Impact of Hybrid Routing on Curriculum-Aware Multi-Collection RAG Architectures

Flaviano M. S. Junior [1] **, Luiz H. C. da Silva** [1] **, José N. S. Combé**[1] **, José T. S. de Oliveira** [1] **, Marcus M. Bezerra** [1] **, Sérgio N. Silva** [1]

[1]Universidade Federal de Campina Grande (UFCG)
Campina Grande – PB – Brasil

{flaviano.junior, luiz.henrique.silva, jose.combe, jose.tayrone.oliveira}@ee.ufcg.edu.br

marcus.bezerra@virtus.ufcg.edu.br, sergionatan@dee.ufcg.edu.br

Abstract. *Retrieval-Augmented Generation (RAG) systems usually rely on a single flat index, disregarding the curricular structure of educational content. This work investigates the impact of hybrid routing in a multi-collection RAG architecture for the Logic Circuits course. The corpus is organized into three curricular collections, and four variants of the same pipeline are compared: routing by* embedding *(R1), by keywords (R2), hybrid* embedding+*keyword routing (R3), and retrieval without routing over a single collection (R4). The hybrid router computes a weighted score that combines cosine similarity between the query* embedding *and the curricular profile of each stage with a keyword-overlap score. The evaluation uses 45 questions annotated and assessed with RAGAS. The results show that hybrid routing (R3) improves faithfulness compared with the flat index without routing (R4) (0.878 vs. 0.855) and achieves the highest answer correctness among the evaluated variants (0.645). Purely lexical routing obtains the highest values for* faithfulness *(0.898) and context recall (0.873), indicating that canonical technical terminology favors the lexical signal in this* benchmark*. Overall, the results suggest that curricular routing, especially in its hybrid form, improves factual grounding and highlights trade-offs between semantic and lexical signals in educational RAG systems.*

Resumo. *Sistemas de Geração Aumentada por Recuperação (RAG) geralmente utilizam um índice único e plano, desconsiderando a estrutura curricular do conteúdo educacional. Este trabalho investiga o impacto do roteamento híbrido em uma arquitetura RAG multi-coleção para a disciplina de Circuitos Lógicos. O corpus é organizado em três coleções curriculares, e comparam-se quatro variantes do mesmo pipeline: roteamento por* embedding *(R1), por palavras-chave (R2), híbrido* embedding+*palavras-chave (R3) e recuperação sem roteamento em coleção única (R4). O roteador híbrido calcula um escore ponderado que combina similaridade cosseno entre o* embedding *da consulta e o perfil curricular de cada estágio com uma pontuação de sobreposição de palavras-chave. A avaliação emprega 45 perguntas anotadas e avaliadas com RAGAS. Os resultados mostram que o roteamento híbrido (R3) melhora a faithfulness em relação ao índice plano sem roteamento (R4) (0,878 vs. 0,855) e alcança a maior correção de resposta entre as variantes avaliadas (0,645). O roteamento*

puramente léxico obtém os maiores valores de faithfulness *(0,898) e revocação de contexto (0,873), indicando que a terminologia técnica canônica favorece o sinal lexical neste* benchmark. *De modo geral, os resultados sugerem que o roteamento curricular, especialmente na forma híbrida, melhora o embasamento factual e evidencia compromissos entre sinais semânticos e léxicos em sistemas RAG educacionais.*

1. Introdução

A integração de Modelos de Linguagem de Grande Escala (LLMs) à educação superior tem despertado interesse crescente, sobretudo pela capacidade de fornecer respostas contextualizadas e adaptadas ao nível do estudante [Abdelghani et al. 2024, Modran et al. 2024]. Entretanto, LLMs utilizados de forma isolada estão sujeitos a alucinações, isto é, respostas factualmente incorretas ou inconsistentes [Ji et al. 2023], um problema que afeta domínios técnicos, como, por exemplo, a Engenharia da Computação, nos quais a precisão conceitual é imprescindível. O paradigma de Geração Aumentada por Recuperação (RAG) [Lewis et al. 2020] aborda essa limitação ao condicionar a geração do LLM a trechos recuperados de uma base documental verificada, ancorando as respostas em material de referência [Gao et al. 2023, Modran et al. 2024].

A maioria das implementações RAG opera sobre um índice vetorial único e plano (*flat index*), no qual todos os documentos compartilham o mesmo espaço de busca. Em contextos educacionais, contudo, o conteúdo apresenta organização curricular hierárquica (unidades, estágios, pré-requisitos) que um índice único não preserva. Termos como "tabela-verdade" (introduzida na álgebra booleana) e "tabela de transição" (usada em circuitos sequenciais) possuem representações vetoriais próximas, levando a recuperações semanticamente plausíveis porém pedagogicamente inadequadas para estudantes em estágios iniciais. A utilização de roteamento baseado unicamente em similaridade de *embeddings* pode, assim, confundir contextos curriculares.

