Algoritmo de Geração Aumentada de Recuperação como Ferramenta de Apoio à Manutenção Aeronáutica

Samir Earvin Miranda Lima[1], Danilo Figueiredo[2]
[1]Parque de Material Aeronáutico do Galeão (PAMA-GL), Rio de Janeiro/RJ – Brasil
[2]Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos/SP – Brasil

***Resumo* – Este artigo apresenta um algoritmo conversacional baseado em geração aumentada por recuperação (*retrieval-augmented generation* - RAG) para apoiar consultas a manuais técnicos de manutenção aeronáutica. A proposta combina extração de documentos, segmentação textual, *embeddings* multilíngues, banco vetorial local, reordenação semântica e geração de respostas por modelo de linguagem executado em ambiente controlado. O protótipo foi avaliado em um experimento com dez militares especialistas na aeronave KC-390, comparando o tempo de localização de respostas em manuais técnicos com e sem apoio da ferramenta. Os resultados indicaram redução média de 60 s por consulta, equivalente a ganho aproximado de 35,4%, além de eliminação dos casos de *timeout* observados na pesquisa exclusiva com o manual. O estudo sugere que sistemas RAG locais podem aumentar a eficiência da consulta técnica, desde que empregados como apoio e não como substituto da documentação aprovada.**

***Palavras-Chave* – Manutenção aeronáutica, inteligência artificial, RAG.**

I. INTRODUÇÃO

A modernização de vetores militares elevou a complexidade dos sistemas aeronáuticos e, por consequência, a exigência sobre os processos de suporte logístico e manutenção. No caso da Força Aérea Brasileira (FAB), iniciativas recentes voltadas à aplicação de técnicas computacionais em dados de falhas indicam um movimento de aproximação entre operação, logística e métodos de inteligência artificial [1]. Esse cenário é particularmente relevante em aeronaves multimissão como o KC-390, cujo suporte depende de consulta frequente a documentação técnica extensa e especializada.

A manutenção aeronáutica exige aderência estrita a manuais aprovados, rastreabilidade da informação e interpretação correta de procedimentos. Ainda assim, a localização de uma informação pontual pode consumir tempo significativo, sobretudo quando a resposta está distribuída entre capítulos, sistemas ou tarefas distintas. Em ambiente operacional, esse tempo impacta a fluidez do trabalho, a alocação de técnicos e, em última instância, a disponibilidade da aeronave.

Modelos de linguagem de grande escala oferecem novas possibilidades de interação com bases textuais, mas sua aplicação direta em domínios críticos é limitada pela possibilidade de respostas imprecisas, alucinações e ausência de lastro documental. A geração aumentada por recuperação (*retrieval-augmented generation* - RAG) busca mitigar esse problema ao combinar recuperação semântica em um corpus controlado com geração de linguagem natural, permitindo que a resposta seja produzida com base em trechos previamente recuperados [2], [3].

S. Lima, samirseml@fab.mil.br; D. Figueiredo, danilodf@ita.br.

Este artigo tem por objetivo apresentar e avaliar um algoritmo conversacional baseado em RAG para apoiar consultas a manuais técnicos de manutenção da aeronave KC-390. O estudo aplica essa arquitetura em ambiente local e controlado, utilizando documentação técnica da aeronave, e avalia experimentalmente seu efeito sobre o tempo de localização de informações por militares especialistas em sua manutenção.

II. FUNDAMENTAÇÃO E ARQUITETURA PROPOSTA

A. Suporte integrado, dados técnicos e recursos computacionais

No enfoque de engenharia logística, recursos computacionais, dados técnicos e treinamento não são elementos isolados. A literatura de suporte ao produto reconhece que software, bancos de dados, documentação técnica e meios de apoio à manutenção compõem a sustentação de sistemas complexos ao longo do ciclo de vida [4]. O conceito de *Integrated Product Support* (IPS) consolidou essa visão ao organizar elementos de suporte, entre eles *Technical Data, Training and Training Support e Computer Resources* [5], [6].

Sob essa perspectiva, uma ferramenta RAG aplicada a manuais de manutenção pode ser enquadrada primariamente como recurso computacional de apoio ao mantenedor. Ao mesmo tempo, ela atua sobre dados técnicos e pode auxiliar a capacitação informal, pois permite ao usuário formular perguntas em linguagem natural e receber uma orientação inicial contextualizada. Essa função, contudo, não altera a primazia da publicação técnica aprovada como fonte normativa para a execução da manutenção.

