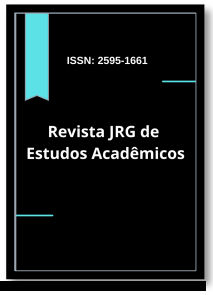
Listas de conteúdos disponíveis em Portal de Periódicos CAPES

ISSN: 2595-1661

ARTIGO

Revista JRG de Estudos Acadêmicos

Página da revista:
https://revistajrg.com/index.php/jrg



A transição do RAG tradicional para o RAG multimodal: desafios na extração semântica e indexação de documentos complexos

From traditional RAG to multimodal RAG: challenges in semantic extraction and indexing of complex documents

DOI: 10.55892/jrg.v9i20.3500
ARK: 57118/JRG.v9i20.3500

Recebido: 12/06/2026 | Aceito: 15/06/2026 | Publicado *on-line*: 16/06/2026

Ítalo Miguel Castor Diniz Pinheiro[1]
Universidade Federal de Campina Grande (UFCG)
E-mail: italomiguelpinheiro@gmail.com

Resumo

Os sistemas Retrieval-Augmented Generation (RAG) consolidaram-se como uma das principais estratégias para ampliar a capacidade dos modelos de linguagem por meio da integração entre recuperação de informações e geração de respostas fundamentadas em fontes externas de conhecimento. Entretanto, as arquiteturas tradicionais de RAG foram desenvolvidas predominantemente para ambientes textuais, apresentando limitações quando aplicadas a documentos complexos compostos por múltiplas modalidades de informação, como tabelas, gráficos, imagens e estruturas visuais. Nesse contexto, o presente estudo teve como objetivo analisar a transição do RAG tradicional para o RAG multimodal, com ênfase nos desafios relacionados à extração semântica, indexação documental e recuperação de informações em documentos visualmente ricos. Trata-se de uma revisão narrativa da literatura baseada na análise de estudos publicados entre 2020 e 2025 que abordaram recuperação aumentada por geração, compreensão multimodal de documentos, modelos visão-linguagem e indexação visual. Os resultados demonstraram que abordagens tradicionais dependentes de OCR e processamento textual tendem a apresentar perdas semânticas decorrentes da incapacidade de preservar adequadamente informações estruturais e visuais. Em contrapartida, modelos multimodais recentes, como LayoutLMv2, LayoutXLM, Donut, ColPali, VisRAG, M3DocRAG e VDocRAG, evidenciam avanços significativos na compreensão documental ao integrar simultaneamente informações textuais, visuais e espaciais. Conclui-se que a evolução para arquiteturas multimodais representa uma mudança paradigmática na recuperação de informações, ampliando a capacidade dos sistemas de inteligência artificial em compreender, indexar e recuperar conhecimento contido em documentos complexos.

Palavras-chave: Retrieval-Augmented Generation. RAG multimodal. Compreensão documental. Modelos visão-linguagem. Recuperação de informações.

[1] Bacharelado em Ciência da Computação

Abstract

Retrieval-Augmented Generation (RAG) systems have become one of the most prominent strategies for enhancing the capabilities of large language models through the integration of information retrieval and response generation grounded in external knowledge sources. However, traditional RAG architectures were primarily designed for text-based environments and present significant limitations when applied to complex documents containing multiple information modalities, such as tables, charts, images, and visual structures. In this context, the present study aimed to analyze the transition from traditional RAG to multimodal RAG, emphasizing the challenges related to semantic extraction, document indexing, and information retrieval in visually rich documents. This study consists of a narrative literature review based on the analysis of publications between 2020 and 2025 addressing retrieval-augmented generation, multimodal document understanding, vision-language models, and visual indexing. The findings indicate that conventional approaches relying on optical character recognition (OCR) and textual processing are prone to semantic information loss due to their limited ability to preserve structural and visual features. In contrast, recent multimodal models, including LayoutLMv2, LayoutXLM, Donut, ColPali, VisRAG, M3DocRAG, and VDocRAG, have demonstrated significant advances in document understanding by jointly integrating textual, visual, and spatial information. It is concluded that the evolution toward multimodal architectures represents a paradigm shift in information retrieval, expanding the ability of artificial intelligence systems to understand, index, and retrieve knowledge embedded in complex documents.