Apesar da importância do roteamento em sistemas RAG multi-coleção, observa-se uma lacuna na literatura científica quanto a estudos quantitativos de ablação que comparem de forma controlada diferentes estratégias de roteamento (puramente semânticas baseadas em *embeddings*, puramente léxicas por palavras-chave e híbridas) no contexto de bases de dados organizadas por currículo. Estudos anteriores concentram-se na qualidade de recuperação em cenários generalistas, sem considerar a estrutura pedagógica da base documental.

Para suprir essa lacuna, este trabalho investiga o impacto de diferentes estratégias de roteamento no TutorIA, um tutor inteligente desenvolvido na Universidade Federal de Campina Grande (UFCG), sobre a precisão conceitual e a fidelidade factual das respostas, visando estruturá-las pedagogicamente e, assim, evitar a mistura de conteúdos e a sobrecarga cognitiva do estudante. O objetivo neste artigo consiste em realizar uma análise do impacto do roteamento híbrido em um pipeline RAG multi-coleção seguindo essa estrutura. Para isso, comparam-se quatro variantes do mesmo pipeline experimental: (R1) apenas *embedding*, (R2) apenas palavras-chave, (R3) híbrido (80% *embedding*, 20% palavras-chave) e (R4) índice plano sem roteamento, empregando o framework RAGAS [Es et al. 2024] e testes estatísticos para analisar os efeitos de cada estratégia. Como caso piloto para avaliação do TutorIA, adotou-se a disciplina de Circuitos Lógicos.

2. Trabalhos Relacionados

O paradigma RAG combina modelos de linguagem com mecanismos de recuperação de informação, permitindo que a resposta gerada seja condicionada por documentos externos ao modelo [Lewis et al. 2020]. Em contextos educacionais, a abordagem de recuperação tem sido explorada para aproximar as respostas dos materiais de apoio das disciplinas e reduzir a ocorrência de alucinações, especialmente em sistemas de tutoria com LLMs [Abdelghani et al. 2024, Modran et al. 2024]. Nesse cenário, a literatura recente mostra que sistemas RAG evoluíram de abordagens ingênuas, baseadas em um único índice documental, para sistemas modulares, com etapas complementares de filtragem, reranqueamento ou incorporação de roteamento de consultas [Gao et al. 2023]. O roteador determina para qual coleção, fonte ou conjunto documental a pergunta deve ser direcionada antes da recuperação, reduzindo o espaço de busca e aumentando a coerência entre a consulta e os documentos esperados.

No que diz respeito às estratégias de recuperação, a abordagem léxica, exemplificada por métodos como BM25 [Robertson and Zaragoza 2009], baseia-se na correspondência explícita de termos e tende a ser eficaz quando a consulta contém o mesmo vocabulário técnico do domínio. Já a recuperação semântica utiliza representações vetoriais densas, como DPR [Karpukhin et al. 2020] e BGE-M3 [Chen et al. 2024], com o intuito de capturar similaridade conceitual mesmo sem correspondência literal entre os termos. Abordagens híbridas buscam equilibrar entre o entendimento contextual dos *embeddings* com a precisão da correspondência léxica, podendo utilizar técnicas de fusão de rankings [Cormack et al. 2009].

Trabalhos recentes aplicaram RAG ao ensino técnico combinando recuperação híbrida [Liu et al. 2025, Lian 2025] e interpretação de código [Lu and Li 2025], sem avaliar o efeito de estratégias de roteamento em bases organizadas por estrutura curricular, lacuna que este trabalho endereça. Apesar dos avanços em RAG educacional, recuperação híbrida e arquiteturas modulares, os estudos que avaliam o efeito do roteamento de consultas em bases educacionais organizadas segundo a estrutura curricular da disciplina ainda são limitados. Este trabalho contribui nesse ponto ao analisar estratégias de roteamento em uma arquitetura RAG multi-coleção aplicada ao ensino de Circuitos Lógicos.

3. Descrição da Proposta

A proposta deste trabalho tem como base o TutorIA, um sistema inteligente desenvolvido na universidade baseado em RAG multi-coleção para apoio à disciplina de Circuitos Lógicos. Diferentemente de arquiteturas RAG baseadas em índice único, o TutorIA organiza a base documental em coleções curriculares, permitindo que a consulta do estudante seja previamente direcionada para o estágio mais adequado da disciplina.

