

Constituição Federal utilizando LLM e Rag

Federal Constitution using LLM and RAG

E. S. Ramos ; C. C. Aquino ; R. P. Leite ; L. S. Lima ; J. F. F.
Neto

¹ Departamento de Computação/Laboratório D02, Universidade Federal de Sergipe,
49100-000, São Cristóvão-Sergipe, Brasil

{edcarlosufs, riquelmewin, cainacastro29, falcao, waynewyn} @ academico.ufs.br

O uso de LLMs (Large Language Models) se tornou cada vez mais normal nos diversos campos dos conhecimentos, ao mesmo tempo que informações potencialmente alucinadas dessas ferramentas começaram a se espalhar. Isso levanta preocupações quanto à confiabilidade das respostas geradas, principalmente em contextos sensíveis como educação, direito e saúde. Apesar disso, os LLMs continuam sendo valiosos aliados na automação de tarefas, geração de conteúdo e suporte à pesquisa, desde que utilizados com senso crítico e validação adequada das informações fornecidas. O desafio atual é encontrar o equilíbrio entre aproveitar o potencial dessas ferramentas e mitigar seus riscos, garantindo o uso ético e responsável da inteligência artificial generativa.

Esse trabalho busca alinhar o conhecimento generalizado de uma llm à uma base de conhecimento especialista na área de direitos constitucionais a fim de divulgar informações confiáveis sobre a base da cidadania no Brasil.

1. INTRODUÇÃO

O desenvolvimento de grandes modelos de linguagem (LLMs) tem sido um campo de pesquisa amplamente explorado na Inteligência Artificial (IA), especialmente após o sucesso dos modelos da OpenAI, a exemplo do ChatGPT, uma inteligência artificial projetada para responder a diversos tópicos de diferentes áreas do conhecimento. Esse avanço impulsionou o surgimento de várias outras LLMs, despertando o interesse do mercado, que passou a adaptar essas ferramentas para suas necessidades específicas, seja no ambiente corporativo, na automação de processos, ou no atendimento ao cliente. Atualmente, essas tecnologias já estão amplamente presentes em diversas aplicações e serviços oferecidos ao público[1].

Além de auxiliarem em tarefas como geração de texto, tradução, resumo automático, e suporte conversacional, essas ferramentas também vêm sendo utilizadas como fontes de informação por milhões de usuários. No entanto, um desafio significativo persiste: a tendência desses modelos gerarem respostas "alucinadas", ou seja, informações falsas, imprecisas ou completamente inventadas[2]. Essa limitação pode ser especialmente crítica em contextos onde a veracidade dos dados é essencial, como em consultas jurídicas, médicas ou educacionais.

No Brasil, esse problema se agrava diante do baixo índice de familiaridade da população com a própria Constituição Federal[3]. Muitos cidadãos recorrem a ferramentas baseadas em IA em busca de orientação sobre direitos e deveres, mas, diante da possibilidade de receberem informações incorretas, tornam-se vulneráveis à desinformação e até mesmo podem contribuir para sua propagação.



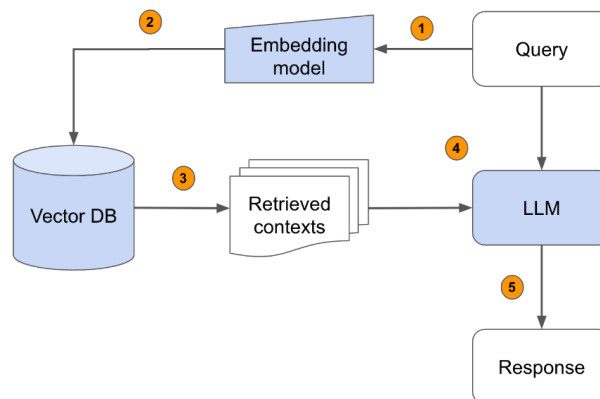
Uma abordagem promissora para mitigar o problema das alucinações é a utilização da técnica conhecida como Retrieval-Augmented Generation (RAG). Essa técnica combina o poder de geração de linguagem dos LLMs com a recuperação de informações verificadas em fontes externas confiáveis, como arquivos PDF, documentos do Word, planilhas CSV ou bancos de dados estruturados. Em vez de depender apenas do conhecimento pré-treinado do modelo, o sistema consulta essas fontes de forma dinâmica durante a geração de respostas, oferecendo maior precisão, rastreabilidade e contextualização nas informações fornecidas.

