

# Progetto GAINS

(Generative AI for Network Sustainability)

Deliverable D1.1

Report sullo Stato dell'Arte su Architetture ad Agenti Basate su Large Language Model (LLM) e Relativo Impatto Energetico

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

Work Package: WP1 Problem Framing ed Analisi dello Stato dell'Arte

Stato: Definitivo | Versione: 1.0

Classificazione: Uso interno di progetto Condivisione controllata con partner su necessità

## Abstract

Questo lavoro sintetizza lo stato dell'arte sulle architetture ad agenti basate su Large Language Model (LLM) applicate al Network Monitoring & Management (NMM) in infrastrutture convergenti (accesso, edge e cloud) e ne inquadra l'impatto energetico. Muovendo dalla letteratura più recente sull'impiego degli LLM nel networking, sui modelli telecom-domain e sulle forme emergenti di gestione autonoma e intent-driven, il report delimita i casi d'uso rilevanti per operation, configurazione, change management e sicurezza, evidenziando i principali failure mode (ambiguità dell'intent, errori di sintesi di configurazioni, drift e disallineamento di versione) e i meccanismi necessari per ottenere affidabilità industriale. L'analisi integra pattern di Retrieval-Augmented Generation (RAG) e retrieval ibrido, con enfasi su filtri forti di metadati e provenienza delle fonti per garantire grounding e auditabilità. Sul piano energetico, la letteratura converge su un punto operativo: nei workflow multi-step il consumo cresce soprattutto con chiamate e token (prefill/decode), quindi lo studio privilegia ottimizzazioni misurabili su chiamate, contesto e retry. Queste evidenze sono sintetiche e dipendono da runtime e hardware con una verifica dell'impatto tramite PoC con misure per richiesta e per token. Sul piano energetico, il documento consolida evidenze secondo cui, nei carichi continui tipici di NMM e nei sistemi agentici multi-step, i driver dominanti del footprint sono il numero di chiamate al modello e la lunghezza del contesto, più che la sola dimensione del modello; ne deriva la priorità di ottimizzazioni misurabili basate su riduzione del contesto, routing e cascata di modelli, speculative decoding, e osservabilità per fase (prefill/decode) e per token (Joule/token). Il deliverable propone infine un impianto sperimentale replicabile con metriche di retrieval (Recall@k, MRR, nDCG), metriche di pipeline (chiamate LLM, token, retry, escalation) e misure energetiche attribuibili, e presenta una valutazione comparativa delle principali tecnologie open source e delle licenze rilevanti per industrializzazione (framework RAG, database vettoriali, workflow/agent builders, standard di integrazione tool e licenze permissive/copyleft/fair-code).

# Indice

## 1. Introduzione

## 2. Metodi

2.1 Impostazione dell'analisi e selezione delle evidenze

2.2 Criteri di maturità e idoneità industriale

2.3 Disegno sperimentale replicabile: dataset, metriche e misure energetiche

## 3. Risultati

3.1 Tassonomia dei casi d'uso e downstream task per NMM

3.2 Evoluzione architetturale: da "LLM che risponde" a "LLM che orchestra"

3.3 RAG e retrieval in domini tecnici: ibrido, filtri forti e provenienza

3.4 Impatto energetico: driver dominanti e ottimizzazioni misurabili

3.5 Analisi integrale dei progetti open source e delle tecnologie emergenti

3.5.1 Standard e framework per integrazione e orchestrazione

3.5.2 Framework RAG e agent builders

3.5.3 Vector database e retrieval ibrido con metadati forti

3.5.4 Piattaforme agentiche orientate a governance e controllo del dato

3.5.5 Tecnologie emergenti: ecosistemi, licenze e vincoli d'uso

3.5.6 Implicazioni di licenza: trasferibilità e compliance

3.6 Direzioni strategiche: agenti "sovrani" e AI factories

## 4. Discussione

## 5. Conclusioni

## Bibliografia