A Figura 1 apresenta o diagrama geral da proposta com a organização multi-coleção do corpus, a etapa de roteamento, a recuperação dos trechos relevantes e a geração da resposta pelo modelo de linguagem. O pipeline é composto pelas seguintes etapas:

1. O módulo de roteamento calcula o escore s_{stage} (Equação 1) para cada coleção e direciona a consulta à coleção de maior escore;
2. São recuperados os $k = 3$ trechos mais relevantes do estágio selecionado;
3. Aplica-se um limiar de confiança $\tau = 0{,}5$ sobre o escore do trecho mais relevante, recusando a resposta quando o contexto recuperado apresenta baixa relevância;

4. Os trechos são embutidos no prompt estruturado (P2), que restringe o escopo da resposta aos dados fornecidos;
5. O modelo Qwen 2.5 7B local gera a resposta em português brasileiro com tom didático.

Os valores de $k = 3$ e $\tau = 0{,}5$ foram definidos empiricamente em testes preliminares, buscando equilibrar a quantidade de contexto recuperado e o critério de confiança, o que permitiu reduzir o número de combinações experimentais e direcionar o foco na estratégia de roteamento.

3.1. Escore de roteamento

No roteamento, a seleção do estágio curricular é realizada a partir da combinação entre um sinal semântico, baseado na similaridade entre o vetor da consulta e o perfil curricular vetorial do estágio, e um sinal léxico, baseado na ocorrência de termos-chave associados a cada estágio. Dessa forma, a mesma formulação permite avaliar o roteamento puramente semântico, o roteamento puramente léxico e o roteamento híbrido. O cálculo desse escore de roteamento pode ser dado por

$$s_{\text{stage}} = w_e \cdot \cos(\mathbf{q}, \mathbf{p}_s) + w_k \cdot \text{kw_score}(\mathbf{q}, K_s), \tag{1}$$

onde w_e e w_k são os pesos dos componentes de *embedding* e palavras-chave, $\mathbf{p}_s$ representa o vetor do perfil curricular do estágio s, e K_s é o conjunto de termos-chave do estágio curricular.

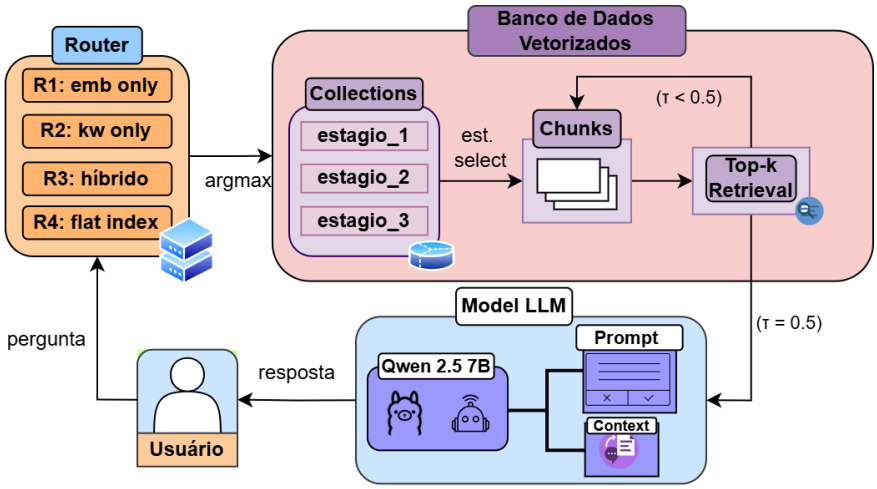


Figura 1. Pipeline do TutorIA. O experimento varia a estratégia de roteamento (bloco destacado em laranja): R1 (*embedding* puro), R2 (palavras-chave), R3 (híbrido 80/20) e R4 (*flat index* sem roteamento).

3.2. Corpus do Trabalho e Organização Curricular

A base documental utilizada inclui notas de aulas sobre conversão de bases, álgebra booleana e circuitos sequenciais [Cambuim sd, Güntzel sd, Hexsel sd, Higuti sd], apostilas de sistemas digitais [Kerschbaumer 2023, Lobo 2005], slides de monitoria sobre álgebra booleana, multiplexadores, circuitos aritméticos e Verilog [Alexandro et al. 2011], os Capítulos 1–8 das notas de aula de Circuitos Digitais de Bertini [Bertini 2017], slides

da disciplina sobre multiplexadores, demultiplexadores, latches e flip-flops [Furtado sd], além de obras de referência em sistemas digitais [Tocci and Widmer 2003, Floyd 2007].

