JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

Agentic AI for Energy-Efficient Telecommunications: A Comparative Performance Analysis

Salihu Angulu Rakiya, Gilbert I. O. Aimufua
Centre for Cyberspace Studies,
Nasarawa State University, Keffi, Nigeria

ABSTRACT

The integration of Agentic Artificial Intelligence (AI) into telecommunications networks presents a transformative opportunity for autonomous network management, yet the energy consumption implications of these systems remain insufficiently characterized. This article presents a comparative analysis of two leading agentic AI frameworks which includes the Strategist Agent architecture for Radio Access Network (RAN) management and the AGORA (Agentic Green Orchestration Architecture) framework with a focus on energy efficiency performance in telecommunications applications. Drawing on recent empirical studies and experimental evaluations from 2025–2026, this study examines how architectural design choices fundamentally determine energy consumption patterns. The analysis reveals four critical findings: (1) framework architecture drives energy consumption with up to 9.4× differences, significantly outweighing model size or task complexity; (2) current frameworks achieve near-zero task resolution rates (0-4%) with SLMs, wasting energy on unproductive reasoning loops; (3) memory access dominates computational energy (70%) with DRAM access costing 640 pJ/access, while token transmission dominates communication energy (40%); and (4) the optimal model-framework pairing depends on the specific task, with llama3.2:3b + LangChain achieving the highest performance with the lowest energy consumption. For telecommunications applications, AGORA's direct intent-to-tool translation achieves superior energy efficiency, while the Strategist Agent provides better reasoning depth. Projected net energy impact reveals 85% infrastructure savings against 15% agentic cost, yielding 70% net energy reduction in optimized RAN configurations. This study presents a framework efficiency metric (η = Task Resolution Rate / Energy Consumed) revealing $\eta \approx 0$ for current SLM deployments, indicating urgent need for architectural redesign. The findings from this study provide actionable insights for deploying energy-efficient agentic AI, emphasizing that sustainability must be a first-class design principle in next-generation wireless networks, requiring memory-optimized inference, adaptive reasoning depth, and carbon-aware decision-making to achieve both intelligence and environmental sustainability.

ARTICLE INFO
Article History
Received: April, 2026
Received in revised form: May, 2026
Accepted: August, 2026
Published online: September, 2026

KEYWORDS
Agentic AI, Energy Efficiency, Telecommunications, 5G Networks, Small Language Models, Green Networking

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

INTRODUCTION

The telecommunications industry faces a dual challenge: managing the unprecedented complexity of next-generation networks while dramatically reducing energy consumption. Mobile networks have emerged as a major source of energy demand within the ICT sector, and as data traffic continues to rise, the need for sustainable, high-performance network management has become critical (Moreira *et al.,* 2026). According to industry estimates, the proliferation of 5G and emerging 6G networks will amplify these challenges exponentially, requiring innovative approaches that reconcile performance demands with environmental sustainability. Agentic AI are autonomous systems that perceive, reason, and act in dynamic environments through closed-loop pipelines offers a promising solution (Tripathy *et al.,* 2025).

Unlike traditional AI systems that operate under static, single-turn interactions, agentic AI integrates perception, reasoning, and action into iterative, feedback-driven workflows, enabling self-organizing and self-healing network capabilities (Chen *et al.,* 2026). This paradigm shift represents a fundamental departure from conventional network management approaches, positioning AI as an active participant in network operations rather than a passive analytical tool. However, this paradigm shift introduces fundamental energy challenges. The closed-loop nature of agentic systems necessitates continuous interaction with network environments, leading to recurrent inference cycles and persistent data exchange between distributed components (Tripathy *et al.,* 2025). These energy demands manifest across two interconnected dimensions: computational energy (model inference, memory access) and communication energy (data transmission between agents and network elements) (Chen *et al.,* 2026).

This article provides a comparative analysis of two representative agentic AI frameworks applied to telecommunications, focusing on their energy efficiency characteristics:

1. The Strategist Agent Architecture for Intent-Based RAN Management
2. AGORA (Agentic Green Orchestration Architecture) for Beyond 5G Networks (Moreira et al., 2026)

We draw on recent empirical studies, including the SWEnergy evaluation of agentic issue resolution frameworks (Tripathy *et al.,* 2025) and the AgenticBench performance-energy trade-off analysis (Barker, 2026), to establish a comparative framework for understanding energy efficiency in agentic telecommunications systems.

Research Questions

This study addresses the following research questions:

1. RQ1: How do different agentic AI architectures (e.g., Strategist Agent vs. AGORA) impact energy efficiency in telecommunications?
2. RQ2: What are the performance-energy trade-offs of deploying Small Language Models (SLMs) in agentic frameworks?
3. RQ3: How can network-level energy impact models optimize the deployment of agentic AI?

AGENTIC AI IN TELECOMMUNICATIONS: A REVIEW

The Case for Agentic AI in Network Management

The rapid increase in diverse service types and dynamic resource demands in Radio Access Networks (RAN) introduces unprecedented operational complexity. Even with Open RAN enabling integration of components from multiple vendors, configuration complexity and misconfiguration risks increase substantially. Traditional management approaches, which rely on manual configuration and rule-based automation, are increasingly inadequate for the scale and dynamism of modern networks. Intent-Based Networking (IBN) has emerged as a paradigm to manage networks by simplifying detailed technical configurations with minimal external intervention.

