INSTITUTO FEDERAL DE CIÊNCIA E TECNOLOGIA DE PERNAMBUCO

Campus Recife

Tecnologia em Análise e Desenvolvimento De Sistemas

NELSON HENRIQUE DE SOUZA NETO

UTILIZAÇÃO DE UM LLM LOCAL COM RAG PARA AUXILIAR A FASE DE EXTRAÇÃO DE DADOS DE UMA REVISÃO SISTEMÁTICA DA LITERATURA

Recife

2026

NELSON HENRIQUE DE SOUZA NETO

UTILIZAÇÃO DE UM LLM LOCAL COM RAG PARA AUXILIAR A FASE DE EXTRAÇÃO DE DADOS DE UMA REVISÃO SISTEMÁTICA DA LITERATURA

Trabalho de Conclusão de Curso apresentado ao curso de Tecnologia em Análise e Desenvolvimento de Sistemas, como parte dos requisitos para a obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas pelo Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco (IFPE), *Campus* Recife.

Orientador: Prof. Dr. Vilmar Santos Nepomuceno

Recife

2026

Ficha catalográfica elaborada pela bibliotecária Danielle Castro da Silva CRB4/1457

S729u
2026 Souza Neto, Nelson Henrique de

Utilização de um LLM local com RAG para auxiliar a fase de extração de dados de uma revisão sistemática da literatura / Nelson Henrique de Souza Neto. --- Recife: O autor, 2026.

38f. il. Color.

Trabalho de Conclusão (Curso Superior Tecnológico em Análise e Desenvolvimento de Sistemas) – Instituto Federal de Pernambuco, Recife, 2025.

Inclui Referências.

Orientador: Prof. Dr. Vilmar Santos Nepomuceno.

1. Inteligência artificial. 2. Large Language Models. 3. Retrieval-Augmented Generation. I. Título. II. Nepomuceno, Vilmar Santos (orientador). III. Instituto Federal de Pernambuco.

CDD 607

UTILIZAÇÃO DE UM LLM LOCAL COM RAG PARA AUXILIAR A FASE DE EXTRAÇÃO DE DADOS DE UMA REVISÃO SISTEMÁTICA DA LITERATURA

Trabalho aprovado. Recife, 17 de março de 2026.

_____

Prof. Dr. Vilmar Santos Nepomuceno

_____

Examinadora interna: Profa. Ma. Renata Freire de Paiva Neves

_____

Examinador externo: Prof. Me. Ivanildo Monteiro de Azevedo

Recife

2026

AGRADECIMENTOS

A Deus, primeiramente, por ter guiado meus passos nesta trajetória.
Ao meu pai, que sempre me incentivou e apoiou em minha educação.
À minha família, que sempre me apoiou e ajudou em todos os momentos.
Aos meus colegas de turma, que me acompanharam e auxiliaram durante o curso.
Ao Prof. Vilmar, pela orientação, atenção e suporte no desenvolvimento deste trabalho.
Aos professores do curso de TADS, que tanto me ensinaram e aconselharam durante esta caminhada.
Ao Instituto Federal de Pernambuco, por prover oportunidades de aprendizado e de crescimento profissional e pessoal.
E a todos os demais que, de alguma forma, me ajudaram nesta jornada.

RESUMO

Revisão Sistemática da Literatura (RSL) é uma metodologia de pesquisa que segue protocolos específicos, muito utilizada em trabalhos acadêmicos para resumir e sintetizar evidências sobre um determinado tópico de estudo, havendo um crescimento de sua aplicação na área de Engenharia de Software. Porém, sua condução é trabalhosa, exigindo muito tempo e recursos humanos. Com os avanços recentes da Inteligência Artificial, ferramentas como os *Large Language Models* (LLMs), *Generative Pre-trained Transformer* (GPT), por exemplo, e *Retrieval-Augmented Generation* (RAG), oferecem oportunidades para diminuir o esforço manual na condução dessas revisões. Este estudo tem como objetivo investigar se a utilização de um LLM local com RAG pode auxiliar na fase de extração de dados de uma revisão sistemática. Para isso, foi utilizado o modelo Llama 3.2 na extração de dados de um estudo de mapeamento sistemático contendo 22 artigos de RSL cujos conteúdos foram fornecidos ao LLM por meio da técnica RAG e as respostas geradas pelo modelo foram comparadas com as já extraídas pelos autores do mapeamento. Esse uso do LLM local com RAG alcançou aproximadamente 42% de respostas corretas, mostrando-se pouco capaz de auxiliar o pesquisador significativamente em relação à fase de extração de dados da RSL, porém a maior parte dos acertos aconteceu sobre dados bibliográficos dos artigos, o que sugere que o modelo pode ser utilizado para obter esses dados com mais facilidade.

Palavras-chave: IA. LLM. RAG. revisão sistemática da literatura. extração de dados.

ABSTRACT

Systematic Literature Review (SLR) is a research methodology that follows specific protocols and is widely used in academic work to summarize and synthesize evidence on a given topic of study, with its application growing in the field of Software Engineering. However, conducting an SLR is laborious, as it requires significant time and human resources. With recent advances in Artificial Intelligence, tools such as Large Language Models (LLMs), Generative Pre-trained Transformer (GPT), for example, and Retrieval-Augmented Generation (RAG) offer opportunities to reduce the manual effort in conducting these reviews. This study aims to investigate whether the use of a local LLM augmented with RAG can assist the data extraction phase of a systematic review. To this end, the Llama 3.2 model was used to extract data from a systematic mapping study containing 22 SLR articles whose contents were provided to the LLM using the RAG technique, and the responses generated by the model were compared with those already extracted by the authors of the mapping. The local LLM augmented with RAG achieved approximately 42% correct answers, demonstrating that it offers limited assistance to the researcher in relation to the data extraction phase of the RSL; however, most of the correct answers concerned bibliographic data from the articles, suggesting that the model can be used to obtain this data more easily.

Keywords: AI. LLM. RAG. systematic literature review. data extraction.

LISTAS DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

APA	American Psychological Association
CLI	Command Line Interface
EMS	Estudo de Mapeamento Sistemático
ES	Engenharia de Software
GPT	Generative Pre-trained Transformer
IA	Inteligência Artificial
LLM	Large Language Model
PDF	Portable Document Format
RAG	Retrieval-Augmented Generation
RSL	Revisão Sistemática da Literatura
S	Study
SLR	Systematic Literature Review
TM	Text Mining
VTM	Visual Text Mining

SUMÁRIO

1 INTRODUÇÃO

Revisões Sistemáticas da Literatura (RSLs) cada vez mais têm sido conduzidas pela comunidade de Engenharia de Software (ES) com o intuito de sintetizar e resumir evidências de variados estudos, evidenciando o estado da arte de um determinado tópico de pesquisa nessa área (Santos *et al*., 2021). No entanto, a execução delas é bastante trabalhosa atualmente (Michelson; Reuter, 2019). A condução dessas revisões exige várias etapas, como o desenvolvimento do protocolo, a seleção de estudos adequados para inclusão, a extração de dados e a síntese, por exemplo (Felizardo; Carver, 2020), consumindo muito tempo e exigindo uma grande quantidade de recursos humanos (Borah *et al*., 2017).

Em um curto espaço de tempo, a Inteligência Artificial (IA) causou um grande impacto no mundo da pesquisa científica, tendo seu uso associado a várias atividades, como analisar dados e imagens, gerar hipóteses, pesquisar e revisar a literatura científica, escrever e editar artigos científicos (Resnik; Hosseini, 2024). Os *Large Language Models* (LLMs) são uma técnica de IA que tem aumentado sua utilização no campo acadêmico atualmente, sendo observado um aumento acentuado de conteúdos modificados por LLMs em textos acadêmicos poucos meses após o lançamento do ChatGPT em 2022, com os artigos da área de Ciência da Computação apresentando o crescimento mais rápido observado (Liang *et al*., 2024).

Os LLMs são modelos estatísticos de linguagem baseados em redes neurais de grande escala (Minaee et al., 2024), treinados a partir de imensa quantidade de texto, ou seja, bilhões de palavras extraídas de artigos, livros e outros conteúdos disponíveis na internet. Esse treinamento faz com que os LLMs aprendam padrões de como as palavras são usadas na linguagem e consigam executar tarefas que envolvam processamento de linguagem natural (Thirunavukarasu et al., 2023), como gerar, resumir, analisar e traduzir textos, por exemplo. Dessa forma, mostrando-se como uma ferramenta com potencial para auxiliar a condução de trabalhos científicos geralmente laborosos, como as RSLs.

Os LLMs, no entanto, são desenvolvidos e pré-treinados a partir de um processo fixo de aprendizagem, ou seja, eles possuem um conhecimento imutável, limitado aos dados aos quais foram expostos durante seus treinamentos. Além disso, esses modelos obtêm seus dados principalmente de fontes de acesso público, o que pode resultar na falta de acesso a informações atualizadas, específicas de domínio ou de fontes de acesso privado (Ng; Matsuba; Zhang, 2025). Em contrapartida, *Retrieval-Augmented Generation* (RAG) é uma técnica que melhora e amplia a capacidade de resposta dos LLMs nesse sentido. O RAG inclui uma fase inicial em que os LLMs buscam informações relevantes em uma fonte externa de dados antes de produzir o texto

de saída, o que diminui as limitações de conhecimento dos modelos ao integrar a recuperação de dados externos, atualizados ou de domínio específico, por exemplo, no processo de gerar respostas, aumentando a precisão e importância delas (Arslan *et al*., 2024).

