**Programa Institucional de Bolsas de Iniciação Científica – PIBIC**

**RELATÓRIO PARCIAL**

IDENTIFICAÇÃO DO PROJETO

Título do Projeto: Um sistema de gerenciamento baseado em intenção para a Open RAN: Automação da configuração de redes via Chatbot

Local de Realização (Unidade/Instituto/Departamento/Laboratorio):\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Endereço:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Bairro: Boa Viagem Cidade: Niterói UF: RJ CEP:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# **Matrícula:**.\_\_\_\_\_\_\_.\_\_\_\_\_\_\_-\_\_\_\_ **C.R.:** \_\_\_\_\_\_\_\_

Curso: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

DADOS DO ORIENTADOR

Nome:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Matrícula Siape:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_CPF:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Endereço:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Bairro:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Cidade:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_UF:\_\_\_\_\_\_\_\_\_\_\_\_\_\_CEP:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_E-mail:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_-\_\_\_\_\_\_\_\_\_ Telefone 1: (\_\_\_\_\_)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Telefone 2: (\_\_\_\_\_)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

DADOS DO BOLSISTA

Nome: Igor Rodrigues Alves

Matrícula: 123031006 CPF: 159.046.507-52 CR: 8.2

Curso/Departamento/Instituto: Ciência da Computação/Departamento de Ciência da Computação/Instituto de Computação

Endereço: Rua Noronha Torrezão, 395, bloco 9, apartamento 403

Bairro: Santa Rosa Cidade: Niterói UF: RJ CEP: 24240-181

E-mail: ig\_rodrigues@id.uff.br

Telefone 1: (21)99142-6202 Telefone 2: (21)96614-7944

##### 

**INTRODUÇÃO**

As redes de acesso via rádio (*Radio Access Network* - RAN) abertas vêm ganhando destaque por permitir interoperabilidade entre equipamentos de diferentes fornecedores, em contraste com as implementações atuais, as quais tendem a ser monolíticas e fechadas, impedindo inovação e interoperabilidade (BONATI et al.,2023). A arquitetura de RAN aberta desacopla *software* e *hardware* e estabelece padrões de interfaces abertas, tornando a RAN mais flexível.

Com a rede de acesso via rádio aberta, a arquitetura tradicional de estação base é desagregada, e controladores inteligentes de RAN (RAN *Intelligent Controllers* - RIC) são utilizados para a otimização do gerenciamento de recursos em tempo real. Em um cenário complexo, o uso de RICs torna a configuração e monitoramento da rede automáticos, permitindo o Gerenciamento Baseado em Intenção (*Intent-Based Management* - IBM). Por meio do IBM, o operador de rede pode realizar a configuração da rede com sentenças imperativas em linguagem natural, chamadas de intenções, as quais serão traduzidas em políticas bem definidas para serem aplicadas e monitoradas, garantindo os resultados desejados (JACOB et al., 2021).

Os Modelos de Linguagem em Larga Escala (*Large Language Models* - LLMs) são modelos avançados de aprendizado profundo que processam e geram linguagem natural com base em grandes quantidades de dados, com o objetivo de entender e criar textos coerentes e relevantes. Como esses modelos demandam muitos recursos computacionais e de memória, a técnica da quantização é promissora, pois essa técnica reduz a precisão numérica na representação dos parâmetros, melhorando a eficiência computacional sem perda significativa de desempenho. No contexto do Gerenciamento Baseado em Intenção, os LLMs são uma interface de linguagem natural que atuam como assistentes inteligentes para interpretar intenções e configurar redes, podendo ser executados em nuvem ou localmente em dispositivos periféricos.

Este trabalho tem como objetivo propor um sistema que automatiza a configuração de políticas na RAN com base em intenções expressas em linguagem natural por meio de um *chatbot*. Este sistema é dividido em dois módulos principais: o Agente Conversacional, um LLM responsável pela extração de informações relevantes, como entidades de rede e métricas, das intenções enviadas pelo usuário; e o sistema operacional de rede ONOS (*Open Network Operating System*), que recebe as intenções pré-processadas pelo Agente Conversacional e as traduz em expressões com parâmetros de baixo nível, as quais são enviadas para a rede por meio do ONOS *Intent Framework*. Para atingir esta meta, foi realizado um estudo da capacidade dos modelos de linguagem em larga escala, incluindo suas versões quantizadas, de interpretar, classificar e captar parâmetros relevantes nas intenções de rede fornecidas.

