Relatório: Small Language Models para a Triagem de Estudos em Revisões Sistemáticas da Literatura: Abordagem *SLM-as-a-Judge*

Ivan Zichtl Santos[1],

[1]Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB)
Caixa Postal 15.064 – 58.015-435 – João Pessoa – PB – Brazil

Resumo. *O crescimento exponencial das publicações científicas impõe desafios significativos à condução de Revisões Sistemáticas da Literatura, tornando a etapa de triagem manual um processo oneroso e sujeito a vieses. Este estudo investiga a viabilidade de Small Language Models para otimizar essa tarefa, propondo uma arquitetura SLM-as-a-Judge que simula o fluxo metodológico de consenso e arbitragem tipicamente realizado por pesquisadores humanos. A pesquisa, de natureza aplicada e experimental, utilizou um corpus de 2.591 estudos para avaliar um pipeline composto por múltiplos agentes artificiais na aplicação de critérios de inclusão e exclusão. Os resultados demonstraram a eficácia da abordagem, com a arquitetura atingindo um Recall superior a 93%, o que garante a segurança necessária para a seleção de estudos relevantes e minimiza drasticamente o risco de perdas informacionais. Conclui-se que a utilização de modelos de linguagem menores e computacionalmente eficientes constitui uma alternativa viável e escalável aos grandes modelos proprietários, validando-se como uma ferramenta assistente capaz de reduzir o esforço humano sem comprometer o rigor metodológico das revisões sistemáticas.*

1. Introdução

O aumento do volume de publicações científicas ao longo do tempo é amplamente documentado. De acordo com levantamento da Fundação Nacional de Ciências dos EUA (NSF), o número de artigos científicos publicados globalmente cresceu 52% entre 2010 e 2021, ultrapassando 2,5 milhões de publicações em 2021 [National Science Foundation (NSF) 2022]. Além disso, dados mostram que, no Brasil, o volume de artigos científicos publicados aumentou nove vezes entre 1996 e 2022, chegando a 74,6 mil publicações [Agência Brasil 2023]. A crescente complexidade e o volume de informações disponíveis na literatura científica têm impulsionado a necessidade de métodos eficientes para a síntese de evidências.

As Revisões Sistemáticas da Literatura (RSL) emergem como uma ferramenta crucial nesse cenário, fornecendo uma visão abrangente e imparcial sobre um tópico específico, além de garantir sua reprodutibilidade. Segundo a estrutura proposta por [Kitchenham 2007], uma RSL desenvolve-se em três ações fundamentais: o planejamento, a condução e o relato. No entanto, a condução de uma RSL é um processo complexo e laborioso, demandando tempo significativo e recursos substanciais, especialmente nas etapas de seleção e triagem, que frequentemente representam o gargalo operacional da pesquisa [Michelson and Reuter 2019, Bullers et al. 2018].

Com o avanço da Inteligência Artificial (IA) e, em particular, do Processamento de Linguagem Natural (PLN), surgem oportunidades para mitigar esses desafios. Embora Modelos de Linguagem de Larga Escala (LLMs) proprietários, como a série GPT, tenham popularizado o uso de IA generativa [Caseli and Nunes 2024], observa-se uma tendência recente voltada para a eficiência computacional: os *Small Language Models* (SLMs). Esses modelos, menores e mais eficientes, buscam democratizar o acesso a IA mantendo capacidades robustas de raciocínio. Segundo [Gu et al. 2025], a capacidade de modelos generativos de imitar processos cognitivos humanos permite que assumam papéis de especialistas, oferecendo uma alternativa escalável e de menor custo para tarefas complexas de avaliação.

A etapa de triagem, conforme delineado por [Kitchenham 2007], inicia-se com a identificação de pesquisas por meio de uma estratégia de busca abrangente e imparcial, visando localizar o maior número possível de estudos primários relevantes para a questão de pesquisa. Após a coleta inicial, a triagem propriamente dita exige a aplicação de critérios explícitos de inclusão e exclusão para avaliar cada estudo primário em potencial. Nesse contexto, a autora aponta que é desejável que essa etapa envolva um número ímpar de pesquisadores, uma configuração estratégica que viabiliza a aplicação do voto majoritário para a resolução de conflitos diante de possíveis divergências sobre a inclusão ou exclusão do estudo na revisão.