Esses materiais foram segmentados e indexados em três coleções distintas no ChromaDB, correspondentes aos estágios curriculares da disciplina:

- **estagio_1:** Fundamentos da lógica digital, sistemas de numeração, portas lógicas, tabela-verdade e álgebra booleana;
- **estagio_2:** Simplificação por Karnaugh, multiplexadores e demultiplexadores, codificadores e decodificadores e circuitos lógicos e aritméticos;
- **estagio_3:** Circuitos sequenciais, flip-flops, registradores, contadores e máquinas de estados.

4. Metodologia

A metodologia experimental foi estruturada como uma avaliação do TutorIA em diferentes configurações de operação. Para cada pergunta do conjunto de avaliação, executa-se um fluxo comum composto por recuperação de contextos, construção do *prompt*, geração da resposta e avaliação automática. A partir desse fluxo, o estudo considera dois níveis de comparação: configurações gerais do sistema, usadas para contextualizar o efeito da recuperação, da organização multi-coleção e do *prompt*; e variantes específicas de roteamento, usadas para analisar o impacto da estratégia de seleção da coleção curricular.

4.1. Configurações experimentais

Foram avaliadas as configurações gerais do TutorIA e de sistemas de referência, conforme apresentado na Tabela 1. Essas configurações permitem observar a diferença entre o uso do LLM sem recuperação, o RAG com índice plano e duas versões multi-coleção do TutorIA com roteamento híbrido, variando o *prompt* utilizado.

Tabela 1. Configurações gerais avaliadas no pipeline experimental.

Conf.	Descrição
C0	LLM puro (Qwen 2.5 7B Instruct), sem recuperação
C1	RAG com índice plano (*flat/single-index*), sem roteamento
C2	TutorIA multi-coleção com roteamento híbrido e *prompt* P2
C3	TutorIA multi-coleção com roteamento híbrido e *prompt* P1

Em seguida foi realizada uma análise do componente de roteamento. Para isso, foram definidas quatro variantes, sumarizadas na Tabela 2. As variantes R1, R2 e R3 utilizam a arquitetura multi-coleção, alterando apenas os pesos do escore de roteamento. A variante R4 utiliza uma coleção única, sem etapa de roteamento, funcionando como referência para a busca em índice plano.

As variantes R1, R2 e R3 operam sobre a arquitetura multi-coleção descrita na Seção 3, variando apenas os pesos w_e e w_k no escore de roteamento (Equação 1). A variante R4 corresponde à configuração sem roteamento, na qual os documentos das três

Tabela 2. Variantes de roteamento e recuperação avaliadas no pipeline experimental.

Cond.	Descrição	w_e	w_k
R1	Apenas *embedding*	1,00	0,00
R2	Apenas palavras-chave	0,00	1,00
R3	Roteamento híbrido proposto	0,80	0,20
R4	Coleção única / sem roteamento	–	–

coleções curriculares são reunidos em uma coleção única, e a recuperação é realizada diretamente sobre esse índice plano. Assim, R4 permite comparar as estratégias de roteamento curricular com uma busca realizada sem seleção prévia de estágio. Os pesos $w_e = 0{,}80$ e $w_k = 0{,}20$ foram definidos de forma empírica de modo a priorizar o sinal semântico e mantendo uma contribuição lexical.

Todas as configurações e variantes utilizam o mesmo conjunto de 45 perguntas, elaboradas pelos membros da equipe de desenvolvimento e auditadas pelos professores da disciplina na instituição, e anotadas quanto ao estágio curricular esperado, distribuídas igualmente entre os três estágios da disciplina e acompanhadas de respostas de referência (*ground truth*). A geração local é realizada com o modelo Qwen 2.5 7B via Ollama. Nas variantes específicas de roteamento (R1–R4), mantém-se o mesmo *prompt* estruturado P2, variando apenas o mecanismo de seleção da coleção ou a ausência dessa etapa.

4.2. Métricas de avaliação

A avaliação emprega o *framework* RAGAS [Es et al. 2024], que fornece as seguintes métricas automáticas para sistemas RAG:

- **Faithfulness**: grau em que a resposta é sustentada pelos contextos recuperados;
- **Answer Correctness**: similaridade entre a resposta gerada e a resposta de referência;
- **Answer Relevancy**: relevância da resposta em relação à pergunta;
- **Context Precision**: precisão dos contextos recuperados em relação aos contextos esperados;
- **Context Recall**: revocação dos contextos esperados nos contextos recuperados;
- **Answer Similarity**: similaridade semântica entre resposta gerada e referência.

