INTELIGÊNCIA ARTIFICIAL GENERATIVA: CONSTRUÇÃO E AVALIAÇÃO DE UM MODELO EM APOIO À AUTORIDADE MARÍTIMA BRASILEIRA

Capitão de Corveta (IM) Márcio Selemen Coelho
Diretoria de Abastecimento da Marinha (DAbM)
Ilha das Cobras - S/N° - Edificio Almirante Gastão Motta, 4° Andar - Centro, Rio de Janeiro – RJ
marcio.selemen@marinha.mil.br

Capitão de Corveta (RM3-T) Ketia Kellen Araújo da Silva
Escola de Guerra Naval (EGN)
Av. Pasteur, 480 - Urca, Rio de Janeiro – RJ
ketia.kellen@marinha.mil.br

RESUMO

Este artigo apresenta o desenvolvimento de um modelo de inteligência artificial generativa com arquitetura de recuperação, ampliação e geração para aplicação de normas da Autoridade Marítima Brasileira. O estudo abordou a construção de uma base vetorial, a partir de normativos selecionados, e a implementação de uma ferramenta usando um grande modelo de linguagem. A avaliação de desempenho do modelo utilizou questões de concurso público da Marinha do Brasil para ingresso no quadro técnico de segurança do tráfego aquaviário, alcançando 75% de acurácia. Métricas complementares mostraram fidelidade entre 50-100%, relevância da resposta entre 80-89% e relevância do contexto entre 5-25%. Os resultados alcançados indicam desempenho robusto da ferramenta, comparável a benchmarks de modelos avançados no mercado. O trabalho discute, ainda, limitações como alucinações, vieses e questões de interpretabilidade. Por fim, conclui-se que o modelo representa um avanço significativo com potencial para automatizar consultas, agilizar interpretações e aprimorar treinamento de pessoal.

PALAVRAS-CHAVE: Inteligência artificial generativa; Grandes modelos de linguagem; Avaliação de desempenho.
Tópico do artigo: Inteligência Artificial; Processamento de Linguagem Natural; Sistemas de Informação.

ABSTRACT

This article presents the development of a generative artificial intelligence model with retrieval-augmented generation architecture for applying Brazilian Maritime Authority regulations. The study addressed the construction of a vector database from selected regulatory documents and the implementation of a tool using a large language model. The model's performance evaluation utilized Brazilian Navy public examination questions for admission to the waterway traffic safety technical corps, achieving 75% accuracy. Supplementary metrics showed faithfulness ranging from 50-100%, answer relevance between 80-89%, and context relevance between 5-25%. The results indicate robust tool performance, comparable to benchmarks of advanced market models. The work also discusses limitations such as hallucinations, biases, and interpretability issues. Finally, it concludes that the model represents a significant advancement with potential to automate queries, streamline regulatory interpretations, and enhance personnel training.

KEYWORDS: Generative artificial intelligence; Large language models; Performance evaluation.
Paper topics: Artificial Intelligence; Natural Language Processing; Information Retrieval.

DOI: 10.59254/sbpo-2025-212623
https://proceedings.science/p/212623?lang=pt-br

1 INTRODUÇÃO

A Inteligência Artificial Generativa (IAG) representa uma nova fronteira no campo da Inteligência Artificial (IA), com potencial para transformar diversas indústrias, incluindo a marítima [Liu et al. 2021]. A crescente aplicação de IAG e Grandes Modelos de Linguagem, ou Large Language Models (LLM), tem impulsionado pesquisas em variados domínios, com um interesse notável no contexto militar para otimizar processos e auxiliar na tomada de decisões [Picinini 2025]. Embora a literatura internacional já apresente exemplos de uso de IAG em forças armadas de outros países [Rivera et al. 2024] [Smith e Jones 2023], a especificidade da aplicação em um contexto nacional como o da Autoridade Marítima Brasileira (AMB) ainda carece de estudos aprofundados.

Nesse cenário, este artigo propõe a construção e avaliação de um modelo de IAG com arquitetura de Recuperação, Ampliação e Geração, ou Retrieval Augmented Generation (RAG), para auxiliar na busca, interpretação e implementação das Normas da Autoridade Marítima (NORMAM) e legislações correspondentes no contexto da Marinha do Brasil (MB). O estudo foi motivado pelos desafios inerentes ao vasto volume de normas marítimas e suas frequentes atualizações, cuja interpretação manual é demorada e sujeita a erros. A utilização desta ferramenta visa otimizar processos, aumentar a assertividade e melhorar a compreensão das normas, promovendo maior eficiência, dinamismo e segurança para os profissionais que operam sob a supervisão da AMB [Cevallos et al. 2023].