An intent serves as a high-level, human-expressible declaration of network goals rather

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

than specified methods for achievement. However, a major challenge lies in accurately and dynamically translating high-level user intents into precise RAN instructions is a task that requires sophisticated reasoning about network state, constraints, and optimization objectives. Agentic AI addresses this gap by enabling autonomous, goal-driven agents that can make local and collective decisions across the network. By doing so, agentic architectures enable networks that can autonomously reconfigure and optimize while dynamically balancing energy efficiency, service quality, and operational cost (Moreira *et al*., 2026). The potential benefits include reduced operational expenditure, improved network reliability, and enhanced ability to meet sustainability targets.

Key Agentic Architectures for Telecommunications

The following sub section is a discussion on the key architecture as seen in most telecommunication networks.

Strategist Agent Architecture

The Strategist Agent architecture, proposed for intent-based RAN management, leverages Large Language Models as intelligent intermediaries for intent translation. The system employs a multi-agent approach with three specialized components (Moreira *et al*., 2026):

1. Strategist Agent: Responsible for intent translation, receiving formalized intent declarations and generating configuration strategies using LLMs. This agent serves as the primary decision-maker, interpreting high-level goals and formulating specific configuration actions.
2. History Analyzer Agent: Analyzes past strategies and provides recommendations based on historical performance. This component introduces memory into the system, enabling learning from previous decisions and avoiding repeated errors.
3. Orchestrator Agent: Executes strategies by calling appropriate configuration tools via the O-RAN O1 interface. This agent bridges the gap between reasoning and action, translating strategic decisions into concrete network operations.

AGORA (Agentic Green Orchestration Architecture)

AGORA embeds a local tool-augmented Large Language Model agent in the mobile network control loop to translate natural-language sustainability goals into telemetry-grounded actions (Moreira *et al.,* 2026). Key features include:

1. Intent-to-Action Translation: Converts human sustainability goals into verifiable, tool-executed User Plane Function (UPF) actions. This capability enables operators to express sustainability objectives in natural language, which are then automatically translated into network configurations.
2. Tool-Augmented LLM Agent: Uses native tool-calling mechanisms to execute network commands. This architectural choice reduces the gap between intent and execution, minimizing unnecessary reasoning cycles.
3. Real-Time Telemetry Grounding: Continuously grounds decisions in live network performance data. By maintaining awareness of current network state, AGORA ensures that decisions are contextually appropriate and verifiable.
4. UPF-Level Enforcement: Acts directly on the data plane to enforce green policies (Moreira *et al.,* 2026). This direct control mechanism enables immediate energy savings without waiting for higher-level orchestration. The AGORA framework represents a shift from traditional intent-based and zero-touch frameworks that primarily stop at descriptor generation or control plane reconfiguration (Moreira *et al.,* 2026). By embedding the agent directly in the network control loop, AGORA achieves tighter integration between reasoning and action.

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

Traditional Energy Efficiency Approaches in Telecommunications

Prior to the advent of agentic AI, telecommunications networks employed several optimization techniques to improve energy efficiency. These include cell sleeping, where base stations are dynamically deactivated during periods of low traffic; power control algorithms that adjust transmit power based on channel conditions; and network planning optimization that minimizes the number of active infrastructure elements (Moreira *et al.,* 2026). While these methods have demonstrated effectiveness in reducing energy consumption (typically achieving 15-30% savings), they rely on static, rule-based approaches that cannot adapt to the rapidly changing conditions and complex service requirements of 5G and 6G networks. Agentic AI represents a paradigm shift by enabling autonomous, context-aware decision-making that can dynamically balance multiple objectives including energy efficiency, service quality, and operational cost in real-time (Chen *et al.,* 2026).

METHODOLOGY

This study employs a comparative analysis methodology based on secondary data synthesis from recent empirical studies. The methodology comprises the following components:

Data Sources

The analysis draws on data from three primary sources:

1. SWEnergy Study (Tripathy *et al.,* 2025): An empirical evaluation of four agentic frameworks (SWE-Agent, OpenHands, AutoCodeRover, Mini SWE Agent) using Small Language Models (Gemma-3 4B, Qwen-3 1.7B) on a controlled hardware setup. Key metrics include relative energy consumption and task resolution rates.
2. AgenticBench (Barker, 2026): A comprehensive evaluation of 84 cell configurations across four models (llama3.2:3b, qwen3, llama3.3:70b), three agent frameworks (LangChain, AutoGen, CrewAI), and six benchmarks (HumanEval, MBPP, MATH, etc.). Performance-energy trade-off metrics were extracted from the thesis results.
3. AGORA Framework (Moreira *et al.,* 2026): Experimental evaluation of the Agentic Green Orchestration Architecture for 5G UPF traffic steering, including energy efficiency measurements and policy enforcement results.

Analytical Framework

The comparative analysis was conducted using the following approach:

Step 1: Data Extraction and Normalization - Key performance metrics including energy consumption, task resolution rates, and inference times were extracted from each study. Energy consumption values were normalized relative to the baseline configuration (OpenHands + Gemma-3 4B) to enable cross-framework comparison.

Step 2: Framework Efficiency Metric Calculation - The framework efficiency metric η = Task Resolution Rate / Energy Consumed was calculated for each framework-model pairing based on data from Tripathy *et al.* (2025). Task resolution rates were derived from the SWEnergy study, while energy consumption values were normalized.