Diante da importância das RSLs para a comunidade científica e do aumento de suas conduções em Engenharia de Software, além do crescimento do uso de IA em atividades acadêmicas, em especial, os LLMs, e do aprimoramento de técnicas que auxiliam essas ferramentas a terem melhores resultados como o RAG, o objetivo geral deste trabalho é responder à seguinte pergunta: “A utilização de um LLM local com RAG pode auxiliar a fase de extração de dados de uma revisão sistemática?”. Os objetivos específicos são:

- Entender como ocorre a execução de uma revisão sistemática da literatura;
- Buscar na literatura sobre o uso de LLMs em revisões sistemáticas da literatura;
- Avaliar quais possíveis LLMs podem ser usados localmente;
- Avaliar quais *software* disponíveis permitem a execução local de um LLM com RAG;
- Selecionar um estudo de revisão sistemática já publicado como base para comparação;
- Obter todos os artefatos necessários do estudo selecionado para avaliar as respostas do modelo local com RAG;
- Avaliar a extração de dados do modelo com a extração já obtida pelos autores do estudo.

Selecionou-se um estudo de mapeamento sistemático intitulado "On the need to update systematic literature reviews" (Nepomuceno; Soares, 2019), que contém 22 estudos que são atualizações de revisões sistemáticas. Em seguida, foi executado localmente o modelo Llama 3.2 (3B, Lightweight Model) (Meta, 2025), utilizando a técnica RAG com o conteúdo textual completo dos 22 artigos. O uso local do modelo é importante para evitar problemas legais de *copyright* dos artigos científicos que podem surgir ao serem usados em modelos disponíveis *online*. Por fim, com o foco na fase de extração de dados de uma RSL, foram elaboradas perguntas ao modelo sobre cada um dos artigos de acordo com o formulário de extração obtido do mapeamento sistemático. As respostas geradas pelo LLM foram então comparadas com os dados já extraídos pelos autores do estudo base. Nessa comparação, contabilizou-se o número de respostas corretas, parcialmente corretas e incorretas. O modelo local com RAG alcançou, aproximadamente, 42% de respostas corretas.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Revisão Sistemática da Literatura (RSL)

Revisão Sistemática da Literatura, ou simplesmente Revisão Sistemática, é um meio de identificar, avaliar e interpretar todas as pesquisas disponíveis importantes para uma pergunta de pesquisa, um tópico ou um fenômeno de interesse particular. Existem vários motivos para se conduzir uma revisão sistemática, tais como: resumir as evidências sobre uma tecnologia ou um tratamento; identificar lacunas na pesquisa atual com o propósito de sugerir campos para investigação; e fornecer uma estrutura para situar adequadamente novas atividades de pesquisa (Kitchenham; Charters, 2007).

Kitchenham e Charters (2007) dividem a execução de uma revisão sistemática em três fases: planejamento da revisão, condução da revisão e relato da revisão. No planejamento, são realizadas atividades como a especificação da pergunta de pesquisa e o desenvolvimento do protocolo da revisão. O protocolo tem a função de especificar os métodos que serão usados na revisão, reduzindo a possibilidade de viés por parte do pesquisador. Na condução da revisão, encontramos tarefas como a seleção dos estudos, a extração de dados, para a qual se planejam formulários que permitem registrar corretamente os dados obtidos dos estudos selecionados, e a síntese de dados, em que se coleta e sintetiza os resultados dos estudos incluídos na revisão. Por fim, o relato da revisão compreende atividades como a especificação de mecanismos de disseminação, ou seja, estratégias para comunicar os resultados da revisão de forma eficaz, e a formatação do relatório principal, que deve seguir um padrão adequado ao seu fim, como um relatório técnico, por exemplo.

Conduzir uma revisão sistemática da literatura demanda uma carga de trabalho muito maior do que elaborar uma revisão de literatura narrativa tradicional (Felizardo; Carver, 2020), o que exige esforço e tempo consideráveis de um pesquisador experiente (Carver *et al*., 2013). No desenvolvimento do protocolo, o autor deve tomar importantes decisões sobre o planejamento da condução da RSL envolvendo a pergunta de pesquisa, critério de seleção e estratégias de busca, por exemplo; na fase de seleção de estudo, o pesquisador precisa analisar uma grande quantidade de estudos; extrair dados fica mais complexo devido à falta de padronização nos formatos dos dados e no *design* dos artigos; sintetizar evidências é difícil porque muitos artigos podem não fornecer todas as informações necessárias (Felizardo; Carver, 2020).

A revisão sistemática compartilha semelhanças metodológicas com o mapeamento sistemático, como a busca e a seleção de estudos, porém ela tem como objetivo sintetizar evidências, já o mapeamento sistemático visa estruturar uma área de pesquisa, fornecendo uma visão geral sobre ela (Petersen; Vakkalanka; Kuzniarz, 2015).

2.2 Inteligência Artificial (IA)

A IA não é uma tecnologia recente, suas origens remontam à década de 1950 e o termo "Inteligência Artificial" foi cunhado em 1956 (Haenlein; Kaplan, 2019). Desde então, esse campo do conhecimento tem passado por contínua expansão, impulsionando o desenvolvimento de várias aplicações inteligentes em diversos setores, tornando-se relevante para qualquer atividade intelectual. Russel e Norving (2021) apontam que a IA é frequentemente destacada em pesquisas como um dos campos mais interessantes e de crescimento acelerado, com um faturamento anual que ultrapassa a marca de um trilhão de dólares. Atualmente, seu escopo abrange um grande número de áreas, desde as mais gerais, como aprendizado, percepção e raciocínio, até campos mais específicos, como jogar xadrez, escrever poesias, provar teoremas matemáticos, diagnosticar doenças ou dirigir veículos de forma autônoma.

Entre os diversos ramos da IA, dois se destacam pela sua importância: o aprendizado de máquina e o aprendizado profundo. O aprendizado de máquina é a técnica que permite a um sistema melhorar seu desempenho ao aprender a partir de um conjunto de dados por meio de métodos computacionais. Seu principal objetivo é desenvolver algoritmos de aprendizado que constroem modelos capazes de fazer previsões sobre novas observações (Zhou, 2021). Já o aprendizado profundo é um subcampo do aprendizado de máquina que possibilita a computadores criar conceitos complexos a partir de conceitos mais simples. Ele permite que computadores aprendam a representar o mundo como uma hierarquia aninhada de representações abstratas, determinadas com base em representações menos abstratas (Goodfellow; Bengio; Courville, 2016).

2.3 Large Language Models (LLMs)

LLMs são modelos de linguagens avançados, caracterizados por uma enorme quantidade de parâmetros e por uma excelente capacidade de aprendizado (Chang *et al*., 2024). Geralmente, referem-se a modelos de linguagens *Transformer*, que contêm centenas de bilhões

de parâmetros ou mais e que são treinados em conjuntos massivos de dados de texto. Modelos como GPT-3 e o Llama são notáveis exemplos que exibem grande capacidade de entender linguagem natural e resolver problemas complexos por meio da geração de texto (Zhao *et al.*, 2023). A Tabela 1 apresenta uma comparação entre o aprendizado de máquina tradicional, o aprendizado profundo e os LLMs.

Tabela 1 - Comparação entre Aprendizado de Máquina Tradicional, Aprendizado Profundo e LLMs

Comparação	**Aprendizado de Máquina Tradicional**	**Aprendizado Profundo**	**LLMs**
Volume de Dados de Treinamento	Elevado	Elevado	Muito Elevado
Engenharia de Atributos	Manual	Automatizada	Automatizada
Complexidade do Modelo	Limitada	Complexa	Muito Complexa
Interpretabilidade	Alta	Baixa	Muito Baixa
Desempenho	Moderado	Alto	Muito Alto
Requisitos de Hardware	Baixo	Elevado	Muito Elevado

Fonte: Adaptado de Chang et al. (2024).

Segundo Chang *et al*. (2024), uma das formas mais comuns de interagir com os LLMs é por meio da Engenharia de *Prompts*, na qual o usuário elabora e fornece instruções textuais específicas (*prompts*) para guiar o modelo na elaboração de respostas e na realização de tarefas específicas. Uma característica fundamental desses modelos é a aprendizagem em contexto, em que o LLM é treinado para gerar respostas com base em um contexto ou *prompt* específico, permitindo a geração de respostas mais coerentes e relevantes, o que torna esses modelos adequados para aplicações interativas e conversacionais.

