Relatório Aula 1

Evandro Felipe Silva[1]

[1]Ciência de Dados – Faculdade Donaduzzi
Paraná – PR – Brasil

evandrofelipesilva@hotmail.com

1. O que é RAG?

Retrieval-Augmented Generation (RAG) é uma técnica que combina um sistema de recuperação (retrieval) de informações com um modelo gerador de linguagem (LLM), permitindo que este acesse e utilize dados externos, além dos seus parâmetros estáticos. Ou seja, antes de gerar uma resposta, o modelo consulta uma base de conhecimento ou documentos relevantes, o que reduz a ocorrência de "alucinações" e permite respostas mais atualizadas e contextualizadas.

2. Aplicação do RAG em sistemas de Pergunta e Resposta

Em sistemas de Perguntas e Respostas (QA), o fluxo típico do RAG é:

- **Retrieval**: busca-se automaticamente documentos ou trechos pertinentes à pergunta.
- **Generation**: com base no conteúdo recuperado, o modelo gera uma resposta coerente e contextualizada.

No contexto de sistemas baseados em Retrieval-Augmented Generation (RAG), a etapa de preparação dos dados tem início com a divisão dos documentos em trechos menores, comumente chamados de chunks. Esses trechos correspondem, geralmente, a parágrafos ou blocos de texto de aproximadamente mil caracteres, podendo conter sobreposições entre si a fim de preservar o contexto semântico integral do conteúdo original (Instituto de Computação, MDPI, Roberto Dias Duarte, Medium).

Cada um desses trechos é submetido a um modelo de embeddings — como Sentence-BERT, MiniLM ou OpenAI text-embedding — que os converte em vetores densos. Esses vetores encapsulam o significado semântico dos textos e os representam numericamente em um espaço vetorial de alta dimensionalidade (Instituto de Computação, Medium, notes.suhaib.in). Uma vez vetorizados, os dados são organizados e armazenados em bancos vetoriais especializados, como FAISS, Pinecone ou Qdrant, que utilizam algoritmos de busca de vizinhança aproximada (ANN), como o HNSW (Hierarchical Navigable Small World), para permitir a recuperação eficiente dos trechos mais similares com base na proximidade vetorial (Reddit, Medium, Andressa Siqueira).

Quando o usuário submete uma pergunta ao sistema, essa pergunta é igualmente vetorizada utilizando o mesmo modelo de embedding aplicado aos documentos indexados. Isso assegura que tanto as perguntas quanto os trechos de texto residam no mesmo espaço semântico, permitindo comparações coerentes e eficazes (Roberto Dias Duarte, Andressa Siqueira, notes.suhaib.in).

A próxima etapa é a busca por similaridade vetorial. Calcula-se, por exemplo, a similaridade do cosseno entre o vetor da pergunta e todos os vetores dos trechos indexados. Os K trechos mais similares — ou seja, aqueles com maior grau de correspondência

semântica com a pergunta — são então selecionados para compor o contexto que será fornecido ao modelo gerador (arXiv, Wikipedia, Roberto Dias Duarte, Andressa Siqueira). Em arquiteturas mais avançadas, aplicam-se técnicas adicionais como MMR (Maximal Marginal Relevance) para garantir diversidade semântica entre os trechos escolhidos, ou mecanismos de reclassificação neural (como cross-encoders) para refinar a relevância dos resultados (Reddit).

Os trechos textuais assim recuperados — e não os vetores — são então agrupados juntamente com a pergunta original e introduzidos no prompt do modelo de linguagem. Esse modelo, por sua vez, utiliza os conteúdos fornecidos como base para gerar uma resposta textual fundamentada, extraída diretamente da base documental (Wikipedia, MDPI, Engineer's Log, Instituto de Computação).

A resposta final, portanto, não é apenas uma produção linguística do modelo, mas sim uma síntese contextualizada dos dados efetivamente presentes nos documentos recuperados. Caso a base recuperada não contenha informações suficientes ou haja ambiguidade, o modelo pode, inclusive, optar por não fornecer uma resposta conclusiva, evitando assim a geração de conteúdo especulativo ou infundado.

