Arquitetura Assíncrona e Escalável para Recuperação de Informação Multimodal com RAG

João P. A. F. Matos[1], **André F. dos S. Caraíba**[1] , **Paulo V. Dos Santos**[1], **Maria C. S. Barreto**[2], **Guilherme C. Dutra** [1], **Sávio S. T. de Oliveira** [1]

[1] Instituto de informática – Universidade Federal de Goiás
Goiania, GO.

[2]ITA – Instituto Tecnológico de Aeronáutica
São José dos Campos, SP.

joao.formiga@discente.ufg.br,andrecaraiba@discente.ufg.br

paulodos@egresso.ufg.br, guilhermecorreiadutra@gmail.com

maria@ccm-ita.org.br

Abstract. *The growing demand for vehicle maintenance has highlighted the challenge of accessing technical information in complex repair manuals. This paper proposes a multimodal Retrieval-Augmented Generation (RAG) architecture to optimize fault diagnosis in the automotive sector. By integrating Vision-Language Models (VLMs), OCR, and vector databases, the system extracts and indexes diagrams, tables, and texts. An event-driven pipeline ensures scalability and resilience, enabling semantic search and intelligent agent support. Initial tests on table extraction validate the approach's effectiveness. Future work includes real-world evaluations and integration with technical assistance systems.*

Resumo. *A expansão da frota veicular intensificou a necessidade de diagnósticos precisos em sistemas automotivos. Este artigo propõe uma arquitetura RAG multimodal para indexar manuais técnicos, utilizando VLMs, OCR e bancos vetoriais. A solução é baseia-se em arquitetura orientada a eventos, garantindo escalabilidade e resiliência no processamento de documentos complexos. O sistema extrai e organiza metadados, tabelas e textos, viabilizando consultas semânticas por agentes inteligentes. Testes iniciais demonstram eficácia na preservação do contexto multimodal. Como próximos passos, propõe-se avaliação quantitativa e integração com fluxos de assistência técnica.*

1. Introdução

A crescente frota veicular tem impulsionado a demanda por serviços de manutenção e reparo no setor automotivo. Com mercado avaliado em 468 bilhões de dólares em 2024 e com projeções de crescimento, a manutenção de veículos, especialmente a identificação de falhas em sistemas complexos, tornou-se o desafio. A precisão do diagnóstico depende diretamente da busca por informações técnicas em manuais e bases de dados [Denton 2020], o que sublinha a importância da capacitação e do acesso a recursos informacionais para a eficiência e precisão do reparo.

Nesse contexto, a Inteligência Artificial Generativa e os Modelos de Linguagem Visual (VLMs) [Zhang et al. 2021] surgem como solução promissora para otimizar o

acesso a informações técnicas. Diferentemente dos modelos de linguagem tradicionais, os VLMs podem interpretar e extrair dados de textos e imagens [Zhang et al. 2024, Wang et al. 2025], automatizando a análise de diagramas e textos técnicos. Essa capacidade é crucial para o setor automotivo, onde a complexidade dos veículos modernos, com maior quantidade de eletrônicos embarcados [Abelein et al. 2012], exige acesso rápido e preciso a manuais técnicos. A aplicação desses modelos oferece potencial significativo para agilizar a identificação de falhas e reduzir custos de reparo.

Entretanto, a aplicação de modelos de IA para otimizar a busca por informações técnicas não é trivial sem que haja preparação para o domínio específico, ou dos dados ou dos modelos [Soudani et al. 2024]. Os manuais de reparo veicular representam desafio significativo para o processamento automático, devido às suas características intrínsecas ou multimodais: são documentos de grande volume, com estruturas heterogêneas que incluem diagramas, tabelas, figuras, textos e gráficos. A simples ingestão desses dados por modelos simples de processamento de linguagem natural ou por modelos simples de LLM, pode resultar em alucinações ou respostas imprecisas [Ji et al. 2023]. Para superar essas limitações, a técnica de Geração Aumentada por Recuperação (RAG) [Lewis et al. 2020] surge como a solução. O RAG aprimora as capacidades de modelos de linguagem, permitindo a recuperação de documentos relevantes a partir de base de dados indexada. Esses documentos são então utilizados como contexto para fundamentar as respostas dos modelos, garantindo maior precisão e confiabilidade.