Para o desenvolvimento do modelo, utilizou-se a linguagem de programação Python para a criação dos scripts, o framework LangChain na construção da arquitetura RAG, o framework RAGAS para avaliação de desempenho, o Gemini 1.0 Pro (Google) como LLM, e o modelo "embedding-001" (Google) com o banco de dados ChromaDB para a "loja de vetores". A pesquisa limita-se, temporalmente, aos documentos vigentes em 2023 e, em escopo, à área de Segurança do Tráfego Aquaviário.

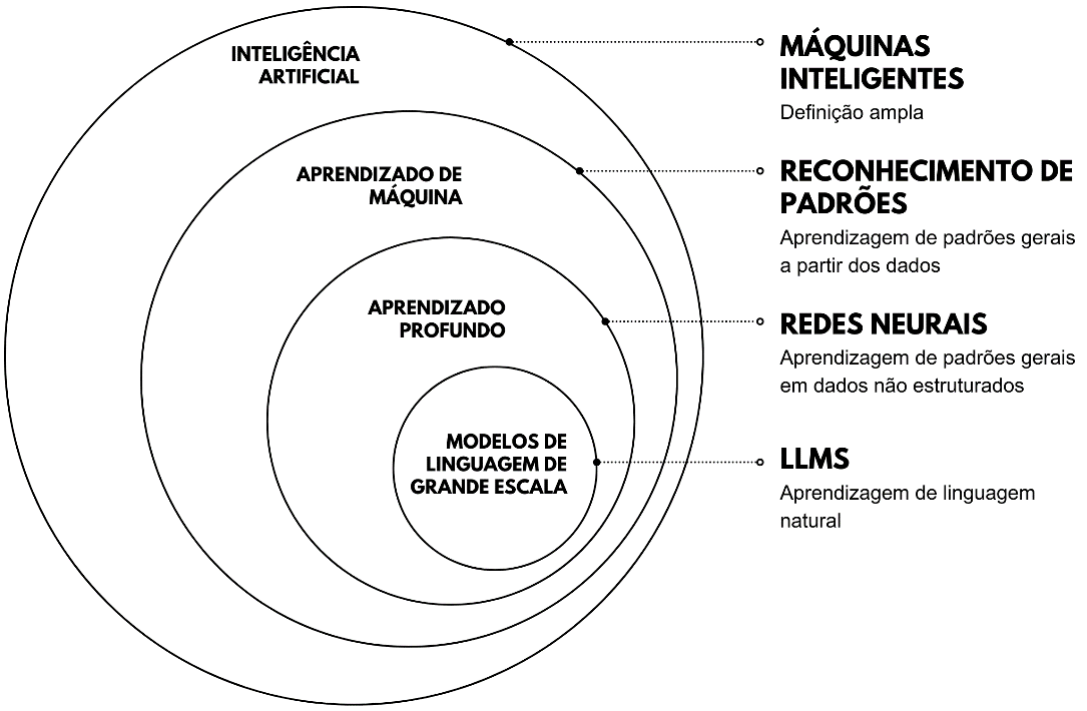
O trabalho se destaca pela aplicação pioneira de um modelo de IA para a interpretação de um vasto e complexo arcabouço normativo. Embora existam estudos sobre IAG em outros contextos militares e governamentais, do Brasil e do mundo, a especificidade do domínio e a avaliação detalhada do desempenho do modelo em um cenário próximo ao real, de aplicação de normas, conferem originalidade e relevância prática a esta pesquisa. A abordagem metodológica e a análise das limitações e desafios têm a intenção de contribuir para o avanço do conhecimento na aplicação de LLMs em ambientes regulatórios críticos, oferecendo um estudo de caso valioso para a MB.

2 REFERENCIAL TEÓRICO

A Inteligência Artificial (IA) representa uma disciplina da ciência da computação dedicada ao desenvolvimento de sistemas capazes de executar tarefas que normalmente exigem inteligência humana [Mogali 2014]. Dentre seus diversos subcampos, destaca-se o Aprendizado de Máquina (Machine Learning), onde algoritmos aprendem a partir de dados sem intervenção humana direta [Janiesch et al. 2021]. Inserido nesse subcampo, o Aprendizado Profundo (Deep Learning) representa uma especialização que utiliza redes neurais complexas para descobrir padrões em grandes volumes de dados [Lecun et al. 2015].

DOI: 10.59254/sbpo-2025-212623
https://proceedings.science/p/212623?lang=pt-br

Figura 1 - Inteligência Artificial



Fonte: Elaborada pelo autor (2025)

Os Modelos de Linguagem de Grande Escala, por sua vez, emergiram como evolução natural da IA, expandindo as fronteiras de interpretação e geração de linguagem. Estes modelos são capazes de compreender nuances complexas da linguagem e responder contextualmente, demonstrando avanços em diálogo automatizado, compreensão textual e criação de conteúdo original [Naveed et al. 2023].