Diante desse cenário, esta pesquisa propõe investigar não apenas se modelos de linguagem podem ler interpretar títulos e resumos, mas se uma arquitetura baseada em SLMs pode simular o fluxo metodológico de consenso e arbitragem humano. A relevância deste estudo reside na construção de um sistema que replica o rigor proposto por [Kitchenham 2007], utilizando múltiplos agentes artificiais para garantir a validação por pares e a resolução de conflitos de forma autônoma.

Para operacionalizar essa abordagem, utiliza-se o conceito de *LLM-as-a-Judge*. Essa técnica define uma arquitetura onde modelos atuam como avaliadores capazes de aplicar critérios explícitos de forma estruturada. Diferentemente de métricas automáticas tradicionais, essa abordagem permite que o modelo realize julgamentos alinhados a regras, aproximando-se do processo decisório humano [Gu et al. 2025]. Ao orquestrar esses modelos em um pipeline que integra avaliadores primários e árbitros especializados, este projeto insere-se na intersecção entre Metodologia de Pesquisa e IA, explorando se SLMs em uma organização de comitê podem oferecer uma triagem automática rigorosa, transparente e alinhada ao estado da arte das RSL.

2. Definição do Problema

A condução de RSLs é amplamente reconhecida como um processo intensivo em tempo, esforço humano e recursos financeiros, com estimativas, em alguns casos, apontando custos decorrentes de milhares de horas de trabalho especializado que podem ser superiores a US$ 140 mil por revisão [Michelson and Reuter 2019, Bullers et al. 2018]. Entre as etapas mais onerosas está a triagem de estudos, que envolve a aplicação de critérios de inclusão e exclusão, além da avaliação de títulos, resumos e textos completos — tarefas que demandam exaustiva dedicação da equipe de pesquisa [Michelson and Reuter 2019, **?**].

Tradicionalmente, essa fase é conduzida manualmente, frequentemente por mais de um revisor humano, a fim de mitigar vieses individuais e garantir maior confiabilidade

nos resultados. No entanto, essa abordagem, embora robusta, está sujeita a limitações humanas intrínsecas, como inconsistência, fadiga, lentidão e, eventualmente, à subjetividade inerente ao perfil do revisor em relação à temática. Esses fatores retardam significativamente a atualização do conhecimento científico e a tomada de decisões baseadas em evidências.

Nesse contexto, e impulsionado pelo crescente interesse demonstrado em trabalhos como os de [Qureshi et al. 2023, Lieberum et al. 2025, Guo et al. 2024], observa-se uma exploração ativa do potencial de modelos de linguagem para a automatização de tarefas complexas e em larga escala que envolvem processamento de linguagem natural. Embora esses modelos demonstrem capacidade de compreender contexto e gerar respostas relevantes, a literatura recente, incluindo a revisão de escopo de [Lieberum et al. 2025], aponta que sua aplicação em etapas críticas de RSLs, como a triagem de estudos, ainda carece de validação rigorosa e de aplicações metodológicas plenamente estabelecidas. Essa lacuna de confiabilidade e padronização justifica a necessidade de investigações aprofundadas como a proposta neste estudo.

3. Trabalhos Relacionados

A condução de revisões sistemáticas é um processo intrinsecamente complexo e demandante, exigindo especialização e tempo consideráveis. Diante dos desafios destacados por [Michelson and Reuter 2019] e [Bullers et al. 2018], como o custo significativo e o tempo prolongado associados à realização de revisões sistemáticas e meta-análises, os autores enfatizam a necessidade de maior envolvimento de tecnologias, como o aprendizado de máquina, para otimizar esse processo. A partir dessa perspectiva, outros estudos foram conduzidos buscando investigar a mitigação dos desafios por meio de tecnologias de IA.

[Qureshi et al. 2023] conduziu um estudo relevante por fornecer uma avaliação inicial e crítica das capacidades dos LLMs em tarefas de RSL, servindo como base para aprofundar a investigação do fenômeno. Enquanto seu enfoque foi uma análise exploratória e conceitual sobre o uso do ChatGPT com ênfase na adequação das respostas a diferentes tipos de prompts, o presente trabalho propõe uma abordagem quantitativa. Em vez de avaliar a utilidade geral dos LLMs por meio de testes pontuais e observacionais, esta pesquisa busca comparar diretamente o desempenho dos SLMs com o de revisores humanos na etapa de triagem de estudos, utilizando métricas objetivas. Dessa forma, o presente estudo avança na direção de avaliar, de forma controlada, a aplicabilidade de modelos em tarefas críticas das revisões sistemáticas, contribuindo para o debate iniciado por [Qureshi et al. 2023] com base em evidências mensuráveis e replicáveis.