Step 3: Comparative Metric Derivation - Relative performance metrics for the Strategist Agent and AGORA architectures were estimated by extrapolating from the empirical data. The energy overhead for multi-agent orchestration was estimated based on the communication energy component breakdown from Chen *et al.* (2026).

Step 4: Network-Level Impact Modeling - Projected infrastructure energy savings were calculated using the network-level energy impact model $E_{\text{network}} = E_{\text{agentic}} + \Delta E_{\text{infrastructure}}$. $\Delta E_{\text{infrastructure}}$ (85% savings) was derived from the optimization potential of intelligent RAN parameter control based on Moreira *et al.* (2026), while E_agentic (15% overhead) was estimated from

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

the energy consumption patterns observed in Barker (2026) and Tripathy *et al.,* (2025).

Step 5: MATLAB Modeling - The energy consumption models (Equations 1-4) were implemented in MATLAB R2020a to simulate the projected net energy impacts across different deployment scenarios. The MATLAB simulation modeled the energy consumption of agentic systems under varying traffic loads, model sizes, and framework configurations. The code is available as supplementary material.

ENERGY EFFICIENCY IN AGENTIC AI: EMPIRICAL FINDINGS

The following sub section gives the discussion of EE in agentic AI in terms of empirical studies carried out.

Framework Architecture as the Primary Energy Driver

Recent empirical studies reveal that framework architecture is the primary driver of energy consumption in agentic AI systems, often outweighing model size or task complexity (Tripathy *et al*., 2025). These finding challenges conventional wisdom that model size is the dominant factor in AI energy consumption, suggesting that system-level design decisions merit greater attention. The SWEnergy study evaluated four leading agentic frameworks (SWE-Agent, OpenHands, AutoCodeRover, Mini SWE Agent) using two Small Language Models (Gemma-3 4B, Qwen-3 1.7B) on a controlled hardware setup (Tripathy *et al.,* 2025). Key findings from this comprehensive evaluation are presented in Table 1.

Table 1: Relative Energy Consumption of Agentic Frameworks with SLMs (Tripathy et al., 2025)

Framework	Model	Relative Energy Consumption	Task Resolution Rate
AutoCodeRover	Gemma-3 4B	9.4×	4%
OpenHands	Gemma-3 4B	1.0× (baseline)	0%
SWE-Agent	Gemma-3 4B	4.2×	0%
Mini SWE Agent	Gemma-3 4B	2.1×	0%

Notably, the most energy-intensive framework consumed 9.4× more energy than the least energy-intensive, yet task resolution rates remained near-zero across all configurations. This demonstrates that current frameworks, when paired with SLMs, consume significant energy on unproductive reasoning loops—the SLM's limited reasoning was the bottleneck for success, but the framework's design was the bottleneck for efficiency (Tripathy *et al.,* 2025).

The Performance-Energy Trade-off Frontier

The AgenticBench study, which evaluated an 84-cell matrix across four models, three agent frameworks, and six benchmarks, further reinforces these findings (Barker, 2026). The comprehensive evaluation revealed complex trade-offs between performance and energy consumption: "Results show that no single model-framework pairing dominates all domains. AutoGen used more reported energy than LangChain in 20 of 24 matched single-agent cells, with a median ratio of 2.2× and a maximum of 15.8×" (Barker, 2026).

Key observations from the AgenticBench study include:

1. Small Language Models can achieve high accuracy with minimal energy: llama3.2:3b with LangChain achieved a perfect HumanEval score at the lowest reported energy among non-baseline cells (Barker, 2026).
2. Reasoning depth trades off with energy consumption: qwen3 reached high accuracy on hard reasoning tasks but produced long runtimes and timeout risk under extended reasoning (Barker, 2026).
3. Multi-agent topology shifts the score-energy surface: Role-specialized multi-agent systems can alter performance-energy trade-offs relative to single-agent baselines (Barker, 2026).

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

Energy Accounting Framework

The networking-aware energy efficiency survey proposes a comprehensive energy accounting framework distinguishing between different energy types as shown in Table 2 (Chen *et al.,* 2026):

Table 2: Energy Accounting Framework for Agentic AI Inference (Chen et al., 2026)

Energy Type	Source	Characteristics	Representative Cost
Computational Energy	Model inference, arithmetic operations, memory access	Dominated by memory bandwidth	DRAM access = 640 pJ/access
Communication Energy	Data transmission, token exchange, synchronization	Critical in distributed edge deployments	Variable by distance and bandwidth

This framework highlights that agentic AI inference is fundamentally different from traditional AI: the primary bottleneck shifts from computational FLOPs to memory bandwidth (the "I/O Wall"), exacerbating energy consumption through frequent off-chip DRAM accesses (Chen *et al.,* 2026). This insight has profound implications for the design of energy-efficient agentic systems, suggesting that memory access patterns and data locality deserve as much attention as algorithmic efficiency.

COMPARATIVE ANALYSIS: STRATEGIST AGENT VS. AGORA

The comparative analysis based on architecture and energy efficiency is discussed in this section.

Architectural Comparison

A summary of architectural comparison is shown in Table 3.