B. Geração aumentada por recuperação

A evolução dos modelos de linguagem está associada à arquitetura *Transformer*, baseada em mecanismos de atenção capazes de relacionar termos em diferentes posições de uma sequência textual [7]. Para recuperação semântica, modelos derivados de representações contextualizadas, como BERT, permitem converter trechos de texto em vetores que preservam relações de significado [8]. Em sistemas RAG, esses vetores são armazenados em bancos vetoriais e comparados à consulta do usuário para selecionar os trechos mais relevantes.

O conceito de RAG foi proposto como a combinação entre um mecanismo de recuperação e um modelo gerador, de modo a produzir respostas fundamentadas em conhecimento externo ao modelo [2]. Revisões recentes ressaltam que a abordagem é útil quando o conhecimento exigido é especializado, privado ou sujeito a atualização constante [3]. Em ambientes militares, a possibilidade de executar o pipeline localmente é especialmente relevante, pois reduz a exposição de manuais técnicos e dados sensíveis a serviços de terceiros [9].

Além de reduzir a dependência do conhecimento parametrizado do modelo, o RAG permite atualização incremental da base documental. Em vez de retreinar o modelo gerador sempre que uma publicação técnica é atualizada, o sistema pode reprocessar os documentos, reconstruir ou complementar o índice vetorial e preservar o mesmo mecanismo conversacional. Essa característica é importante em manutenção aeronáutica, na qual revisões, boletins e instruções técnicas podem alterar procedimentos específicos.

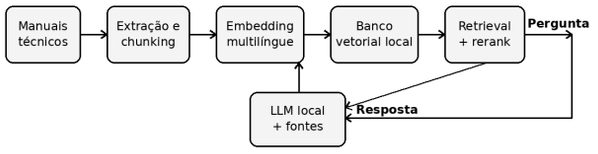


Fig. 1. Fluxo lógico do algoritmo RAG proposto.
Fonte: elaboração própria.

C. Pipeline implementado

A arquitetura proposta, sintetizada na Fig. 1, foi estruturada como um pipeline modular. O usuário formula a pergunta, o sistema recupera inicialmente os 30 (trinta) trechos com maior similaridade semântica, reordena os candidatos por meio do modelo de *rerank* e envia ao modelo gerador apenas os 3 (três) documentos mais relevantes. Ao final, uma etapa de pós-processamento acrescenta a indicação das fontes consultadas, reforçando a necessidade de validação no manual técnico.

O protótipo utilizou capítulos do AMM Part II; segmentação em *chunks* de 1200 caracteres com *overlap* de 120 caracteres; modelo intfloat/multilingual-e5-base para *embeddings*; Chroma como banco vetorial local; *rerank* com BAAI/bge-reranker-v2-m3; e Llama 3.1 8B, via Ollama, como modelo gerador. A configuração buscou equilibrar desempenho computacional, capacidade multilíngue e possibilidade de execução em ambiente controlado.

Os parâmetros de segmentação e recuperação foram definidos empiricamente a partir de testes preliminares com o corpus utilizado. O tamanho de 1.200 caracteres, com *overlap* de 120, buscou preservar contexto suficiente dos procedimentos sem aumentar excessivamente a redundância entre segmentos. De forma análoga, a recuperação inicial de 30 candidatos foi adotada buscando ampliar a cobertura da busca antes da etapa de *rerank*, mantendo o volume de documentos submetidos à reordenação em nível computacionalmente adequado.

A Tabela I sintetiza os principais componentes, ferramentas e parâmetros adotados na implementação do pipeline RAG, permitindo uma visão consolidada da arquitetura utilizada neste estudo.

TABELA I. SÍNTESE DA ARQUITETURA DO PIPELINE

Componente	***Ferramenta/modelo***	***Parâmetro principal***
Segmentação	Chunking	1.200 caracteres; overlap 120
Embedding	multilingual-e5-base	Multilíngue
Banco vetorial	Chroma	Local; similaridade de cosseno, 30 recuperações
Rerank	bge-reranker-v2-m3	Reordenação semântica, top 3
Geração	Llama 3.1 8B / Ollama	Execução local
Pós-processamento	Rotina do protótipo	Indicação das fontes

O modelo gerador recebeu instruções para responder em português, preservar terminologia técnica e evitar extrapolações não sustentadas pelos documentos recuperados. Essa estratégia não elimina completamente o risco de erro, mas reduz a probabilidade de respostas desconectadas do corpus e facilita a auditoria humana, uma vez que a saída é acompanhada pelas fontes documentais utilizadas.