**METODOLOGIA**

A seguinte análise foi realizada utilizando modelos personalizados da OpenWebUI, uma plataforma de código aberto e extensível para a interação entre com modelos de inteligência artificial, especialmente LLMs. A plataforma disponibiliza um *frontend* próprio e intuitivo, capaz de receber intenções através de uma interface no formato de um assistente conversacional. Também oferece algumas funcionalidades para compor um *backend* personalizado, como a possibilidade de escolher o modelo base; a descrição do *prompt*, isto é, a instrução básica do modelo, a qual define textualmente a tarefa a ser desempenhada por ele, a qualidade do *prompt* influencia diretamente a qualidade das respostas do modelo; o desenvolvimento de ferramentas, *scripts* Python que expandem as capacidades do modelo, permitindo a execução de rotinas específicas ao longo de um diálogo, sem forçar seu uso. Neste trabalho, foram utilizados os modelos Llama, produzidos pela Meta e Qwen, produzidos pelo grupo Alibaba, bem como suas versões quantizadas de 4 e 8 bits. A Figura 1 mostra o fluxograma composto pelas quatro etapas principais executadas por cada variação dos modelos base.

Na Etapa de Ingestão, múltiplas intenções inicialmente armazenadas em uma base de dados são inseridas na interface *web* do agente de conversação.

Na Etapa de Interpretação, cada intenção é processada individualmente. Independentemente do modelo base e da quantização aplicada, as intenções são submetidas a três ferramentas desenvolvidas internamente na plataforma OpenWebUI. A primeira, denominada Classificador de Intenções, categoriza cada intenção em uma das três classes definidas, considerando o contexto da intenção, sem se limitar ao uso de palavras-chave específicas. Em seguida, as intenções classificadas são analisadas pelo Extrator de Parâmetros, responsável por identificar e armazenar os elementos essenciais para a interpretação da intenção. Os parâmetros extraídos são organizados em uma estrutura JSON que inclui informações como: categoria, ação solicitada, entidades envolvidas, magnitude, unidade de medida e os carimbos de data e hora de início e fim da validade da intenção. Esse processo também remove ruídos textuais, desconsiderando estruturas gramaticais e ortográficas irrelevantes, como artigos, pronomes e pontuação.

Na Etapa de Avaliação Textual, os parâmetros extraídos e as classificações fornecidas pelo modelo personalizado são comparados com os valores obtidos a partir de uma anotação manual prévia. Com base nessa comparação, métricas, como F1-score, *recall*, precisão e acurácia, podem ser calculadas. A acurácia é calculada em relação aos parâmetros das categorias, a fim de mensurar a capacidade do modelo de acertar os parâmetros de uma categoria dado que ele previu corretamente essa categoria. Já as demais métricas têm relação com a capacidade de prever a categoria de intenção.

Finalmente, na Etapa de Avaliação de Custo, os parâmetros extraídos são enviados ao controlador ONOS. Nessa etapa, verifica-se o consumo de memória e de processamento durante a execução dos modelos. A avaliação realizada busca validar a capacidade dos modelos personalizados na compreensão de intenções textualmente complexas.