Os LLMs são denominados "grandes" devido ao seu vasto número de parâmetros, que podem alcançar bilhões, permitindo modelagem linguística excepcionalmente rica [Brown et al. 2020]. Um exemplo notável de LLM é o Generative Pre-trained Transformer (GPT), cuja arquitetura revolucionou o Processamento de Linguagem Natural (PLN). O termo "Generative" indica a capacidade de prever a próxima palavra em uma sequência, gerando texto coerente e contextualmente relevante [Yenduri et al. 2023]. "Pre-trained" refere-se ao treinamento inicial em vastos conjuntos textuais antes da especialização em tarefas específicas, enquanto "Transformer" remete à arquitetura que utiliza mecanismos de atenção para modelar interdependências de longo alcance no texto [Vaswani et al. 2017].

Apesar de suas funcionalidades avançadas, uma limitação significativa dos LLMs, entretanto, é que sua capacidade de gerar respostas precisas está restrita ao conhecimento incorporado durante seu treinamento [Kandpal et al. 2022]. Para superar esta restrição, técnicas como Ajuste Fino (Fine-tuning) e Geração Aumentada por Recuperação (RAG) são frequentemente empregadas [Soong et al. 2023].

O Ajuste Fino aprimora os parâmetros de um modelo pré-treinado em conjunto de dados específico, melhorando significativamente seu desempenho em tarefas particulares. Já a técnica RAG permite que o modelo consulte bases de dados externas para enriquecer suas respostas, combinando capacidades generativas com informações recuperadas em tempo real [Lakatos et al. 2024].

DOI: 10.59254/sbpo-2025-212623
https://proceedings.science/p/212623?lang=pt-br

Para criar uma aplicação RAG eficiente, primeiramente carregam-se os dados em formato adequado para manipulação por meio de carregadores de documentos que suportam diversos formatos como PDF, HTML e JSON [Auffarth 2023] [Gao et al. 2023]. Em seguida, divide-se o conteúdo em partes menores (chunks), mantendo trechos semanticamente relevantes juntos e estabelecendo sobreposições para preservar conexões lógicas entre as partes [Alan et al. 2024].

Após a divisão, indexam-se os trechos utilizando embeddings, que são representações numéricas que permitem que textos semanticamente semelhantes possuam vetores similares [Pilehvar e Camacho-Collados 2020], e armazenam-se em vetores de armazenamento, facilitando posterior recuperação [Ke et al. 2024]. A busca por similaridade (busca semântica) identifica partes de texto semanticamente semelhantes, permitindo recuperar informações relevantes mesmo quando a consulta não coincide exatamente com os documentos armazenados [Moradi et al. 2024].

Posteriormente, a avaliação de LLMs sob arquitetura RAG visa melhorar a precisão e relevância das respostas geradas. Nessa avaliação, um desafio crucial é garantir que os componentes de recuperação e geração funcionem harmoniosamente [Huang e Huang 2024]. Nesse sentido, o framework RAGAS (Retrieval Augmented Generation Assessment) propõe métricas que podem ser aplicadas sem necessidade de anotações humanas, focando em três aspectos principais: fidelidade da resposta, relevância da resposta e relevância do contexto [Es et al. 2023].

A fidelidade avalia se a resposta está fundamentada no contexto fornecido, evitando alucinações. A relevância da resposta mensura se a resposta gerada aborda apropriadamente a pergunta. Por fim, a relevância do contexto mede se o contexto recuperado é focado e contém mínimas informações irrelevantes. Estas métricas são essenciais para garantir que modelos RAG funcionem adequadamente, especialmente em aplicações específicas como a de apoio à Autoridade Marítima Brasileira.

No contexto da AMB, a complexidade e o volume do arcabouço normativo representam um desafio significativo. A legislação marítima é vasta, dinâmica e abrange diversas áreas, desde segurança da navegação e proteção ambiental até fiscalização de embarcações e formação de pessoal. A interpretação e aplicação correta dessas normas são cruciais para garantir a conformidade regulatória e a segurança das operações no ambiente aquaviário. A necessidade de uma ferramenta que auxilie na navegação e compreensão desse corpo normativo justifica a relevância da aplicação de tecnologias como a IAG neste domínio.

