a persistently stressed edge node, increasing the likelihood of revealing energy penalties and QoS degradation. Conversely, any observed stability improvements when migration occurs can be attributed to

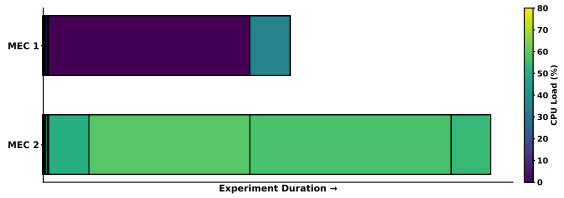


FIGURE 13. CPU load timeline applied to MEC 1 and MEC 2.

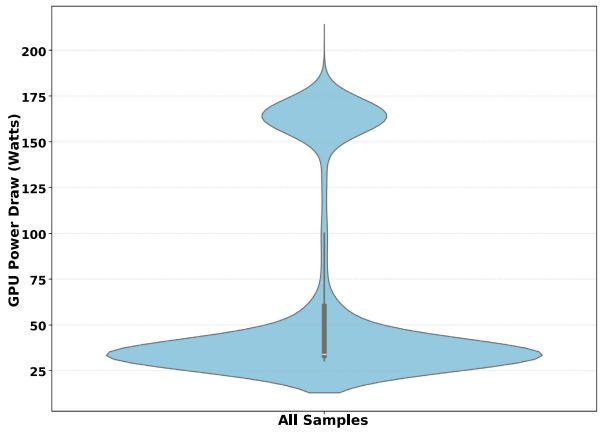


FIGURE 14. GPU power draw distribution over the experiment.

routing decisions made under measurable stress rather than uniform background conditions.

GPU power dynamics: bimodality and bursty execution. Figure 14 shows the distribution of GPU power draw during the experiment. The shape is strongly bimodal, with one dominant low-power band centered at 30-40 W and a second high-power band centered at 160-175 W. This pattern is consistent with bursty inference: the system alternates between an idle or lightly utilized state and a saturated state. An important implication is that the average power can obscure operational reality: substantial time may be spent in low-power mode while still reaching high-power peaks during inference bursts, which is vital for thermal constraints and energy-aware orchestration policies.

GPU operational states: utilization versus memory utilization. Figure 15 complements the power distribution by mapping the GPU utilization against the memory utilization with a density representation. The observations were concentrated near the bottom-left corner (approximately 0% utilization and 0% memory utilization) and formed a smaller cluster near the high-utilization region. The two clusters indicate that the accelerator frequently remained quiescent and transitioned to a high-activity state during inference. This separation supports scheduling strategies that reduce unnecessary high-power residency, for example, by consolidating requests and minimizing warm, underutilized periods.

Latency-energy coupling during decision loops. Finally, Figure 16 directly relates the AGORA inference time to the corresponding active MEC energy. The positive trend indicates that longer decision loops translate into higher infrastructure-energy expenditures. Model clusters further

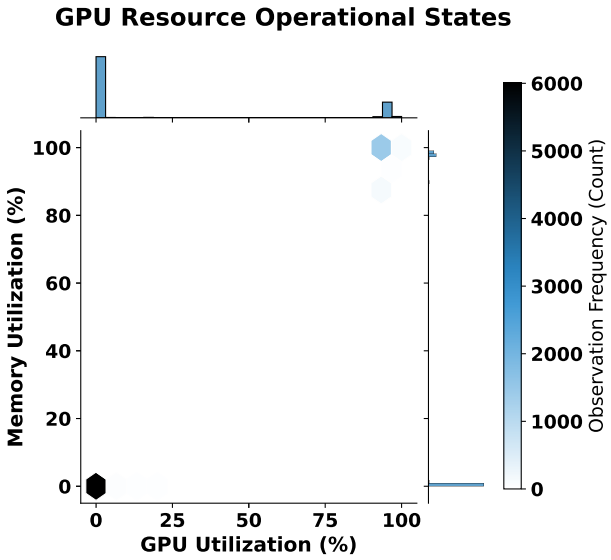


FIGURE 15. Density of GPU utilization versus GPU memory utilization.

