

APLICAÇÃO DE RETRIEVAL AUGMENTED GENERATION PARA ANÁLISE AUTOMÁTICA DE PROCESSOS JURÍDICOS

Luis Fernando Alves de Oliveira¹, João Henrique Simionato Domingues², Ana Carolina Gonçalves Ribeiro³, Cristiane Yae Mi Imamura⁴

¹Centro Universitário das Faculdades Associadas de Ensino - Unifae, São João da Boa Vista-SP, Brasil (luis.oliveira@sou.fae.br)

^{2, 3, 4}Centro Universitário das Faculdades Associadas de Ensino - Unifae, São João da Boa Vista-SP, Brasil

Resumo: Tendo em vista o poder das ferramentas de Inteligência Artificial Generativa, pode-se usá-las na análise de contratos judiciais. Para evitar o risco de alucinação, neste trabalho, são descritos os resultados obtidos no desenvolvimento de uma aplicação com Naive Retrieval-Augmented Generation com HyDE. Essa aplicação visa apoiar a tomada de decisões sobre documentos jurídicos relacionados ao Direito do Consumidor. Os resultados apontam que a aplicação devolve pareceres mais detalhados quando comparados ao de Inteligência Artificial Generativa.

Palavras-chave: Inteligência Artificial Generativa; Retrieval Augmented Generation; Direito do Consumidor.

INTRODUÇÃO

O Poder Judiciário brasileiro enfrenta uma alta litigiosidade, com mais de 80 milhões de processos pendentes (CNJ, 2025, p. 20). Em 2024, o Direito do Consumidor registrou mais de 533 mil ações movidas (Silva, 2025). O elevado volume e a complexidade dessas demandas tornam a análise manual de documentos e fundamentos jurídicos uma tarefa excessivamente demorada, evidenciando a necessidade de técnicas de automatização.

Nesse contexto, a Inteligência Artificial (IA) e o Processamento de Linguagem Natural (PLN), especialmente os *Large Language Models* (LLM), surgem como soluções promissoras. Contudo, a aplicação de LLMs apresenta alguns problemas, como a propensão a alucinações e o acesso restrito a bases de dados atualizadas (Paranhos et al., 2024).

Como alternativa a esse problema, destaca-se a técnica de *Retrieval Augmented Generation* (RAG), que fornece à IA uma base de conhecimento atualizada e válida (Lewis et al., 2020). Entre as arquiteturas disponíveis, podemos destacar o *Naive RAG with HyDE* (*Hypothetical Document Embeddings*) que otimiza a busca semântica ao gerar um documento de resposta hipotético a partir da consulta do usuário. Isso resulta em recuperações mais detalhadas e coerentes do que a busca tradicional (*Naive RAG*) (Paranhos et al., 2024; Gao et al. 2022).

Diante desse cenário, este trabalho propõe a construção de um sistema baseado em *RAG with HyDE* para a análise de documentos ligados ao direito do consumidor. Desta forma, espera-se que a aplicação possa ser útil para pessoas comuns ou profissionais da área do direito na sua tomada de decisão.

MATERIAL E MÉTODOS

Para estruturar a arquitetura da aplicação proposta, foi observada a descrição da arquitetura RAG descrita por Dubey (2024), conforme vista na Figura 1.

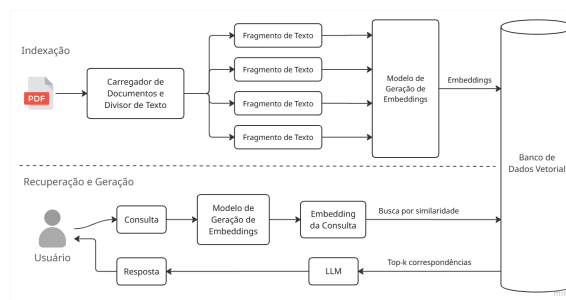
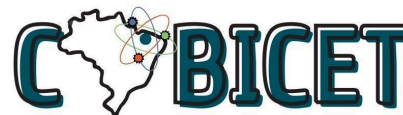


Figura 1. Visão Geral da arquitetura de sistemas baseados em RAG. Fonte: adaptada de Dubey (2024).

De acordo com o autor, primeiro, os documentos externos são inseridos na base de dados, passando pelo estágio de pré-processamento, que envolve a



limpeza textual, normalização e segmentação em pequenas partes, facilitando a manipulação.

Após essa etapa, pedaços desses textos são transformados em embeddings, para posterior indexação para busca semântica, um processo que permite que o sistema encontre informações relevantes ainda que a consulta não combine com a forma das consultas.