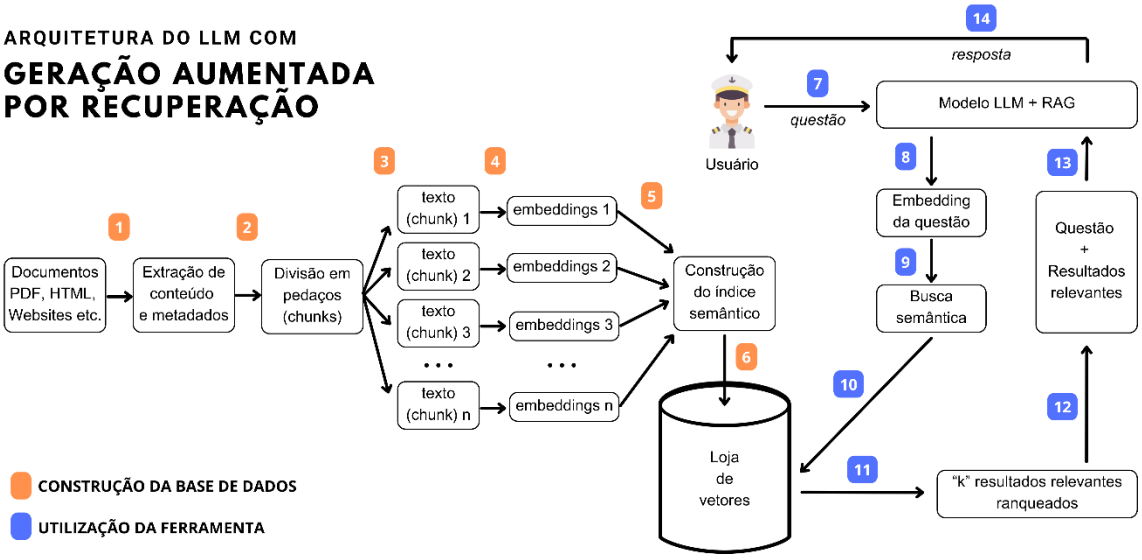
3 METODOLOGIA

Para o desenvolvimento do modelo de IAG com arquitetura de RAG, inicialmente foram selecionados 34 documentos regulatórios baseados na bibliografia oficial do concurso público CP-T/STA (2023) para ingresso no quadro técnico da Marinha do Brasil, incluindo normas da Autoridade Marítima, leis e decretos que regulamentam a segurança do tráfego aquaviário no Brasil.

A construção da base de dados e a utilização da ferramenta seguiram um processo estruturado conforme detalhado na Figura 2, que ilustra a arquitetura proposta do LLM com geração aumentada por recuperação.

DOI: 10.59254/sbpo-2025-212623
https://proceedings.science/p/212623?lang=pt-br

Figura 2 – Arquitetura proposta do LLM com RAG



Fonte: Elaborada pelo autor (2025)

A tabela, a seguir, descreve as etapas ilustradas acima, fornecendo os respectivos detalhamentos.

Tabela 1 - Etapas de construção da base de dados e de utilização da ferramenta

Etapa	Descrição	Detalhamento
1	Ingestão dos documentos	Carregamento de 34 documentos regulatórios (PDFs) da bibliografia oficial do CP-T/STA (2023).
2	Extração de conteúdo e metadados	Extração de textos e metadados (título, autor, data) dos documentos.
3	Divisão em chunks	Divisão do conteúdo em pedaços de até 1.000 caracteres, com uma sobreposição de 200 caracteres entre eles. A segmentação utiliza a estratégia "RecursiveCharacterTextSplitter", que tenta manter a coesão do texto ao priorizar a divisão por parágrafos, frases e espaços.
4	Geração de embeddings	Conversão de cada chunk em vetores numéricos (embeddings) utilizando o modelo "embedding-001" do Google, que é otimizado para a tokenização de texto em português e outros idiomas.
5	Construção do índice semântico	Indexação dos embeddings em uma estrutura otimizada para busca semântica (ChromaDB). O índice é construído estaticamente para ser reutilizado em múltiplas consultas.
6	Armazenamento dos vetores	Armazenamento dos embeddings e do índice semântico no ChromaDB.
7	Formulação da questão pelo usuário	O usuário insere uma questão.
8	Transformação em embedding vetorial	A questão do usuário é transformada em um embedding vetorial pelo mesmo modelo "embedding-001".
9 e 10	Busca semântica	Realização de busca semântica na loja de vetores (ChromaDB).
11	Recuperação dos resultados relevantes	Recuperação dos cinco (k=5) resultados mais relevantes da loja de vetores, com base na similaridade semântica com a questão do usuário.
12 e 13	Combinação com a questão original	Os resultados recuperados são combinados com a questão original.
14	Geração da resposta final	O LLM (Gemini 1.0 Pro) gera a resposta final, com temperatura definida em zero para respostas determinísticas.

