

Influência da Quantidade e Abrangência em Sistemas RAG para Gestão do Espectro

David Vinícius F. Moreira, Leandro da Silva Pires e Rodrigo S. Couto

Resumo—A gestão do espectro de radiofrequências exige a análise de vasta documentação, o que desafia a eficiência dos processos regulatórios. Este trabalho investiga a implementação de um assistente de IA baseado em Geração Aumentada por Recuperação (RAG) no contexto da regulamentação de telecomunicações brasileira. O foco é analisar como a quantidade e a abrangência dos dados impactam a qualidade das respostas, testando a resiliência de uma arquitetura RAG tradicional com o incremento de informações correlatas e ruído temático. Os resultados comprovam a estabilidade da qualidade das respostas ao acréscimo de dados e ruídos, validando a robustez da arquitetura RAG testada.

Palavras-Chave—Educação em Comunicações, LLM.

Abstract—Radio frequency spectrum management requires analyzing vast documentation, challenging regulatory efficiency. As a solution, this work investigates the implementation of an AI assistant based on Retrieval-Augmented Generation (RAG) within the Brazilian telecommunications regulatory context. The focus is on analyzing how the quantity and scope of data inserted into the RAG impact the quality of the responses provided. Thus, the resilience of a traditional RAG architecture is tested with the addition of correlated information and thematic noise. The results demonstrate the stability of response quality to the addition of data and noise, validating the robustness of the tested RAG architecture.

Keywords—Education in Communications, LLM.

I. INTRODUÇÃO

Como uma das competências da Agência Nacional de Telecomunicações (Anatel), a gestão do espectro compreende as atividades de planejamento e controle das radiofrequências para assegurar a eficiência técnica e a coexistência de serviços. Para viabilizar esse objetivo e proporcionar estabilidade no setor, a Agência estabelece um arcabouço normativo focado na redução de falhas operacionais e na prevenção de ambiguidades nos processos de consignação. Uma das normas desse conjunto é o Ato nº 915, que regula requisitos técnicos para o Serviço Limitado Privado (SLP). Entretanto, a disparidade nos perfis dos usuários, que variam de grandes operadoras a radioamadores individuais, frequentemente resulta em uma lacuna entre as exigências regulatórias e a capacidade de compreensão do público, gerando erros processuais e incertezas técnicas que consomem recursos da infraestrutura administrativa da Agência.

Nesse cenário, os Grandes Modelos de Linguagem (LLMs) surgem como ferramentas promissoras para o auxílio

na análise e interpretação de textos técnicos [1]. Para garantir precisão técnica e mitigar "alucinações" em setores especializados como a gestão de espectro [2], [3], utiliza-se a Geração Aumentada por Recuperação (RAG). O RAG atua como um filtro de precisão, fundamentando as respostas do modelo em documentos regulatórios específicos.

Nesse sentido, implementa-se um sistema de chatbot especializado na resolução de dúvidas relacionadas a procedimentos e aspectos regulatórios da Anatel. Esse assistente possui potencial para auxiliar em atividades de consulta regulatória, interpretação de requisitos técnicos, apoio à fiscalização, análise documental e disseminação de conhecimento institucional, contribuindo para maior eficiência no acesso à informação e na tomada de decisão no ecossistema de gestão do espectro.

Este trabalho investiga se a qualidade de um assistente baseado em RAG varia conforme a quantidade e diversidade dos dados de entrada, visando mensurar a influência da escala de dados e do ruído temático no desempenho do sistema.

O restante deste trabalho está organizado da seguinte forma: A Seção II apresenta a arquitetura do sistema implementado. A Seção III descreve a metodologia dos experimentos demonstrando os resultados e a Seção IV apresenta as considerações finais e direções futuras.

II. ARQUITETURA

A arquitetura do sistema implementa uma estrutura tradicional de RAG. O sistema baseia-se em fluxos de trabalho orquestrados pelo LangFlow, operando em um modelo cliente-servidor, no qual a lógica de processamento e as integrações são centralizadas no servidor. A arquitetura é dividida em dois fluxos principais e complementares: o incremento de dados e a inferência.