Keywords: *Retrieval-Augmented Generation. Multimodal RAG. Document understanding. Vision-language models. Information retrieval.*

INTRODUÇÃO

A crescente evolução dos modelos de linguagem de grande escala (Large Language Models – LLMs) tem promovido transformações significativas na forma como sistemas computacionais processam, recuperam e geram informações. O avanço dessas arquiteturas possibilitou o desenvolvimento de aplicações capazes de realizar tarefas complexas de compreensão textual, síntese de conhecimento e geração de conteúdo em linguagem natural. Entretanto, apesar do elevado desempenho apresentado por esses modelos, sua dependência exclusiva do conhecimento armazenado durante o treinamento limita a atualização contínua das informações e favorece a ocorrência de alucinações, fenômeno caracterizado pela geração de respostas plausíveis, porém factualmente incorretas. Nesse contexto, surgiram estratégias voltadas à integração entre modelos generativos e fontes externas de conhecimento, buscando ampliar a confiabilidade e a rastreabilidade das informações produzidas pelos sistemas inteligentes (LEWIS et al., 2020).

Dentre essas estratégias, o Retrieval-Augmented Generation (RAG) consolidou-se como uma das abordagens mais relevantes para a combinação entre recuperação de informações e geração de linguagem natural. Proposto por Lewis et al. (2020), o modelo fundamenta-se na recuperação prévia de documentos relevantes em bases externas para subsidiar o processo de geração de respostas. Essa arquitetura permitiu reduzir limitações associadas à memória paramétrica dos modelos de linguagem, favorecendo maior precisão factual e capacidade de atualização do conhecimento. Desde então, o RAG passou a ocupar posição central em aplicações de busca inteligente, sistemas de perguntas e respostas, assistentes virtuais e mecanismos de apoio à tomada de decisão baseados em inteligência artificial. Contudo, as implementações iniciais foram concebidas

predominantemente para ambientes compostos por documentos textuais estruturados, nos quais a recuperação depende da extração e indexação de conteúdo em formato textual (LEWIS et al., 2020).

Embora eficazes em coleções essencialmente textuais, os sistemas tradicionais de RAG apresentam limitações importantes quando aplicados a documentos complexos e visualmente ricos, como relatórios corporativos, artigos científicos, apresentações, formulários, tabelas, gráficos e documentos digitalizados. Nesses cenários, a recuperação de informações geralmente depende de pipelines compostos por reconhecimento óptico de caracteres (OCR), segmentação de layout, extração textual e técnicas de chunking semântico. Tais processos podem introduzir perdas de informação, distorções contextuais e dificuldades na preservação das relações espaciais presentes nos documentos originais. Estudos recentes demonstram que elementos visuais, como tabelas, figuras, diagramas, tipografia e organização espacial, frequentemente carregam informações semânticas relevantes que não são adequadamente capturadas por abordagens baseadas exclusivamente em texto. Essas limitações motivaram o desenvolvimento de novos modelos voltados à compreensão multimodal de documentos e à preservação simultânea de informações textuais, visuais e estruturais (XU et al., 2021a; XU et al., 2021b; KIM et al., 2022).

Nesse contexto, modelos como LayoutLMv2, LayoutXLM e Donut representaram avanços significativos na compreensão documental multimodal ao integrar informações textuais, visuais e de layout em arquiteturas unificadas. Posteriormente, o desenvolvimento de modelos visão-linguagem especializados impulsionou uma nova geração de sistemas capazes de realizar indexação e recuperação diretamente a partir da representação visual dos documentos, reduzindo a dependência de etapas intermediárias de OCR e processamento textual. Trabalhos recentes, como ColPali, VisRAG, M3DocRAG e VDocRAG, demonstram que a utilização direta de representações visuais pode preservar melhor a semântica dos documentos, melhorar a recuperação de informações em conteúdos multimodais e ampliar a capacidade dos sistemas RAG em cenários envolvendo múltiplas páginas, múltiplos documentos e diferentes modalidades de evidência, incluindo textos, tabelas, gráficos e imagens (FAYSSE et al., 2025; YU et al., 2025; CHO et al., 2024; TANAKA et al., 2025).