Fonte: Elaborada pelo autor (2025)

DOI: 10.59254/sbpo-2025-212623
https://proceedings.science/p/212623?lang=pt-br

Para materializar a arquitetura proposta, o pseudocódigo, a seguir, traduz as etapas do diagrama e da tabela em uma estrutura de lógica computacional, detalhando as funções desde a construção da base de conhecimento até a interação com o usuário.

Figura 3 – Pseudocódigo da construção e utilização do modelo

```
ALGORITMO ConstrucaoEUtilizacaoModeloIAG_RAG

// Fase de Construção da Base de Conhecimento
// Esta função processa os documentos e os armazena no banco de dados vetorial.
FUNCAO ConstruirBaseDeConhecimento(documentos_normativos):
    PARA CADA documento EM documentos_normativos:
        conteudo_texto = ExtrairTexto(documento)
        metadados = ExtrairMetadados(documento)

        chunks = DividirEmChunks(conteudo_texto, tamanho_chunk=1000, overlap=200)

        PARA CADA chunk EM chunks:
            embedding = GerarEmbedding(chunk, modelo='embedding-001')
            Armazenar(embedding, chunk, metadados, em='ChromaDB')
        FIM PARA
    FIM PARA
FIM FUNCAO


// Fase de Utilização da Ferramenta
// Esta função recebe a pergunta do usuário, busca o contexto relevante e gera a
resposta final.
FUNCAO UtilizarFerramenta(questao_usuario):
    embedding_questao = GerarEmbedding(questao_usuario, modelo='embedding-001')
    resultados_relevantes = BuscarSimilaridade(embedding_questao, em='ChromaDB',
k_vizinhos=5)
    contexto_recuperado = CombinarChunks(resultados_relevantes)
    resposta_final = GerarResposta(questao_usuario, contexto_recuperado,
modelo='Gemini-1.0-Pro', temperatura=0)

    RETORNAR resposta_final
FIM FUNCAO


// Fluxo Principal da Aplicação
INICIO
    documentos = CarregarDocumentosNormativos()
    ConstruirBaseDeConhecimento(documentos)

    ENQUANTO usuario_ativo:
        pergunta = ObterQuestaoDoUsuario()
        resposta = UtilizarFerramenta(pergunta)
        ExibirResposta(resposta)
    FIM ENQUANTO
FIM
```

Fonte: Elaborada pelo autor (2025)

Diante da construção e utilização da ferramenta, sua avaliação foi realizada, então, por meio de questões da prova do CP-T/STA (2023), abordando quatro métricas principais: grau de acurácia (GA), fidelidade da resposta (FR), relevância da resposta (RR) e relevância do contexto (RC).

DOI: 10.59254/sbpo-2025-212623
https://proceedings.science/p/212623?lang=pt-br

Como resultado, o modelo alcançou um grau de acurácia (GA) de 75%, acertando 36 questões, errando 8 e respondendo "Não sei a resposta" em 4 casos, superando significativamente o mesmo modelo sem RAG, que obteve apenas 46% de acurácia. Destaca-se que 35% das questões foram acertadas exclusivamente pelo modelo com RAG.

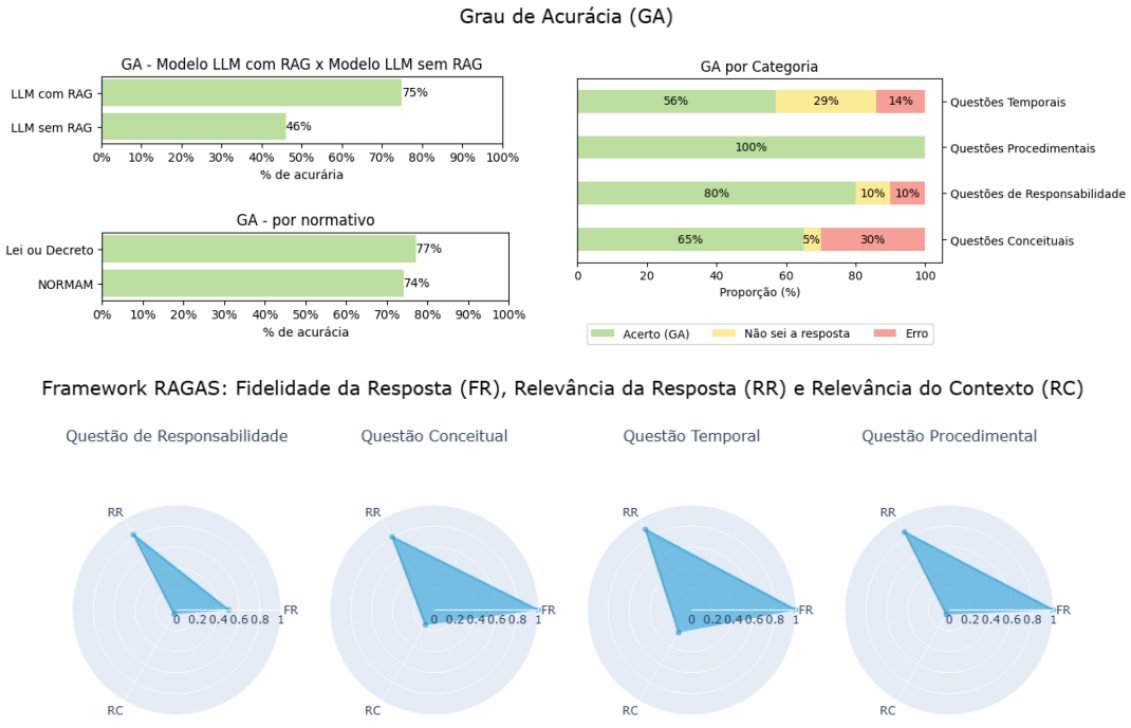
Nos resultados por categoria de questões, o modelo demonstrou variações de desempenho: 56% nas questões temporais, 100% nas procedimentais, 80% nas de responsabilidade e 65% nas conceituais. Quanto aos resultados por normativos, obteve-se 77% em questões baseadas em leis ou decretos e 74% de acurácia em questões baseadas nas NORMAM.