[Qureshi et al. 2023] exploraram a utilidade do ChatGPT em RSL, analisando a adequação e aplicabilidade de suas respostas a prompts relacionados à metodologia de RSL. O estudo tem caráter exploratório e conceitual, sendo baseado em uma demonstração prática realizada durante um webinar promovido pelos desenvolvedores do PICO Portal. Nessa atividade, os autores submeteram o ChatGPT a uma série de tarefas típicas de uma RSL, formulação de perguntas estruturadas, definição de critérios de elegibilidade, elaboração de estratégias de busca e sumarização de conteúdo. Foram avaliados qualitativamente os resultados obtidos. As respostas geradas foram variadas: o modelo produziu uma seleção preliminar de títulos, mas apresentou erros na estratégia de busca

para o PubMed e, ao apresentar referências, listou citações não verificáveis, evidenciando alucinações. Os autores reconhecem o potencial dos LLMs, mas alertam que a tecnologia exige desenvolvimento substancial e verificação criteriosa humana.

O estudo de [Qureshi et al. 2023] fornece uma base sobre as capacidades e limitações dos modelos generativos. No entanto, sua abordagem difere fundamentalmente da proposta deste trabalho em escopo e método. Enquanto [Qureshi et al. 2023] adota uma análise qualitativa e ampla sobre diversas etapas da RSL utilizando um modelo generalista via chat, esta pesquisa foca especificamente na etapa de triagem, adotando uma abordagem quantitativa sistemática. O presente trabalho avança ao incrementar a interação por chat para um pipeline estruturado e replicável de *LLM-as-a-Judge*, onde modelos não apenas geram texto, mas atuam como classificadores binários guiados por instruções complexas. Além disso, busca-se mitigar a ”necessidade de verificação”apontada por [Qureshi et al. 2023] através da implementação de um sistema de arbitragem automática para casos de incerteza.

[Guo et al. 2024] investigaram o desempenho de LLMs proprietários (*OpenAI GPT-3.5 e GPT-4*) na identificação de títulos e resumos relevantes em revisões clínicas do mundo real. O estudo comparou o desempenho desses modelos com o de revisores humanos em seis artigos de revisão, totalizando mais de 24.000 registros. A abordagem envolveu a criação de um fluxo de trabalho via API para aplicar critérios de triagem em linguagem natural.

Os resultados indicaram uma acurácia geral de 0,91 e concordância substancial com decisões humanas, Kappa = 0,96. Os autores concluíram que modelos GPT têm potencial para otimizar a revisão clínica, atuando como auxílio aos pesquisadores. Este estudo é próximo da presente pesquisa, compartilhando o objetivo de automatizar a triagem e o uso de métricas como Precisão, Sensibilidade e *F1-score*. Contudo, existem distinções cruciais que evidenciam a contribuição inédita da metodologia proposta nesse estudo. [Guo et al. 2024] baseiam-se em modelos proprietários massivos ([**?**]), que implicam custos e dependência de API. Em contraste, esta pesquisa investiga a viabilidade de SLMs de código aberto, focando em eficiência computacional e reprodutibilidade científica acessível.

A metodologia de [Guo et al. 2024] utiliza uma abordagem de classificação direta. A presente pesquisa inova ao propor uma arquitetura de consenso e arbitragem multimodelo. Diferentemente de uma classificação única, nosso método implementa: (i) triagem por consenso entre modelos distintos para reduzir viés individual; e (ii) um comitê de árbitros especializados para resolver divergências. Enquanto [Guo et al. 2024] medem o desempenho do modelo isolado, esta pesquisa avalia a eficácia de um sistema de decisão colegiada, simulando o fluxo real de revisores humanos independentes e um terceiro árbitro, algo não explorado no trabalho citado.