Registram-se ainda a taxa de recusa (*refusal rate*), proporção de perguntas para as quais o sistema recusa responder, e a latência média de geração.

A avaliação automatizada emprega o juiz externo (`qwen-2.5-72b-instruct` via OpenRouter). De forma geral, o funcionamento do LLM juiz pode ser interpretado como composto por três etapas: recebimento das entradas necessárias para a métrica avaliada, realização do julgamento por meio de prompts, estruturados pelo próprio RAGAS, e agregação dos resultados em uma pontuação final que depende de qual métrica está sendo avaliada.

Para verificar diferenças entre as condições de roteamento, aplica-se o teste de Friedman [Friedman 1937] (natureza pareada e não paramétrica das métricas RAGAS), com $\alpha = 0{,}05$.

4.3. Ambiente computacional

Todos os experimentos de inferência local e medição de consumo de recursos foram conduzidos em uma estação de trabalho com GPU NVIDIA GeForce RTX 4060 com 8 GB de VRAM dedicada, 64 GB de RAM DDR5, processador Intel Core i9 e sistema operacional Windows 11. Os modelos foram executados sob a plataforma Ollama v0.1.48.

5. Resultados e Discussão

A avaliação das configurações propostas permitiu investigar o impacto das diferentes estratégias de recuperação e roteamento tanto sobre a qualidade das respostas quanto sobre a consistência do embasamento fornecido pelo sistema. Os resultados correspondentes ao primeiro nível de comparação descrito na metodologia são apresentados na Tabela 3, enquanto a Figura 2 fornece uma comparação visual das métricas de *faithfulness* e *context recall* entre as configurações C0–C3. Adicionalmente, a Tabela 4 resume os resultados do teste de Friedman aplicado às métricas RAGAS para as quatro configurações gerais do pipeline.

Tabela 3. Resultados RAGAS e métricas operacionais das configurações gerais do pipeline ($n = 45$).

Métrica	C0	C1	C2	C3
Faithfulness	0,058	0,855	**0,874**	0,776
Answer Correctness	**0,672**	0,623	0,629	0,555
Answer Relevancy	**0,836**	0,737	0,732	0,574
Context Precision	0,000	**0,812**	0,698	0,713
Context Recall	0,067	**0,808**	0,740	0,752
Answer Similarity	**0,850**	0,835	0,824	0,781
Refusal Rate (%)	0,0	8,9	11,1	33,3
Latência (s)	12,27	6,57	13,24	5,30

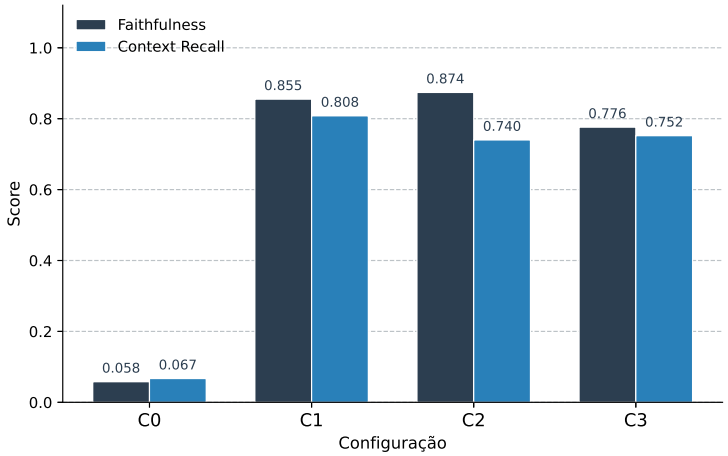


Figura 2. Comparação de *faithfulness* e *context recall* entre as configurações gerais do pipeline (C0–C3).

Os resultados indicam que existem diferenças estatisticamente significativas entre as configurações gerais para três métricas: *faithfulness*, *context precision* e *context recall* (todas com $p < 0{,}001$). Em contrapartida, as métricas de qualidade de resposta (*answer*

Tabela 4. Resultados do teste de Friedman ($n = 32$–45, conforme pares completos disponíveis por métrica) para as métricas RAGAS nas configurações C0–C3.

Métrica	χ^2	p	Significância
Faithfulness	61,39	$< 0{,}001$	Sim ($p < 0{,}001$)
Answer Correctness	4,57	0,206	Não
Answer Relevancy	3,78	0,287	Não
Context Precision	53,48	$< 0{,}001$	Sim ($p < 0{,}001$)
Context Recall	49,62	$< 0{,}001$	Sim ($p < 0{,}001$)
Answer Similarity	3,11	0,375	Não

correctness, *answer relevancy* e *answer similarity*) não apresentam diferenças significativas, sugerindo que a estratégia de recuperação afeta primariamente o embasamento factual, sem comprometer a qualidade percebida das respostas.