Diante desse cenário, observa-se uma transição progressiva dos sistemas de recuperação baseados exclusivamente em texto para arquiteturas multimodais capazes de compreender documentos em sua estrutura original. Essa mudança representa não apenas uma evolução tecnológica dos mecanismos de recuperação de informações, mas também uma redefinição dos processos de extração semântica e indexação documental em ambientes complexos. Assim, o presente estudo tem como objetivo analisar a transição do RAG tradicional para o RAG multimodal, discutindo os principais desafios relacionados à extração semântica, preservação de contexto, indexação visual e recuperação de informações em documentos complexos, bem como os avanços recentes que vêm redefinindo o estado da arte na integração entre inteligência artificial generativa e compreensão documental multimodal.

METODOLOGIA

Trata-se de uma revisão narrativa da literatura desenvolvida com o objetivo de analisar a evolução dos sistemas Retrieval-Augmented Generation (RAG), desde as arquiteturas tradicionais baseadas em recuperação textual até os modelos multimodais voltados à compreensão e indexação de documentos complexos.

Foram selecionados artigos de referência publicados entre 2020 e 2025 em periódicos e conferências de alto impacto nas áreas de Inteligência Artificial, Processamento de Linguagem Natural, Visão Computacional e Compreensão Documental. A seleção dos estudos considerou sua relevância científica para a discussão sobre recuperação de informações, compreensão multimodal de documentos, modelos visão-linguagem, extração semântica, indexação visual e aplicações de RAG em documentos visualmente ricos.

A amostra final foi composta por oito estudos considerados fundamentais para a compreensão da evolução conceitual e tecnológica da área, incluindo os trabalhos de Lewis et al. (2020), responsáveis pela proposição da arquitetura Retrieval-Augmented Generation; Xu et al. (2021a) e Xu et al. (2021b), que contribuíram para o avanço da compreensão multimodal de documentos por meio dos modelos LayoutLMv2 e LayoutXLM; Kim et al. (2022), que introduziram uma abordagem OCR-free para compreensão documental com o modelo Donut; Cho et al. (2024), responsáveis pela proposta do M3DocRAG para recuperação multimodal em múltiplos documentos e páginas; Faysse et al. (2025), que apresentaram o modelo ColPali para indexação visual de documentos; Yu et al. (2025), autores do VisRAG; e Tanaka et al. (2025), que desenvolveram o framework VDocRAG para recuperação aumentada por geração em documentos visualmente ricos.

A análise dos estudos concentrou-se na identificação dos principais avanços relacionados à recuperação de informações baseada em documentos complexos, com ênfase na transição dos sistemas dependentes de OCR e indexação textual para arquiteturas capazes de explorar simultaneamente informações textuais, visuais e estruturais. Foram examinados aspectos relacionados aos mecanismos de extração semântica, preservação de layout, indexação multimodal, recuperação de informações, utilização de modelos visão-linguagem e integração entre recuperação e geração em ambientes documentais complexos.

Por fim, os estudos foram analisados de forma descritiva e interpretativa, buscando compreender as limitações das abordagens tradicionais de RAG, os fatores que impulsionaram o desenvolvimento das arquiteturas multimodais e as perspectivas futuras para sistemas de recuperação e geração capazes de operar sobre documentos visualmente ricos, compostos por múltiplas modalidades de informação.

RESULTADOS E DISCUSSÃO

3.1 FUNDAMENTOS DO RAG TRADICIONAL E LIMITAÇÕES DA RECUPERAÇÃO BASEADA EM TEXTO

O desenvolvimento dos modelos de linguagem de grande escala ampliou significativamente a capacidade dos sistemas de inteligência artificial em tarefas de compreensão e geração de linguagem natural. Entretanto, a dependência exclusiva do conhecimento armazenado nos parâmetros desses modelos evidenciou limitações relacionadas à atualização de informações, rastreabilidade das fontes utilizadas e ocorrência de alucinações. Com o objetivo de superar essas restrições, foi proposta a arquitetura Retrieval-Augmented Generation (RAG), baseada na integração entre modelos generativos e mecanismos externos de recuperação de informações. Nessa abordagem, o sistema consulta uma base documental externa antes da geração da resposta, combinando memória paramétrica e memória não paramétrica em um único fluxo de processamento. Tal estratégia permitiu aumentar a precisão factual das

respostas e tornou possível a incorporação dinâmica de novos conhecimentos sem a necessidade de reentreinamento completo do modelo (LEWIS et al., 2020).