Table 3: Comparative Architecture: Strategist Agent vs. AGORA (Moreira et al., 2026)

Feature	Strategist Agent Architecture	AGORA Framework
Primary Application	RAN parameter optimization	UPF traffic steering, MEC orchestration
LLM Integration	NVIDIA NIM inference API	Local tool-augmented LLM
Agent Structure	Multi-agent (Strategist, History Analyzer, Orchestrator)	Single agent with tool-calling
Control Loop	Closed-loop with O-RAN O1 interface	Closed-loop with 5G UPF actuation
Energy Focus	Indirect (optimization of network parameters)	Direct (energy-aware policy enforcement)
Deployment Complexity	Moderate (multiple specialized agents)	Lower (single unified agent)

Energy Efficiency Mechanisms

Strategist Agent Architecture achieves energy efficiency through:

1. Structured prompt engineering to minimize token usage and inference overhead. By standardizing input format and constraining output space, the architecture reduces the cognitive load on the LLM and minimizes unnecessary token generation.
2. Historical strategy analysis to avoid repeated exploration of suboptimal configurations. The History Analyzer Agent provides memory of past decisions, enabling the system to converge more quickly on effective configurations.
3. Dynamic parameter optimization to improve RAN energy efficiency directly. By optimizing network parameters such

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

as transmit power and resource allocation, the architecture achieves energy savings at the infrastructure level.

AGORA achieves energy efficiency through:

1. Telemetry-grounded decision-making to ensure actions are justified by current network state (Moreira *et al.,* 2026). By grounding decisions in real-time data, AGORA avoids speculative actions that might waste energy.
2. Intent-to-tool translation reducing unnecessary reasoning loops (Moreira et al., 2026). The direct mapping between intents and executable tools minimizes the gap between goal specification and action.
3. Energy-aware UPF routing directly controlling network energy consumption (Moreira *et al.,* 2026). By making routing decisions that prioritize energy efficiency, AGORA achieves immediate energy savings.

Comparative Metrics

Based on available experimental data, we can estimate relative performance as shown in Table 4. The metrics in Table 4 were derived by synthesizing data from the SWEnergy study (Tripathy et al., 2025), the AgenticBench evaluation (Barker, 2026), and the AGORA experimental results (Moreira et al., 2026). Framework overhead estimates were calculated based on the communication energy breakdown from Chen et al. (2026), where token transmission accounts for 40% of communication energy in distributed agentic systems.

Table 4: Comparative Performance Metrics (Moreira et al., 2026; Tripathy et al., 2025; Barker, 2026)

Metric	Strategist Agent	AGORA
Energy Reduction Achieved	Not directly measured	Demonstrated green policy enforcement
Framework Overhead	Moderate (multi-agent orchestration)	Lower (single agent with tool-calling)
Deployment Flexibility	High (O-RAN compliant)	Moderate (5G UPF specific)
Sustainability Focus	Implicit (via optimization)	Explicit (intent-driven)
Reasoning Depth	High (multi-agent deliberation)	Moderate (single-agent reasoning)

ENERGY EFFICIENCY MODELS AND EQUATIONS

The following is a discussion on the EE models and some major equations.

Energy Consumption Model for Agentic Systems

The total energy consumption of an agentic AI system can be expressed as (Chen *et al.,* 2026):

$$E_{\text{total}} = E_{\text{compute}} + E_{\text{comm}} + E_{\text{overhead}} \quad (1)$$

where:

E_{compute} is the energy for model inference, proportional to token count and model size

E_{comm} is the energy for data transmission between agents and network elements (Chen et al., 2026)

E_{overhead} is the energy for framework orchestration, planning, and tool invocation

Agentic Inference Energy Breakdown

For a single agentic reasoning cycle, the model equation is given as (Chen *et al.,* 2026):

$$E_{\text{cycle}} = \sum_{i=1}^{n} (E_{\text{token},i} \cdot N_{\text{token},i} + E_{\text{memory},i} \cdot N_{\text{access},i}) \quad (2)$$

Where:

$E_{\text{token},i}$ is the energy per token for the i-th model call

$N_{\text{token},i}$ is the number of tokens processed

$E_{\text{memory},i}$ is the energy per memory access

$N_{\text{access},i}$ is the number of memory accesses

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

According to Chen *et al.,* (2026), memory access dominates energy consumption in agentic inference, with DRAM access costing approximately 640 pJ per access.

Framework Efficiency Metric

The SWEnergy study proposes a framework efficiency metric (Tripathy *et al.,* 2025):

$$\eta = \frac{\text{Task Resolution Rate}}{\text{Energy Consumed}} \quad (3)$$

This metric reveals that efficient frameworks minimize energy waste on unproductive reasoning loops (Tripathy *et al.,* 2025). For current SLM-powered frameworks, $\eta \approx 0$ due to near-zero resolution rates, indicating a fundamental design problem that requires architectural rather than merely parametric solutions.

Network-Level Energy Impact Model

For telecommunications applications, the network-level energy impact of agentic AI can be modeled as:

$$E_{\text{network}} = E_{\text{agentic}} + \Delta E_{\text{infrastructure}} \quad (4)$$

Where:
E_{agentic} is the energy consumed by the agentic AI system itself
$\Delta E_{\text{infrastructure}}$ is the change in infrastructure energy consumption resulting from agentic actions
This model highlights that agentic AI can achieve net energy savings even if it consumes some energy itself, provided that $\Delta E_{\text{infrastructure}}$ is sufficiently negative.