O comportamento observado na Tabela 3, com *answer relevancy* elevada e *faithfulness* praticamente nula, em C0, é consistente com o fenômeno de alucinação extrínseca descrito por [Ji et al. 2023] no qual é discutido que LLMs são capazes de produzir respostas linguisticamente fluentes e semanticamente pertinentes à consulta, independentemente da existência de embasamento em evidências.

A comparação entre C2 (multi-coleção com roteamento híbrido) e C1 (índice plano) revela um padrão interessante. O roteamento híbrido proporciona ganho em *faithfulness* (0,874 vs. 0,855), indicando que a restrição do espaço de busca a uma coleção curricular específica resulta em contextos mais coerentes com a resposta gerada. Contudo, observa-se redução na *context precision* (0,698 vs. 0,812) e na *context recall* (0,740 vs. 0,808).

Esse resultado pode ser explicado pela rigidez do roteamento, que ao restringir a busca a uma única coleção elimina trechos potencialmente relevantes presentes em outras coleções. Quando o roteamento direciona corretamente, os contextos recuperados são altamente coerentes, elevando a *faithfulness*. Contudo, quando ocorre *misrouting* (erro de direcionamento), a *precision* cai, pois os contextos recuperados da coleção incorreta são irrelevantes para a pergunta. A duplicação de *chunks* em fronteiras curriculares atenua parcialmente esse efeito, mas não o elimina.

A taxa de recusa de C2 (11,1%) é superior à de C1 (8,9%), refletindo o comportamento conservador do *prompt* P2, que instrui o modelo a recusar respostas quando os contextos recuperados são insuficientes. Esse comportamento é pedagogicamente desejável, pois evita a disseminação de informações incorretas, alinhando-se a princípios de confiabilidade em tutores digitais [Abdelghani et al. 2024].

A latência de C2 (13,24s) é superior à de C1 (6,57s), mas a comparação direta entre essas duas configurações não isola o custo do roteamento. Por exemplo, C3 que também aplica roteamento multi-coleção, apresenta a menor latência do experimento (5,30s), indicando que a etapa de roteamento não é o fator determinante. A diferença mais expressiva entre C2 e C3 reside no *prompt* utilizado, sendo que C2 emprega P2, que produz respostas mais extensas e, consequentemente, maior tempo de geração. Assim, a latência observada em C2 reflete principalmente o custo de geração associado ao *prompt*,

e não uma penalização inerente à arquitetura multi-coleção.

A Tabela 5 apresenta os resultados das variantes de roteamento, correspondentes ao segundo nível de comparação descrito na Seção 4. Essa análise exibe o comportamento do roteador ao comparar os diferentes tipos de roteamentos do estudo e revela que o roteamento puramente semântico (R1) apresenta o pior desempenho geral, com a mais alta taxa de recusa (24,4%) e a menor precisão de contexto (0,570), sugerindo que *embeddings* sozinhos falham em capturar jargões específicos de circuitos lógicos: termos ambíguos entre estágios (p.ex., "tabela", associada tanto a tabelas-verdade quanto a tabelas de transição) favorecem o *misrouting*.

Tabela 5. Resultados RAGAS e métricas operacionais das variantes de roteamento (R1–R4).

Métrica	**R1** (emb)	**R2** (kw)	**R3** (híbrido)	**R4** (flat)
Faithfulness	0,730	0,898	0,878	0,855
Answer Correctness	0,530	0,620	0,645	0,623
Answer Relevancy	0,609	0,760	0,741	0,737
Context Precision	0,570	0,791	0,702	0,812
Context Recall	0,666	0,873	0,734	0,808
Answer Similarity	0,747	0,826	0,824	0,835
Refusal Rate (%)	24,4	8,9	11,1	8,9
Latência (s)	5,48	5,69	5,80	6,57

Diferentemente das configurações gerais (C0–C3), nas quais o teste de Friedman foi aplicado para verificar o efeito global da recuperação, a análise das variantes de roteamento (R1–R4) é conduzida de forma descritiva. Essa escolha se justifica pelo fato de que as quatro variantes compartilham o mesmo prompt, o mesmo modelo e o mesmo conjunto de perguntas, diferindo exclusivamente nos pesos do escore de roteamento. Nesse contexto, as diferenças observadas nas métricas refletem diretamente o efeito da estratégia de seleção da coleção, sem confundimento com outras variáveis do pipeline.