Dessa forma, o usuário faz uma consulta, o sistema processa a entrada, busca os documentos mais relevantes encontrados na base, e utiliza o modelo de linguagem para gerar uma resposta baseada nas informações recuperadas.

Para a proposta deste trabalho, a parte de inserção do HyDE dá-se no aspecto que a consulta é enviada para um gerador LLM de hipótese que gera um documento hipotético antes de criar o *embedding* de busca.

Os documentos considerados para a geração dos *embeddings* da base de conhecimento são os da legislação de Defesa do Consumidor: a Lei nº 8.078/1990 (Brasil, 1990) e o Decreto nº 7.962/2013 (Brasil, 2013).

As ferramentas selecionadas para desenvolvimento do projeto foram:

- **Gemini 1,3 Flash:** Empregado como o sistema de LLM para gerar o documento hipotético para a fase de HyDE e também para gerar a resposta final a ser devolvida e apresentada pelo usuário. Foi escolhido pela eficiência e simples utilização (Google DeepMind, 2023).
- **Ollama, versão 0.6.3 com o LLM nomic-embed-text:** Utilizada para prover o modelo de *embeddings* de código aberto (Ollama, 2023).
- **SQLite:** Aplicado como sistema de gerenciamento de banco de dados (SGBD), oferecendo leveza e facilidade de integração permitindo execução local (SQLite, 2000).
- **Node.js:** Adotado por seu modelo de I/O não bloqueante (Node.js Foundation, 2009), o que o torna ideal para gerenciar as chamadas com latência para APIs externas do Gemini e Ollama.
- **React:** Utilizado para o desenvolvimento da interface Web da aplicação, permitindo a criação de componentes reutilizáveis e uma interação dinâmica com o usuário (React Foundation, 2013).
- **React Native:** Utilizado para o desenvolvimento da aplicação mobile, com a criação de componentes reutilizáveis e uma interação dinâmica com o usuário (React Foundation, 2013).

RESULTADOS E DISCUSSÃO

Boa parte deste trabalho se deu na implementação de integrações entre ferramentas responsáveis pelos modelos de LLMs.

A Figura 2 mostra como a arquitetura da aplicação funcionará.

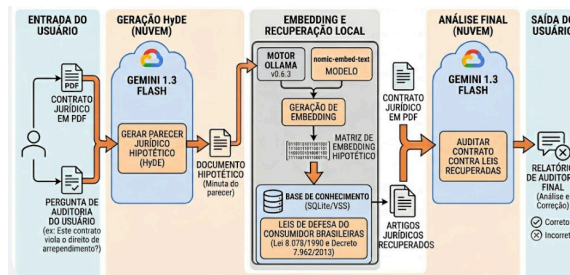


Figura 2. Fluxo principal da execução do sistema proposto. Fonte: Construído pelos autores com auxílio da Ferramenta Nano Banana 2.

Neste primeiro momento para testar a aplicação, um documento foi gerado e o prompt passado por enquanto é estático e sempre o mesmo: analisar o documento completamente.

Na Figura 3 é possível observar parte da resposta obtida através da aplicação.

```
response: [
  {
    "secao": "Qualificação das Partes",
    "analise": "As partes são qualificadas como Contratante (João da Silva, pessoa física) e Contratado (Empresa XYZ Ltda., pessoa jurídica). Conforme o Art. 2º da Lei nº 8.078/1990, se João da Silva adquire ou utiliza o serviço como destinatário final, ele se enquadra como consumidor. A Empresa XYZ Ltda., ao desenvolver atividade de prestação de serviços, enquadra-se como fornecedor (Art. 3º da Lei nº 8.078/1990). Assumindo a condição de destinatário final do serviço para o Contratante, as normas do Código de Defesa do Consumidor são aplicáveis ao presente contrato, conforme Art. 1º da Lei nº 8.078/1990.",
    "classified": "Seguro"
  },
  {
    "secao": "CLÁUSULA 1ª - OBJETO E CLÁUSULA 2ª - ESCOPO DOS SERVIÇOS",
    "analise": "As cláusulas definem o objeto do contrato como prestação de serviços de desenvolvimento, manutenção e suporte de sistemas de software, e o escopo das atividades do Contratado. O conteúdo é informativo e não apresenta termos que conflitem com os direitos básicos do consumidor, como a informação adequada e clara sobre o serviço (Art. 6º, III da Lei nº 8.078/1990).",
    "classified": "Seguro"
  },
  {
    "secao": "CLÁUSULA 3ª - PRAZO",
    "analise": "A cláusula estabelece o prazo de execução do contrato e a possibilidade de prorrogação mediante acordo entre as partes. Não há termos problemáticos à luz do conhecimento disponível sobre Direito do Consumidor.",
    "classified": "Seguro"
  },
  {
    "secao": "CLÁUSULA 4ª - RESCISÃO",
    "analise": "A cláusula estabelece as condições para a rescisão do contrato por ambas as partes, incluindo a multa rescisória. Os termos são claros e não violam os direitos do consumidor.",
    "classified": "Seguro"
  }
]
```