3. Vantagens e limitações do RAG — comparado à IA tradicional

Em comparação com abordagens tradicionais de inteligência artificial, a técnica de Retrieval-Augmented Generation (RAG) apresenta vantagens expressivas, especialmente em tarefas que exigem conhecimento factual atualizado e contextualização precisa. Um dos principais benefícios do RAG é a redução significativa das chamadas "alucinações" — respostas fabricadas ou imprecisas por parte do modelo gerador — uma vez que a geração textual é ancorada em documentos efetivamente recuperados e fornecidos como contexto. Além disso, o RAG proporciona flexibilidade ao permitir a atualização de seu conhecimento sem necessidade de reprocessamento ou retreinamento do modelo base, bastando a inclusão de novos documentos na base vetorial. Isso se traduz em menor custo computacional para adaptação a domínios específicos.

Outro diferencial relevante é a explicabilidade. Como as respostas são derivadas de trechos documentais explícitos, é possível rastrear as fontes utilizadas, o que promove maior transparência e confiabilidade — aspecto crítico em domínios regulados, como o jurídico e o médico.

No entanto, essa arquitetura também impõe certas limitações. Em primeiro lugar, há um aumento considerável na complexidade técnica e infraestrutural do sistema, que agora deve englobar mecanismos robustos de chunking, vetorização, indexação e recuperação eficiente. A latência tende a ser maior, dado que cada pergunta implica uma consulta à base vetorial antes da geração da resposta. Ademais, a qualidade da resposta está diretamente condicionada à qualidade da base documental: se os documentos recuperados forem irrelevantes, desatualizados ou imprecisos, o modelo poderá gerar respostas igualmente comprometidas.

Outro desafio consiste na curadoria e manutenção contínua da base de conhecimento. Diferentemente de modelos tradicionais, cujo conhecimento é fixado durante o treinamento, os sistemas RAG requerem uma base documental confiável, limpa e atualizada para manter sua eficácia. Há ainda riscos relacionados à segurança e privacidade

da informação, especialmente em ambientes sensíveis: documentos internos podem ser expostos inadvertidamente por meio da recuperação ou da geração textual, exigindo mecanismos rigorosos de controle de acesso e monitoramento de integridade dos dados.

Em síntese, o RAG representa uma evolução significativa no campo da IA, sobretudo para tarefas que exigem precisão factual e transparência. No entanto, seus ganhos vêm acompanhados de desafios técnicos e operacionais que exigem planejamento criterioso e governança sólida para que seu potencial seja plenamente realizado.

4. Exemplos de uso de IA no campo jurídico

Jurimetria: análise estatística de decisões, tempo médio de processos, taxas de êxito etc., com IA para prever padrões judiciais e subsidiar estratégias.

Chatbots jurídicos: assistentes automatizados que respondem a dúvidas legais, auxiliam no preenchimento de documentos ou orientam usuários, baseando-se em legislação, jurisprudência e normas internas.

Jurisprudência automatizada: sistemas que recuperam acórdãos e precedentes relevantes para fundamentar decisões ou elaborar pareceres de forma mais ágil.

5. Desafios da IA aplicada ao Direito: ética, responsabilidade e confiabilidade

A aplicação da inteligência artificial no campo jurídico levanta desafios significativos relacionados à ética, responsabilidade e confiabilidade. Do ponto de vista ético, um dos principais riscos é a reprodução de vieses históricos presentes nos dados utilizados para treinar os modelos, o que pode perpetuar desigualdades e injustiças no processo decisório. Além disso, a transparência das decisões automatizadas é essencial, especialmente em temas sensíveis, exigindo que os sistemas permitam a rastreabilidade das informações utilizadas na geração das respostas.

Quanto à responsabilidade, é fundamental estabelecer claramente quem responde por eventuais erros cometidos por sistemas de IA — se os desenvolvedores, as instituições que os implementam ou os profissionais que os utilizam como apoio. A ausência de regulamentação específica sobre esse ponto pode gerar insegurança jurídica e dificultar a adoção segura dessas tecnologias.

Por fim, a confiabilidade dos sistemas depende da qualidade, integridade e atualização constante das bases de dados utilizadas. Informações incompletas, desatualizadas ou mal interpretadas podem comprometer a acurácia das respostas. Portanto, a adoção de IA no Direito exige não apenas rigor técnico, mas também uma governança sólida, pautada em princípios de justiça, transparência e responsabilidade social.