A Figura 1 apresenta a dinâmica do fluxo de incremento de dados. Esse fluxo é responsável pela alimentação e atualização contínua da base de conhecimento. O processo é iniciado com uma requisição que fornece uma URL do portal da Anatel por meio da interação 1. Um componente de extração realiza a raspagem dos dados, possibilitando a transformação das tabelas em formatos estruturados como *Markdown* ou Valores Separados por Vírgulas (*Comma-Separated Values*, CSV). Após a interação 2, o componente de segmentação divide o conteúdo em blocos (*chunks*) de tamanho predefinido para otimizar a busca semântica. Esses blocos são enviados ao componente de *embeddings* (interação 3), que os transforma em vetores numéricos, finalizando na interação 4 com o armazenamento no banco de dados vetorial para permitir consultas futuras.

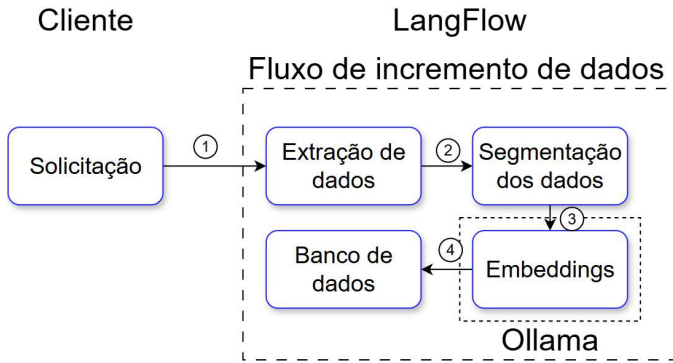


Fig. 1. Diagrama da arquitetura responsável por realizar a ingestão de dados no sistema RAG.

O fluxo de inferência gerencia o processamento das consultas em tempo real. A dinâmica desse fluxo é observada na Figura 2, onde o número presente na interação entre os blocos representa a ordem de execução e a letra é usada para diferenciar interações entre blocos que ocorrem simultaneamente. Ao receber uma pergunta, o sistema aciona simultaneamente os módulos de memória (1c), que recuperam o histórico de interações, e de *embeddings* (1a), que vetoriza a dúvida atual. O banco de dados recebe os vetores através da interação 2, realiza uma busca de similaridade e encaminha os fragmentos de texto mais relevantes ao formulador de *prompt* na interação 3. Este componente organiza a pergunta (1b), o histórico (1d) e o contexto recuperado (3) em uma estrutura única enviada à inferência (LLM) por meio da interação 4. A resposta gerada é então transmitida ao destinatário (5a) e armazenada na memória para preservar o contexto da conversação (5b).

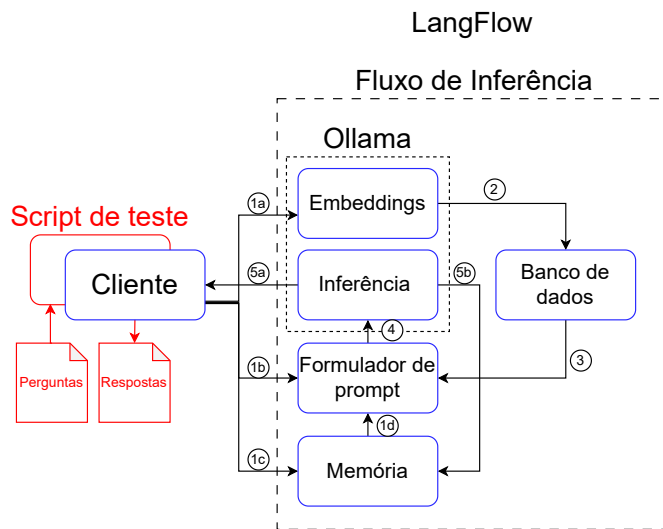


Fig. 2. Diagrama da arquitetura responsável por realizar a inferência do sistema RAG.