A literatura atual, exemplificada por [Qureshi et al. 2023] e [Guo et al. 2024], demonstra uma evolução clara: de testes exploratórios de utilidade para validações de desempenho de grandes modelos proprietários na triagem de estudos. Embora esses trabalhos confirmem o potencial da IA, eles deixam lacunas significativas quanto à confiabilidade e à acessibilidade da tecnologia.

A principal lacuna identificada reside na dependência de grandes modelos co-

merciais e na ausência de mecanismos para lidar com a incerteza dos modelos como, alucinações ou dúvidas na classificação, sem intervenção humana direta. A presente pesquisa preenche essa lacuna ao propor um framework metodológico que não apenas avalia o desempenho, mas introduz uma camada de estabilidade e arbitragem. Ao utilizar SLMs em um arranjo de comitê, o estudo avança para além da simples comparação IA vs Humano, investigando se um pipeline de agentes autônomos cooperativos pode simular o rigor metodológico de uma triagem de estudos. Dessa forma, a pesquisa contribui para a democratização da RSL automatizada via modelos *open-source* e para a engenharia de sistemas de IA mais resilientes a erros. Um resumo comparativo é apresentado na (Tabela 1).

Tabela 1. Comparativo Trabalhos Relacionados - Fonte: Os autores (2025)

Estudo	Objetivo Principal	Escopo Metodológico	Modelos e Custo	Tarefa RSL	Contribuição
Qureshi et al. (2023) Are ChatGPT and large language models 'the answer' to bringing us closer to systematic review automation?	Explorar conceitualmente o uso do ChatGPT em tarefas de RSL	Demonstrações práticas em webinar; tarefas típicas de RSL	ChatGPT, API paga	Várias tarefas de RSL	- Primeira análise conceitual sistemática sobre LLMs em RSL - Discussão conceitual pioneira
Guo et al. (2024) Automated paper screening for clinical reviews using large language models: data analysis study	Avaliar GPT/GPT-4 na triagem de artigos clínicos	Pipeline automatizado via API; triagem de títulos e resumos	GPT-3.5 e GPT-4, API paga	Triagem inicial de títulos/resumos	- Evidência quantitativa robusta e replicável -Pipeline com dados reais
Estudo Proposto	Avaliar SLMs open-source comparados a Ground Truth anotado por especialistas na etapa de triagem em RSL	Pipeline completo com avaliadores, arbitragem e avaliação formal	Gemma, Qwen Auto-J, M-Prometheus, Phi3, Sem custo de Api.	Triagem inicial de títulos/resumos	- Análise com pipeline multimodelo - Uso de modelos abertos - Uso de comitê avaliativo. - Entrega da ferramenta/pipeline.

4. Objetivos

Como uma arquitetura baseada em *LLMs-as-a-Judge* multimodelo se compara, em termos de precisão, recall e alinhamento com julgamentos humanos *Ground Truth*, as abordagens tradicionais de classificação única na triagem de estudos em Revisões Sistemáticas da Literatura?

4.1. Objetivos Específicos

1. Medir a estabilidade dos modelos avaliativos A e B através do grau de concordância entre as três execuções, utilizando métricas de similaridade inter-modelo.
2. Analisar a concordância inter-anotador entre os três modelos árbitros (J1, J2 e J3) para validar a coerência e consistência das decisões dentro do comitê de arbitragem.
3. Avaliar e comparar a eficácia das diferentes estratégias de triagem em relação ao *Ground Truth* humano, mensurando o ganho de desempenho entre:
    - Modelos Isolados: O desempenho individual dos modelos A e B (após consolidação de suas rodadas internas);
    - Pipelines de Árbitro Único: A combinação das decisões consensuais ($A \cap B$) com a decisão individual de cada árbitro (J1, J2 ou J3) nas divergências;
    - Pipeline de Comitê: A combinação das decisões consensuais ($A \cap B$) com a decisão majoritária do comitê de árbitros nas divergências.

5. Metodologia

Este estudo é motivado pela necessidade de resolver problemas concretos, com finalidade prática, sendo, portanto, de natureza aplicada, conforme classificação de [Gil 2002]. Os procedimentos adotados são caracterizados como experimentais e comparativos. Quanto à abordagem do problema, a pesquisa é classificada como quantitativa. Ainda [Gil 2002], os objetivos caracterizados nessa pesquisa são exploratórios, tendo em vista sua busca por compreender um fenômeno ainda pouco conhecido, *LLMs-as-a-Judge*, como apontado por [Gu et al. 2025].