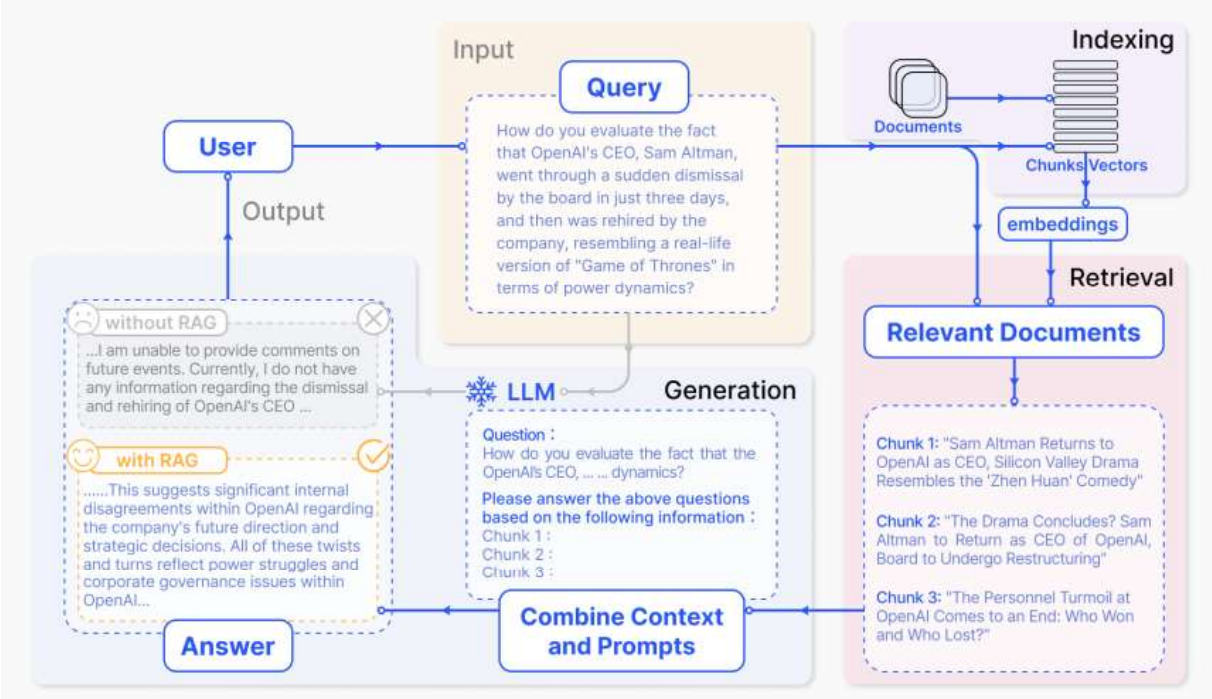
2.4 Retrieval-Augmented Generation (RAG)

Apesar de os LLMs mostrarem um grande poder, ainda encontramos vários desafios que dificultam o seu desenvolvimento, tais como: o problema de alucinações, tendência dos LLMs de gerar respostas coerentes e fluentes, mas incorretas; a dificuldade de atualização do conhecimento, é necessário retreinar ou ajustar os modelos com novos dados para atualizar o conhecimento da sua memória interna; e a falta de perícia em domínios específicos, treinar o LLM para uma área específica exige uma mão de obra considerável para coleta de dados. Diante

dessas limitações, foi proposta a utilização de uma base de conhecimento externa, uma técnica chamada *Retrieval-Augmented Generation* (Wu *et al*., 2024).

RAG é uma técnica de IA que aprimora os LLMs a partir da recuperação de trechos de documentos importantes de uma base externa de conhecimento, utilizando cálculo de similaridade semântica. Essa referência ao conhecimento externo ajuda a reduzir a produção de conteúdos incorretos pelo modelo, principalmente em tarefas de domínios específicos ou que exigem conhecimento intensivo. O processo RAG funciona em 3 passos: *Indexing*, dividir documentos em trechos, codificá-los em vetores (*embeddings*) e armazená-los em um banco de dados vetorial; *Retrieval*, recuperar os *Top* k trechos mais relevantes para a pergunta de acordo com a similaridade semântica; *Generation*, fornecer a pergunta original e os trechos recuperados juntos ao LLM para gerar a resposta final (Gao *et al*., 2023). A figura 1 ilustra o processo RAG quando um usuário realiza uma pergunta ao modelo.

Figura 1 - Processo RAG aplicado à resposta a perguntas



Fonte: Gao *et al.* (2023).

3 TRABALHOS RELACIONADOS

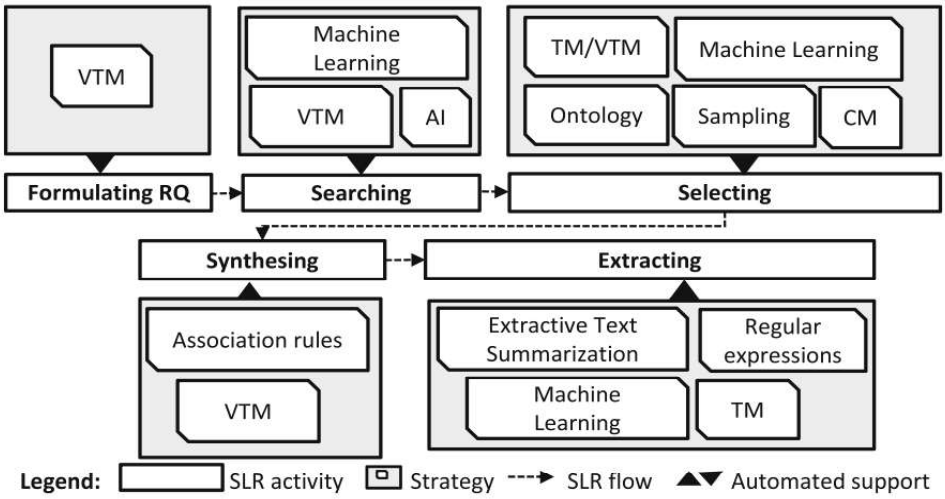
Felizardo e Carver (2020) conduziram um estudo com o objetivo de oferecer uma visão geral das estratégias desenvolvidas por pesquisadores para automatizar o processo de Revisão Sistemática da Literatura na área de Engenharia de Software. Como metodologia, utilizaram uma busca sistemática por meio das técnicas de *backward* e *forward snowballing* em 19 artigos científicos sobre automação de RSL em ES. Os autores identificaram cinco atividades de RSL com maior potencial para automatização: desenvolver o protocolo, buscar evidências, selecionar estudos relevantes, extrair dados e sintetizar as evidências.

As estratégias de automações encontradas pelos autores para cada atividade são as seguintes:

1. Desenvolvimento do Protocolo: automações direcionadas à formulação da pergunta de pesquisa, usando *Visual Text Mining* (VTM).
2. Busca por Evidências: automações para a criação e ajuste de *strings* de busca, utilizando VTM e IA (algoritmo de *hill climbing*); e uso de aprendizado de máquina para ajudar na execução da própria busca.
3. Seleção de Estudos Relevantes: estratégias com *Text Mining* (TM) e VTM para filtrar estudos relevantes na primeira fase de seleção (leitura de títulos e *abstracts*); e aprendizado de máquina para identificação e classificação de estudos relevantes.
4. Extração de Dados: procedimentos como *Extractive Text Summarization* para dividir o documento em tópicos e escolher as frases mais importantes para cada um deles; uso de expressões regulares para extração direta de texto do documento; e aplicação de TM e aprendizado de máquina para identificar cabeçalho de seções relevantes do documento por meio de análise estatística.
5. Síntese de Evidências: estratégias como VTM para categorização e classificação em um mapeamento sistemático; e regras de associação para extração de múltiplos padrões entre os atributos de uma coleção de estudos.

A Figura 2 ilustra as estratégias encontradas para cada atividade.

Figura 2 - Estratégias de automação encontradas para auxiliar as atividades de uma RSL



Fonte: Felizardo e Carver (2020, p. 344)

O mesmo estudo aponta que o desenvolvimento de ferramentas de automação para RSL em ES ainda tem sido vagaroso e fragmentado. Para atingir o potencial de automatização, é necessário um esforço colaborativo entre pesquisadores e as ferramentas de automação precisam ser capazes de serem interoperáveis e de trocar dados. Nesse sentido, considerando os avanços recentes dos modelos de aprendizado de máquina e, em especial, a disponibilidade de LLMs de código aberto, este trabalho tem como objetivo utilizar um LLM local com RAG para auxiliar na fase de extração de dados de uma RSL.

Burguer *et al*. (2023) realizaram um estudo discutindo a atual importância da IA em pesquisas acadêmicas e como essa tecnologia pode melhorar vários métodos de pesquisa. Para ilustrar isso, os autores conduziram um estudo de caso prático de uma RSL, utilizando o modelo de LLM privado *Generative Pre-trained Transformer* (GPT), na versão 3, da empresa OpenAI. Como resultado, propuseram um guia para a utilização de IA em revisões da literatura, contendo três principais fases:

1. Preparação da IA: esta fase estabelece a necessidade de criar *snapshots* (registros) da interação com a IA. Cada *snapshot* deve documentar integralmente os *prompts* utilizados, as respostas geradas pelo modelo, a data de uso, a versão da ferramenta e os parâmetros configurados. Especificamente, o guia recomenda a criação de três *snapshots* distintos para IA sensível ao contexto: IA Breve, conhece a pergunta da pesquisa, mas não tem acesso aos dados; IA Informada, tem acesso aos dados, porém não realizou uma síntese prévia; e IA Sintetizada, realizou uma síntese dos dados.