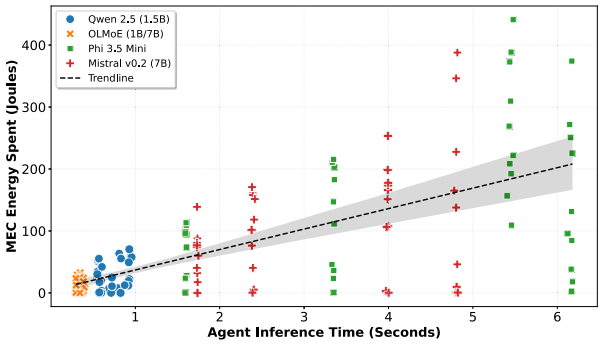


FIGURE 16. Inference time versus active MEC energy.

clarify this coupling: lightweight models occupy the low-latency/low-energy region, whereas larger models span a wider range of latencies and consume higher energy. The dispersion among larger models suggests sensitivity to run-to-run variability, such as caching behavior, tool invocation timing, and transient system conditions. Overall, this relationship supports the earlier conclusion that model choice impacts both responsiveness and energy footprint in closed-loop tool-augmented orchestration.

VI. CONCLUDING REMARKS

In this paper, we propose an agentic orchestration architecture for energy-efficient Beyond-5G operation that enables sustainability-driven control of the mobile network data plane. Most importantly, AGORA operationalizes sustainability intents through telemetry-grounded, verifiable UPF actuation, partially closing the gap between human sustainability goals and executable B5G data-plane control. While prior work advances intent-based and zero-touch management, it rarely operationalizes sustainability objectives as first-class constraints at the 5G core, particularly at the UPF.

Our approach integrates local LLMs into a telemetry-grounded, tool-augmented control loop that translates

natural-language intents into executable function calls and UPF routing actions. We note that LLM-driven control is most beneficial when intent flexibility is required; for fully fixed policies, non-generative scripts remain the most efficient option. Conversely, we demonstrate that an agentic approach becomes indispensable when dealing with ambiguous, high-level sustainability goals that require translation into verifiable data-plane actions, a scenario where simple automation scripts lack the necessary semantic understanding and adaptability. Experimental results show a strong latency-energy coupling in tool-driven decision loops and indicate that compact models can achieve low energy footprints while still enabling correct policy execution, including non-zero migration behavior under stressed MEC conditions.

In future work, we will extend the architecture along three directions. First, we will broaden the model space by evaluating additional open LLMs and multi-agent configurations under larger intent suites and more diverse traffic patterns, including multi-slice and multi-tenant scenarios. Second, we will generalize tool integration beyond ad hoc wrappers by adopting standardized interfaces such as MCP and Coral Protocol, enabling safer tool invocation, richer observability connectors, and portable agent deployments across platforms. Third, we will strengthen sustainability awareness by incorporating carbon intensity signals, risk-aware policies, and longer-horizon objectives, and by validating the approach on larger-scale testbeds with heterogeneous MECs and real-world workload traces. We also note that our current evaluation is limited by the scale and workload diversity of a controlled testbed, motivating validation under larger, multisite settings with heterogeneous hardware and operator-like traffic mixes.

REFERENCES

[1] S. E. Trevlakis, M. Belesioti, H. Koumaras, A.-A. A. Boulogeorgos, I. P. Chochliouros, and T. A. Tsiftsis, "An innovative architectural blueprint towards sustainable 6G systems," in *Proc. IEEE 29th Int. Workshop Comput. Aided Model. Design Commun. Links Netw. (CAMAD)*, Apr. 2024, pp. 1–8, doi: 10.1109/CAMAD62243.2024.10943010.

[2] Ö. U. Akgül, A. Varvara, A. de la Oliva, P. Charatsaris, M. Diamanti, P. G. Burguera, M. Ericson, S. Wänstedt, M. Ziółkowski, H. Tarasiuk, H. Hellaoui, S. Papavassiliou, V. Tsekenis, S. Barmpounakis, P. Demestichas, B. M. Khorsandi, and H. Harkous, "Sustainable 6G architecture: An organic evolution of 5G networks," *Comput. Netw.*, vol. 271, Oct. 2025, Art. no. 111629. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128625005961

[3] L. F. Rodrigues Moreira, R. Moreira, B. A. N. Travençolo, and A. R. Backes, "An artificial intelligence-as-a-service architecture for deep learning model embodiment on low-cost devices: A case study of COVID-19 diagnosis," *Appl. Soft Comput.*, vol. 134, Feb. 2023, Art. no. 110014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494623000327

[4] L. F. R. Moreira, R. Moreira, F. D. O. Silva, and A. R. Backes, "Towards cognitive service delivery on B5G through AIaaS architecture," in *Proc. Anais do IV Workshop de Redes 6G*, 2024, pp. 1–8. [Online]. Available: https://sol.sbc.org.br/index.php/w6g/article/view/29773

[5] R. Moreira, L. F. R. Moreira, and F. D. O. Silva, "Unleashing AI-empowered slices on mobile networks for natively cognitive service delivery," *IEEE Access*, vol. 13, pp. 147757–147771, 2025.