A arquitetura também permite a substituição do cliente convencional por um *script* de automação em Python, representado também na Figura 2 com a cor vermelha. Esse *script* atua como orquestrador das interações ao consumir

arquivos no formato *JavaScript Object Notation* (JSON), contendo as perguntas estruturadas e a especificação do modelo de linguagem a ser utilizado. Esse *script* insere as consultas de forma sistemática no fluxo de inferência e registra as saídas em arquivos de resultados, vinculando cada resposta ao seu respectivo emissor para eliminar a variabilidade da interação manual. Tal estrutura torna os dados coletados autossuficientes para a extração de métricas referente a qualidade das respostas.

III. EXPERIMENTOS

Esta seção detalha os experimentos realizados com o sistema RAG, voltados à análise de desempenho da ferramenta sob diferentes condições de e recuperação. A discussão aborda, inicialmente, o experimento que verifica a influência da quantidade de dados inseridos no sistema sobre a qualidade das respostas fornecidas. Na sequência, examina-se o impacto da abrangência desses dados, avaliando como a presença dessas informações na base de conhecimento afeta a qualidade da resposta do assistente regulatório.

A. Experimento da quantidade de dados no sistema RAG

Este experimento busca observar a influência da quantidade de dados na qualidade das respostas fornecidas pelo assistente. O objetivo central é compreender como o acréscimo de informações relacionadas, porém não idênticas, afeta a capacidade de recuperação e a assertividade do sistema. Os parâmetros do sistema RAG são fixados em blocos (*chunks*) de 1.000 caracteres com sobreposição (*overlap*) de 200 caracteres. A técnica de busca empregada é a *Maximal Marginal Relevance* (MMR), configurada para recuperar no máximo 10 blocos por consulta, visando otimizar a diversidade das informações recuperadas e reduzir a redundância no contexto enviado ao modelo. Além disso, o modelo usado para inferência é o Mistral-small 3.2: 24b, que possui o melhor desempenho nos testes realizados anteriormente.

O conjunto de dados do sistema RAG é composto pelo documento alvo e por outros documentos que são relacionados com o alvo. O documento alvo é o Ato nº 915 da Anatel referente ao SLP, que contém as respostas para todas as perguntas do teste. O conjunto de documentos relacionados é composto pelas Resoluções nº 773, nº 671 e nº 772, as quais abordam, respectivamente: as condições de uso de radiofrequências; o Plano de Atribuição, Destinação e Distribuição de Faixas de Frequências (PDFF); e uso do espectro de radiofrequências. Essas resoluções pertencem ao conjunto de dados relacionados devido ao elevado volume de caracteres que apresentam.

Para entender o impacto da diluição da informação, as avaliações ocorrem em quatro configurações distintas de proporção entre o documento alvo e os dados relacionados: 100% de presença do documento alvo (atuando como linha de base), seguido por composições de 75%, 50% e 25%. Essas porcentagens são calculadas a partir da quantidade total de caracteres, na qual o documento alvo possui 138.017 caracteres. A transição entre as configurações é realizada com o acréscimo das informações das Resoluções nº 773, 671 e

772, sem a retirada de informações referentes ao documento alvo.

Buscando mitigar a influência da escolha do conteúdo presente em cada conjunto (75%, 50% e 25%), são criadas diferentes versões para cada configuração, de modo que todas as fontes de dados sejam contempladas ao longo dessas variações. Dessa forma, a configuração de 100% consiste exclusivamente no Ato nº 915 (138.017 caracteres), constituindo uma versão única. A proporção de 75% é composta por três versões, nas quais todas incluem o Ato nº 915 acrescido de 45.805 caracteres de uma única resolução: a primeira versão incorpora trechos da Resolução nº 773, a segunda da Resolução nº 772 e a terceira da Resolução nº 671.

No cenário de 50%, são consideradas quatro versões, todas contendo o Ato nº 915, diferenciando-se pela combinação dos demais documentos. Na primeira versão, são adicionadas integralmente as Resoluções nº 773 (56.117 caracteres) e nº 671 (66.710 caracteres), além de parte da Resolução nº 772 (14.594 caracteres). A segunda versão inclui a totalidade da Resolução nº 773 (56.117 caracteres) e parte da Resolução nº 772 (81.304 caracteres). A terceira versão agrega integralmente a Resolução nº 671 (66.710 caracteres) e parte da Resolução nº 772 (71.307 caracteres). Já a quarta versão contém apenas uma parte da Resolução nº 772 (138.017 caracteres).