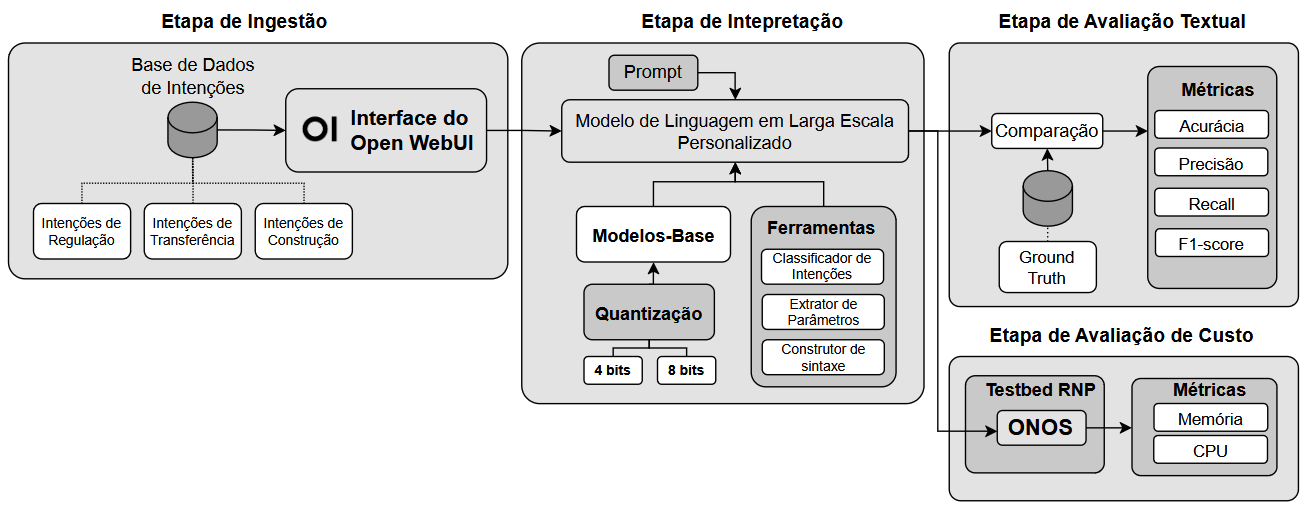


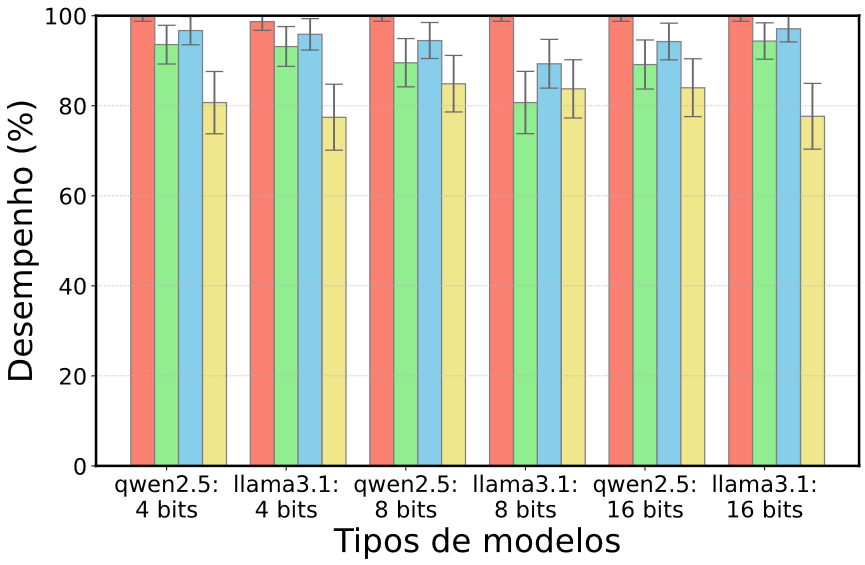
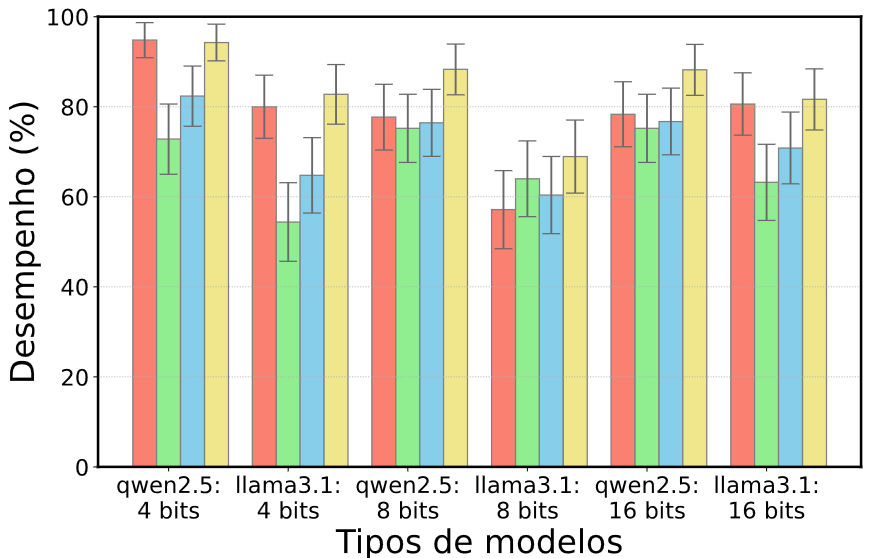
Figura 1. Fluxograma da análise proposta consiste nas quatro etapas sequenciais descritas.