Para validar que a precisão das respostas não foi acidental, utilizou-se o framework RAGAS para calcular métricas adicionais em quatro questões representativas. Como resultado, a fidelidade da resposta (FR) variou entre 50% e 100%, a relevância da resposta (RR) ficou entre 80% e 89%, enquanto a relevância do contexto (RC) apresentou valores entre 5% e 25%.

4 ANÁLISE DOS RESULTADOS

Os resultados obtidos demonstram a eficácia do modelo desenvolvido na interpretação e aplicação dos normativos da Autoridade Marítima Brasileira, especialmente quando comparado ao desempenho do mesmo LLM sem a arquitetura RAG. A figura abaixo consolida as métricas de avaliação a fim de discutir o desempenho da ferramenta.

Figura 4 - Resultados consolidados da avaliação de desempenho do modelo



Fonte: Elaborada pelo autor (2025)

Conforme apresentado, o modelo de IAG implementado com arquitetura RAG alcançou um grau de acurácia (GA) de 75% na resolução de questões do CP-T/STA (2023). Este resultado é comparável a benchmarks de modelos de linguagem comerciais, como o Gemini Pro que atingiu 79,13% no MMLU (Massive Multitask Language Understanding), enquanto o GPT-3.5 obteve 70% [Google 2024].

DOI: 10.59254/sbpo-2025-212623
https://proceedings.science/p/212623?lang=pt-br

O desempenho também se mostrou expressivo quando comparado à média de 81,50% dos candidatos aprovados no concurso CP-T/STA (2023) [Brasil 2024]. Em contraste, o modelo sem a arquitetura RAG obteve apenas 46% de acurácia, evidenciando um ganho de 29 pontos percentuais com a implementação desta técnica.

Na avaliação pelo framework RAGAS, o modelo apresentou resultados satisfatórios em geração de texto, com alta fidelidade (FR) às informações normativas recuperadas e boa relevância (RR) contextual das respostas. Entretanto, a capacidade de recuperação de contexto (RC) mostrou variações significativas (5% a 25%), indicando a necessidade de otimização dos parâmetros utilizados.

O baixo desempenho na recuperação de contexto implica que, no estado atual, o modelo pode, em certas situações, gerar respostas com base em informações incompletas ou parcialmente relevantes, aumentando o risco de alucinações ou respostas imprecisas. Nesse sentido, a melhoria desse indicador é fundamental para a adoção plena da ferramenta em cenários operacionais.

Tal performance, aquém da desejada, pode estar relacionada a múltiplos fatores técnicos que potencialmente comprometeram a eficácia do sistema RAG. Entre os principais aspectos vislumbrados, destacam-se uma configuração não otimizada dos hiperparâmetros (particularmente no que se refere ao tamanho dos chunks de texto), sobreposição entre segmentos e número de extratos recuperados, que podem estar fragmentando ou perdendo informações contextuais importantes.

Adicionalmente, o modelo de embedding utilizado pode não ter capturado adequadamente a semântica específica do domínio marítimo-jurídico, onde termos técnicos e conceitos especializados demandam representações vetoriais mais precisas. As estratégias de recuperação semântica empregadas também podem ter apresentado deficiências na identificação e ranqueamento dos trechos mais relevantes, sugerindo a necessidade de abordagens híbridas que combinem busca semântica com métodos baseados em palavras-chave. Adicionalmente, problemas na qualidade e pré-processamento dos dados normativos, incluindo inconsistências de formatação, caracteres especiais e variações de encoding, possivelmente impactaram diretamente a formação dos embeddings e, consequentemente, a precisão da recuperação de informações relevantes.