A proporção de 25% é composta por duas versões: a primeira inclui parte da Resolução nº 772 (414.051 caracteres), enquanto a segunda combina a totalidade das Resoluções nº 773 (56.117 caracteres) e nº 671 (66.710 caracteres), além de parte da Resolução nº 772 (291.224 caracteres). A existência dessas variações permite comparar, de forma sistemática, como a adição de normas secundárias influencia a qualidade das respostas, sem que a análise seja afetada pela ordem de inserção dos documentos.

A formulação das perguntas para o teste segue um método de geração sintética automatizada, fundamentado no trabalho de Finardi et al. [4]. O documento alvo (Ato nº 915) é segmentado em blocos de 1.000 caracteres sem sobreposição (overlap = 0) para minimizar a redundância de informações entre os blocos e as chances de serem geradas perguntas parecidas. Posteriormente, é utilizado o modelo GPT-OSS com um *prompt* adaptado para elaborar uma questão cuja resposta esteja obrigatoriamente contida em um bloco de referência, denominado de "*Prompt* de formulação das perguntas" III.1. O comando exige que a pergunta seja detalhada, clara, que a resposta seja uma reprodução fiel de um texto presente no bloco de referência e o retorno seja no formato JSON. Essa etapa do modelo de gerar o par pergunta e resposta é repetida até todos os blocos do documento alvo terem sido contemplados como bloco de referência. Desse modo, é obtido um banco de 139 pares de perguntas e respostas em formato JSON, garantindo uma base de dados para os testes robusta, padronizada e verificável.

A avaliação do desempenho é realizada por meio de um método automático, no qual o LLM identificado com o melhor desempenho nos experimentos anteriores (Mistral-small) também assume o papel de avaliador além do de

Prompt III.1: *Prompt* de formulação das perguntas

Texto de referência:
{Texto}

O texto de referência acima é retirado de um documento normativo da Anatel que fala sobre gestão do espectro. Elabore somente uma questão sobre esse texto, usarei essa pergunta criada para ser respondida considerando todo o documento, então a pergunta deve ser clara e detalhada para que exista somente uma única resposta.

Elabore a pergunta obedecendo às seguintes condições:

1. A pergunta deve obrigatoriamente ter a sua resposta contida no texto de referência.
2. A resposta da pergunta deve ser uma reprodução fiel do texto de referência.
3. A pergunta deve ser detalhada, contendo mais do que 4 palavras.
4. Evite perguntas genéricas.

avaliado. Para cada uma das 139 perguntas, o avaliador recebe um *prompt* estruturado, denominado "*Prompt* de avaliação" III.2, que contém a pergunta, a resposta gerada pelo sistema sob teste, uma escala de pontuação de 1 a 10 e exemplos para que o LLM entenda a escala. Os critérios variam desde a ausência de relação com a pergunta, que é classificada como nota 1, até a precisão e alinhamento totais, classificada com nota 10. Quando todas as 139 respostas são avaliadas para uma configuração de dados específica, é calculada sua pontuação média com a finalidade de realizar uma análise comparativa.

Para aumentar a robustez do método de avaliação e reduzir a sensibilidade a possíveis alucinações do modelo, adota-se um procedimento de redundância nas etapas de geração e avaliação. Para cada pergunta, são geradas três respostas independentes. Cada uma dessas respostas é, por sua vez, avaliada três vezes pelo modelo avaliador, sendo utilizada a média dessas três avaliações como a pontuação final daquela resposta específica. Em seguida, calcula-se a média das pontuações das três respostas associadas à mesma pergunta, resultando na nota final do respectivo par pergunta-resposta. Esse procedimento é repetido para todas as 139 perguntas do conjunto de teste. Posteriormente, calcula-se a média das notas obtidas para cada versão avaliada, gerando uma pontuação representativa da versão considerada. Por fim, realiza-se a média entre as pontuações das diferentes versões pertencentes a uma mesma configuração de dados (75%, 50% e 25%), obtendo-se, assim, a nota final de cada configuração. Esse processo permite uma análise comparativa mais robusta, ao mitigar variações decorrentes da geração estocástica do modelo e tornar a avaliação menos sensível a flutuações pontuais de desempenho.

