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