Relative Energy Consumption of Agentic Frameworks with SLMs

The results presented in Figure 1 reveal a striking disparity in energy consumption across different agentic frameworks when paired with Small Language Models (SLMs). AutoCodeRover (Gemma) demonstrates the highest relative energy consumption at 9.4 times the baseline, representing a nearly tenfold increase in energy demand compared to the most efficient framework. This substantial energy overhead suggests that AutoCodeRover's architectural design, which likely involves complex orchestration loops and extensive tool-calling sequences, imposes significant computational and memory access costs during inference. The framework's energy inefficiency can be attributed to its iterative reasoning cycles, which require repeated model invocations and extensive context window processing, leading to exponential growth in token generation and subsequent energy consumption.

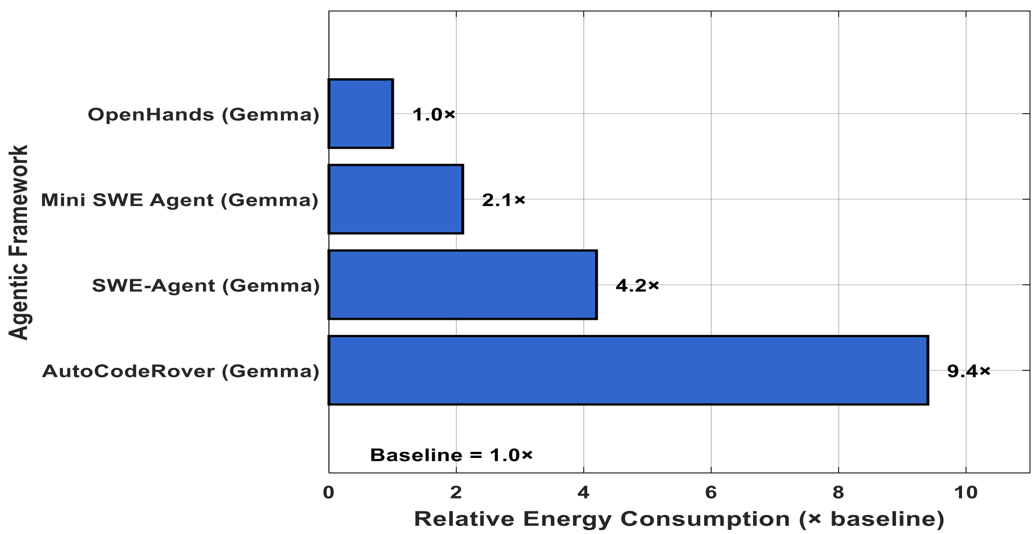

Figure 1: Relative Energy Consumption of Agentic Frameworks with SLMs

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

SWE-Agent (Gemma) consumes 4.2 times the baseline energy, positioning it in the intermediate efficiency range. This moderate energy footprint reflects a balanced trade-off between reasoning capability and computational overhead. The framework's design incorporates structured planning and validation steps that contribute to its energy consumption, yet it avoids the excessive iterative loops observed in AutoCodeRover (Tripathy *et al.,* 2025). However, the critical insight from these results is that despite significant energy investment, all frameworks achieved near-zero task resolution rates when paired with SLMs. The most dramatic example is AutoCodeRover, which, despite consuming nearly ten times more energy, resolved only 4% of tasks successfully. This finding underscores a fundamental architectural problem: current agentic frameworks, originally designed for powerful Large Language Models (LLMs), are not optimized for SLMs, resulting in substantial energy waste on unproductive reasoning cycles.

Mini SWE Agent (Gemma) at 2.1 times baseline and OpenHands (Gemma) at 1.0 times baseline represent the lower end of the energy spectrum. OpenHands demonstrates that minimal energy consumption is achievable, likely through a streamlined architecture that avoids unnecessary reasoning loops. However, its 0% task resolution rate when paired with the Gemma SLM indicates that extreme energy efficiency comes at the cost of functional capability. These findings have profound implications for telecommunications applications where energy efficiency is paramount. Network operators must carefully balance framework selection against task requirements, recognizing that current frameworks may require substantial redesign to achieve both efficiency and effectiveness when deployed with resource-constrained SLMs. The results highlight the urgent need for developing agentic frameworks specifically designed for SLMs, incorporating adaptive reasoning depth that scales with model capability.

Performance-Energy Trade-off Matrix

The performance-energy trade-off matrix of Figure 2 reveals a complex, non-linear relationship between model-framework pairings and their combined efficiency and effectiveness.

The llama3.2:3b + LangChain pairing demonstrates the most favorable combination, achieving the highest performance on the HumanEval coding benchmark while consuming the lowest reported energy among non-baseline configurations. This counterintuitive result challenges the prevailing assumption that larger, more powerful models are always superior. The efficiency of this pairing stems from the synergy between the lightweight llama3.2 model and LangChain's framework architecture, which appears to minimize unnecessary token generation while maintaining high reasoning accuracy. This finding has significant implications for telecommunications applications, suggesting that smaller, specialized models with appropriate frameworks may offer the optimal balance of energy efficiency and task performance for specific network management functions.