No sentido de avaliar o impacto das diferentes composições da base de dados, é utilizada a métrica de degradação da pontuação média (PM), uma adaptação da metodologia de [4]. Essa métrica representa o quanto a performance do sistema reduz em relação à configuração ideal (linha de base), no qual o documento alvo representa 100% no conjunto de dados presente no sistema RAG. A Equação 1 descreve o cálculo utilizado:

Prompt III.2: Prompt de avaliação

Você é um avaliador automático de qualidade de respostas. Sua tarefa é avaliar a relevância e alinhamento de uma resposta em relação a uma pergunta, atribuindo uma pontuação inteira de 1 a 10, conforme os critérios abaixo.

Escala de Avaliação

Pontuação 1: A resposta não tem nenhuma relação com a pergunta.

Pontuação 3: A resposta tem pouca relevância e não está alinhada com a pergunta.

Pontuação 5: A resposta tem relevância moderada, mas apresenta imprecisões ou informações incompletas.

Pontuação 7: A resposta está majoritariamente alinhada com a pergunta, mas apresenta pequenas omissões ou falta de detalhes.

Pontuação 10: A resposta é completamente precisa, correta e perfeitamente alinhada com a pergunta.

Use valores intermediários (2, 4, 6, 8, 9) se considerar necessário, desde que respeitem a progressão lógica da escala.

Par de Textos a Serem Avaliados

Pergunta:
{pergunta}

Resposta:
{resposta}

Exemplos de Avaliação

Pontuação 1

Pergunta: "Tudo é lindo e maravilhoso?"

Resposta: "Só os loucos sabem."

Pontuação 3

Pergunta: "Existe algum carro azul com 5 rodas?"

Resposta: "Um carro tem 4 rodas."

Pontuação 5

Pergunta: "Hermione tem vários amigos incluindo 3 homens?"

Resposta: "Harry Potter é amigo de Hermione e Snape."

Pontuação 7

Pergunta: "Hermione é amiga de alguns bruxos incluindo um que usa óculos?"

Resposta: "Harry Potter é amigo de Hermione."

Pontuação 10

Pergunta: "O Brasil é pentacampeão mundial?"

Resposta: "O Brasil ganhou 5 Copas do Mundo da FIFA no futebol."

Instrução Final (MUITO IMPORTANTE)

Retorne somente o número inteiro correspondente à pontuação da avaliação.

Não explique o motivo.

Não retorne texto adicional.

Não use formatação.

Saída esperada (exemplo):

7

$$\text{Degradação da PM} = 1 - \left(\frac{\text{PM da Configuração Atual}}{\text{PM da Linha de Base}} \right).$$
 (1)

Os resultados deste experimento estão presentes na Tabela I, que correlaciona as diferentes composições do banco de dados com a eficácia do sistema. A primeira coluna identifica as quatro configurações de porcentagem do documento alvo, conforme detalhado anteriormente nesta seção, seguida pela média das avaliações das respostas em uma escala de 1 a 10 pontos. A última coluna apresenta o percentual de degradação dessa média em relação à linha de base, que se encontra destacada em vermelho, para evidenciar as variações

de desempenho entre os cenários testados.

TABELA I
COMPARAÇÃO DA QUALIDADE DAS RESPOSTAS COM DIFERENTES QUANTIDADES DE DOCUMENTOS RELACIONADOS NO SISTEMA RAG.

Conteúdo Alvo (%)	PM (1-10)	Degradação da PM(%)
100	8,91	0
75	8,71	2,3
50	8,74	2,0
25	8,60	3,5

De acordo com os valores apresentados, a quantidade de dados relacionados ao tema de interesse não exerce influência significativa na qualidade final da resposta. Observa-se que a arquitetura RAG tradicional utilizada, em conjunto com o modelo Mistral-small, demonstra competência em retornar os blocos de interesse e gerar a resposta adequada independentemente do volume de informações secundárias testado neste experimento. Esse comportamento sugere que o mecanismo de recuperação semântica é robusto o suficiente para filtrar ruídos provenientes de documentos do mesmo domínio regulatório, mantendo a precisão do assistente mesmo com o incremento da base de conhecimento.