A implementação da arquitetura baseada em RAG para manuais automotivos requer infraestrutura robusta, capaz de processar e indexar grandes volumes de documentos, extraindo seu conteúdo e preservando o contexto por meio de técnicas como o Reconhecimento Óptico de Caracteres (OCR). O desafio reside na criação de modelos de vetores que compreendam as estruturas complexas das páginas para a indexação precisa em bases de dados vetoriais, como demonstrado por modelos como o Colpali [Faysse et al. 2024]. Diante disso, este trabalho propõe a arquitetura orientada a eventos para o processamento e indexação de manuais automotivos, utilizando modelos multimodais para extrair metadados, tabelas e textos de cada página. O objetivo é indexar esse conteúdo em bancos de dados vetoriais, otimizando o acesso e a recuperação de informações essenciais para diagnóstico e reparo veicular por meio de agentes inteligentes.

2. Métodos

2.1. Arquitetura Proposta

Para lidar com o desafio de processar grandes volumes de documentos multimodais de forma eficiente, propõe-se arquitetura desacoplada, resiliente e horizontalmente escalável. A natureza dos dados é composta por documentos desestruturados, impondo o desafio de *big data* que exige rastreabilidade e confiabilidade desde o início. Falhas na indexação de uma única página pode comprometer a integridade da base de conhecimento, tornando a robustez do *pipeline* um dos requisitos fundamentais para que o sistema de RAG seja confiável.

Atender aos requisitos estabelecidos, necessita que a solução seja fundamentada aos conceitos da Arquitetura Orientada a Eventos (EDA - *Event-Driven Architecture*). Neste paradigma, módulos independentes comunicam-se através de *brokers* de mensagens, utilizando protocolos bem definidos com metadados. Conforme ilustrado na

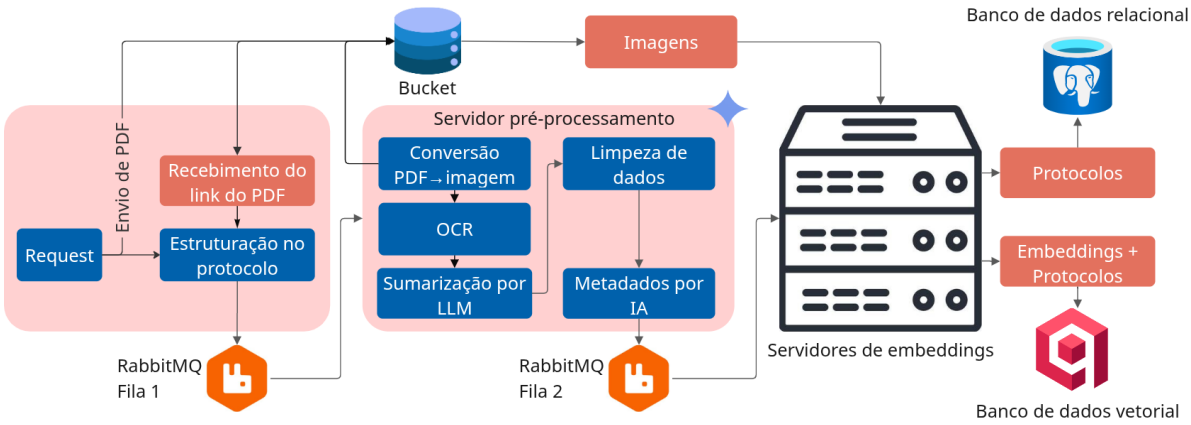

Figure 1. Arquitetura Proposta

Figura 1, a abordagem otimiza o fluxo evitando a transmissão direta de dados como: imagens, vídeos ou qualquer outra mídia digital, garantindo assim, a agilidade e principalmente a entrega das informações a serem processadas.