III. MATERIAL E MÉTODO

A. Corpus e tratamento dos dados

A pesquisa utilizou como corpus capítulos selecionados do *Aircraft Maintenance Manual* (AMM) Part II - *Maintenance Practices and Procedures* da aeronave KC-390. Foram empregados capítulos referentes a *servicing, electrical power, hydraulic power, navigation, cargo and accessory compartments, doors, powerplant e surveillance*. A seleção buscou reunir sistemas com relevância operacional, variedade de especialidades e riqueza suficiente de detalhes técnicos para a avaliação do algoritmo.

Os arquivos foram carregados com extração de texto e metadados por página. Em seguida, o conteúdo foi dividido em *chunks*, vetorizado pelo modelo de *embedding* e persistido no banco vetorial local. A comparação entre consulta e trechos armazenados foi baseada em similaridade semântica; em particular, a similaridade de cosseno é adequada por privilegiar a direção dos vetores e reduzir o efeito da magnitude do texto comparado [10].

Após a recuperação inicial, os documentos candidatos foram reavaliados por um modelo de *rerank*. Essa etapa aumenta a precisão ao comparar diretamente pares formados pela pergunta e por cada trecho recuperado, embora exija maior custo computacional [11]. Como o rerank opera apenas sobre uma lista já filtrada, preserva-se uma relação aceitável entre desempenho e qualidade da resposta.

B. Delineamento do teste de campo

A avaliação foi conduzida com dez militares especialistas em manutenção da aeronave KC-390. Os participantes foram distribuídos em dois grupos, A e B, para reduzir o efeito de diferença de dificuldade entre as questões. O questionário continha dez perguntas técnicas, separadas em duas fases com cinco perguntas cada. Em uma fase, o participante consultava exclusivamente o manual técnico; na outra, podia utilizar a ferramenta de inteligência artificial e, obrigatoriamente, confirmar a resposta no manual indicado pelo sistema.

Por se tratarem de questões objetivas de localização de informações, com respostas unívocas e diretamente verificáveis nos manuais técnicos empregados como corpus, não foi realizada validação de conteúdo por painel independente de especialistas. As questões e seus respectivos gabaritos foram elaborados e verificados pelo autor, oficial de manutenção com experiência na área, quanto à correspondência com a documentação técnica e à pertinência para o contexto de manutenção da aeronave.

O desempenho foi medido pelo tempo, em segundos, entre a apresentação da questão e a localização da resposta. Foi estabelecido um *timeout* de 300 s por item. Quando o participante não localizava a resposta nesse intervalo, a consulta era encerrada e o valor de 300 s era registrado e incluído no cálculo das médias. Não se avaliou erro ou acerto como métrica de estudo, pois o objetivo não era testar o conhecimento individual do técnico, ou correção da resposta do algoritmo,

mas comparar o tempo necessário para localizar a informação em condições distintas de consulta.

As fases correspondiam a conjuntos fixos de questões, sendo alternada entre os grupos apenas a condição de consulta associada a cada uma delas, conforme apresentado na Tabela II.

TABELA II. CONTRABALANCEAMENTO DAS CONDIÇÕES DE CONSULTA

Grupo	***Fase 1***	***Fase 2***
A	Manual	IA + manual
B	IA + manual	Manual

Esse contrabalanceamento buscou mitigar vieses de especialidade, dificuldade, ordem e familiarização. Como cada técnico respondeu parte das questões com o método tradicional e parte com apoio da ferramenta, a análise pôde considerar tanto a comparação por questão quanto a comparação por participante. Essa estrutura é adequada para uma amostra reduzida, embora não substitua estudos posteriores com maior número de participantes e repetição controlada das tarefas.

As médias individuais obtidas por cada participante nas duas condições de consulta foram comparadas por meio do teste de postos sinalizados de Wilcoxon, adequado ao delineamento pareado e à reduzida dimensão da amostra. A dispersão das diferenças foi descrita pela mediana e pelo intervalo interquartílico, e a magnitude do efeito foi estimada pela correlação bisserial por postos.