A arquitetura original do RAG foi concebida para operar predominantemente sobre coleções textuais estruturadas, utilizando técnicas de recuperação baseadas em embeddings semânticos e índices vetoriais. Nesse cenário, documentos são convertidos em representações textuais e segmentados em unidades menores para facilitar a recuperação de informações relevantes durante a etapa de busca. Embora essa estratégia tenha demonstrado resultados expressivos em tarefas de perguntas e respostas, recuperação de conhecimento e geração de conteúdo, sua eficácia depende diretamente da qualidade das informações extraídas e indexadas previamente. Dessa forma, o desempenho do sistema pode ser comprometido quando os documentos apresentam estruturas complexas que extrapolam o conteúdo textual puro (LEWIS et al., 2020).

As limitações tornam-se particularmente evidentes em documentos visualmente ricos, como relatórios corporativos, artigos científicos, apresentações, formulários, tabelas e gráficos. Nesses casos, os sistemas tradicionais de recuperação dependem de etapas intermediárias de reconhecimento óptico de caracteres (OCR), identificação de layout e segmentação textual. A simples extração do texto frequentemente não é suficiente para representar adequadamente o significado dos documentos, uma vez que informações relevantes também estão associadas à organização espacial dos elementos, às relações entre blocos de conteúdo e à disposição visual das informações na página. Documentos visualmente ricos exigem a integração simultânea de informações textuais, visuais e estruturais para que a compreensão ocorra de forma consistente em diferentes contextos e idiomas (XU et al., 2021a; XU et al., 2021b).

Além das limitações relacionadas ao layout, a dependência de OCR introduz desafios adicionais decorrentes de erros de reconhecimento, perda de contexto visual e propagação de falhas para etapas posteriores do processamento. Pipelines tradicionais baseados em OCR tendem a apresentar elevado custo computacional, baixa flexibilidade para diferentes tipos documentais e vulnerabilidade a erros acumulativos durante a extração de informações. Como consequência, tabelas, gráficos, diagramas e outros componentes visuais podem ter seu conteúdo parcialmente perdido ou distorcido durante a conversão para texto. Essas limitações passaram a impulsionar o desenvolvimento de abordagens multimodais capazes de preservar a representação original dos documentos, constituindo o principal fator que motivou a transição dos sistemas de recuperação exclusivamente textuais para arquiteturas multimodais voltadas à compreensão documental mais abrangente e semanticamente fiel (KIM et al., 2022).

3.2 EVOLUÇÃO DA COMPREENSÃO MULTIMODAL DE DOCUMENTOS

As limitações observadas nos sistemas tradicionais de recuperação e compreensão documental impulsionaram o desenvolvimento de modelos capazes de processar simultaneamente diferentes modalidades de informação presentes em documentos digitais. Nesse contexto, surgiu o conceito de compreensão multimodal de documentos, fundamentado na integração de informações textuais, visuais e estruturais em uma única arquitetura de processamento. Diferentemente das abordagens convencionais, que tratam o texto como principal fonte de informação, os modelos multimodais reconhecem que elementos como posicionamento espacial, formatação, imagens, tabelas e gráficos desempenham papel fundamental na construção do significado dos documentos. Essa perspectiva tornou-se particularmente relevante diante da crescente digitalização de documentos corporativos, acadêmicos e governamentais, caracterizados por layouts cada vez mais complexos e heterogêneos (XU et al., 2021a).

Entre os principais avanços nessa área destaca-se o desenvolvimento do LayoutLMv2, modelo que ampliou significativamente a capacidade de compreensão documental ao integrar texto, layout e imagem em uma arquitetura multimodal unificada. O modelo introduziu mecanismos específicos para capturar relações espaciais entre diferentes elementos presentes na página, permitindo que a compreensão do conteúdo considerasse não apenas as palavras extraídas, mas também sua posição relativa e contexto visual. Os resultados apresentados pelos autores demonstraram ganhos expressivos em tarefas de classificação documental, extração de informações e resposta a perguntas em documentos visualmente ricos, evidenciando a importância da modelagem conjunta das diferentes modalidades de informação (XU et al., 2021a).

A evolução desse paradigma também motivou o desenvolvimento do LayoutXLM, concebido para expandir as capacidades da compreensão documental multimodal para ambientes multilíngues. Além de incorporar informações textuais, visuais e espaciais, o modelo foi treinado para lidar com documentos produzidos em diferentes idiomas, permitindo maior generalização em cenários globais. Os resultados demonstraram que a combinação entre múltiplas modalidades e múltiplas línguas contribui para melhorar a robustez dos sistemas documentais, especialmente em aplicações envolvendo formulários, documentos administrativos e processos corporativos internacionais. Esses avanços reforçaram a ideia de que a compreensão documental não pode ser reduzida apenas à análise textual, exigindo abordagens capazes de capturar a complexidade estrutural dos documentos modernos (XU et al., 2021b).