[6] R. C. Bello, N. Slamnik–Krieštorac, and J. M. Márquez-Barja, "Zero-touch service management for 6G verticals: Smart traffic management case study," in *Proc. IEEE 21st Consum. Commun. Netw. Conf. (CCNC)*, Mar. 2024, pp. 582–585, doi: 10.1109/CCNC51664.2024.10454808.

[7] H. Sun, Y. Liu, A. Al-Tahmeesschi, A. Nag, M. Soleimanpour, B. Canberk, H. Arslan, and H. Ahmadi, "Advancing 6G: Survey for explainable AI on communications and network slicing," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 1372–1412, 2025.

[8] D. G. S. Pivoto, T. T. Rezende, M. S. P. Facina, R. Moreira, F. D. O. Silva, K. V. Cardoso, S. L. Correa, A. V. D. Araujo, W. R. S. E. Silva, H. S. Neto, G. R. D. L. Tejerina, and A. M. Alberti, "A detailed relevance analysis of enabling technologies for 6G architectures," *IEEE Access*, vol. 11, pp. 89644–89684, 2023.

[9] Y. Xiao, Z. Ye, M. Wu, H. Li, M. Xiao, M. Alouini, A. Al-Hourani, and S. Cioni, "Space-air-ground integrated wireless networks for 6G: Basics, key technologies, and future trends," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 12, pp. 3327–3354, Dec. 2024, doi: 10.1109/JSAC.2024.3492720.

[10] R. Moreira, J. S. B. Martins, T. C. M. B. Carvalho, and F. D. O. Silva, "On enhancing network slicing life-cycle through an AI-native orchestration architecture," in *Proc. Adv. Inf. Netw. Appl.*, 2023, pp. 124–136.

[11] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, D. Niyato, and O. A. Dobre, "Large language model enhanced multi-agent systems for 6G communications," *IEEE Wireless Commun.*, vol. 31, no. 6, pp. 48–55, Aug. 2024, doi: 10.1109/MWC.016.2300600.

[12] S. Çımen, S. N. Karahan, D. Karhan, M. Güllü, and M. S. Osmanca, "The integration of agentic AI in 6G wireless networks: State-of-the-art, challenges, and future perspectives," in *Proc. Innov. Intell. Syst. Appl. Conf. (ASYU)*, Sep. 2025, pp. 1–6.

[13] S. A. Alex, V. G. Menon, M. Adhikari, and S. Mumtaz, "Zero-touch self-organizing networks in 6G: Integrating edge computing and artificial intelligence," *IEEE Commun. Standards Mag.*, early access, Oct. 14, 2025, doi: 10.1109/MCOMSTD.2025.3614978.

[14] P. Dazzi, "The Internet of AI agents (IAIA): A new frontier in networked and distributed intelligence," *Int. J. Networked Distrib. Comput.*, vol. 13, no. 1, p. 16, Mar. 2025, doi: 10.1007/s44227-025-00057-0.

[15] A. Mekrache, A. Ksentini, and C. Verikoukis, "DMO-GPT: An intent-driven framework for distributed 6G management and orchestration," *IEEE Commun. Mag.*, vol. 64, no. 1, pp. 48–54, Oct. 2025, doi: 10.1109/MCOM.001.2500258.

[16] X. Cheng, J. Zhang, N. Ding, N. Li, Y. Li, T. Wu, W. Xu, J. Zhang, and Q. Sun, "Integrated AI and communications: A two-way catalysis toward 6G and beyond," *J. Commun. Inf. Netw.*, vol. 10, no. 3, pp. 191–200, 2025.

[17] D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey," *IEEE Access*, vol. 13, pp. 18912–18936, 2025.

[18] Y. Wang, S. Guo, Y. Pan, Z. Su, F. Chen, T. H. Luan, P. Li, J. Kang, and D. Niyato, "Internet of Agents: Fundamentals, applications, and challenges," *IEEE Trans. Cognit. Commun. Netw.*, vol. 12, pp. 4476–4501, 2025, doi: 10.1109/TCCN.2025.3623369.

[19] Y. Njah, A. Leivadeas, and M. Falkner, "An AI-driven intent-based network architecture," *IEEE Commun. Mag.*, vol. 63, no. 4, pp. 146–153, Apr. 2024, doi: 10.1109/MCOM.001.2400143.