Quanto aos desafios organizacionais para implementação de IAG, o estudo identificou também limitações importantes em relação a soluções de IAG. Segundo a Gartner (2024), a principal barreira para adoção dessas ferramentas é a dificuldade em estimar e demonstrar seu valor real, com 49% das empresas enfrentando desafios nessa mensuração. Outros desafios incluem alucinações, que comprometem a confiabilidade das respostas [Cambridge, 2024], a natureza não determinística dos modelos, vieses, baixa interpretabilidade dos sistemas de IA, e questões de segurança e privacidade.

Considerando o contexto da Marinha do Brasil, a integração de modelos de IAG em ambientes institucionais militares apresenta desafios específicos que devem ser cuidadosamente avaliados. Para a plena implementação nesse ambiente, é fundamental considerar aspectos como a interoperabilidade com outros sistemas, explicabilidade, efeitos de informações incorretas, manipulação de informações classificadas/sigilosas e questões relacionadas a vieses éticos e responsabilidade corporativa.

Em relação às limitações metodológicas desta pesquisa, a limitada quantidade de questões e métricas utilizadas representa a principal restrição do estudo, podendo o trabalho não ter capturado a complexidade dos desafios em contextos reais. Ademais, a evolução tecnológica constante oferece métodos mais avançados que supera a "Busca Semântica" simples utilizada neste estudo. Finalmente, questiona-se a relevância futura da arquitetura RAG diante da crescente capacidade dos LLMs em gerenciar contextos extensos, embora ainda seja essencial para melhorar a eficiência operacional e a precisão nas respostas.

DOI: 10.59254/sbpo-2025-212623
https://proceedings.science/p/212623?lang=pt-br

5 CONCLUSÃO

Este estudo explorou a utilização de Inteligência Artificial Generativa como ferramenta de apoio para busca, interpretação e implementação de normativos da Autoridade Marítima Brasileira. Foi desenvolvido um modelo com arquitetura RAG aplicado às NORMAM e legislações correlatas, e o seu desempenho foi avaliado por meio de métricas específicas.

O modelo apresentou resultados robustos, demonstrando potencial significativo como ferramenta auxiliar na conjuntura do poder marítimo. Contudo, reconhece-se a necessidade de cautela quanto às repercussões de sua implementação, dadas as limitações tecnológicas dos modelos de linguagem e dos frameworks disponíveis no momento deste trabalho, que ainda precisam ser superadas.

Contudo, esta pesquisa representa um passo importante para a aplicação da IAG em apoio à Autoridade Marítima Brasileira. Com a contínua evolução dos LLMs e do ecossistema de IA como um todo, vislumbra-se que os modelos RAG poderão automatizar e agilizar consultas e interpretações normativas, além de servir como ferramenta de treinamento para novos Oficiais e Praças da Marinha. Sua capacidade de gerar respostas precisas também poderá auxiliar na gestão de incidentes e tomadas de decisão estratégicas.

As possibilidades de aplicação serão limitadas apenas pela criatividade humana, e o avanço tecnológico certamente abrirá novas oportunidades. Essas poderão transformar profundamente a atuação da autoridade marítima nacional, resultando em desempenho mais eficiente, ágil e seguro dos profissionais da MB, contribuindo para a conformidade regulatória e o fortalecimento do poder marítimo brasileiro.

Para trabalhos futuros, com base nas limitações e desafios identificados neste estudo, recomenda-se o aprimoramento contínuo da ferramenta por meio da otimização comparativa de modelos LLM, técnicas e parâmetros RAG para maximizar a acurácia, além da realização de estudos de impacto e viabilidade considerando aspectos operacionais, éticos e de segurança para implementação em larga escala.

DOI: 10.59254/sbpo-2025-212623
https://proceedings.science/p/212623?lang=pt-br

REFERÊNCIAS

Alan, Ahmet Yusuf, Karaarslan, Enis e Aydin, Omer. A RAG-based Question Answering System Proposal for Understanding Islam: MufassirQAS LLM. SSRN Electronic Journal, 2024.

Auffarth, Ben. Generative Al with LangChain: Build large language model (LLM) apps with Python, ChatGPT, and other LLMs. Birmingham, England: Packt Publishing, 2023.

Brasil. Marinha do Brasil. Serviço de Seleção do Pessoal da Marinha. Disponível em: https://www.inscricao.marinha.mil.br/marinha/NotasPOCP-T-2023.pdf?id file=7725. Acesso em: 5 jul. 2024.