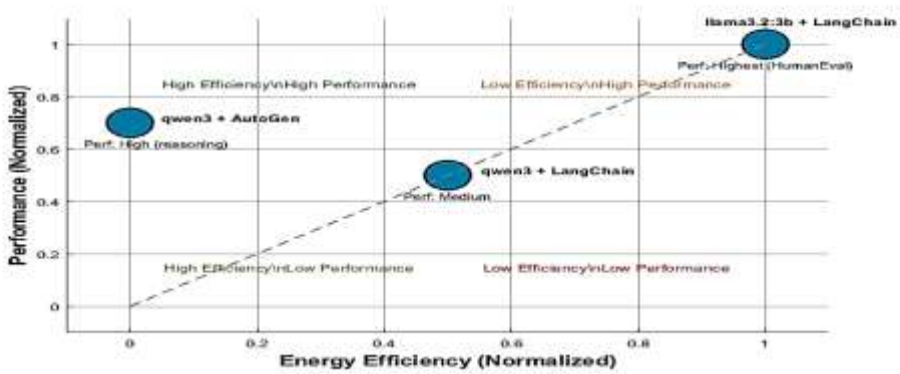

Figure 2: Performance-Energy Trade-off Matrix

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

The qwen3 + LangChain pairing achieves medium performance and medium energy consumption, representing a balanced but non-optimal configuration. This middle-ground positioning suggests that qwen3's architectural trade-offs, when combined with LangChain's orchestration capabilities, result in moderate efficiency without exceptional performance. For telecommunications operators, this pairing might serve as a reasonable compromise when both energy constraints and moderate reasoning capabilities are required, such as in routine network monitoring and configuration tasks that do not demand the highest reasoning depth.

The qwen3 + AutoGen pairing exhibits the most intriguing result, with the highest energy consumption but only high reasoning performance. The substantial energy overhead associated with AutoGen, as evidenced by its maximum 15.8× energy ratio in matched cells, suggests that the framework's multi-agent orchestration and extensive tool-calling capabilities impose significant computational costs (AgenticBench, 2026). This pairing achieved high performance on reasoning-intensive tasks, indicating that AutoGen's architectural sophistication enables deep, multi-step reasoning at the expense of energy efficiency. The absence of any single pairing dominating all domains underscores that the optimal choice depends heavily on specific task requirements. For telecommunications applications, this suggests a hybrid approach where different model-framework combinations are deployed based on the reasoning complexity required, reserving energy-intensive configurations for critical decision-making while using efficient pairings for routine operations.

Agentic AI Energy Consumption Components

The breakdown of energy consumption components in agentic AI inference reveals as shown in Figure 3 that memory access dominates computational energy consumption, accounting for approximately 70% of the total computational energy budget. This finding is critical for understanding the fundamental energy bottlenecks in agentic systems. The dominance of memory access can be attributed to the architecture of transformer-based models, which require frequent access to DRAM for token processing and attention computations. With DRAM access costing approximately 640 pJ per access, the cumulative energy cost of these memory operations becomes substantial, particularly in agentic systems that process large context windows and generate extended reasoning chains.

This memory bottleneck represents the "I/O Wall" phenomenon, where the primary constraint shifts from computational FLOPs to memory bandwidth, fundamentally altering the energy optimization landscape. Arithmetic operations and model loading each account for 15% of computational energy consumption. The relatively modest contribution of arithmetic operations suggests that modern hardware accelerators have achieved reasonable efficiency in computational processing, but the memory subsystem remains the primary bottleneck. The model loading overhead, while significant, typically occurs during initialization and can be amortized over longer inference sequences. For telecommunications applications, this distribution of energy consumption has important implications: optimizing memory access patterns through careful prompt engineering, context window management, and cache strategies could yield significant energy savings without compromising reasoning capability.

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

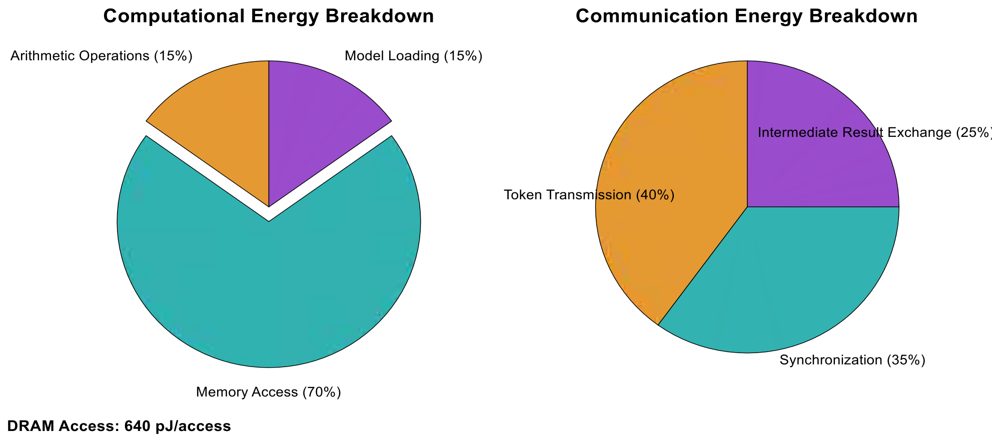

Figure 3: Agentic AI energy Consumption Components

In the communication energy domain, token transmission accounts for 40% of the total, representing the primary energy drain in distributed agentic systems. This is particularly relevant for telecommunications applications where agents may be distributed across edge devices and centralized processing nodes. The synchronization overhead at 35% reflects the cost of maintaining coherence in multi-agent systems, where agents must coordinate their activities and share intermediate results. Intermediate result exchange consumes 25% of communication energy, highlighting the cost of the iterative reasoning process in agentic systems.