Um resultado inesperado é o desempenho do roteamento puramente léxico (R2), que obtém a maior *faithfulness* (0,898), a maior *context recall* (0,873) e a maior *context precision* entre as condições multi-coleção (0,791), com taxa de recusa equivalente à do índice plano (8,9%). Esse achado indica que, em um domínio com vocabulário técnico fortemente discriminativo como Circuitos Lógicos, a correspondência de palavras-chave é um sinal de roteamento altamente eficaz, pois as perguntas do *benchmark*, formuladas pela equipe de desenvolvimento, tendem a empregar a terminologia canônica de cada estágio.

O roteador híbrido (R3) mantém-se competitivo, atingindo a melhor *answer correctness* (0,645) e *faithfulness* superior ao índice plano (0,878 vs. 0,855 de R4), com latência inferior (5,80 s vs. 6,57 s) por buscar em um subconjunto menor e mais focado de documentos. A vantagem esperada do componente semântico, robustez a consultas coloquiais ou sem termos técnicos, não é exercitada por este *benchmark*, o que sugere cautela na generalização: para consultas de estudantes reais, com linguagem menos canônica, a combinação híbrida tende a ser mais robusta que o roteamento puramente léxico. A

confirmação dessa hipótese requer um conjunto de avaliação com perguntas do dia a dia.

Este estudo apresenta quatro limitações. O *benchmark* de 45 perguntas, revisada por professores com terminologia canônica, favorece o roteamento léxico e limita o poder estatístico dos testes. As medições de latência refletem estimativas operacionais sujeitas a variações do ambiente local. A avaliação automatizada via RAGAS não conta com validação humana, e o modelo juiz (`qwen-2.5-72b-instruct`) pertence à mesma família do modelo avaliado, o que pode introduzir viés de afinidade. Por fim, a análise das variantes de roteamento (R1–R4) é descritiva, sem testes de significância dedicados.

6. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma análise comparativa de estratégias de roteamento de consultas em uma arquitetura RAG multi-coleção aplicada ao apoio pedagógico na disciplina de Circuitos Lógicos. Nas configurações gerais do pipeline, o roteamento híbrido (*embedding* + palavras-chave) proporcionou a maior *faithfulness* (0,874) entre as configurações avaliadas, superando tanto o LLM puro (0,058) quanto o RAG com índice plano (0,855). Testes de Friedman indicaram diferenças estatisticamente significativas para *faithfulness* ($\chi^2 = 61{,}39$, $p < 0{,}001$), *context precision* ($\chi^2 = 53{,}48$, $p < 0{,}001$) e *context recall* ($\chi^2 = 49{,}62$, $p < 0{,}001$), enquanto métricas de qualidade da resposta permanecem estatisticamente equivalentes.

A análise específica das variantes de roteamento acrescenta uma observação adicional, uma vez que o roteamento puramente léxico (R2) superou o híbrido em *faithfulness* (0,898 vs. 0,878) e revocação de contexto (0,873 vs. 0,734) neste *benchmark* com terminologia canônica, enquanto o híbrido preservou a melhor correção de resposta (0,645). Assim, os resultados sugerem que a escolha da estratégia de roteamento deve considerar o perfil linguístico esperado das consultas dos estudantes.

A principal contribuição do trabalho está em mostrar que a organização curricular da base documental, combinada a estratégias explícitas de roteamento, afeta diretamente o embasamento factual das respostas em sistemas RAG educacionais. Esse achado é relevante para o projeto de tutores inteligentes em domínios técnicos, nos quais a confiabilidade das respostas é prioritária [Abdelghani et al. 2024].

Como trabalhos futuros, destacam-se a avaliação com consultas coloquiais de estudantes reais para testar a hipótese de que o componente semântico do roteador híbrido supera o roteamento léxico na ausência de terminologia canônica, a aplicação de testes *post-hoc* pareados entre as variantes de roteamento; a construção de uma matriz de confusão de roteamento para quantificar a taxa de *misrouting* por condição, a otimização sistemática dos pesos (w_e, w_k) via *grid search* ou otimização bayesiana, a comparação com abordagens mais recentes de roteamento para RAG e um estudo com usuários reais para avaliar o impacto do roteamento na aprendizagem efetiva.