Figura 3. Análise da IA utilizando RAG with HyDE

Para testar se o comportamento do sistema se alterava, a partir de seu uso, o mesmo documento foi passado como entrada uma segunda vez. Parte da saída obtida está descrita na Figura 4.

Comparando as duas figuras, é possível observar que a resposta da aplicação foi mais superficial na Figura 3 e que a segunda vez gerou uma análise mais completa. Isso provavelmente se deve ao fato de que o conhecimento do sistema é alterado também de acordo com as análises anteriores feitas. Mas não significa que passar o mesmo documento no sistema, por mais vezes, fará com que o sistema gere análises melhores sempre.

```

Parsed response: {
  contract_analysis: {
    {
      secao: 'Qualificação das Partes',
      analise: 'Identifica as partes Contratante (pessoa física) e Contratado (pessoa jurídica). Assumindo que o Contratante, João da Silva, adquire o serviço como destinatário final, o contrato pode ser regido pelas normas de proteção e defesa do consumidor, conforme o Art. 2º e Art. 3º da Lei nº 8.078/1990. Os dados essenciais para identificação das partes estão presentes.',
      classified: 'Seguro'
    },
    {
      secao: 'CLÁUSULA 1ª - OBJETO',
      analise: 'Define o objeto do contrato como prestação de serviços de desenvolvimento, manutenção e suporte de sistemas de software, de forma clara e adequada, conforme o direito básico à informação (Art. 6º, III da Lei nº 8.078/1990).',
      classified: 'Seguro'
    },
    {
      secao: 'CLÁUSULA 2ª - ESCOPO DOS SERVIÇOS',
      analise: 'Detalha as atividades incluídas no escopo dos serviços. A clareza na definição do escopo é essencial para a informação adequada do consumidor (Art. 6º, III da Lei nº 8.078/1990).',
      classified: 'Seguro'
    }
  }
}

```

Figura 4. Análise da IA utilizando RAG with HyDE sobre um documento já analisado.

Também é importante ressaltar que quando comparado com o ChatGPT (GPT-5.5 Instant), o presente sistema que utiliza o RAG apresentou um resultado consideravelmente mais satisfatório em termos de relevância e conhecimento da lei, se for observada na Figura 5, em comparação com os dois resultados anteriores.

```

sections": [
  {
    "secao": "Identificação das Partes",
    "classified": "Seguro",
    "analysis": "A seção apresenta identificação básica das partes contratantes, incluindo CPF, CNPJ e endereço. Não há termos abusivos ou cláusulas de risco elevado."
  },
  {
    "secao": "Objeto do Contrato",
    "classified": "Seguro",
    "analysis": "Define claramente a prestação de serviços de desenvolvimento, manutenção e suporte de software, sem ambiguidades relevantes."
  },
  {
    "secao": "Escopo dos Serviços",
    "classified": "Seguro",
    "analysis": "O escopo descreve atividades previstas como reunidas, documentação e suporte pós-implantação. Não há obrigações excessivamente desequilibradas."
  }
]

```

Figura 5. Análise do documento gerada com ChatGPT (GPT-5.5 Instant) através da sua própria interface.

CONCLUSÃO

O trabalho desenvolvido, mesmo que ainda em fase de desenvolvimento, já conseguiu atingir resultados expressivos, demonstrando que a implementação do método *RAG* em conjunto com a técnica de HyDE, conseguem lapidar os resultados gerados por uma inteligência artificial, além de diminuir a ocorrência de alucinações.

Essa demonstração pode ser convalidada após as análises realizadas, e por isso espera-se que a implementação da aplicação para este cenário jurídico resulte em uma análise superior à das LLMs convencionais, em virtude do seu repositório flexível e da maior profundidade de conhecimento proporcionada pela técnica.

Para atingir os objetivos propostos neste trabalho, a aplicação será alterada para permitir o recebimento de prompt com pergunta de auditoria por parte do usuário, como entrada.