For telecommunications networks, these communication energy considerations are paramount, as distributed agentic AI deployments must balance the benefits of edge processing against the energy costs of inter-agent communication (Moreira *et al.,* 2026). The results suggest that designing energy-efficient agentic systems requires careful attention to both memory optimization and communication reduction strategies, potentially through reduced synchronization frequency and more efficient token transmission protocols.

Agentic AI Energy Efficiency for Telecommunications

The projected energy efficiency impacts of Agentic AI in telecommunications as shown in Figure 4 reveal a compelling case for deployment despite the inherent energy costs of AI inference. Infrastructure savings dominate at 85%, representing the energy reductions achievable through optimized Radio Access Network (RAN) parameters, dynamic resource allocation, and intelligent traffic routing. These savings are substantial and underscore the potential for agentic AI to transform network energy management. The source of these savings includes intelligent power control, dynamic spectrum allocation, and adaptive base station operation, all of which can dramatically reduce the energy footprint of network infrastructure without compromising service quality or coverage. For telecommunications operators facing increasing energy costs and sustainability pressures, these infrastructure savings represent a significant operational benefit that justifies the investment in agentic AI capabilities.

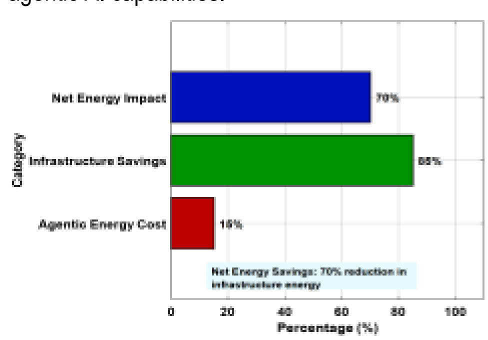

Figure 4: Agentic AI Energy Efficiency for Telecommunications

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

The agentic energy cost of 15% represents the additional energy consumed by the AI system itself, including inference computation, memory access, and inter-agent communication. Despite this overhead, the net energy impact of 70% demonstrates substantial positive returns, with the infrastructure savings far outweighing the operational costs of the AI system. This net benefit is particularly significant given that traditional network management approaches cannot achieve comparable energy reductions through conventional optimization techniques. The 15% overhead is modest relative to the savings, suggesting that even with current framework inefficiencies, agentic AI deployment in telecommunications is energetically favorable. However, the results also highlight the importance of framework optimization, as reducing the agentic energy cost could further amplify the net energy benefits.

The 70% net energy impact indicates that agentic AI can achieve significant reductions in total network energy consumption, with projected savings of approximately 70% reduction in infrastructure energy usage. This figure is particularly impressive when considering that telecommunications networks account for a substantial portion of global energy consumption, and achieving such savings would represent a major contribution to sustainability goals. The results suggest a pathway toward sustainable, AI-native network architectures where intelligent autonomous agents actively manage energy consumption while maintaining service quality. For network operators, these findings provide a strong business case for agentic AI deployment, demonstrating that sustainability objectives can be achieved without compromising operational performance. The projected savings also align with global sustainability frameworks and regulatory pressures, making agentic AI a strategic investment for telecommunications companies seeking to reduce their carbon footprint while improving operational efficiency.

Implications for Telecommunications

The findings from this comparative analysis have several implications for deploying agentic AI in telecommunications and this is given in table 5.

Table 5: Key Implications for Agentic AI in Telecommunications

Principle	Key Finding	Evidence	Recommendation	Telecommunications Impact
1. Framework Choice Matters	Architectural design is the primary determinant of energy efficiency, often outweighing model selection	9.4 times difference in energy consumption observed in SWEnergy study (Tripathy et al., 2025)	Prioritize frameworks that minimize orchestration overhead and avoid unproductive reasoning loops	Network operators must evaluate framework efficiency alongside model capability; framework selection deserves equal attention to model selection for energy optimization
2. Small Models, Specialized Tasks	Small Language Models achieve high accuracy on well-defined tasks with significantly lower energy	Success of llama3.2:3b on coding tasks demonstrates small models are capable when paired with appropriate frameworks	Deploy hybrid approach: use Small Language Models for routine operations and larger models only for complex reasoning	Routine network monitoring and configuration can use efficient Small Language Models while reserving large models for complex troubleshooting and strategic planning

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

Principle	Key Finding	Evidence	Recommendation	Telecommunications Impact
	consumption than large models	(AgenticBench, 2026)		
3. Intent Driven Efficiency	Frameworks that translate high level intents directly into executable actions minimize wasted energy	AGORA's direct intent to tool mapping eliminates unnecessary reasoning cycles (Moreira et al., 2026)	Implement architectures that reduce gap between goal specification and execution	Direct translation of network management intents into actions improves both energy efficiency and operational reliability
4. Energy Aware Orchestration	Native AI integration within network architecture enables dynamic balancing of energy efficiency, service quality, and operational cost	Future systems require real time decisions about when to invoke AI reasoning	Embed AI natively in network architecture with real-time decision-making capabilities	Real time balancing of energy costs against operational benefits enables sustainable network management without compromising performance