Este processo foi realizado em um conjunto de dados rotulados manualmente, contendo 500 intenções em linguagem natural e seus respectivos parâmetros. As intenções possuem três categorias distintas, sendo elas “Construção”, “Regulação” e “Transferência”. Em geral, as intenções da categoria “Construção” têm o objetivo de estabelecer conexões entre aplicativos ou dispositivos. As ações nesta categoria são: “Conectar”, que conecta os alvos especificados; “Descobrir”, a qual busca outros aplicativos fornecidos por determinados dispositivos; e “Anunciar”, que visa divulgar um serviço disponível capaz de interagir com outros aplicativos. As ações da categoria “Regulação” manipulam o tráfego de rede de acordo com as especificações de entrada. A ação “Priorizar” é usada para garantir que um alvo tenha maior prioridade na rede. A ação “Bloquear” é empregada para bloquear o tráfego de certos alvos, impedindo sua interação com outros alvos. Já a ação “Limitar” define uma restrição a um requisito de rede específico, garantindo que não exceda o valor expresso na entrada. Por fim, as intenções na categoria “Transferência” permitem que as aplicações executem ações como enviar (*push*) e receber (*pull*) conteúdos de aplicações e servidores capazes de atender a essas requisições.

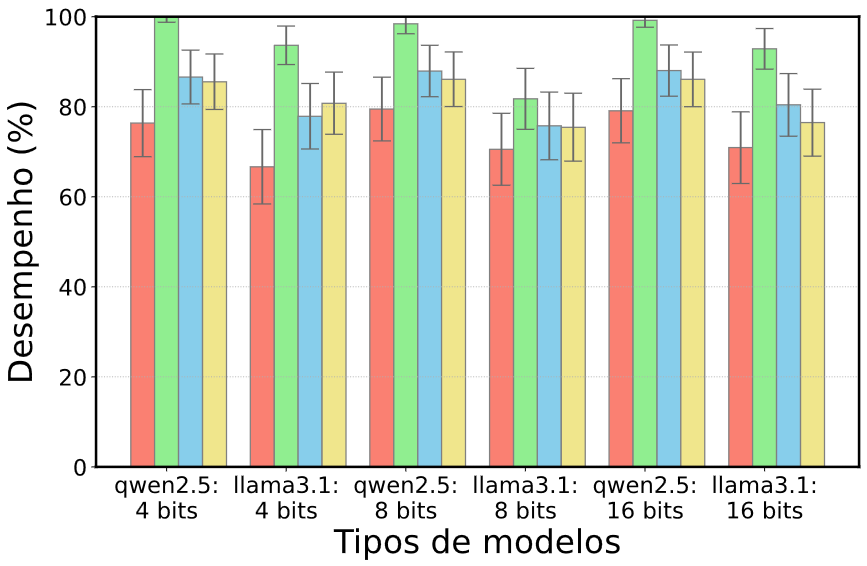
**RESULTADOS**

Foram utilizados dois modelos para avaliação: Llama 3.1:8B, LLM da Meta treinado com 8 bilhões de parâmetros, e Qwen2.5:7B, LLM do grupo chinês Alibaba, treinado com 7 bilhões de parâmetros. Estes modelos foram avaliados nas suas versões com pesos em precisões de ponto flutuante de 16 bits (FP16), números inteiros de 8 bits (INT8) e números inteiros de 4 bits (INT4). A versão FP16 serve como referência para a análise realizada, pois é a versão sem qualquer tipo de quantização. A versão quantizada INT4 do modelo Llama 3.1:8B segue a quantização de 4 bits *K-means*, uma técnica que utiliza o algoritmo de mesmo nome para agrupar os valores dos pesos do modelo em *clusters*. Cada *cluster* é representado por um valor central e os pesos são aproximados para o centróide mais próximo. A versão quantizada INT8 do modelo Llama 3.1:8B e as versões INT4 e INT8 do modelo Qwen2.5:7B seguem uma técnica de quantização direta, na qual os valores dos pesos são diretamente mapeados para um conjunto fixo de valores.

A avaliação dos efeitos da quantização sobre os modelos de larga escala é realizada através da comparação dos modelos Llama3.1:8B e Qwen2.5:7B em diferentes versões de quantização (FP16, INT8 e INT4), considerando métricas de precisão, *recall* e *F1-score*. O objetivo é verificar o impacto da quantização no desempenho dos modelos, utilizando a versão FP16, modelo base, como referência para análise comparativa. A máquina utilizada para os experimentos é equipada com um processador Intel Core i7-13700KF (13ª geração), contendo 16 núcleos e 24 threads, com frequência base de 800 MHz e máxima de 5300 MHz. A máquina possui 128 GB de memória RAM DDR5. Para a execução dos modelos de linguagem em larga escala foram utilizadas duas unidades de processamento gráfico NVIDIA GeForce RTX 4060 Ti com 16 GB de memória em cada. O armazenamento é realizado em um SSD NVMe de 2 TB.