Agradecimentos

Este trabalho foi apoiado em parte pela Universidade Federal de Campina Grande (UFCG) e em parte pelo Projeto Base iSOP apoiado pelo Centro de Competência Embrapii Virtus em Hardware Inteligente Para Indústria (VIRTUS-CC), por meio do Programa Prioritário (PPI) HardwareBR do Ministério da Ciência, Tecnologia e Inovação (MCTI), firmado

com a Empresa Brasileira de Pesquisa e Inovação Industrial (EMBRAPII), sob concessão 055/2023.

Referências

Abdelghani, R., Wang, Y.-H., Yuan, X., Wang, T., Lucas, P., Sauzéon, H., and Oudeyer, P.-Y. (2024). GPT-3-driven pedagogical agents to train children's curious question-asking skills. *International Journal of Artificial Intelligence in Education*, 34:483–518.

Alexandro, D., Delgado, R., and Ogg, V. (2011). Sistemas digitais. Slides de monitoria da disciplina de Sistemas Digitais, Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE). Aulas sobre Álgebra Booleana, Multiplexadores, Demultiplexadores e Circuitos Aritméticos.

Bertini, L. (2017). Circuitos digitais. Notas de aula da disciplina RCM 00011 – Circuitos Digitais, Universidade Federal Fluminense (UFF). Capítulos 1–8.

Cambuim, L. ([s.d.]). Aula 2: Conversão de bases e aritmética computacional. Notas de aula da disciplina Arquitetura e Organização de Computadores. Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE). Aula 02. Acesso em: jun. 2026.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (2024). M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Cormack, G. V., Clarke, C. L. A., and Büttcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. ACM.

Es, S., James, J., Espinosa-Anke, L., and Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL) - System Demonstrations*, pages 150–158.

Floyd, T. L. (2007). *Sistemas Digitais: Fundamentos e Aplicações*. Bookman, Porto Alegre, 9 edition.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701.

Furtado, D. A. ([s.d.]). Sistemas digitais: Módulos 9 e 10. Apresentação de slides da disciplina Sistemas Digitais. Graduação em Sistemas de Informação, Faculdade de Computação, Universidade Federal de Uberlândia (UFU). Módulo 9: Multiplexadores e Demultiplexadores; Módulo 10: Circuitos Sequenciais: Latches e Flip-Flops. Acesso em: jun. 2026.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Güntzel, J. L. ([s.d.]). Introdução a sistemas digitais: Álgebra booleana e circuitos lógicos. Apostila da disciplina INE 5406 - Introdução a Sistemas Digitais. Departamento de Informática e Estatística (INE), Universidade Federal de Santa Catarina (UFSC). Álgebra Booleana e Circuitos Lógicos. Acesso em: jun. 2026.

Hexsel, R. A. ([s.d.]). Capítulo 8: Circuitos sequenciais síncronos. Notas de aula da disciplina CI 068 - Circuitos Digitais. Departamento de Informática, Universidade Federal do Paraná (UFPR). Capítulo 8. Acesso em: jun. 2026.

Higuti, R. T. ([s.d.]). ELE 1065 - circuitos digitais I: Lógica booleana. Notas de aula da disciplina ELE 1065 Circuitos Digitais I.Departamento de Engenharia Elétrica, Faculdade de Engenharia de Ilha Solteira (FEIS),Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP). Lógica Booleana. Acesso em: jun. 2026.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Kerschbaumer, R. (2023). Sistemas digitais. Apostila didática. Instituto Federal Catarinense (IFC) - *Campus* Luzerna.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Lian, Y. (2025). Machine assistant with reliable knowledge: Enhancing student learning via rag-based retrieval. *arXiv preprint arXiv:2506.23026*.

Liu, Z., Agrawal, P., Singhal, S., Madaan, V., Kumar, M., and Verma, P. K. (2025). LPITutor: An LLM-based personalized intelligent tutoring system using RAG and prompt engineering. *PeerJ Computer Science*, 11:e2991.

Lobo, V. (2005). Circuitos sequenciais síncronos. Folhas de estudo da disciplina Hardware e Software. Departamento de Armas e Electrónica, Escola Naval de Portugal / NOVA IMS - Universidade Nova de Lisboa.

Lu, J. and Li, J. (2025). A novel framework for educational q&a: Leveraging rag and code interpreters for knowledge retrieval and logical computation. *PloS one*, 20(12):e0337361.

Modran, H., Bogdan, I. C., Ursuţiu, D., Samoila, C., and Modran, P. L. (2024). LLM intelligent agent tutoring in higher education courses using a RAG approach. *Preprints*.

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 4(1-2):1–174.

Tocci, R. J. and Widmer, N. S. (2003). *Sistemas Digitais: Princípios e Aplicações*. Prentice Hall, São Paulo, 8 edition.