IV. RESULTADOS E DISCUSSÃO

Os resultados indicaram redução do tempo de consulta em todas as dez questões. A média agregada com consulta exclusiva ao manual foi de 169,46 s, enquanto a média com apoio da ferramenta foi de 109,46 s. A diferença média foi, portanto, de 60 s por consulta, correspondendo a ganho aproximado de 35,4%. Além disso, ocorreram seis timeouts na condição de consulta exclusiva ao manual e nenhum timeout quando a ferramenta foi utilizada.

Os seis *timeouts* ocorreram exclusivamente na condição de consulta ao manual e foram registrados pelo valor-limite de 300 s. Por se tratarem de observações censuradas à direita, esse procedimento estabelece um limite inferior para o tempo que seria necessário à localização da resposta nesses casos, tornando conservadora a estimativa do tempo médio da consulta manual. Como análise de sensibilidade, excluindo-se essas seis observações, a média da consulta exclusiva ao manual foi de 151,66 s, ainda superior aos 109,46 s observados com apoio da ferramenta, correspondendo a uma redução de 27,8% no tempo médio de localização.

A dispersão dos tempos também foi consistentemente menor na condição com apoio da ferramenta. Considerando a análise por questão, os desvios-padrão variaram entre 11,9 s e 80,7 s na consulta exclusiva ao manual e entre 6,11 s e 22,2 s na consulta com apoio da IA, indicando maior estabilidade dos tempos de localização nesta última condição.

A Fig. 2 mostra que o ganho não se concentrou em um único caso. Mesmo nas questões em que a diferença absoluta foi menor, a utilização da ferramenta reduziu o tempo médio. Nas questões de maior dificuldade, o ganho foi mais expressivo, como indicam os Δ de 107,8 s e 134,8 s nas questões Q5 e Q8, respectivamente. Esse comportamento sugere que a ferramenta é especialmente útil quando a informação está menos evidente ou depende de navegação mais precisa dentro do manual.

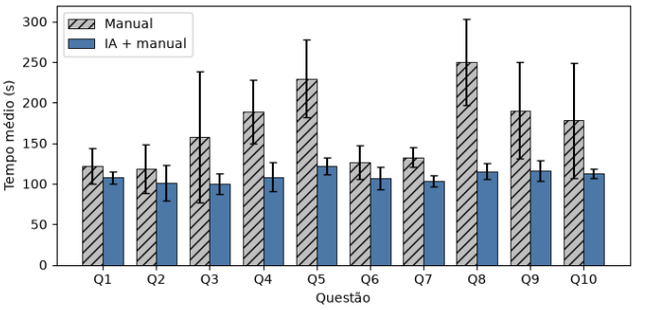


Fig. 2. Comparação dos tempos médios de localização por questão. As barras de erro representam ±1 desvio padrão.
Fonte: elaboração própria.

A análise por participante também apresentou efeito consistente. Os ganhos individuais variaram de 24,7% a 54%, sempre favoráveis ao uso da ferramenta. Observou-se ainda menor dispersão dos tempos quando a consulta foi apoiada pela IA, indicando desempenho mais estável entre técnicos de diferentes especialidades. Esse ponto é operacionalmente relevante, pois a manutenção de aeronaves envolve equipes heterogêneas, com graus distintos de familiaridade com cada sistema.

Na comparação intraindividual, a redução mediana do tempo foi de 56,0 s (Q1 = 50,25 s; Q3 = 60,8 s). O teste de postos sinalizados de Wilcoxon indicou diferença estatisticamente significativa entre as condições ($W = 0$; $p = 0{,}002$), com correlação bisserial por postos de 1,00. Esse valor reflete o fato de que todos os participantes apresentaram redução do tempo médio na condição com apoio da ferramenta.

O protocolo adotado torna o resultado conservador. Como a resposta fornecida pelo algoritmo deveria ser confirmada no manual técnico correspondente, o tempo da condição com IA inclui não apenas a interação com a ferramenta, mas também a validação documental. Mesmo com essa exigência, a redução média permaneceu significativa. Isso reforça o papel pretendido do sistema: acelerar a localização e a contextualização da informação, sem dispensar a consulta à fonte aprovada.

Do ponto de vista de implantação, o principal benefício está na combinação entre busca semântica e resposta contextualizada. A ferramenta permite que o usuário faça perguntas em linguagem natural e receba uma indicação inicial de procedimento, sistema ou tarefa associada. Também pode reduzir diferenças entre turnos, apoiar pessoal menos experiente e absorver picos de demanda em cenários de maior pressão operacional.