Outro marco relevante nesse processo foi a introdução do modelo Donut, que propôs uma ruptura com os pipelines tradicionais dependentes de OCR. Diferentemente das abordagens anteriores, o Donut realiza a compreensão documental diretamente a partir da imagem do documento, eliminando etapas intermediárias de reconhecimento textual. Essa estratégia reduz a propagação de erros associada ao OCR e permite preservar informações visuais frequentemente perdidas durante os processos de extração textual. Os resultados apresentados pelos autores demonstraram que modelos OCR-free podem alcançar desempenho equivalente ou superior às abordagens convencionais, ao mesmo tempo em que simplificam o pipeline de processamento documental. Esse avanço consolidou a compreensão multimodal como uma das principais direções de pesquisa na área, estabelecendo as bases para o desenvolvimento de sistemas de recuperação capazes de operar diretamente sobre representações visuais dos documentos (KIM et al., 2022).

3.3 INDEXAÇÃO VISUAL E RECUPERAÇÃO MULTIMODAL EM DOCUMENTOS COMPLEXOS

Os avanços obtidos pelos modelos de compreensão documental multimodal evidenciaram que grande parte das limitações dos sistemas tradicionais de recuperação estava associada não apenas à etapa de interpretação dos documentos, mas também aos mecanismos de indexação utilizados para armazenar e recuperar informações. Historicamente, os sistemas de recuperação documental foram construídos a partir da conversão dos documentos em representações textuais, geralmente obtidas por meio de OCR, extração de texto nativo ou segmentação em blocos semânticos. Embora eficientes em documentos predominantemente textuais, essas estratégias frequentemente negligenciam informações presentes em elementos visuais, como gráficos, tabelas, diagramas, fórmulas matemáticas e características espaciais do layout. Como consequência, parte relevante do significado dos documentos pode ser perdida antes mesmo da etapa de recuperação da informação (FAYSSE et al., 2025).

Nesse cenário, o surgimento dos modelos visão-linguagem (Vision-Language Models – VLMs) abriu novas possibilidades para a indexação documental. Diferentemente dos métodos convencionais, esses modelos são capazes de processar simultaneamente informações visuais e textuais, gerando representações vetoriais que preservam aspectos semânticos e estruturais dos documentos originais. A partir dessa perspectiva, Faysse et al. (2025) desenvolveram o ColPali, modelo projetado para realizar indexação diretamente a partir das imagens das páginas documentais. Em vez de depender de pipelines complexos envolvendo OCR, extração textual, segmentação e descrição de imagens, o ColPali produz embeddings multimodais capazes de representar o conteúdo integral da página, preservando relações visuais frequentemente descartadas nos métodos tradicionais. Os resultados apresentados demonstraram desempenho superior em tarefas de recuperação documental visualmente rica, além de maior simplicidade operacional e redução da complexidade dos processos de indexação (FAYSSE et al., 2025).

Uma abordagem semelhante foi proposta pelo VisRAG, que amplia os princípios do Retrieval-Augmented Generation para ambientes compostos por documentos multimodais. Os autores argumentam que a conversão prévia dos documentos em texto pode introduzir perdas significativas de informação, comprometendo tanto a recuperação quanto a geração de respostas. Para contornar esse problema, o VisRAG utiliza modelos visão-linguagem tanto na etapa de recuperação quanto na etapa de geração, permitindo que os documentos sejam processados diretamente em seu formato visual original. Essa estratégia favorece a preservação de informações relacionadas ao layout, imagens, tabelas e demais componentes gráficos, resultando em melhorias expressivas no desempenho global do sistema quando comparado às arquiteturas tradicionais baseadas exclusivamente em texto (YU et al., 2025).