2. Uso de IA na Iniciação da Pesquisa: nesta fase, a IA é empregada para auxiliar na formulação precisa e inequívoca da pergunta de pesquisa. Recomenda-se testar a compreensão da IA sobre essa pergunta para validá-la. Em seguida, é criada a *string* de busca seguindo procedimentos metodológicos padrão para estudos de revisão. O protocolo de pesquisa, nesse contexto, é definido como a sequência exata dos *prompts* fornecidos à IA, a qual deve ser refinada até que os resultados obtidos em uma amostra de dados sejam os esperados.
3. Uso de IA na Análise de Dados: esta fase é adaptada à metodologia específica. No contexto de RSLs, o processo envolve:
    a. Seleção de Artigos Candidatos: a IA pode ser usada para realizar uma leitura completa de cada artigo candidato.
    b. Extração de Dados: primeiramente, a IA é utilizada para gerar metadados e informações sobre os artigos (periódico, número de citações, data de publicação, metodologia, palavras-chave, resultados). Posteriormente, perguntas específicas são feitas à IA sobre esses mesmos dados, e suas respostas são comparadas.
    c. Síntese de Dados: inicialmente, um modelo é estabelecido para a síntese de acordo com a literatura e, com base nele, os dados são sintetizados.

Em contraste com o modelo proprietário de IA usado no estudo de Burger *et al*. (2023), neste trabalho é utilizado localmente o modelo de código aberto Llama 3.2, disponível pela empresa Meta (2025), para a fase de extração de dados de uma RSL. Alinhando-se às recomendações do guia citado, os prompts e as respostas do modelo são documentados em uma planilha eletrônica disponibilizada como arquivo externo através do link https://zenodo.org/records/17968883. Além disso, seguindo a orientação para a etapa de extração de dados, os dados extraídos manualmente pelos autores de um conjunto de artigos de RSL foram comparados com os dados gerados pelo modelo de IA a partir desses mesmos artigos.

Felizardo *et al*. (2024a) conduziram um estudo com o objetivo de avaliar a acurácia do uso do ChatGPT–4.0 na fase de seleção de estudos (títulos, *abstracts* e palavras-chave) de uma RSL. Os autores utilizaram duas RSLs previamente publicadas como amostra, elas tinham todos os dados e detalhes necessários para a realização do estudo, como critérios de inclusão e exclusão e a lista completa de estudos retornados, aceitos e rejeitados. Nesse experimento, o ChatGPT–4.0 foi empregado na fase de triagem de estudos nas duas RSLs. Os resultados da seleção feita pelo modelo foram então comparados com os resultados obtidos pelos

pesquisadores (estudos incluídos e excluídos) em cada uma das revisões. A acurácia do modelo foi calculada utilizando a seguinte fórmula:

$$\frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

Para o cálculo, foram adotadas as seguintes definições:

- Verdadeiros Positivos (TP): número de artigos que o ChatGPT classificou para inclusão, em concordância com a classificação dos pesquisadores.
- Verdadeiros Negativos (TN): número de artigos que o ChatGPT classificou para exclusão, em concordância com a classificação dos pesquisadores.
- Falsos Positivos (FP): número de artigos que o ChatGPT classificou para inclusão, mas foram excluídos pelos pesquisadores.
- Falsos Negativos (FN): número de artigos que o ChatGPT classificou para exclusão, mas foram incluídos pelos pesquisadores.

Como resultado, o modelo atingiu uma acurácia de 75,3% para a primeira RSL (TP=48, TN=53, FP=17, FN=16) e 86,1% para a segunda RSL (TP=113, TN=273, FP=27, FN=35). Os autores concluíram que o ChatGPT pode não fornecer classificações suficientemente precisas para a seleção de estudos, destacando alguns problemas como classificações incorretas, perda de evidências com a exclusão de artigos importantes (falsos negativos) e fragilidade de *prompts*, uma vez que os LLMs são sensíveis às mudanças sutis na formatação das entradas.

De modo semelhante ao estudo de Felizardo *et al*. (2024a), este trabalho também emprega um LLM, porém local e com RAG, para auxiliar no processo de uma RSL já publicada, a qual contém todos os dados e detalhes disponíveis para este trabalho, como o formulário de extração e os artigos usados. Contudo, enquanto o trabalho anterior focou na fase de seleção de estudos, o enfoque deste trabalho está na fase de extração de dados. Além disso, as respostas do modelo sobre os artigos de RSL do mapeamento também são comparadas com os dados já extraídos pelos autores.

Felizardo *et al*. (2024b) realizaram um estudo com o objetivo de avaliar e fornecer evidências introdutórias sobre como o uso do ChatGPT–4.0 pode auxiliar a fase de extração de dados de um Estudo de Mapeamento Sistemático (EMS) em ES. Os autores realizaram um estudo de prova de conceito estruturado em três etapas: planejamento, coleta de dados e interpretação. Na etapa de planejamento, os autores definiram a meta do estudo, selecionaram

o EMS que serviria de base e definiram a estratégia de coleta de dados e o modo de calcular a acurácia. Na etapa de coleta de dados, os pesquisadores prepararam os dados do EMS para serem processados pelo ChatGPT–4.0, elaboraram e executaram os *prompts*, documentaram os dados e, por fim, calcularam a acurácia do modelo com a seguinte fórmula:

$$\frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (2)$$

Para o cálculo, foram adotadas as seguintes definições:

- Verdadeiros Positivos (TP): número de dados que o ChatGPT–4.0 extraiu corretamente.
- Verdadeiros Negativos (TN): número de dados que o modelo identificou, corretamente, como indisponíveis no texto completo do artigo.
- Falsos Positivos (FP): número de dados que foram incorretamente gerados pelo modelo (alucinações).
- Falsos Negativos (FN): número de dados que estavam presentes nos artigos, mas não foram identificados pelo ChatGPT–4.0.

Por fim, os pesquisadores realizaram uma análise comparando os resultados produzidos pelo ChatGPT com os produzidos manualmente. Como resultado, o modelo alcançou uma acurácia de 87,83%, sendo os falsos positivos (alucinações) o tipo de erro mais frequente. Apesar das evidências iniciais indicarem um bom potencial para o uso do ChatGPT para a fase de extração de dados semiautomatizada, os pesquisadores não recomendam a substituição completa da extração feita manualmente pela extração feita com o modelo.

Assim como o estudo de Felizardo *et al*. (2024b), este trabalho emprega um LLM para auxiliar na fase de extração de dados de uma RSL já publicada com os dados de extração dos artigos já dispostos pelos autores. No entanto, este trabalho adota o modelo de código aberto Llama 3.2, que é mais leve (suporta apenas texto), possui menos parâmetros que o GPT–4 e é executado localmente com a aplicação da técnica RAG. Também, neste trabalho, é realizada uma comparação das respostas geradas pelo LLM com os dados já obtidos manualmente pelos pesquisadores para a obtenção do número de respostas corretas, parcialmente corretas e incorretas.

4 METODOLOGIA

As etapas para o desenvolvimento deste estudo foram as seguintes: selecionar um estudo de revisão sistemática; executar um LLM de código aberto localmente utilizando os textos completos dos artigos do estudo selecionado como dados; realizar perguntas para o modelo de IA de acordo com o formulário de extração do estudo feito pelos pesquisadores; avaliar as respostas obtidas pelo modelo comparando com os dados já obtidos de cada artigo pelos pesquisadores do estudo.

4.1 Seleção do estudo

Por conveniência de obter todos os artefatos necessários para a execução deste trabalho, foi selecionado como estudo base o mapeamento sistemático "*On the need to update systematic literature reviews*" (Nepomuceno; Soares, 2019). Neste trabalho, os autores apresentam um panorama de como as RSLs estão sendo atualizadas e o que seus pesquisadores pensam sobre atualizações de RSLs. Para isso, os autores realizaram um estudo de mapeamento sistemático sobre atualizações de RSLs e um *survey* com pesquisadores em Engenharia de Software Baseada em Evidências que publicaram suas RSLs entre 2011 e 2015. Como resultado, foram inseridos 22 artigos no mapeamento sistemático e obtidas 28 respostas no *survey*.

Para a condução deste trabalho, um dos autores do mapeamento disponibilizou todos os materiais necessários: o formulário de extração de dados dos artigos com as respostas já obtidas por eles e o *corpus* completo dos 22 artigos em formato PDF (Portable Document Format).

4.2 Execução de um LLM local com os dados dos artigos

Por adequação aos recursos restritos de hardware disponíveis para execução local do modelo, ou seja, necessidade de um LLM que consumisse pouco poder computacional, ser um modelo de código aberto e ter uma arquitetura mais leve e especializada apenas em processamento de texto, optou-se pelo LLM Llama 3.2 (3B*, Lightweight Model*) como o modelo a ser usado para a extração dos dados dos artigos do estudo. Além disso, selecionou-se o software Ollama (2023), também gratuito e de código aberto, para *download* e execução local do modelo Llama 3.2 em um sistema operacional Windows. A escolha de executar o modelo localmente foi devido a evitar questões de *copyright* dos artigos envolvidos no estudo ao usá-los em um LLM *online*, pois há o risco de se utilizar esses artigos protegidos por direitos

autorais no treinamento de modelos de IA disponíveis *online* sem consentimento dos autores, gerando problemas legais.