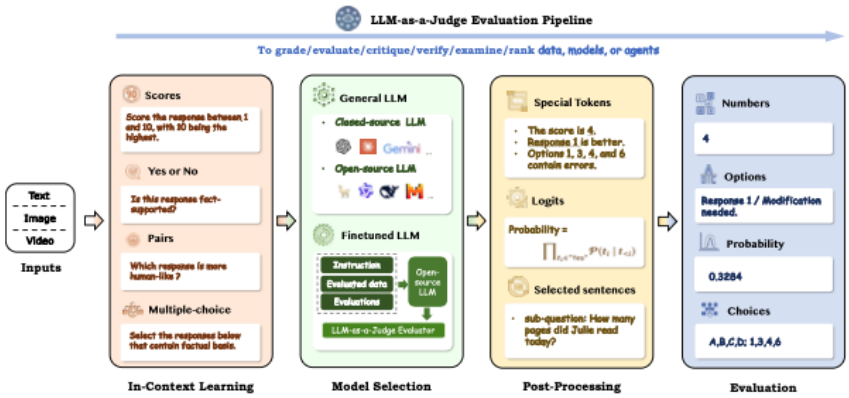


Figura 1. *LLM-as-a-Judge* evaluation pipelines - Fonte: [Gu et al. 2025]

Na construção da metodologia, esta pesquisa utilizou como embasamento o pipeline sistematizado por [Gu et al. 2025], conforme mostrado na (Figura 1), para construção de sistemas confiáveis de *LLM-as-a-Judge*, estruturado em quatro componentes principais: (i) definição explícita de critérios via *In-Context Learning*, que envolve o design de prompts com instruções claras para guiar o raciocínio do modelo alinhado a regras pré-definidas; (ii) seleção e diversificação de modelos de julgamento, que abrange a escolha de múltiplos modelos para mitigar vieses, considerando diversidade em capacidades e origens; (iii) integração e consolidação de múltiplas fontes de decisão, por meio de pós-processamento como votação majoritária para combinar saídas e reduzir variabilidade; e (iv) avaliação formal baseada em alinhamento humano, compara o pipeline e sua concordância com o *Ground Truth*.

Nesse contexto, a metodologia adotada nesta pesquisa segue um pipeline em cinco etapas: (i) pré-processamento do *corpus*, para padronização, limpeza e preparação das instâncias referentes à seleção dos estudos; (ii) aplicação dos critérios de inclusão e exclusão por dois avaliadores SLMs distintos, executada em três rodadas independentes para mitigar variabilidade inerente; (iii) pós-processamento e consolidação, unificando as instâncias em uma decisão final por modelo via voto majoritário para decisões binárias e seleção da melhor explicação por similaridade semântica; (iv) arbitragem por SLMs, com três árbitros analisando exclusivamente as instâncias em que houve divergência entre os avaliadores primários, escolhendo entre a melhor explicação e consolidando os três votos em um voto final; e (v) avaliação formal, analisando similaridades intra e inter-modelo, concordância com decisões de referência e alinhamento geral do pipeline ao *Ground Truth* humano na etapa de triagem dos estudos.

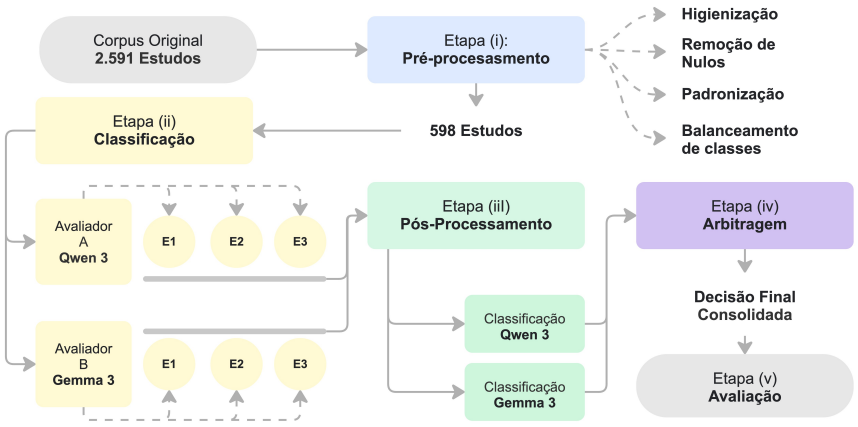


Figura 2. Fluxo Metodológico - Fonte: Os autores (2025)