A relevância dessas abordagens torna-se ainda mais evidente diante da crescente utilização de documentos complexos em ambientes corporativos, acadêmicos e científicos. Relatórios financeiros, artigos científicos, apresentações institucionais e documentos regulatórios frequentemente apresentam informações distribuídas entre diferentes modalidades, exigindo mecanismos de recuperação capazes de compreender simultaneamente texto, imagem e estrutura visual. Nesse contexto, a indexação visual e a recuperação multimodal representam uma mudança de paradigma na área de recuperação de informações, substituindo gradualmente a lógica centrada em texto por modelos capazes de explorar a riqueza semântica presente nos documentos em sua forma original. Tal transformação constitui um dos principais fundamentos tecnológicos que sustentam a evolução do RAG tradicional para o RAG multimodal (FAYSSE et al., 2025; YU et al., 2025).

3.4 DESAFIOS ATUAIS E PERSPECTIVAS FUTURAS DO RAG MULTIMODAL

Embora os avanços recentes tenham demonstrado o potencial das abordagens multimodais para superar limitações históricas dos sistemas de recuperação baseados exclusivamente em texto, diversos desafios ainda permanecem para a consolidação do RAG multimodal em aplicações de larga escala. Um dos principais obstáculos está relacionado ao elevado custo computacional necessário para processar documentos em sua representação visual completa. Diferentemente dos sistemas textuais tradicionais, que operam sobre sequências de caracteres relativamente compactas, os modelos multimodais precisam processar simultaneamente imagens, estruturas espaciais e conteúdo textual, demandando maior capacidade de armazenamento, processamento e gerenciamento de índices vetoriais complexos. À medida que o volume documental

cresce, torna-se necessário desenvolver estratégias capazes de equilibrar desempenho, precisão e eficiência computacional (YU et al., 2025; TANAKA et al., 2025).

Outro desafio relevante refere-se à escalabilidade dos mecanismos de recuperação em ambientes compostos por múltiplos documentos e múltiplas páginas. Em cenários reais, informações necessárias para responder a uma consulta frequentemente encontram-se distribuídas em diferentes documentos ou espalhadas ao longo de extensos conjuntos documentais. Nesse contexto, Cho et al. (2024) destacam que sistemas convencionais de Document Visual Question Answering apresentam limitações significativas quando confrontados com tarefas que exigem raciocínio sobre múltiplas fontes de informação. O desenvolvimento do M3DocRAG representa uma tentativa de solucionar esse problema por meio da integração entre recuperação multimodal e modelos visão-linguagem, permitindo localizar evidências relevantes em diferentes documentos sem comprometer a preservação das informações visuais. Ainda assim, a recuperação eficiente em grandes coleções documentais permanece como uma das principais fronteiras de pesquisa da área (CHO et al., 2024).

Além das questões relacionadas à escalabilidade, observa-se a necessidade de estabelecer métricas e benchmarks mais abrangentes para avaliação de sistemas multimodais. Grande parte dos conjuntos de dados historicamente utilizados para treinamento e validação foi concebida para tarefas específicas de OCR, classificação documental ou recuperação textual, não refletindo adequadamente a complexidade dos cenários multimodais contemporâneos. Trabalhos recentes, como VDocRAG e M3DocRAG, evidenciam a importância de ambientes de avaliação capazes de considerar simultaneamente recuperação, compreensão visual, integração multimodal e geração de respostas. A ausência de padrões amplamente consolidados dificulta comparações diretas entre diferentes arquiteturas e limita a mensuração precisa dos avanços alcançados pelos modelos mais recentes (TANAKA et al., 2025; CHO et al., 2024).

As perspectivas futuras apontam para o desenvolvimento de sistemas cada vez mais integrados, capazes de compreender documentos em sua totalidade sem depender de etapas intermediárias de conversão para texto. A evolução dos modelos visão-linguagem, associada ao aperfeiçoamento dos mecanismos de indexação visual e recuperação multimodal, sugere que futuras arquiteturas de RAG poderão operar diretamente sobre documentos complexos preservando integralmente informações textuais, visuais e estruturais. Nesse cenário, espera-se a consolidação de agentes inteligentes especializados em ambientes documentais, aptos a navegar por grandes repositórios de informações, realizar análises contextuais avançadas e gerar respostas fundamentadas em múltiplas evidências. Dessa forma, o RAG multimodal tende a representar não apenas uma evolução incremental dos sistemas de recuperação de informações, mas uma mudança estrutural na forma como documentos digitais serão compreendidos, indexados e explorados pelas futuras aplicações de inteligência artificial (FAYSSE et al., 2025; YU et al., 2025; TANAKA et al., 2025).