Dessa forma, o intuito deste trabalho é complementar o conhecimento de uma LLM pré-treinada com uma base de dados especializada — neste caso, o texto da Constituição Federal Brasileira — por meio da técnica de RAG. O objetivo final é criar um chatbot inteligente, hospedado em um perfil de usuário no WhatsApp, que seja capaz de responder dúvidas relacionadas aos artigos e leis presentes na Constituição. Essa solução tem o potencial de democratizar o acesso ao conhecimento jurídico, combatendo a desinformação e promovendo a cidadania consciente, ao mesmo tempo que explora as capacidades avançadas da inteligência artificial aplicada a um contexto socialmente relevante.

2. METODOLOGIA

2.1 Arquitetura da Aplicação

- a. **Frontend de consulta** (usuário final interage via WhatsApp).
- b. **Módulo RAG**: *engine* de recuperação semântica + geração:
 1. Transformação de arquivo pdf da constituição para *markdown* via Docling.
 2. Divisão por cabeçalhos.
 3. Redivisão inteligente para dividir *chunks* muito grandes e concatenar *chunks* pequenas.
 4. Embedding das *chunks* utilizando “*text-embedding-3-small*”
 5. Armazenamento em banco de dados vetorial.
- c. **Base de dados vetorial**: responsável pelo armazenamento e busca dos embeddings.
 1. Armazena todos os *embeddings*.
 2. Retorna os *embeddings* utilizando busca por similaridade.
 3. Acrescenta retorno da busca ao *prompt* antes de enviar para a llm.



- d. **Resposta ao usuário.**

2.2 Pseudocódigo dos Algoritmos Centrais

- Leitura do documento.

```
FUNCTION process_pdf_to_markdown(pdf_path):
  raw_text = Docling.extract_text(pdf_path)
  markdown = add_markdown_headers(raw_text)
  RETURN markdown
```

- Divisão em segmentos menores.

```
FUNCTION split_by_headers(markdown):
  splitter = MarkdownHeaderTextSplitter(
    headers_to_split_on=[("###", "Section")]
  )
  chunks = splitter.split_text(markdown)
  RETURN chunks
FUNCTION rebalance_chunks(chunks):
  final_chunks = []
  skip_next = False
  FOR EACH chunk IN chunks:
    IF skip_next:
      skip_next = False
      CONTINUE
    IF chunk.length > 2000:
      split_chunks = RecursiveTextSplitter(
        chunk_size=1100,
        separators=["\n\n", "\n", "."],
        keep_separator=True
      ).split_text(chunk.content)
      FOR split_chunk IN split_chunks:
        final_chunks.append({
          "content": split_chunk,
          "metadata": chunk.metadata,
          "length": len(split_chunk)
        })
    ELIF chunk.length < 500:
      IF NOT is_last_chunk(chunk, chunks):
        merged_chunk = merge_with_next(chunk, chunks)
        final_chunks.append(merged_chunk)
      ELSE:
        final_chunks.append(chunk)
    ELSE:
      final_chunks.append(chunk)
  RETURN final_chunks
FUNCTION merge_with_next(current_chunk, all_chunks):
  next_index = all_chunks.index(current_chunk) + 1
  next_chunk = all_chunks[next_index]
  RETURN {
    "content": f"{current_chunk.content}\n\n{next_chunk.content}",
```

```

    "metadata": combine_metadata(current_chunk.metadata,
next_chunk.metadata),
    "length": current_chunk.length + next_chunk.length
    skip_next = True
}

```

- Geração dos embeddings com modelo transformer.

```

FUNCTION generate_embeddings(final_chunks):
    embedder = OpenAIEmbeddings(model = "text-embedding-3-small")
    texts = [chunk.content FOR chunk IN final_chunks]
    vectors = embedder.embed_documents(texts)
    metadata = [
        {
            "header": chunk.metadata.Section,
            "length": chunk.length
        }
        FOR chunk IN final_chunks
    ]
    RETURN vectors, metadata

```

- Armazenamento no banco vetorial com metadados.

```

FUNCTION store_in_vectordb(vectors, metadata, texts):
    vectordb = ChromaDB(
        persist_path = "path/to/storage",
        collection_name = "constitution"
    )
    vectordb.add(
        embeddings = vectors,
        documents = texts,
        metadata = metadata
    )
    RETURN vectordb

```

2.3. Datasets Utilizados

- **Fonte de dados:**
 - Constituição Federal Brasileira [4]
- **Estrutura do corpus:**
 - Divisão em chunks maiores que 500 e menores que 2000 caracteres para compatibilidade.