[20] F. Guim, T. Metsch, H. Moustafa, T. Verrall, D. Carrera, N. Cadenelli, J. Chen, D. Doria, C. Ghadie, and R. G. Prats, "Autonomous lifecycle management for resource-efficient workload orchestration for green edge computing," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 1, pp. 571–582, Mar. 2022.

[21] S. Barrachina-Muñoz, F. Rezazadeh, L. Blanco, S. Kukliński, E. Zeydan, A. Chawla, L. Zanzi, F. Devoti, V. Vlahodimitropoulou, I. P. Chochliouros, A.-M. Bosneag, S. Cherrared, L. Vettori, and J. Mangues-Bafalluy, "Empowering beyond 5G networks: An experimental assessment of zero-touch management and orchestration," *IEEE Access*, vol. 12, pp. 182752–182762, 2024.

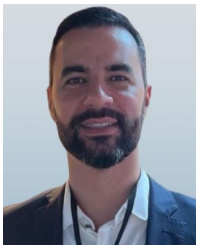
[22] A. Dalgkitsis, L. A. Garrido, F. Rezazadeh, H. Chergui, K. Ramantas, J. S. Vardakas, and C. Verikoukis, "SCHE2MA: Scalable, energy-aware, multidomain orchestration for beyond-5G URLLC services," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 7, pp. 7653–7663, Jul. 2023.

[23] A. Dandoush, V. Kumarskandpriya, M. Uddin, and U. Khalil, "Large language models meet network slicing management and orchestration," 2024, *arXiv:2403.13721*.

[24] 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 [Invited]," *J. Opt. Commun. Netw.*, vol. 17, no. 1, pp. A112–A123, 2025.
[25] A. Mekrache, A. Ksentini, and C. Verikoukis, "OSS-GPT: An LLM-powered intent-driven operations support system for 6G networks," in *Proc. IEEE 11th Int. Conf. Netw. Softwarization (NetSoft)*, Jul. 2025, pp. 155–163, doi: 10.1109/NetSoft64993.2025.11080632.
[26] I. Chatzistefanidis, A. Leone, and N. Nikaein, "Maestro: LLM-driven collaborative automation of intent-based 6G networks," *IEEE Netw. Lett.*, vol. 6, no. 4, pp. 227–231, Dec. 2024.
[27] N. Molner, L. Rosa, F. Risso, K. Samdanis, D. A. Guillen, R. Smets, T. Taleb, and D. Gómez-Barquero, "AIORA: An AI-native multi-stakeholder orchestration architecture for 6G continuum," *IEEE Netw.*, early access, Aug. 6, 2025, doi: 10.1109/MNET.2025.3596343.
[28] D. Brodimas, A. Birbas, D. Kapolos, and S. Denazis, "Intent-based infrastructure and service orchestration using agentic-AI," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 7150–7168, 2025.
[29] M. Elkael, S. D'Oro, L. Bonati, M. Polese, Y. Lee, K. Furueda, and T. Melodia, "AgentRAN: An agentic AI architecture for autonomous control of open 6G networks," 2025, *arXiv:2508.17778*.
[30] I. Chatzistefanidis and N. Nikaein, "Symbiotic agents: A novel paradigm for trustworthy AGI-driven networks," *Comput. Netw.*, vol. 273, 2025, Art. no. 111749. [Online]. Available: https://doi.org/10.1016/j.comnet.2025.111749
[31] I. Chatzistefanidis, A. Leone, A. Yaghoubian, M. Irazabal, S. Nassim, L. Bariah, M. Debbah, and N. Nikaein, "MX-AI: Agentic observability and control platform for open and AI-RAN," 2025, *arXiv:2508.09197*.
[32] I. Chatzistefanidis, N. Nikaein, A. Leone, A. Maatouk, L. Tassiulas, R. Morabito, I. Pitsiorlas, and M. Kountouris, "Agoran: An agentic open marketplace for 6G RAN automation," *Comput. Netw.*, vol. 275, Feb. 2025, Art. no. 111927. [Online]. Available: https://doi.org/10.1016/j.comnet.2025.111927
[33] M. A. Habib, P. E. Iturria Rivera, Y. Ozcan, M. Elsayed, M. Bavand, R. Gaigalas, and M. Erol-Kantarci, "LLM-based intent processing and network optimization using attention-based hierarchical reinforcement learning," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2025, pp. 1–6.
[34] M. A. Habib, P. E. Iturria-Rivera, Y. Ozcan, M. Elsayed, M. Bavand, R. Gaigalas, and M. Erol-Kantarci, "Harnessing the power of LLMs, informers and decision transformers for intent-driven RAN management in 6G," *IEEE Trans. Netw. Sci. Eng.*, vol. 13, pp. 4187–4206, 2026.
[35] H. Chergui, F. Rezazadeh, M. Bennis, and M. Debbah, "LLM-based agentic negotiation for 6G: Addressing uncertainty neglect and tail-event risk," 2025, *arXiv:2511.19175*.
[36] K. Dev, S. A. Khowaja, E. Zeydan, K. Singh, and M. Debbah, "Advanced architectures integrated with agentic AI for next-generation wireless networks," *IEEE Commun. Standards Mag.*, early access, Nov. 20, 2025, doi: 10.1109/MCOMSTD.2025.3632205.
[37] D. F. Pedroso, L. Almeida, L. E. G. Pulcinelli, W. A. A. Aisawa, I. Dutra, and S. M. Bruschi, "Anomaly detection and root cause analysis in cloud-native environments using large language models and Bayesian networks," *IEEE Access*, vol. 13, pp. 77550–77564, 2025.
[38] *System Architecture for the 5G System (5gs); Stage 2*, document (TS) 23.501, 3GPP, 2024.
[39] R. Pires, H. Abonizio, T. S. Almeida, and R. Nogueira, "Sabiá: Portuguese large language models," in *Intelligent Systems*. Cham, Switzerland: Springer, 2023, pp. 226–240.
[40] I. Baldin, A. Nikolich, J. Griffioen, I. I. S. Monga, K.-C. Wang, T. Lehman, and P. Ruth, "FABRIC: A national-scale programmable experimental network infrastructure," *IEEE Internet Comput.*, vol. 23, no. 6, pp. 38–47, Nov. 2019.
[41] A. Yang et al., "Qwen2 technical report," 2024, *arXiv:2407.10671*.
[42] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. Singh Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. Renard Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed, "Mistral 7B," 2023, *arXiv:2310.06825*.
[43] M. Abdin et al., "Phi-3 technical report: A highly capable language model locally on your phone," 2024, *arXiv:2404.14219*.
[44] N. Muennighoff et al., "OLMoE: Open mixture-of-experts language models," 2024, *arXiv:2409.02060*.
[45] C. Mesh. (2021). *Chaos Mesh: A Powerful Chaos Engineering Platform for Kubernetes*. [Online]. Available: https://github.com/chaos-mesh/chaos-mesh