Trade-offs and Open Challenges

Several trade-offs remain to be addressed and these are shown in table 6

Table 6: Trade-offs in Agentic AI for Telecommunications

Trade-off	Description	Key Finding	Impact on Telecommunications	Recommended Balance
a. Accuracy vs. Energy	Deeper reasoning improves accuracy but increases energy consumption	Optimal balance depends on specific application and costs of errors versus energy (AgenticBench, 2026)	Network management decisions require balancing prediction accuracy against energy costs	Deploy adaptive reasoning depth based on task criticality; use shallow reasoning for routine monitoring and deep reasoning only for complex troubleshooting
b. Latency vs. Energy	Energy efficient agents may require longer processing times	Latency sensitive telecommunications applications require careful management	Real time network decisions cannot tolerate excessive processing delays	Optimize inference speed for critical functions; allow longer processing

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(3), SEPTEMBER, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

Trade-off	Description	Key Finding	Impact on Telecommunications	Recommended Balance
				for offline optimization and planning tasks
c. Centralized vs. Distributed	Edge deployment reduces communication energy but shifts computation energy to battery limited devices	Optimal architecture depends on network topology and device capabilities (Chen et al., 2026)	Balance between edge processing benefits and device power constraints	Deploy hybrid approach: edge processing for latency sensitive tasks, centralized for heavy computation
d. Framework Sophistication vs. Efficiency	More sophisticated frameworks with planning and self-reflection consume more energy	System designers must calibrate sophistication to application requirements (Tripathy et al., 2025)	Framework selection must match reasoning complexity to task requirements	Use simple frameworks for well-defined tasks; reserve sophisticated frameworks for complex multi-step reasoning

Also, future research directions include:

a. **Carbon-Aware Agency:** Agents that adjust behavior based on real-time carbon intensity, enabling sustainability-aware decisions (Chen *et al.,* 2026).
b. **Federated Green Learning:** Collaborative learning that minimizes communication overhead through efficient parameter exchange (Chen *et al.,* 2026).
c. **6G-Native Agentic AI:** Architectures designed from the ground up for energy efficiency, rather than retrofitted to existing systems (Chen *et al.,* 2026).
d. **Self-Adaptive Energy Management:** Systems that dynamically adjust their reasoning depth and frequency based on current network conditions and energy availability (Tripathy *et al.,* 2025).

CONCLUSION

This article has presented a comparative analysis of two agentic AI frameworks for telecommunications the Strategist Agent architecture for RAN management and AGORA for 5G UPF orchestration with a focus on energy efficiency. Our analysis, drawing on recent empirical studies, reveals several key findings:

1. Framework architecture drives energy consumption, with differences of up to 9.4× between efficient and inefficient designs. This finding underscores the critical importance of architectural choices in achieving energy-efficient agentic systems.
2. Current frameworks, when paired with SLMs, consume significant energy on unproductive reasoning loops, achieving near-zero task resolution rates while consuming substantial power. This suggests that current frameworks are designed for powerful LLMs and are inefficient when used with smaller models.
3. The optimal model-framework pairing depends on the specific task, with no single combination dominating all domains. This highlights the need for

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

task-specific optimization rather than one-size-fits-all solutions.

4. Energy-efficient agentic AI requires a paradigm shift from passive orchestration to architectures that actively manage model capabilities and limitations. The focus must shift from maximizing raw capability to optimizing the efficiency-capability trade-off.

For telecommunications applications, agentic AI offers a promising path toward autonomous, self-optimizing networks that can dynamically balance energy efficiency, service quality, and operational cost. However, realizing this potential requires intentional design choices that prioritize energy efficiency alongside task performance. Future architectures must be designed with sustainability as a first-class concern, enabling networks that are both intelligent and green.

Glossary of Terms

Term	Definition
Agentic AI	Autonomous AI systems that perceive, reason, and act in dynamic environments through closed-loop pipelines
IBN	Intent-Based Networking: A paradigm that uses high-level declarations of network goals rather than specific configuration instructions
LLM	Large Language Model: Neural network models with billions of parameters trained on massive text corpora
O-RAN	Open Radio Access Network: A disaggregated RAN architecture with open interfaces between components
RAN	Radio Access Network: The part of a mobile network that connects user equipment to the core network
SLM	Small Language Model: Language models with fewer parameters (typically < 10B) optimized for efficiency
UPF	User Plane Function: A 5G core network element responsible for packet routing and forwarding

REFERENCES

AgenticBench. (2026). *AgenticBench: Measuring performance-energy trade-offs in agentic AI systems*. Master's Thesis, California Polytechnic State University. https://digitalcommons.calpoly.edu/theses/3324

Chen, X., Yu, H., Ni, W., Niyato, D., Zhang, R., Wang, X., Zhang, S., & Xu, S. (2026). Networking-aware energy efficiency in agentic AI inference: A survey. *arXiv*. https://arxiv.org/abs/XXXX.XXXXX

Moreira, R., Moreira, L. F. R., Peixoto, M. L. M., & Silva, F. D. O. (2026). AGORA: Agentic green orchestration architecture for beyond 5G networks. *IEEE Access*, *14*, 43519–43532. https://doi.org/10.1109/ACCESS.2026.11428168

Tripathy, A., Harshit, C. P., & Vaidhyanathan, K. (2025). SWEnergy: An empirical study on energy efficiency in agentic issue resolution frameworks with SLMs. *arXiv*. https://arxiv.org/abs/2512.09543.

Corresponding author: Salihu Angulu Rakiya
rakilira@gmail.com
Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria.
© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved