Project GAINS

(Generative AI for Network Sustainability)

Deliverable D1.1

State of the Art Report on Agent Architectures Based on Large Language Model (LLM) and Its Energy Impact

Giordano, A. – McGann, O. – Fruncillo, D. – Haider, B.

Work Package: WP1 Problem Framing and Analysis of the State of the Art

Status: Final | Version: 1.0

Classification: Internal project use; controlled sharing with partners as needed.

Abstract

This work synthesizes the state of the art on agent architectures based on Large Language Models (LLM) applied to Network Monitoring & Management (NMM) in converging infrastructures (access, edge, and cloud) and frames its energy impact. Starting from the most recent literature on the use of LLMs in networking, telecom-domain models, and emerging forms of autonomous and intent-driven management, the report delineates relevant use cases for operations, configuration, change management, and security, highlighting the main failure modes (intent ambiguity, configuration synthesis errors, drift, and version misalignment) and the necessary mechanisms to achieve industrial reliability. The analysis integrates patterns of Retrieval-Augmented Generation (RAG) and hybrid retrieval, with an emphasis on strong metadata filters and source provenance to ensure grounding and auditability. In terms of energy, the literature converges on an operational point: in multi-step workflows, consumption increases primarily with calls and tokens (prefill/decode), thus the study prioritizes measurable optimizations on calls, context, and retries. These findings are concise and depend on runtime and hardware with an impact verification through PoC with measures per request and per token. In terms of energy, the document consolidates evidence that, in typical continuous loads of NMM and multi-step agent systems, the dominant drivers of the footprint are the number of calls to the model and the length of the context, rather than just the model size; this leads to the priority of measurable optimizations based on context reduction, routing, and model cascading, speculative decoding, and observability per phase (prefill/decode) and per token (Joule/token). Finally, the deliverable proposes a replicable experimental setup with retrieval metrics (Recall@k, MRR, nDCG), pipeline metrics (LLM calls, tokens, retries, escalations), and attributable energy measures, and presents a comparative evaluation of the main open-source technologies and relevant licenses for industrialization (RAG frameworks, vector databases, workflow/agent builders, tool integration standards, and permissive/copy-left/fair-code licenses).

Table of Contents