Há, entretanto, riscos a serem controlados. O uso frequente de sistemas de IA pode induzir dependência cognitiva e preferência por respostas rápidas em detrimento de análise técnica cuidadosa [12], [13]. Em manutenção aeronáutica, esse risco é crítico. Por isso, o sistema deve apresentar fontes, manter linguagem prudente, limitar-se ao *corpus* recuperado e ser incorporado a um processo que preserve a autoridade do manual técnico e a responsabilidade do mantenedor.

V. LIMITAÇÕES E CONTROLES OPERACIONAIS

A avaliação realizada deve ser interpretada como prova de conceito aplicada, e não como validação definitiva para emprego institucional. A amostra de dez participantes permitiu observar tendência clara de redução de tempo, mas estudos posteriores devem ampliar o número de técnicos, repetir

rodadas em diferentes turnos e incluir especialistas de maior variedade de sistemas. Também é recomendável separar efeitos de familiaridade com a interface, experiência prévia no manual e dificuldade intrínseca de cada questão.

O delineamento também não contemplou métricas formais de acerto da recuperação, correção ou completude das respostas e incidência de extrapolações, de modo que os resultados demonstram ganho de eficiência na localização da informação, mas não permitem, nesta etapa, inferir a confiabilidade técnica do sistema a partir dessas dimensões.

Outra limitação está no escopo documental. O corpus utilizado reuniu capítulos selecionados do AMM Part II, suficientes para testar a arquitetura, mas ainda distantes da complexidade de uma biblioteca técnica completa. Em um ambiente real, o sistema teria de lidar com diferentes tipos de publicação, revisões, boletins, ilustrações, tabelas, diagramas, procedimentos de troubleshooting e documentos produzidos durante a operação. A simples inserção de mais arquivos não garante melhoria automática; é necessário governar a qualidade da extração, a granularidade dos *chunks* e a preservação dos metadados.

A etapa de extração textual merece atenção específica. Manuais aeronáuticos apresentam elementos estruturais que podem ser perdidos em uma conversão direta, como notas, advertências, *cautions*, *warnings*, tabelas, referências cruzadas e legendas de figuras. Caso esses elementos não sejam preservados, o recuperador pode selecionar trechos incompletos ou o gerador pode perder condicionantes relevantes do procedimento. Por isso, uma versão operacional deve adotar *pipeline* documental auditável, com testes de regressão sempre que houver nova revisão dos manuais.

No campo de segurança da informação, a execução local é condição necessária, mas não suficiente. A arquitetura deve ser acompanhada por controle de acesso, registro de consultas, versionamento do banco vetorial, segregação de bases por sensibilidade, política de retenção de dados e procedimentos para atualização autorizada do *corpus*. Esses controles são coerentes com o tratamento de dados técnicos sensíveis e com a recomendação de reduzir a exposição de bases privadas em fluxos de RAG [9].

Por fim, há uma dimensão humana e doutrinária. A interface deve evidenciar que a resposta é um auxílio de localização e contextualização, não uma autorização de manutenção. Recomenda-se exibir as fontes, destacar que a execução deve seguir a publicação técnica vigente e, quando possível, mostrar trechos recuperados antes da resposta final. Treinamento específico dos usuários também é necessário para evitar dependência excessiva da ferramenta, risco já observado em estudos sobre delegação cognitiva a sistemas de inteligência artificial [12], [13].

Uma implantação institucional deve ser incremental. Em vez de disponibilizar a ferramenta diretamente como assistente amplo, recomenda-se iniciar com uso supervisionado em tarefas de baixa criticidade, com base documental delimitada e perguntas de referência previamente validadas por especialistas. Esse procedimento permitiria medir ganhos reais, identificar falhas de recuperação e ajustar *prompts*, *thresholds* e parâmetros de *chunking* antes de qualquer expansão para atividades mais sensíveis.

Também é recomendável estabelecer uma bateria permanente de testes. Cada alteração no *corpus*, no modelo de *embedding*, no *reranker* ou no modelo gerador deve ser acompanhada por perguntas-padrão e respostas esperadas, verificando se o sistema mantém coerência, recupera os mesmos trechos críticos e não introduz orientação divergente. Esse tipo de *benchmark* interno é particularmente importante em sistemas de apoio à manutenção, nos quais uma melhoria aparente de fluidez textual não significa, necessariamente, melhoria de confiabilidade técnica.