RODRIGO MOREIRA (Member, IEEE) received the B.S. degree from the Federal University of Viçosa, in 2014, the M.Sc. degree from the Federal University of Uberlândia, Brazil, in 2017, and the Ph.D. degree from the University of Uberlândia (UFU), in 2021. He is currently a Professor with the Federal University of Viçosa. He has several papers published and has presented at conferences. His research interests include the future of the internet, quality of service, cloud computing, network function virtualization, software-defined networking, computational intelligence, and edge computing.

LARISSA FERREIRA RODRIGUES MOREIRA (Member, IEEE) received the B.Sc. degree in computer information systems and the M.Sc. degree in computer science from the Federal University of Viçosa, Brazil, in 2016 and 2018, respectively, and the Ph.D. degree in computer science from the Federal University of Uberlândia, Brazil, in 2024. She is currently a Professor with the Federal University of Viçosa. Her research interests include artificial intelligence, computer vision, and image processing.



MAYCON LEONE MACIEL PEIXOTO received the master's and Ph.D. degrees in computer science from the University of São Paulo (USP), Brazil, in 2008 and 2012, respectively. He was a Postdoctoral Researcher at the University of Campinas (UNICAMP), in 2020. He was a Visiting Scholar at the University of Toronto, Canada, from 2023 to 2024. He is currently an Associate Professor with the Institute of Computing, Federal University of Bahia (UFBA). He serves on the leading network and distributed system committees. Over the years, he has participated and coordinated several projects funded by public agencies and the private sector, with research interests spanning edge–cloud continuum, artificial intelligence and federated learning, vehicular and the IoT networks, quantum computing, and data-driven digital transformation.

FLÁVIO DE OLIVEIRA SILVA (Senior Member, IEEE) received the Ph.D. degree from the University of São Paulo (USP), in 2013. He is currently a Professor with the Department of Informatics (DI), School of Engineering, University of Minho, Braga, Portugal, and a Researcher with the ALGORITMI Centre. He has published and presented several papers at conferences worldwide. His research interests include future networks, the IoT, network softwarization (SDN and NFV), future intelligent applications and systems, cloud computing, and software-based innovation. He is a member of ACM and SBC. He is a reviewer of several journals and a member of the TPC at several IEEE conferences.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - ROR identifier: 00x0ma614