Com o propósito de fornecer o conteúdo textual completo dos 22 artigos do mapeamento sistemático ao modelo e realizar consultas sobre cada um deles, foi desenvolvido um *chatbot* utilizando a linguagem Java e o *framework* Spring Boot (2025), cujo código está disponível *online* no endereço https://github.com/nhsneto/tcc-nelson-neto. Nesse sistema, todos os artigos em formato PDF são primeiramente convertidos em arquivos de texto (formato .txt). Em seguida, esses textos são fornecidos ao modelo de IA por meio da técnica RAG. Essa técnica permite que os LLMs acessem e usem informações de bases de dados externas, ou seja, dados que não fizeram parte de seu treinamento original.

Para a implementação da técnica RAG e para a interação com o modelo, foi utilizada a biblioteca LangChain4j (2025). No processo de RAG, o conteúdo textual de todos os arquivos (em formato .txt) foi convertido em *embeddings*, representações vetoriais de dados amplamente utilizadas em modelos de aprendizado de máquina. Essa conversão foi realizada pelo codificador de texto nomic-embed-text (2025), disponibilizado pelo Ollama. Os *embeddings* gerados foram então armazenados em um banco de dados SQLite (2025) para posterior consulta do LLM. Em relação à interação com o modelo, a biblioteca Langchain4j fornece uma interface de alto nível que abstrai a comunicação com o Ollama, facilitando a interação com o LLM, a qual é feita através da arquitetura cliente-servidor.

Para a execução do *chatbot*, é necessário que o Ollama esteja em execução no endereço de rede local padrão (http://localhost:11434/) e que o modelo Llama 3.2 já esteja instalado e disponível nele. Ao ser iniciado, o c*hatbot* executa automaticamente as seguintes etapas: converte todos os arquivos PDF localizados na pasta designada (./artigos) para arquivos de texto (formato .txt); gera os *embeddings* de todo o conteúdo textual convertido, utilizando o codificador nomic-embed-text; e armazena os *embeddings* gerados em um banco de dados SQLite para posterior consulta do modelo.

Após a inicialização, o programa configura o valor do parâmetro de temperatura do modelo para 0.1. Esse parâmetro do LLM é responsável por controlar a aleatoriedade das respostas geradas por ele. Quanto maior o valor, mais criativas e diversas são as respostas, portanto, visando a obtenção de respostas mais determinísticas e concisas neste trabalho, foi escolhido o valor 0.1 para esse parâmetro. Em seguida, o programa fornece ao LLM a seguinte instrução inicial (*system prompt* ou *system message*) para definir seu comportamento: “Você é um assistente inteligente especializado em responder a perguntas com base no conteúdo de artigos científicos fornecidos. Você possui 22 artigos científicos de revisão sistemática e seu

objetivo é responder questões sobre eles de forma precisa e concisa". Por fim, a *Command Line Interface* (CLI) do *chatbot* é apresentada, permitindo que o usuário formule perguntas ao LLM local e receba suas respostas diretamente no terminal.

4.3 Questionamento ao LLM de acordo com o formulário de extração

Os *prompts* utilizados foram elaborados com base nas perguntas contidas no formulário de extração de dados do estudo de Nepomuceno e Soares (2019). Esse formulário contém 16 perguntas para cada um dos 22 artigos presentes no estudo. A Tabela 1 apresenta a relação entre essas perguntas do formulário e os *prompts* iniciais estruturados para consulta ao LLM. Antes de cada consulta, a palavra "Título" contida no *prompt* era substituída pelo título específico do artigo em questão.

Tabela 2 - Perguntas do formulário e os prompts iniciais de cada artigo

Pergunta	**Prompt inicial**
Título da Publicação	Informado no prompt
Ano da publicação	Sobre o artigo "Título", qual é o ano da sua publicação?
Lista dos autores (separados por vírgula)	Sobre o artigo "Título", quais são seus autores? Liste-os em uma frase separados por vírgulas.
País	Sobre o artigo "Título", qual é o país de origem desse artigo?
Qual o tipo de abordagem sistemática?	Sobre o artigo "Título", qual o tipo de abordagem sistemática utilizada? Responda apenas o nome da abordagem sistemática utilizada, sem descrevê-la.
A atualização mudou algum resultado do estudo original?	Sobre o artigo "Título", a atualização mudou algum resultado do estudo original? Responda apenas sim ou não.
Qual foi o tempo decorrido entre a atualização e o estudo original correspondente (publicação)?	Sobre o artigo "Título", qual foi o tempo decorrido entre a atualização e o estudo original correspondente (publicação)?
Os artefatos foram alterados em relação ao estudo anterior?	Artefatos de uma RSL (Revisão Sistemática da Literatura) são a documentação utilizada para elaborar e conduzir a RSL, o conjunto de resultados de estudos primários obtidos a partir do procedimento de busca, quaisquer formulários de extração utilizados para armazenar dados dos estudos primários (nas etapas de avaliação da qualidade e extração de dados), e os relatórios da RSL. Sobre o artigo "Título", os artefatos desse artigo

	foram alterados em relação ao seu estudo anterior? Responda apenas sim ou não.
Quais foram os artefatos alterados? (NA, se nenhum.)	Artefatos de uma RSL (Revisão Sistemática da Literatura) são a documentação utilizada para elaborar e conduzir a RSL, o conjunto de resultados de estudos primários obtidos a partir do procedimento de busca, quaisquer formulários de extração utilizados para armazenar dados dos estudos primários (nas etapas de avaliação da qualidade e extração de dados), e os relatórios da RSL. Sobre o artigo "Título", quais foram os artefatos alterados? Responda "NA", se nenhum.
O autor relata dificuldades em encontrar os artefatos do estudo original?	Sobre o artigo "Título", o autor relata dificuldades em encontrar os artefatos do estudo original? Responda apenas sim ou não.
Quais dificuldades são relatadas (NA, se nenhuma.)	Sobre o artigo "Título", quais são as dificuldades relatadas em encontrar os artefatos do estudo original? Responda apenas "NA" se nenhuma dificuldade foi relatada.
O autor disponibilizou os artefatos da atualização? (disponibilizou online)	Sobre o artigo "Título", o autor disponibilizou os artefatos da atualização online?
Quem realizou o estudo original?	Sobre o artigo "Título", quem realizou o estudo original?
Houve algum contato com o autor original?	Sobre o artigo "Título", houve algum contato com o autor original?
Quais são os motivos para atualizar?	Sobre o artigo "Título", quais são os motivos para atualizar?
Referência do estudo original	Sobre o artigo "Título", qual é a referência do estudo original? Responda usando as normas APA.

Fonte: do próprio autor.

Todos os *prompts* utilizados e as respectivas respostas geradas pelo modelo foram documentados em uma planilha eletrônica, disponível publicamente no repositório de pesquisas Zenodo (2013) através do endereço: https://zenodo.org/records/17968883. Para cada um dos 22 artigos, foi criada uma planilha incluindo: as 16 perguntas do formulário; os respectivos *prompts* de acordo com cada pergunta; as respostas obtidas pelo LLM; e, se existir, os *prompts* de iteração e as respostas correspondentes. Todas as interações com o modelo foram conduzidas por meio da CLI do *chatbot*.

As consultas ao LLM seguiram a estrutura e a ordem estabelecidas no formulário de extração de dados do estudo. Nesse formulário, cada um dos artigos foi identificado como *Study*

(S) e recebeu uma numeração sequencial de S01 a S22. Dessa forma, as consultas foram feitas seguindo tanto a ordem dos artigos, isto é, do S01 ao S22, quanto a ordem da sequência das 16 perguntas contidas no formulário.

Para cada um dos 22 artigos, foi seguido o mesmo procedimento padrão: o *chatbot* era iniciado; as perguntas (e eventuais iterações) eram submetidas ao LLM; todos os *prompts* e as respostas correspondentes eram documentados na planilha eletrônica; e, ao final, o *chatbot* era encerrado e reinicializado para dar início ao processamento do artigo seguinte.

4.4 Avaliação das respostas obtidas pelo modelo

Ao final do processo de coleta, foi realizada uma avaliação das respostas obtidas pelo modelo em relação a cada um dos 22 artigos. Para cada uma das perguntas, foi feita uma comparação das respostas obtidas pelos pesquisadores do estudo com as obtidas pelo LLM. A classificação das respostas do modelo seguiu três categorias: corretas, parcialmente corretas e incorretas. As respostas corretas correspondem às que são idênticas ou semanticamente equivalentes às dos autores do mapeamento; as parcialmente corretas correspondem às que contêm ao menos um elemento ou afirmação correta em relação às dos autores; e as incorretas correspondem às que não apresentam semelhança ou equivalência com as dos autores. Para facilitar a visualização e análise nas planilhas, as células contendo as respostas do modelo foram destacadas com cores de fundo, de acordo com a seguinte legenda: verde claro para respostas corretas, amarelo claro para parcialmente corretas e vermelho claro para incorretas.

5 RESULTADOS

Foi obtido o total de 330 respostas do modelo: 137 corretas (41,52%), 28 parcialmente corretas (8,48%) e 165 incorretas (50%). A tabela 3 resume a contagem de respostas corretas, parcialmente corretas e incorretas para cada um dos 22 artigos. A tabela 4 apresenta a contagem de respostas corretas, parcialmente corretas e incorretas para cada uma das 15 perguntas do formulário. A tabela 5 indica quais artigos receberam respostas corretas do modelo em cada pergunta. A tabela 6 mostra quais artigos receberam respostas parcialmente corretas do LLM em cada uma das perguntas. A tabela 7 indica quais artigos receberam respostas incorretas do modelo em cada pergunta. Todos esses resultados estão disponíveis em uma planilha eletrônica no endereço https://zenodo.org/records/17968883.