O registro das consultas pode gerar benefício adicional de gestão do conhecimento. Perguntas recorrentes indicam pontos do manual de difícil navegação, temas que demandam treinamento adicional ou sistemas com maior incidência de dúvidas operacionais. Assim, além de auxiliar o técnico durante a consulta, a ferramenta pode produzir indicadores úteis para treinamento, revisão de procedimentos internos e priorização de ações de suporte logístico.

A atualização do *corpus* deve seguir lógica de configuração. Cada revisão de manual, boletim ou documento inserido precisa estar associada a versão, data, origem e escopo de validade. Sem esse controle, a ferramenta pode recuperar conteúdo obsoleto ou misturar orientações de revisões distintas. A rastreabilidade documental, portanto, é tão relevante quanto a qualidade do modelo de linguagem utilizado. A ampliação do corpus deve ainda observar eventuais restrições de direitos autorais, licenciamento e controle de exportação aplicáveis à documentação técnica e às demais fontes utilizadas.

Finalmente, a adoção operacional exige delimitação de responsabilidade. A resposta gerada não deve ser registrada como fundamento autônomo de execução, mas como meio auxiliar de localização da publicação aplicável. O procedimento técnico continua dependente da leitura, interpretação e aplicação da documentação aprovada pelo fabricante e pelas autoridades competentes. Essa distinção deve ser explícita na interface, no treinamento e nas normas internas de uso.

VI. CONCLUSÃO

Este artigo apresentou e avaliou um algoritmo RAG executado localmente para apoio à consulta de manuais técnicos de manutenção aeronáutica. O protótipo integrou carregamento documental, *chunking, embeddings* multilíngues, banco vetorial local, *rerank*, geração por LLM e pós-processamento com indicação de fontes. Essa combinação permitiu transformar uma base documental extensa em uma interface conversacional orientada à recuperação de informação técnica.

A avaliação com dez técnicos de manutenção do KC-390 indicou redução média de 60 s por consulta, ganho aproximado de 35,4% e eliminação dos *timeouts* observados na consulta exclusiva ao manual. Os resultados apontam que a arquitetura proposta pode contribuir para a eficiência operacional, especialmente em situações nas quais a informação está dispersa ou quando o usuário não domina plenamente o capítulo técnico consultado.

A ferramenta, contudo, deve ser tratada como apoio à decisão e não como substituta da documentação aprovada. Trabalhos futuros devem ampliar o corpus para outros manuais e aeronaves, aumentar o tamanho da amostra, avaliar métricas de precisão e completude das respostas, cuja avaliação será necessária antes de qualquer utilização da ferramenta além do caráter de prova de conceito. Além disso, será necessário desenvolver interface de usuário, implementar controle de versão documental e aprofundar requisitos de segurança da informação para uso institucional em ambiente restrito.

REFERÊNCIAS

[1] Y. Almeida, Use of Machine-Learning Techniques Based on Python Language Code to Classify Failure Data from the Brazilian Air Force Database. Monterey, CA, USA: Naval Postgraduate School, 2023.
[2] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.
[3] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2023.
[4] B. S. Blanchard, Logistics Engineering and Management, 6th ed. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2004.
[5] United States Department of Defense, Defense Acquisition University, Integrated Product Support (IPS) Elements Guidebook, 2024.
[6] L. Vaskic and K. Paetzold, "A critical review of the integrated logistics support suite for aerospace and defence programmes," in Proc. 22nd Int. Conf. Engineering Design, Delft, The Netherlands, 2019, pp. 3541-3550.
[7] A. Vaswani et al., "Attention is all you need," in Proc. Advances in Neural Information Processing Systems, Long Beach, CA, USA, 2017, pp. 5998-6008.
[8] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[9] S. Zeng, J. Liu, X. Yu, W. Chen et al., "Exploring privacy issues in retrieval-augmented generation (RAG)," in Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, 2024.
[10] Y. Li, J. Wang, B. Pullman, N. Bandeira and Y. Papakonstantinou, "Index-based, high-dimensional, cosine threshold querying with optimality guarantees," arXiv preprint arXiv:1812.07695, 2018.
[11] M. Xu et al., "MM-R5: MultiModal reasoning-enhanced reranker via reinforcement learning for document retrieval," arXiv preprint arXiv:2506.12364, 2025.
[12] M. Gerlich, "AI tools in society: Impacts on cognitive offloading and the future of critical thinking," Societies, vol. 15, no. 1, p. 6, 2025.
[13] C. Zhai, S. Wibowo and L. D. Li, "The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review," Smart Learning Environments, vol. 11, no. 1, p. 28, 2024.