Tendo em vista que outras áreas do direito também podem ser beneficiadas com aplicações deste tipo, pretende-se considerar futuramente os documentos referentes às leis trabalhistas brasileiras. Isso poderia beneficiar a verificação de contratos de prestação de serviços que tentam camuflar relações de consumo, ou contratos de consumo que tentam esconder vínculos empregatícios.

Vale ressaltar que a ferramenta proposta não visa substituir o profissional do direito e sim apoiá-lo para realização de suas tarefas com mais qualidade. Neste sentido, para avaliar melhor a ferramenta e os resultados que ela gera, pretende-se contar com especialistas da área do direito também.

Diante dos resultados, mesmo que iniciais, considera-se que aplicações como a proposta neste trabalho são importantes, pois, considerando o atual cenário jurídico brasileiro, a automatização de determinadas funções é um caminho plausível para auxiliar no desengasgo dos processos pendentes.

REFERÊNCIAS

BRASIL. Lei nº 8.078, de 11 de setembro de 1990. Código de Defesa do Consumidor. Disponível em: <https://www.jusbrasil.com.br/legislacao/91585/codigo-de-defesa-do-consumidor-lei-8078-90>. Acesso em: 23 mai. 2026.

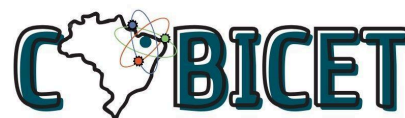
BRASIL. Decreto nº 7.962, de 15 de março de 2013. Regulamenta a Lei nº 8.078, de 11 de setembro de 1990, para dispor sobre a contratação no comércio eletrônico. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2013/decreto/d7962.htm. Acesso em: 23 mai. 2026.

CONSELHO NACIONAL DE JUSTIÇA. Justiça em Números. Brasília: CNJ, <https://www.cnj.jus.br/wp-content/uploads/2025/11/justica-em-numeros-2025.pdf>. 2025. Acesso em: 13 jul. 2026

DUBEY, Ashish. RAG Architecture: Build scalable and efficient RAG systems. TrueFoundry, 2024. Disponível em: <https://www.truefoundry.com/blog/rag-architecture>. Acesso em: 24 maio 2026.

GAO, Luyu et al. Precise Zero-Shot Dense Retrieval without Relevance Labels. [s.l.]: arxiv, <https://arxiv.org/abs/2212.10496>. 2022. Acesso em: 14 mai. 2026

GOOGLE DEEPMIND. Gemini: a family of multimodal large language models. Londres: Google DeepMind, 2023. Disponível em: <https://arxiv.org/pdf/2312.11805>. Acesso em: 23 mai. 2026.



LEWIS, Patrick et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, 33, 9459-9474, 2020. Acesso em: 27 mar 2026

OLLAMA. Ollama: get up and running with large language models locally. [S.l.]: Ollama Inc., 2023. Disponível em: <https://ollama.com>. Acesso em: 23 mai. 2026.

PARANHOS, S. L. et al. Avaliação do Impacto de Diferentes Padrões de Arquiteturas RAG em Domínios Jurídicos. *Anais da XII Escola Regional de Informática de Goiás (ERI-GO 2024)*, 99-108, 2024. Acesso em: 24 mar 2026

REACT FOUNDATION. React: a JavaScript library for building user interfaces. [S.l.]: React Foundation, 2013. Disponível em: <https://react.dev>. Acesso em: 23 mai. 2026.

REACT FOUNDATION. React Native: an open-source UI software framework. [S.l.]: React Foundation, 2015. Disponível em: <https://reactnative.dev>. Acesso em: 23 mai. 2026.

SILVA, Maria Luiza Christina Ferreira da. A importância do código de defesa do consumidor na gestão comercial. Trabalho de conclusão de curso (Tecnologia em gestão comercial) - Campus Pedreiras, Universidade Estadual do Maranhão, <https://repositorio.uema.br/bitstream/123456789/5380/1/TCC%20-%20MARIA%20LUIZA%20CHRISTINA%20FERREIRA%20DA%20SILVA%20%e2%80%93%20TEC.%20EM%20GEST%C3%83O%20COMERCIAL%20PEDREIRAS%202025.pdf>, 2024. Acesso: em 14 mai 2026

SQLITE CONSORTIUM. SQLite: a C-language library that implements a small, fast, self-contained SQL database engine. [S.l.]: SQLite, 2000. Disponível em: <https://sqlite.org>. Acesso em: 23 mai. 2026.