B. Experimento da diversidade de dados no sistema RAG

Este experimento investiga se a proximidade semântica dos dados inseridos influencia a precisão do sistema. Enquanto o experimento anterior foca em documentos do mesmo domínio regulatório, esta etapa busca verificar o impacto do ruído proveniente de contextos temáticos distintos. Para garantir a comparabilidade dos resultados e o isolamento da variável de análise, as configurações técnicas do sistema RAG são mantidas, isto é blocos de 1000 caracteres, sobreposição de 200 caracteres, técnica de busca MMR, mesmas perguntas, modelo de inferência (Mistral-small) e critérios de avaliação do experimento anterior.

O diferencial deste experimento é a substituição do conteúdo relacionado pelo não relacionado, que é composto por obras clássicas da literatura brasileira e de língua portuguesa, visando estabelecer um alto distanciamento semântico em relação ao vocabulário técnico da Anatel. O conjunto de dados inclui obras como Dom Casmurro, Os Lusíadas, O Guarani, Memórias Póstumas de Brás Cubas, poesias, Helena e O Cortiço. A substituição de normas técnicas por textos literários permite observar se a natureza do dado inserido facilita ou dificulta a tarefa de recuperação do sistema, já que são mantidas as mesmas proporções de divisão do conjunto de dados, 75%, 50% e 25%.

A configuração de 100% permanece inalterada, sendo utilizada como linha de base para comparação. Para as demais proporções, adota-se uma estratégia de versionamento semelhante à do experimento anterior, com o objetivo de mitigar possíveis vieses decorrentes da seleção específica dos dados. No cenário de 75%, são consideradas sete versões, cada uma incorporando uma fonte distinta de texto literário ao conjunto de dados. De forma análoga, a configuração de 50% também é composta por sete versões. Já na proporção de

25%, são definidas quatro versões: a primeira inclui parte da obra O Guarani; a segunda combina Dom Casmurro e parte da Os Lusíadas; a terceira reúne Helena e parte da O Cortiço; e a quarta contempla Memórias Póstumas de Brás Cubas em conjunto com poesias.

Além disso, com o propósito de avaliar o sistema em um cenário de estresse, o experimento amplia a diluição do documento alvo. Nesta etapa final, é testada uma composição de versão única com 4,7% de conteúdo referente ao Ato nº 915 e 95,3% de documentos não relacionados, composta por todas as fontes de dados, totalizando quase 3 milhões de caracteres.

Os resultados referentes à diversidade temática dos dados estão na Tabela II. O formato de apresentação é semelhante ao da seção anterior, com a particularidade de que valores negativos na coluna de degradação representam uma pontuação média superior à linha de base estipulada.

TABELA II
COMPARAÇÃO DA QUALIDADE DAS RESPOSTAS COM DIFERENTES QUANTIDADES DE DOCUMENTOS NÃO RELACIONADOS NO SISTEMA RAG.

Conteúdo Alvo (%)	PM (1-10)	Degradação da PM(%)
100	8,91	0
75	8,90	0,2
50	8,90	0,2
25	8,93	-0,1
4,7	8,95	-0,6

A observação da tendência crescente na pontuação média sugere que a distância semântica entre os dados presentes no sistema RAG influencia positivamente a qualidade das respostas. Os valores de degradação negativa de -0,1 e -0,6 para as composições de 25% e 4,7%, respectivamente, indicam uma pequena melhora no desempenho com o acréscimo de informações semanticamente distantes do conteúdo de interesse. Essa característica é reafirmada pela comparação entre os resultados das Tabelas I e II, cujas diferenças de degradação variam entre 4,1% e 1,8%. Esses resultados podem ser explicados pela maior facilidade do mecanismo de recuperação em distinguir contextos semanticamente distintos, bem como pela eficiência da técnica MMR em selecionar fragmentos que equilibram relevância e diversidade para a geração das respostas. Dessa forma, mesmo na presença de ruídos literários, o sistema não apenas preserva sua capacidade de fornecer respostas assertivas, como também apresenta ganhos marginais de desempenho.