Tabela 3 - Número de respostas corretas, parcialmente corretas e incorretas por artigo

Artigo	N° de corretas	Nº de parcialmente corretas	Nº de incorretas
S01	10	1	4
S02	4	2	9
S03	4	0	11
S04	5	1	9
S05	4	2	9
S06	9	2	4
S07	8	0	7
S08	7	0	8
S09	7	3	5
S10	5	1	9
S11	5	3	7
S12	8	2	5
S13	8	0	7
S14	7	4	4
S15	9	1	5
S16	6	1	8
S17	8	0	7
S18	5	1	9
S19	2	2	11
S20	4	1	10
S21	7	1	7
S22	5	0	10
Total	137	28	165
Total geral	330		

Fonte: do próprio autor.

Tabela 4 - Número de respostas corretas, parcialmente corretas e incorretas por pergunta

Pergunta	**Nº de corretas**	**Nº de parcialmente corretas**	**Nº de incorretas**
Ano da publicação	7	2	13
Lista dos autores (separados por vírgula)	18	2	2
País	13	7	2
Qual o tipo de abordagem sistemática?	14	2	6
A atualização mudou algum resultado do estudo original?	11	0	11
Qual foi o tempo decorrido entre a atualização e o estudo original correspondente (publicação)?	0	5	17
Os artefatos foram alterados em relação ao estudo anterior?	10	0	12
Quais foram os artefatos alterados? (NA, se nenhum.)	6	0	16
O autor relata dificuldades em encontrar os artefatos do estudo original?	14	0	8
Quais são as dificuldades relatadas? (NA, se nenhuma.)	18	0	4
O autor disponibilizou os artefatos da atualização?	7	0	15
Quem realizou o estudo original?	4	5	13
Houve algum contato com o autor original?	2	2	18
Quais são os motivos para atualizar?	11	0	11
Referência do Estudo Original	2	3	17
Total	137	28	165

Total geral	330

Fonte: do próprio autor.

Tabela 5 - Artigos que o LLM respondeu corretamente por pergunta

Pergunta	**Artigos**
Ano da publicação	S01, S03, S04, S10, S16, S17, S20
Lista dos autores (separados por vírgula)	S01, S04, S05, S06, S07, S08, S09, S10, S11, S12, S13, S14, S15, S16, S17, S18, S21, S22
País	S01, S02, S06, S07, S08, S12, S13, S14, S16, S17, S18, S21, S22
Qual o tipo de abordagem sistemática?	S01, S04, S05, S06, S07, S08, S09, S10, S12, S13, S15, S17, S21, S22
A atualização mudou algum resultado do estudo original?	S01, S02, S04, S06, S08, S09, S10, S11, S15, S16, S17
Qual foi o tempo decorrido entre a atualização e o estudo original correspondente (publicação)?	
Os artefatos foram alterados em relação ao estudo anterior?	S01, S03, S06, S09, S14, S15, S16, S17, S18, S21
Quais foram os artefatos alterados? (NA, se nenhum.)	S07, S08, S11, S12, S13, S22
O autor relata dificuldades em encontrar os artefatos do estudo original?	S01, S02, S04, S06, S07, S08, S12, S13, S14, S15, S18, S19, S20, S21
Quais são as dificuldades relatadas? (NA, se nenhuma.)	S01, S02, S03, S05, S06, S07, S08, S10, S11, S12, S13, S14, S15, S17, S18, S19, S20, S21
O autor disponibilizou os artefatos da atualização?	S07, S09, S11, S13, S14, S15, S22
Quem realizou o estudo original?	S01, S07, S12, S13
Houve algum contato com o autor original?	S01, S09
Quais são os motivos para atualizar?	S03, S05, S06, S09, S12, S14, S15, S16, S17, S20, S21
Referência do Estudo Original	S06, S15

Fonte: do próprio autor.

Tabela 6 - Artigos que o LLM respondeu parcialmente corretamente por pergunta

Pergunta	Artigos
Ano da publicação	S11, S12
Lista dos autores (separados por vírgula)	S02, S20
País	S04, S05, S09, S10, S11, S15, S19
Qual o tipo de abordagem sistemática?	S16, S19
A atualização mudou algum resultado do estudo original?	
Qual foi o tempo decorrido entre a atualização e o estudo original correspondente (publicação)?	S01, S06, S12, S14, S21
Os artefatos foram alterados em relação ao estudo anterior?	
Quais foram os artefatos alterados? (NA, se nenhum.)	
O autor relata dificuldades em encontrar os artefatos do estudo original?	
Quais são as dificuldades relatadas? (NA, se nenhuma.)	
O autor disponibilizou os artefatos da atualização?	
Quem realizou o estudo original?	S02, S09, S11, S14, S18
Houve algum contato com o autor original?	S06, S14
Quais são os motivos para atualizar?	
Referência do Estudo Original	S05, S09, S14

Fonte: do próprio autor.

Tabela 7 - Artigos que o LLM respondeu incorretamente por pergunta

Pergunta	Artigos
Ano da publicação	S02, S05, S06, S07, S08, S09, S13, S14, S15, S18, S19, S21, S22
Lista dos autores (separados por vírgula)	S03, S19
País	S03, S20
Qual o tipo de abordagem sistemática?	S02, S03, S11, S14, S18, S20
A atualização mudou algum resultado do estudo original?	S03, S05, S07, S12, S13, S14, S18, S19, S20, S21, S22
Qual foi o tempo decorrido entre a atualização e o estudo original correspondente (publicação)?	S02, S03, S04, S05, S07, S08, S09, S10, S11, S13, S15, S16, S17, S18, S19, S20, S22
Os artefatos foram alterados em relação ao estudo anterior?	S02, S04, S05, S07, S08, S10, S11, S12, S13, S19, S20, S22
Quais foram os artefatos alterados? (NA, se nenhum.)	S01, S02, S03, S04, S05, S06, S09, S10, S14, S15, S16, S17, S18, S19, S20, S21
O autor relata dificuldades em encontrar os artefatos do estudo original?	S03, S05, S09, S10, S11, S16, S17, S22
Quais são as dificuldades relatadas? (NA, se nenhuma.)	S04, S09, S16, S22

O autor disponibilizou os artefatos da atualização?	S01, S02, S03, S04, S05, S06, S08, S10, S12, S16, S17, S18, S19, S20, S21
Quem realizou o estudo original?	S03, S04, S05, S06, S08, S10, S15, S16, S17, S19, S20, S21, S22
Houve algum contato com o autor original?	S02, S03, S04, S05, S07, S08, S10, S11, S12, S13, S15, S16, S17, S18, S19, S20, S21, S22
Quais são os motivos para atualizar?	S01, S02, S04, S07, S08, S10, S11, S13, S18, S19, S22
Referência do Estudo Original	S01, S02, S03, S04, S07, S08, S10, S11, S12, S13, S16, S17, S18, S19, S20, S21, S22

Fonte: do próprio autor.

6 DISCUSSÕES

A concepção inicial do *chatbot* priorizava facilitar o processamento de um grande volume de artigos, característico de revisões sistemáticas. Para isso, optou-se por converter todos os PDFs para texto de uma única vez e realizar as consultas em sequência, inserindo o título do artigo em cada *prompt* para fornecer o contexto adequado, ao invés de converter e consultar cada artigo de forma isolada, o que seria uma abordagem mais demorada.

Entretanto, durante a execução, observou-se que o LLM começou a apresentar um problema: em suas respostas, o modelo incluía informações de outros artigos, gerando dados inconsistentes com o artigo sendo consultado no momento. Para mitigar esse problema, o procedimento foi modificado. Decidiu-se encerrar e reinicializar o *chatbot* antes de cada nova consulta a um artigo diferente. Essa medida reduziu significativamente a ocorrência de respostas confusas por parte do modelo, porém limitou a facilidade de uso, uma vez que é necessária a intervenção manual a cada ciclo.

Neste trabalho, adotou-se uma abordagem direta na interação com o LLM. O contexto fornecido ao modelo foi intencionalmente restrito, limitando-se a informações gerais, como a natureza dos documentos (artigos científicos de revisão sistemática) e sua quantidade (22), ambos informados na instrução inicial, e artefatos, informados em dois dos *prompts* iniciais. A intenção era fornecer um conhecimento básico para contextualizar as perguntas, sem detalhar informações específicas, como o título do estudo original que cada artigo atualiza.

Apesar dessa restrição contextual poder aumentar a dificuldade da tarefa para o modelo, ele continuou a gerar respostas incorretas mesmo quando os dados solicitados eram fornecidos explicitamente nos *prompts* de interação. Em 14 das 19 iterações em que houve fornecimento direto de dados, o modelo mostrou-se incapaz de gerar uma resposta correta, indicando uma limitação que vai além da simples falta de contexto.