2.4. Frameworks, Bibliotecas e Ferramentas

- **LLM e Embeddings:**
 - LLM : gpt-4o-mini

- Embedding : text-embedding-3-small
 - **Banco de dados vetorial:**
 - ChromaDB
 - **Linguagem e ambiente de desenvolvimento:**
 - Python (versão 3.10+).
 - Principais Frameworks e Bibliotecas de apoio: LangChain, Transformers , Docling e Flask.
 - Todos os requirements estão descritos no arquivo requirements.txt do repositório.
 - **Interface de comunicação:**
 - API oficial do Whatsapp, Openai.
-

2.5. Hardware

- **Hardware:**
 - CPU: AMD Ryzen 5 5500U with Radeon Graphics × 6
 - RAM: 20 GB DDR4
 - PLACA DE VÍDEO: Advanced Micro Devices, Inc. [AMD/ATI] Lucienne

3. TESTES REALIZADOS

1 Q : “O Presidente da República pode determinar a intervenção das Forças Armadas em operações de segurança pública de um estado sem solicitação do governo estadual, em caso de 'crise prolongada' não declarada como estado de defesa?”

R(Gpt 4 Mini) : “No Brasil, a intervenção das Forças Armadas em questões de segurança pública, sem a solicitação do governo estadual, é um tema complexo e depende do contexto legal. De acordo com a Constituição Federal, para a intervenção das Forças Armadas em uma situação de segurança pública, normalmente é necessário que haja um pedido do governo estadual ou que esta ação esteja fundamentada em situações de grave perturbação da ordem pública.

No entanto, a Constituição também prevê que o Presidente da República pode, em caso de "crise extrema" ou situações de emergência, ordenar intervenções. Isso geralmente é aplicado a casos onde há uma ameaça à ordem pública que não seja adequadamente resolvida pelas autoridades locais.

Portanto, a atuação das Forças Armadas em operações de segurança pública sem o pedido do estado pode ocorrer, mas deve ser em situações excepcionais e fundamentadas em normas específicas, levando em conta a complexidade dos aspectos legais e constitucionais envolvidos. A falta de uma declaração formal de estado de defesa ou de estado de sítio complicaria a situação e poderia gerar debates sobre a legalidade da ação. Em resumo, embora ao Presidente da República haja essa possibilidade em casos extremos, a questão é juridicamente delicada e deve seguir os preceitos legais estabelecidos.

R(Rag + Gpt 4 Mini) : “De acordo com os documentos fornecidos, o Presidente da República não pode determinar a intervenção das Forças Armadas em operações de segurança pública de um estado sem a solicitação do governo estadual, em caso de "crise prolongada" não declarada como estado de defesa. A Constituição prevê, no Art. 136, que o Presidente pode decretar estado de defesa para preservar a ordem pública, mas isso requer a declaração formal e a comunicação com o Congresso Nacional. Para situações que se enquadram como crise, mas não configuram um estado de defesa ou de sítio, não há previsão legal para a intervenção sem solicitação do estado.

Referência: Constituição, Art. 136 (do Estado de Defesa).”

2 Q : “Um servidor público pode recusar-se a emitir certidões de casamento para casais homoafetivos, invocando objeção de consciência religiosa, sem violar a Constituição?”

R(Gpt 4 Mini) : “A recusa de um servidor público em emitir certidões de casamento para casais homoafetivos, invocando objeção de consciência religiosa, pode gerar conflitos com princípios constitucionais, principalmente com o direito à igualdade e não discriminação, assegurados na Constituição Brasileira.

O Supremo Tribunal Federal (STF) já se posicionou a favor do reconhecimento da união estável e do casamento entre pessoas do mesmo sexo, considerando que a orientação sexual não deve ser motivo para a exclusão de direitos. A recusa em prestar um serviço público essencial, como a emissão de certidões de casamento, pode caracterizar discriminação e violar o dever do servidor de atuar com imparcialidade e respeito aos direitos fundamentais.

Embora a objeção de consciência seja um direito reconhecido em determinadas situações, a sua aplicação em serviços públicos, especialmente em casos que envolvem direitos civis fundamentais, é complexa e deve ser ponderada. Em geral, a proteção dos direitos de grupos historicamente marginalizados, como a população LGBT, tende a prevalecer sobre objeções pessoais de consciência que poderiam resultar em discriminação.

Assim, a recusa pode ser considerada inconstitucional e contrária ao dever do servidor público, que é garantir o acesso igualitário a todos os cidadãos. Portanto, a objeção de consciência, nesse contexto, não pode ser utilizada como justificativa válida para a discriminação.“

R(Rag + Gpt 4 Mini): “ Um servidor público não pode recusar-se a emitir certidões de casamento para casais homoafetivos invocando objeção de consciência religiosa, pois tal recusa violaria princípios constitucionais. A Constituição Federal, conforme o Art. 226, garante que o casamento é civil e que a celebração é gratuita, sem discriminação em relação à orientação sexual dos seus participantes. Além disso, a vedação de recusa em prestar serviços públicos, conforme o Art. 19, II, reforça que a administração pública deve agir de maneira imparcial e com respeito aos direitos de todos os cidadãos. “

3 Q : “Civis que cometem crimes contra militares durante uma Garantia da Lei e da Ordem (GLO) podem ser julgados pela Justiça Militar, mesmo sem vínculo com as Forças Armadas?”