A implementação se materializa nos seguintes componentes principais:

- **Orquestração e Filas (RabbitMQ):** Atua como o núcleo da comunicação assíncrona, gerenciando o fluxo de tarefas de forma desacoplada. Garante que falhas em um componente não interrompam o sistema e não resultem em perda de dados, assegurando a resiliência do processo.
- **Pré-processamento e Anotação:** Módulo responsável pela inteligência multimodal. Utiliza Modelos de Linguagem de Visão (VLMs) para interpretar o conteúdo visual (imagens, diagramas) para extrair e contextualizar tabelas e textos em cada página. Utilizando o modelo gemini-2.5-pro com temperatura ajustada em $0,1$.
- **Geração de Embeddings e Indexação:**Recebe as anotações multimodais e as converte em vetores dinâmicos. Utiliza o modelo ColQwen2, com 2 bilhões de parâmetros, para gerar embeddings de cada imagem, empregando uma placa de vídeo RTX 2070 Max-Q com 8 GB de memória — o tempo de processamento varia entre 20 e 30 segundos. Esses embeddings são posteriormente armazenados e indexados em bancos vetoriais, como o Qdrant, viabilizando a busca semântica e propiciando a eficiente recuperação das informações contidas nessa base de dados de manuais de veículos.

Com a operação da esteira de processamento, conforme descrito na Figura 1, é possível validar a proposta em cenários controlados. A escolha recai sobre a extração e organização de tabelas, que são elementos críticos em documentos técnicos automotivos, a fim de aferir a capacidade do sistema em preservar tanto o contexto semântico do documento, quanto a estrutura visual necessária para consultas posteriores.

3. Conclusão

Este trabalho aborda o desafio de estruturar e realizar consultas em grandes volumes de documentos técnicos multimodais, caracterizados por alta heterogeneidade e ausência de padronização, como os manuais de veículos disponíveis nos sites das montadoras, que

compõem a base de dados desta abordagem, fundamentandando-se em um arquitetura RAG multimodal orientada a eventos.

A metodologia proposta integra etapas de OCR, anotação multimodal LLMs e, posteriormente, geração de embeddings com o modelo de VLM ColPali, armazenando os resultados do processo de geração de vetores no banco vetorial (Qdrant). Permitindo consumo imediato para agentes em tarefas *question answering* técnico.

Para validar a metodologia, foi realizada uma análise exploratória inicial como prova de conceito, visando justificar a expansão futura do dataset. Nessa análise 16 arquivos foram usados (totalizando 41 páginas de documentos diversos do âmbito automotivo), nas quais foram identificadas 31 páginas com tabelas. A arquitetura proposta foi capaz de extrair a informação tabular contida no teste. Adicionalmente, foi identificados grupos de erros como problema de tradução, de formatação, de conteúdo e de não identificação de tabela. Para os próximos passos, propõe-se a avaliação quantitativa em cenários reais de recuperação, com métricas de precisão e revocação

References

Abelein, U., Lochner, H., Hahn, D., and Straube, S. (2012). Complexity, quality and robustness-the challenges of tomorrow's automotive electronics. In *2012 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 870–871. IEEE.

Denton, T. (2020). *Advanced automotive fault diagnosis: automotive technology: vehicle maintenance and repair*. Routledge.

Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., and Colombo, P. (2024). Colpali: Efficient document retrieval with vision language models.

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., and Fung, P. (2023). Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Soudani, H., Kanoulas, E., and Hasibi, F. (2024). Fine tuning vs. retrieval augmented generation for less popular knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 12–22.

Wang, Q., Ding, R., Chen, Z., Wu, W., Wang, S., Xie, P., and Zhao, F. (2025). Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*.

Zhang, J., Huang, J., Jin, S., and Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.