Brown, Tom B. et al. Language Models are Few-Shot Learners. 2020. Disponível em: http://arxiv.org/abs/2005.14165.

Cambridge University Press & Assessment. ‘Hallucinate' is Cambridge Dictionary's Word of the Year 2023. Disponível em: https://www.cambridge.org/news-and-insights/hallucinate-is-cambridge-word-of-the-year-2023. Acesso em: 09 jul. 2024.

Cevallos, Adrian et al. Tech Report Generative Al. 2023. https://doi.org/10.18235/0005105. Acesso em: 21 mar. 2024.

Es, Shahul et al. RAGAS: Automated Evaluation of Retrieval Augmented Generation. 2023. Disponível em: http://arxiv.org/abs/2309.15217.

Gao, Yunfan et al. Retrieval-Augmented Generation for Large Language Models: A Survey. 2023. Disponível em: http://arxiv.org/abs/2312.10997.

Gartner. Gartner Survey Finds Generative Al Is Now the Most Frequently Deployed Al Solution in Organizations. 2024. Disponível em: https://www.gartner.com/en/newsroom/press-releases/2024-05-07-gartner-survey-finds-generative-ai-is-now-the-most-frequently-deployed-ai-solution-in-organizations. Acesso em: 08 jul. 2024.

Google. Gemini: A Family of Highly Capable Multimodal Models. 2024. Disponível em: https://storage.googleapis.com/deepmind-media/gemini/gemini 1 report.pdf. Acesso em: 5 jul. 2024.

Huang, Yizheng e Huang, Jimmy. A Survey on Retrieval-Augmented Text Generation for Large Language Models. 2024. Disponível em: http://arxiv.org/abs/2404.10981.

Janiesch, Christian e Zschech, Patrick e Heinrich, Kai. Machine learning and deep learning. Electronic Markets, v. 31, n. 3, p. 685–695, 2021.

Kandpal, Nikhil et al. Large Language Models Struggle to Learn Long-Tail Knowledge. 2022. Disponível em: http://arxiv.org/abs/2211.08411.

Ke, Yu He et al. Development and Testing of Retrieval Augmented Generation in Large Language Models-A Case Study Report. [s.l: s.n.].

DOI: 10.59254/sbpo-2025-212623
https://proceedings.science/p/212623?lang=pt-br

Lakatos, Robert et al. Investigating the performance of Retrieval-Augmented Generation and fine-tuning for the development of AI-driven knowledge-based systems. 2024. Disponível em: http://arxiv.org/abs/2403.09727.

Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature. Maio: Nature Publishing Group, 2015. v. 27

Liu, Ryan Wen et al. An enhanced CNN-enabled learning method for promoting ship detection in maritime surveillance system. Maritime Safety. 2021.

Mogali, S.; Shivayogi E, M.; Shivaranjini, S. Artificial Intelligence and its applications in Libraries. [s.l: s.n.].

Moradi, M. et al. Exploring the landscape of large language models: Foundations, techniques, and challenges. 2024. Disponível em: http://arxiv.org/abs/2404.11973.

Naveed, Humza et al. A Comprehensive Overview of Large Language Models. 2023. Disponível em: http://arxiv.org/abs/2307.06435.

Picinini, A. M. Inteligência Artificial: principais possibilidades de aplicação nas operações militares do Exército Brasileiro. 2025. Disponível em: https://bdex.eb.mil.br/jspui/bitstream/123456789/14121/1/MO%207075%20-%20Alexandre%20Medeiros%20PICININI.pdf.

Pilehvar, M. T.; Camacho-Collados, J. Embeddings in natural language processing: Theory and advances in vector representations of meaning. Synthesis lectures on human language technologies, v. 13, n. 4, p. 1–175, 2020.

Rivera, Juan P. et al. Escalation Risks from LLMs in Military and Diplomatic Contexts. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), p. 1-12, 2024. Disponível em: https://dl.acm.org/doi/10.1145/3630106.3658942.

Smith, David e Jones, Michael. On Large Language Models in National Security Applications. Air Force Institute of Technology, 2023. Disponível em: https://scholar.afit.edu/cgi/viewcontent.cgi?article=2572&context=facpub.

Soong, David et al. Improving accuracy of GPT-3/4 results on biomedical data using a retrieval-augmented language model. 2023. Disponível em: http://arxiv.org/abs/2305.17116.

Vaswani, Ashish et al. Attention Is All You Need. 2017. Disponível em: http://arxiv.org/abs/1706.03762.

Yenduri, Gokul et al. Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. 2023. Disponível em: http://arxiv.org/abs/2305.10435.

DOI: 10.59254/sbpo-2025-212623
https://proceedings.science/p/212623?lang=pt-br