IV. CONCLUSÕES

Este trabalho analisa a influência da quantidade e da abrangência dos dados de um sistema RAG na qualidade das respostas de um assistente de IA voltado à gestão do espectro de radiofrequência. Para realizar a análise, implementa-se uma arquitetura RAG tradicional e realizam-se experimentos controlados que verificam a degradação das respostas de acordo com o aumento do volume de informações, tanto em contextos alinhados à temática regulatória quanto em cenários de distanciamento semântico. Neste trabalho, são utilizados métodos de avaliação subjetiva e automática, para avaliar a qualidade das respostas geradas pelo modelo.

Os resultados demonstram insensibilidade do sistema ao acréscimo de dados relacionados à gestão do espectro e uma tendência positiva com ganhos marginais de desempenho para dados compostos por um ruído temático. Essa estabilidade justifica-se pela eficiência do mecanismo de recuperação e da técnica MMR em distinguir contextos semanticamente distintos e selecionar fragmentos que equilibram relevância e diversidade, garantindo que o assistente preserve sua capacidade de fornecer respostas precisas e contextualizadas.

Como perspectiva para trabalhos futuros está a avaliação de diferentes arquiteturas de RAG, abrangendo técnicas tradicionais como a busca por similaridade, que seleciona fragmentos de texto pela proximidade semântica com a consulta, e avançadas como a recuperação em multicamadas, além da avaliação de técnicas de engenharia de *prompt*. A recuperação em multicamadas aproveita a organização hierárquica do texto para otimizar a extração de informações, o que se revela particularmente promissor para as normas da Anatel, dada sua estruturação lógica em títulos, capítulos e artigos. O propósito dessa análise inicial é verificar o ganho de desempenho que tais abordagens oferecem ao cenário regulatório. Adicionalmente, projeta-se comparar o desempenho do sistema RAG com o processo de *fine-tuning* do modelo, a fim de verificar se o RAG representa o melhor caminho para a realização de um assistente voltado à gestão de espectro. O objetivo geral é identificar as soluções que apresentam o melhor desempenho para a assistência técnica no setor.

AGRADECIMENTOS

Os autores gostariam de agradecer o apoio técnico e o fomento à pesquisa do Centro de Altos Estudos em Comunicações Digitais e Inovações Tecnológicas (Ceadi), da Agência Nacional de Telecomunicações (Anatel), por meio de Termo de Execução Descentralizada celebrado entre a Agência Reguladora e a Universidade Federal do Rio de Janeiro, no contexto do Projeto nº 4/2024 (53500.060216/2024-35), “Soluções de Inteligência Artificial (IA) voltadas à gestão do espectro”. Também são gratos pelo financiamento da Capes, CNPq, FAPERJ, e Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), auxílios no. 2015/24494-8 e 2015/24490-2.

REFERÊNCIAS

[1] Movva, R., Balachandar, S., Peng, K., Agostini, G., Garg, N. e Pierson, E. (2024). To- pics, authors, and institutions in large language model research: Trends from 17K arXiv papers. Em Duh, K., Gomez, H. e Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, p. 1223–1243, Mexico City, Mexico. Association for Computational Linguistics.

[2] Kernycky, A., Coleman, D., Spence, C. e Das, U. (2024). *Evaluating the Performance of LLMs on Technical Language Processing Tasks*, p. 75–85. Springer Nature Switzerland.

[3] Bansal, A., Muriga, V., Li, J., Duan, L. e Kumar, S. (2025). Can we make FCC Experts out of LLMs? Em *Proceedings of the 26th International Workshop on Mobile Computing Systems and Applications*, p. 85–90. Association for Computing Machinery.

[4] Finardi, P., Avila, L., Castaldoni, R., Gengo, P., Larcher, C., Piau, M., Costa, P. e Caridá, V. (2024). The chronicles of rag: The retriever, the chunk and the generator.