R(Gpt 4 Mini) : Sim, civis que cometem crimes contra militares durante uma Garantia da Lei e da Ordem (GLO) podem ser julgados pela Justiça Militar. Isso ocorre porque, durante uma GLO, as Forças Armadas são mobilizadas para a manutenção da ordem pública, e os crimes que envolvem militares no exercício de suas funções podem ser considerados de competência da Justiça Militar, conforme estabelece a legislação brasileira. A jurisprudência e a interpretação do direito corroboram que, no contexto de atuação das Forças Armadas, a Justiça Militar pode ter jurisdição sobre os atos praticados por civis contra militares.

R (Rag + Gpt 4 Mini) : Com base nos documentos fornecidos, a resposta é não. Civis que cometem crimes contra militares durante uma Garantia da Lei e da Ordem (GLO) não podem ser julgados pela Justiça Militar, pois a competência da Justiça Militar se restringe ao processamento e julgamento de crimes militares definidos em lei, sem previsão para o julgamento de civis nesses casos.

A informação foi extraída do § 4 e § 5, que abordam a competência da Justiça Militar em relação a militares e as ações judiciais contra atos disciplinares militares, ressaltando a exclusão da competência do júri quando a vítima for civil.

4. RESULTADOS

- Redução de Alucinações
 - O RAG eliminou respostas incorretas em questões sensíveis (ex.: Justiça Militar).
 - Sem RAG, o LLM errou em 1/3 dos casos (33% de imprecisão grave).
- Fundamentação Jurídica Sólida
 - Todas as respostas com RAG citam artigos específicos da Constituição.
 - Sem RAG, as respostas são genéricas e sem referências.
- Clareza e Autoridade
 - O RAG padroniza respostas com:
 - Citação direta da lei.
 - Estrutura lógica (ex.: "De acordo com o Art. X...").
 - Sem RAG, o LLM tende a divagar e incluir suposições.

Métricas LLM	Com Rag	Sem Rag
Citação de artigos	100%	0%
Precisão jurídica	100%	66%

5. CONCLUSÃO:

Apesar de ter fornecido uma resposta correta em relação à constituição de 1988, em 2023 o Supremo Tribunal Federal (STF), em 9 de novembro de 2023, validou o julgamento de civis pela Justiça Militar da União em casos de crimes cometidos contra militares durante operações de Garantia da Lei e da Ordem (GLO).[5]

Então mesmo utilizando um documento de expertise no assunto devido à situações muito específicas a LLM com Rag acertou no que diz referente às

informações possuídas pela llm e pelo rag, mas devido a acontecimentos não registrados na base de informações a resposta está desatualizada e alucinada. Essa limitação evidencia um dos principais desafios do uso de LLMs com RAG: a dependência da atualidade e abrangência da base de dados utilizada para recuperação de informações.

6. REFERÊNCIAS BIBLIOGRÁFICAS

1. PORTAL DA INDÚSTRIA. Como as empresas têm usado a inteligência artificial e qual o impacto no mercado de trabalho? Notícias, Educação, [S. l.], [2025]. Disponível em: <https://noticias.portaldaindustria.com.br/noticias/educacao/como-as-empresas-tem-usado-a-inteligencia-artificial-e-qual-o-impacto-no-mercado-de-trabalho/>. Acesso em: 8 abr. 2025.
2. TECNOBLOG. Pesquisa mostra que chatbots estão errados em 60% do tempo. Notícias, [S. l.], 2025. Disponível em: <https://tecnoblog.net/noticias/pesquisa-mostra-que-chatbots-estao-errados-em-60-do-tempo/>. Acesso em: 7 ABR. 2025.
3. SENADO FEDERAL. Pesquisa DataSenado mostra que poucos conhecem realmente a Constituição do país. Notícias, Brasília, 25 out. 2013. Disponível em: <https://www12.senado.leg.br/noticias/materias/2013/10/25/pesquisa-datasenado-mostra-que-poucos-conhecem-realmente-a-constituicao-do-pais>. Acesso em: 7 ABR. 2025
4. BRASIL. Constituição da República Federativa do Brasil de 1988. Brasília, DF: Supremo Tribunal Federal, 2024. Disponível em: https://www2.senado.leg.br/bdsf/bitstream/handle/id/518231/CF88_Livro_EC91_2016.pdf. Acesso em: 7 abr. 2025.
5. AGÊNCIA BRASIL. Supremo valida julgamento de civil pela Justiça Militar. Justiça, 9 nov. 2023. Disponível em: <https://agenciabrasil.ebc.com.br/justica/noticia/2023-11/supremo-valida-julgamento-de-civil-pela-justica-militar>. Acesso em: 8 abr. 2025.