CONCLUSÃO

A evolução dos modelos de linguagem e dos sistemas de recuperação de informações impulsionou o desenvolvimento de arquiteturas cada vez mais sofisticadas para lidar com tarefas intensivas em conhecimento. Nesse contexto, o Retrieval-Augmented Generation (RAG) consolidou-se como uma estratégia eficaz para integrar modelos generativos a fontes externas de informação, reduzindo limitações relacionadas à memória paramétrica e contribuindo para a melhoria da precisão factual das respostas. Entretanto, a presente revisão evidenciou que os modelos tradicionais de RAG,

fundamentados predominantemente em recuperação textual, apresentam limitações significativas quando aplicados a documentos complexos compostos por múltiplas modalidades de informação.

Os estudos analisados demonstraram que a dependência de OCR, segmentação textual e técnicas convencionais de indexação pode resultar em perdas semânticas importantes, especialmente em documentos que contêm tabelas, gráficos, diagramas, imagens e estruturas espaciais relevantes para a compreensão do conteúdo. Nesse cenário, modelos como LayoutLMv2, LayoutXLM e Donut contribuíram para ampliar a capacidade de compreensão documental ao integrar informações textuais, visuais e estruturais em arquiteturas multimodais, estabelecendo as bases para uma nova geração de sistemas de recuperação documental.

A revisão também evidenciou que abordagens mais recentes, representadas por ColPali, VisRAG, M3DocRAG e VDocRAG, vêm promovendo uma mudança de paradigma ao possibilitar a indexação e a recuperação diretamente a partir das representações visuais dos documentos. Esses modelos demonstraram maior capacidade de preservar informações semânticas complexas, reduzir perdas decorrentes da conversão para texto e melhorar o desempenho em tarefas envolvendo documentos visualmente ricos, múltiplas páginas e múltiplas fontes documentais.

Dessa forma, conclui-se que a transição do RAG tradicional para o RAG multimodal representa uma evolução necessária para atender às demandas contemporâneas de recuperação e compreensão documental. Embora desafios relacionados à escalabilidade, eficiência computacional e padronização de métricas de avaliação ainda persistam, as evidências analisadas indicam que as arquiteturas multimodais tendem a desempenhar papel central no futuro dos sistemas de recuperação de informações. A capacidade de integrar simultaneamente informações textuais, visuais e estruturais amplia significativamente o potencial das aplicações baseadas em inteligência artificial, contribuindo para o desenvolvimento de sistemas mais precisos, contextualizados e alinhados à complexidade dos documentos digitais modernos.

REFERÊNCIAS

CHO, Jaemin et al. M3DOCRAG: Multi-modal Retrieval is What You Need for Multi-page Multi-document Understanding. arXiv:2411.04952, 2024. Disponível em: https://arxiv.org/abs/2411.04952.

FAYSSE, Manuel et al. ColPali: Efficient Document Retrieval with Vision Language Models. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), 2025, Singapore. Singapore: ICLR, 2025. Disponível em: https://openreview.net/forum?id=ogjBpZ8uSi.

KIM, Geewook et al. OCR-free Document Understanding Transformer. In: EUROPEAN CONFERENCE ON COMPUTER VISION (ECCV), 17., 2022, Tel Aviv. Cham: Springer, 2022. p. 498-517. DOI: 10.1007/978-3-031-19815-1_29.

LEWIS, Patrick et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS), 34., 2020, Vancouver. Vancouver: NeurIPS, 2020. p. 9459-9474. Disponível em: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

TANAKA, Ryota et al. VDocRAG: Retrieval-Augmented Generation over Visually-Rich Documents. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2025, Nashville. Nashville: IEEE/CVF, 2025. Disponível em:

https://openaccess.thecvf.com/content/CVPR2025/html/Tanaka_VDocRAG_Retrieval-Augmented_Generation_over_Visually-Rich_Documents_CVPR_2025_paper.html.

XU, Yang et al. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL), 59., 2021, Bangkok. Bangkok: Association for Computational Linguistics, 2021. p. 2579-2591. Disponível em: https://aclanthology.org/2021.acl-long.201/.

XU, Yiheng et al. LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding. arXiv:2104.08836, 2021. Disponível em: https://arxiv.org/abs/2104.08836.

YU, Shi et al. VisRAG: Vision-based Retrieval-Augmented Generation on Multi-Modality Documents. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), 2025, Singapore. Singapore: ICLR, 2025. Disponível em: https://openreview.net/forum?id=zG459X3Xge.