Conforme os dados da Tabela 3, o modelo alcançou 137 respostas corretas. Pode-se inferir que a maior parte desses acertos ocorreu em perguntas que demandavam respostas objetivas, tais como confirmações binárias (sim/não) ou a extração de dados bibliográficos diretamente citados nos textos. Essa tendência é confirmada pela Tabela 4, onde encontramos as seguintes perguntas com os maiores números de acertos: “Lista dos autores (separados por vírgula)” 18 acertos; “País” 13 acertos; “Qual o tipo de abordagem sistemática?” 14 acertos; “O autor relata dificuldades em encontrar os artefatos do estudo original?” 14 acertos; e “Quais são as dificuldades relatadas? (NA, se nenhuma.)” 18 acertos. É importante destacar que as respostas da pergunta sobre o tipo de abordagem sistemática foram influenciadas pela instrução

inicial (*system message*) fornecida ao modelo, que explicitamente informava que todos os 22 artigos tratavam de revisões sistemáticas.

A Tabela 4 revela que a única pergunta para a qual o modelo não gerou nenhuma resposta correta foi: "Qual o tempo decorrido entre a atualização e o estudo original correspondente (publicação)?". Esse resultado evidencia uma grande limitação do LLM na tarefa proposta. A incapacidade de acertar sequer uma instância dessa pergunta sugere grandes dificuldades em: identificar corretamente o estudo original ao qual um artigo de atualização se refere; extrair e relacionar as datas de publicação de ambos os estudos; e executar o cálculo temporal simples decorrente dessas informações. Portanto, para questões que envolvem esse tipo de raciocínio relacional e matemático básico entre os estudos, o modelo local com RAG não se mostrou um auxílio confiável para o pesquisador que conduz a revisão.

A Tabela 5 indica que o modelo forneceu a referência correta do estudo original em apenas dois artigos: S06 e S15. No entanto, mesmo nesses casos de acerto, as respostas não seguiram o formato padrão APA (American Psychological Association) solicitado no *prompt*: para o artigo S06, retornou Beecham et al. (2008); para o S15, retornou [13]. Em tentativas subsequentes de iteração, nas quais se solicitou explicitamente o título dos estudos originais, o LLM local com RAG mostrou-se incapaz de extrair essa informação corretamente. Diante da inconsistência e da falta de conformidade com o padrão solicitado, foi necessária a intervenção manual para localizar e verificar as referências nos textos dos artigos S06 e S15. Essa necessidade de intervenção e validação humana representa uma falha na automatização da tarefa, que era o objetivo a ser facilitado pelo modelo.

O modelo gerou 28 respostas consideradas parcialmente corretas. Conforme a Tabela 6, as perguntas que mais receberam respostas com essa classificação foram: “País” 7 parcialmente corretas; “Qual foi o tempo decorrido entre a atualização e o estudo original correspondente (publicação)?” 5 parcialmente Corretas; e “Quem realizou o estudo original?” 5 parcialmente corretas. Uma análise mais detalhada das respostas parcialmente corretas sobre “País” revela padrões de erro consistentes: omissão de países, em 4 das 7 respostas, o modelo deixou de citar todos os países de afiliação dos autores, principalmente em artigos escritos por colaboradores internacionais; inclusão incorreta (alucinação), em 2 das 7 respostas, o modelo informou países que não estavam associados ao estudo; confusão entre países similares, em uma resposta específica (artigo S09), houve a confusão entre "Áustria" e "Austrália", demonstrando uma falha na compreensão ou recuperação precisa do texto.

Para a pergunta “Qual foi o tempo decorrido entre a atualização e o estudo original correspondente (publicação)?”, o modelo conseguiu inferir ou identificar corretamente o ano

de publicação do estudo original em 4 das 5 respostas. No entanto, apresentou inconsistências: calculou erroneamente a duração de 3 dessas 4 respostas e identificou incorretamente o ano de publicação do próprio artigo de atualização em 4 das 5 respostas. Esse padrão sugere que o modelo local com RAG teve dificuldade em integrar e relacionar informações temporais presentes no texto para realizar um cálculo simples, mesmo quando parte dos dados (ano do estudo original) é identificada corretamente. Em relação à pergunta “Quem realizou o estudo original?”, ele demonstrou dificuldade em distinguir autores de estudos diferentes. Em 3 das 5 respostas classificadas como parcialmente corretas, ele incluiu autores que não pertenciam ao estudo original (alucinação) ou repetiu os autores do artigo de atualização.

Conforme as Tabelas 3 e 4, o LLM produziu um total de 165 respostas incorretas, o que corresponde à metade (50%) de todas as respostas geradas. As perguntas que apresentaram o maior volume de respostas incorretas foram: “Qual o tempo decorrido entre a atualização e o estudo original correspondente (publicação)?” 17 incorretas; “Houve algum contato com o autor original?” 18 inc.; “Quais foram os artefatos alterados?” (NA, se nenhum.) 16 inc.; e “Referência do estudo original” 17 incorretas.

O modelo local com RAG demonstrou uma grande limitação de identificar o estudo original ao qual o artigo de atualização se refere. Mesmo em iterações em que o título do estudo original foi explicitamente fornecido, ele foi incapaz de utilizar essa informação para calcular corretamente o intervalo de tempo entre os estudos. A pergunta "Houve algum contato com o autor original?" apresentou, como esperado, um alto índice de respostas incorretas (18). Isso ocorreu porque a informação necessária não estava contida nos textos dos artigos, mas sim obtida pelos autores do mapeamento em entrevista (um dado externo e indisponível para o modelo).

Portanto, o experimento serviu para testar os limites da capacidade de inferência e o risco de alucinação do LLM local com RAG em situações de informação ausente. Dos 22 artigos, o modelo acertou 2 (S01 e S09), acertou parcialmente 2 (S06 e S14) e produziu alucinações ou confundiu os autores da atualização com os do estudo original nos 18 restantes. Esse resultado sugere que, para questões cujas respostas dependem exclusivamente de conhecimento externo não fornecido, o modelo local com RAG não foi uma ferramenta confiável.

Para a pergunta “Quais foram os artefatos alterados?”, o LLM local com RAG mostrou-se completamente ineficaz. Ele foi incapaz de identificar corretamente os artefatos em qualquer um dos 15 artigos que reportaram alterações. Além disso, no artigo S05, que não tinha alterações, o modelo informou a existência de artefatos modificados (alucinação), indicando

uma limitação na tarefa de extração desses dados específicos. A respeito da pergunta "Referência do estudo original", ele demonstrou grande dificuldade em diferenciar entre referências dentro do texto. Seus erros (17 no total) seguiram dois padrões principais: alucinação com referências incorretas, a maior parte das respostas incorretas foi composta por citação de outras referências presentes no artigo, mas que não correspondiam ao estudo original; confusão de contexto, em 5 das 17 incorretas, o modelo repetiu a referência do próprio artigo de atualização, evidenciando novamente a dificuldade em isolar a entidade "estudo original" do contexto geral. Diante dessas limitações, o LLM local com RAG não se mostrou uma ferramenta útil para auxiliar o pesquisador na recuperação dessas informações.

Foram realizados 42 inícios de iterações de *prompting* com o LLM local com RAG, nas quais foram alcançadas 10 respostas corretas (23,81%), 4 parcialmente corretas (9,52%) e 28 incorretas (66,67%). Essas interações podem ser categorizadas de acordo com o seu objetivo principal: fornecimento e esclarecimento de dados (21 iterações), fornecer dados ausentes ou detalhar informações contidas em respostas anteriores do modelo, na tentativa de guiá-lo para uma extração correta; confirmação e ajuste (13 iterações), confirmar respostas, eliminar dúvidas do LLM sobre algumas perguntas e solicitar para que ele reformulasse a resposta (13 iterações); e definição e explicação de termos (8 iterações), definir ou explicar termos ou conceitos presentes na pergunta, visando melhorar a compreensão do LLM. O fato de que dois terços (66,67%) das iterações resultaram em respostas incorretas, mesmo após intervenção manual, evidencia a limitação da abordagem iterativa para corrigir os erros do LLM local com RAG, ou seja, não foi suficiente para corrigir a maioria das respostas incorretas.

Um caso particular das limitações do modelo ocorreu com os artigos S20 ("*Empirical evidence about the UML: a systematic literature review*") e S08 ("*What Is the Further Evidence about UML? - A Systematic Literature Review*"). O LLM local com RAG demonstrou grande dificuldade de diferenciar esses dois artigos, que tratam do mesmo tópico (UML) e possuem títulos muito semelhantes. Em suas respostas iniciais, confundiu os conteúdos, atribuindo informações do artigo S08 como respostas para as perguntas sobre o artigo S20.

Para resolver esse problema, foram realizadas iterações em todas as perguntas do artigo S20, nas quais os prompts foram modificados para incluir os nomes de dois de seus autores, visando especificar o artigo-alvo para o modelo. Apesar dessa especificação, o modelo continuou respondendo incorretamente em 3 das 15 perguntas, continuando a fornecer dados do artigo S08. Esse caso indica que, mesmo com o fornecimento dos autores nos *prompts* para identificar o artigo-alvo, o LLM local com RAG pode apresentar dificuldades em isolar e recuperar informações de documentos muito similares.