a) Intenções de Construção b) Intenções de Regulação



c) Intenções de Transferência

**Figura 2. Análise de desempenho de LLMs não quantizados (FP16) e com diferentes níveis de quantização (INT8 e INT4). Os modelos Llama 3.1:8B e Qwen2.5:7B são analisados quanto a precisão, *recall e* F1-*score* na inferência de intenções do tipo Construção, Regulação e Transferência.**

De maneira geral, esses resultados indicam que a quantização pode, em certas situações, proporcionar desempenhos comparáveis à versão não quantizada dos modelos-base. O modelo Qwen2.5 se destaca na extração de intenções relacionadas à categoria Transferência e demonstra uma melhor precisão média na categoria Construção. Por outro lado, os modelos Llama3.1 e Qwen2.5 apresentam dificuldades na categoria Construção, possivelmente devido à menor diversidade de ações possíveis nessa classe. Ademais, os padrões explícitos das intenções de Regulação podem ter facilitado a categorização, contribuindo para um melhor desempenho nessa categoria.

**PRODUÇÃO TÉCNICO-CIENTÍFICA**

Durante o desenvolvimento desta pesquisa, foi possível gerar um subproduto técnico-científico na forma de um relatório intitulado "Análise do Impacto da Quantização em Modelos de Linguagem em Larga Escala Aplicados ao Gerenciamento de Redes Baseadas em Intenção". O trabalho foi submetido ao WPerformance 2025 (*Workshop* em Desempenho de Sistemas Computacionais e de Comunicação) e está em processo de avaliação e aguarda o resultado da revisão.

**CONCLUSÕES**

Os resultados demonstraram que a técnica de quantização reduz o custo de processamento sem comprometer de maneira significativa o desempenho dos modelos, evidenciando a viabilidade de uso de modelos quantizados para o gerenciamento de redes baseadas em intenção. Para trabalhos futuros, são consideradas técnicas para ampliar a precisão dos modelos no domínio de redes, como a geração aumentada por recuperação (*Retrieval Augmented Generation* - RAG) e o ajuste fino, com o objetivo de tornar os LLMs mais robustos e adaptados a cenários reais.

**AUTOAVALIAÇÃO DO ALUNO**

No decorrer deste projeto, procurei demonstrar proatividade e dedicação na execução das atividades propostas, buscando aplicar meus conhecimentos teóricos e práticos na área de estudo, além de aprender com novas situações e colegas mais experientes. Aprendi a organizar melhor meu tempo, a lidar com desafios metodológicos e a buscar soluções criativas para os problemas que surgiam. Acredito que houve crescimento tanto no âmbito acadêmico quanto no desenvolvimento de habilidades como autonomia, pensamento crítico e capacidade de análise. No entanto, identifico que ainda posso melhorar em aspectos como a gestão de prazos e a comunicação mais assertiva dos resultados parciais. Estou motivado a continuar evoluindo e contribuindo para o avanço da pesquisa, aplicando os feedbacks recebidos e mantendo o foco nos objetivos finais do projeto.

**REFERÊNCIAS BIBLIOGRÁFICAS**

BONATI, L.; POLESE, M.; D’ORO S.; BASAGNI, S., MELODIA, T. OpenRAN Gym: AI/ML development, data collection, and testing for O-RAN on PAWR platforms, *Computer Networks*, vol. 220, p. 109502, 2023. [Online]. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1389128622005369>

JACOBS, A. S.; PFITSCHER, R. J.; RIBEIRO, R. H.; FERREIRA, R. A.; GRANVILLE, L. Z.; WILLINGER, W.; RAO, S. G. Hey, Lumi! Using natural language for {intent-based} network management. In: 2021 USENIX Annual Technical Conference (USENIX ATC 21), 2021, pp. 625–639.

MARTINI, B.; GHARBAOUI, M.; CASTOLDI, P. “Intent-based network slicing for SDN vertical services with assurance: Context, design and preliminary experiments”, *Future Generation Computer Systems*, vol.

142, pp. 101–116, 2023. [Online]. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167739X2200437X>

YANG, J.; JIN, H.; TANG, R.; HAN, X.; FENG, Q.; JIANG, H.; ZHONG, S.; YIN, B.; HU, X. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *ACM Transaction on Knowledge Discovery from Data*, 18(6).