É importante mencionar algumas limitações deste trabalho. Os resultados obtidos são baseados em um único estudo de mapeamento sistemático contendo apenas 22 artigos de revisões sistemáticas. Além disso, só um modelo de IA foi utilizado, Llama3.2 (3B*, Lightweight Model*), localmente, em uma versão mais simples e com menos parâmetros para se adequar aos recursos de hardware disponíveis. Ademais, é reconhecido que as respostas emitidas pelo LLM, considerando sua natureza estocástica, podem variar e ter influenciado os resultados obtidos. Por fim, foram utilizados dados extraídos por humanos como referência para as comparações das respostas do modelo, o que pode apresentar erros e viés, afetando a precisão dos resultados, porém essa questão foi considerada minimizada, pois dois pesquisadores realizaram a extração de dados do mapeamento sistemático, Soares e Nepomuceno (2019).

7 CONCLUSÃO

Os resultados obtidos deste trabalho indicam que a arquitetura utilizada, através de um LLM local com RAG, não foi capaz de auxiliar o pesquisador significativamente em relação à fase de extração de dados da RSL, tendo êxito em 41,52% das respostas. Porém, ele alcançou o seu maior número de acertos em informações bibliográficas como lista de autores, país e abordagem sistemática do artigo, o que sugere que o LLM local com RAG pode ser utilizado para obter esses dados com mais facilidade na fase de extração de dados em revisões sistemáticas da literatura, seguido de revisão e refinamento humano, diminuindo o esforço do pesquisador e garantindo que os dados extraídos sejam confiáveis.

Dadas as limitações deste estudo, é importante evidenciar que outros trabalhos como este sejam feitos com a finalidade de encontrar resultados com maior índice de acerto do LLM, usando hardware de alto desempenho e utilizando modelos diferentes que sejam mais atuais, poderosos e com maior número de parâmetros, além de prover uma contextualização mais detalhada ao modelo sobre a revisão sistemática escolhida como objeto de estudo e de comparações.

REFERÊNCIAS

ARSLAN, M. et al. A Survey on RAG with LLMs. **Procedia Computer Science**, [S.l.], v. 246, p. 3781-3790, 2024. Disponível em: https://doi.org/10.1016/j.procs.2024.09.178. Acesso em: 26 nov. 2025.

BORAH, R. et al. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. **BMJ Open**, [s.l.], v. 7, n. 2, e012545, 2017. Disponível em: https://doi.org/10.1136/bmjopen-2016-012545. Acesso em: 24 nov. 2025.

BURGER, B. et al. On the use of AI-based tools like ChatGPT to support management research. **European Journal of Innovation Management**, v. 26, n. 7, p. 233-241, 2023. Disponível em: https://doi.org/10.1108/EJIM-02-2023-0156. Acesso em: 12 jul. 2025.

CARVER, J. C. et al. Identifying Barriers to the Systematic Literature Review Process. In: ACM / IEEE INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT, 2013. **Proceedings [...]**. [S. l.]: IEEE, 2013. p. 203-212. DOI: https://doi.org/10.1109/ESEM.2013.28.

CHANG, Yupeng et al. A survey on evaluation of large language models. **ACM transactions on intelligent systems and technology**, v. 15, n. 3, p. 1-45, 2024.

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH; OPENAIRE. **Zenodo**. [S. l.]: CERN, 2013. Disponível em: https://www.zenodo.org/. Acesso em: 17 dez. 2025. DOI: https://doi.org/10.25495/7GXK-RD71.

FELIZARDO, K. R. et al. ChatGPT application in Systematic Literature Reviews in Software Engineering: an evaluation of its accuracy to support the selection activity. In: ACM/IEEE INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT, 18., 2024, Barcelona, Espanha. **Proceedings [...]**. New York, NY, USA: Association for Computing Machinery, 2024a. p. 25-36. Disponível em: https://doi.org/10.1145/3674805.3686666. Acesso em: 16 jul. 2025.

FELIZARDO, K. R. et al. Data extraction for systematic mapping study using a large language model - a proof-of-concept study in software engineering. In: ACM/IEEE INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT, 18., 2024, Barcelona, Espanha. **Proceedings [...]**. New York, NY, USA: Association for Computing Machinery, 2024b. p. 407-413. Disponível em: https://doi.org/10.1145/3674805.3690743. Acesso em: 17 jul. 2025.

FELIZARDO, K. R.; CARVER, J. C. Automating Systematic Literature Review. In: FELDERER, M.; TRAVASSOS, G. (eds.). **Contemporary Empirical Methods in Software Engineering**. Cham: Springer, 2020. p. 327-355. Disponível em: https://doi.org/10.1007/978-3-030-32489-6_12. Acesso em: 11 jul. 2025.

GAO, Yunfan et al. Retrieval-augmented generation for large language models: A survey. **arXiv preprint arXiv:2312.10997**, v. 2, n. 1, 2023.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. Cambridge: MIT Press, 2016. Disponível em: http://www.deeplearningbook.org.

HAENLEIN, Michael; KAPLAN, Andreas. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. **California Management Review**, [S. l.], v. 61, n. 4, p. 5-14, 2019. Disponível em: https://doi.org/10.1177/0008125619864925. Acesso em: 18 dez. 2025.

KITCHENHAM, B. A.; CHARTERS, S. **Guidelines for performing systematic literature reviews in software engineering**. Keele, UK: School of Computer Science and Mathematics, Keele University, 2007. Disponível em: https://legacyfileshare.elsevier.com/promis_misc/525444systematicreviewsguide.pdf. Acesso em: 4 set. 2025.

LANGCHAIN4J. **Langchain4j**. [Biblioteca de software]. 2025. Disponível em: https://docs.langchain4j.dev/. Acesso em: 28 jul. 2025.

LIANG, Weixin et al. Mapping the increasing use of LLMs in scientific papers. **arXiv preprint arXiv:2404.01268**, 2024.

META. **Llama 3.2**. 2025. Disponível em: https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/. Acesso em: 23 jul. 2025.

MICHELSON, M.; REUTER, K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. **Contemporary Clinical Trials Communications**, [s.l.], v. 16, p. 100443, 2019. Disponível em: https://doi.org/10.1016/j.conctc.2019.100443. Acesso em: 24 nov. 2025.

MINAEE, Shervin et al. Large language models: A survey. **arXiv preprint arXiv:2402.06196**, 2024.

NEPOMUCENO, V.; SOARES, S. On the need to update systematic literature reviews. **Information and Software Technology**, v. 109, p. 40-42, 2019. Disponível em: https://doi.org/10.1016/j.infsof.2019.01.005. Acesso em: 18 jul. 2025.

NG, K. K. Y.; MATSUBA, I.; ZHANG, P. C. RAG in Health Care: A Novel Framework for Improving Communication and Decision-Making by Addressing LLM Limitations. **NEJM AI**, [S.l.], v. 2, n. 1, p. AIra2400380, 2025. Disponível em: https://doi.org/10.1056/AIra2400380. Acesso em: 26 nov. 2025.

OLLAMA. **nomic-embed-text**. [Modelo de embedding]. 2025. Disponível em: https://ollama.com/library/nomic-embed-text. Acesso em: 28 jul. 2025.

OLLAMA. **Ollama**. 2023. Disponível em: https://ollama.com/. Acesso em: 23 jul. 2025.

PETERSEN, Kai; VAKKALANKA, Sairam; KUZNIARZ, Ludwik. Guidelines for conducting systematic mapping studies in software engineering: An update. **Information and Software Technology**, v. 64, p. 1-18, 2015. Disponível em: https://doi.org/10.1016/j.infsof.2015.03.007. Acesso em: 5 set. 2025.

RESNIK, D. B.; HOSSEINI, M. The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool. **AI and Ethics**, [s.l.], v. 5, n. 2, p. 1499-1521, 2024. Disponível em: https://doi.org/10.1007/s43681-024-00493-8. Acesso em: 25 nov. 2025.

RUSSELL, Stuart J.; NORVIG, Peter. **Artificial Intelligence: a modern approach**. 4. ed. Boston: Pearson, 2021.

SANTOS, V. dos et al. Towards Sustainability of Systematic Literature Reviews. In: INTERNATIONAL SYMPOSIUM ON EMPIRICAL SOFTWARE ENGINEERING AND MEASUREMENT, 15., 2021, Bari, Itália. **Proceedings [...]**. New York: Association for Computing Machinery, 2021. p. 34. Disponível em: https://doi.org/10.1145/3475716.3484192. Acesso em: 24 nov. 2025.

SPRING BOOT. **Spring Boot**. Versão 3.3.4. 2024. Disponível em: https://spring.io/projects/spring-boot. Acesso em: 28 jul. 2025.

SQLITE. **SQLite**. 2025. Disponível em: https://www.sqlite.org/. Acesso em: 28 jul. 2025.

THIRUNAVUKARASU, A. J. et al. Large language models in medicine. **Nature Medicine**, [S.l.], v. 29, n. 8, p. 1930-1940, 2023. Disponível em: https://doi.org/10.1038/s41591-023-02448-8. Acesso em: 26 nov. 2025.

WU, Shangyu et al. Retrieval-augmented generation for natural language processing: A survey. **arXiv preprint arXiv:2407.13193**, 2024.

ZHAO, Wayne Xin et al. A survey of large language models. **arXiv preprint arXiv:2303.18223**, v. 1, n. 2, 2023.

ZHOU, Zhi-hua. **Machine Learning**. Singapore: Springer, 2021. Disponível em: https://doi.org/10.1007/978-981-15-1967-3. Acesso em: 14 ago. 2025.