5.1. Dataset

O corpus deste estudo baseia-se em dados de uma Revisão Sistemática da Literatura (RSL) sobre Engenharia de Requisitos em Projetos Ágeis, originalmente conduzida por [Medeiros et al. 2015]. A escolha deste conjunto de dados justifica-se pelo seu rigor metodológico, aderente às diretrizes de [Kitchenham 2007]. Embora os artefatos da RSL original não sejam publicamente disponibilizados, o autor responsável pelo estudo forneceu diretamente aos autores deste trabalho todos os arquivos necessários e autorizou sua divulgação no repositório associado a esta pesquisa. O conjunto inclui 2.591 registros (títulos e resumos) provenientes de seis bases indexadoras de alto impacto (*IEEE Xplore, Compendex, Scopus, ACM, SpringerLink e ScienceDirect*). O dataset contém uma classificação binária manual (*Ground Truth*) validada por especialistas, que aplicaram critérios explícitos de inclusão, exclusão e qualidade para selecionar 24 estudos primários ao final do estudo. Essa estrutura rotulada permite uma comparação direta e quantitativa entre as decisões automatizadas pela arquitetura de arbitragem proposta e o julgamento humano, servindo como linha de base confiável para a validação experimental.

5.2. Pré-processamento

A partir do corpus original de 2.591 estudos, conduzimos um processo de pré-processamento orientado à preparação do material para posterior avaliação automática. Inicialmente, realizou-se a higienização estrutural do *corpus*, removendo colunas auxiliares irrelevantes ao experimento e padronizando o campo *STATUS* por meio da normalização textual das classes *INCLUDED* e *EXCLUDED*. Na sequência, foram excluídas as instâncias associadas aos critérios EC2 (duplicatas) e EC8 (artigos indisponíveis para acesso institucional), conforme a diretriz de que tais critérios não podem ser inferidos pelo modelo a partir apenas de título e resumo e também foram removidos todos os registros cujo *TITLE* ou *ABSTRACT* estivesse ausente, assegurando que cada instância contenha informação mínima necessária para o julgamento automático.

Após essas etapas, aplicou-se um balanceamento estratificado das classes por amostragem aleatória, com random_state = 42 para garantir reprodutibilidade e obter

um conjunto proporcional entre estudos incluídos e excluídos, reduzindo possíveis vieses de classe no processo de avaliação por SLMs. Por fim, o corpus resultante foi embaralhado e dividido em blocos de 100 instâncias, possibilitando execução distribuída e controle operacional durante a etapa de julgamento automatizado. Esse pipeline de pré-processamento, preparou o conjunto de dados descrito por [Medeiros et al. 2015] para uso no experimento.

5.3. Classificação

A etapa de classificação constituiu na aplicação dos critérios de triagem por dois SLMs distintos, designados como avaliadores primários. A seleção desses modelos, avaliador A *Qwen3-4b-Instruct* [Yang et al. 2025] e avaliador B *gemma-3-4b-it* [Kamath et al. 2025] tem base na necessidade de diversificação de suas metodologias de treinamento, visando mitigar vieses algorítmicos singulares [Gu et al. 2025]. O processo de triagem é realizado via *In-Context Learning*, utilizando um prompt estruturado do tipo *Chain-of-Thought* [Caseli and Nunes 2024] [Gu et al. 2025]. O prompt classificador incorpora a definição explícita dos Critérios de Inclusão e Critérios de Exclusão originalmente usados por [Medeiros et al. 2015], traduzidos para língua inglesa, em seguida o título e resumo das instâncias, adicionado de instruções claras e necessárias para guiar o raciocínio dos SLMs no alinhamento com as regras pré-definidas da revisão sistemática (Listagem 1). Cada avaliador foi instruído a retornar sua classificação em um formato JSON estrito, que obriga a saída a conter o resultado binário (*INCLUDED*/*EXCLUDED*) e uma justificativa para a decisão. Buscando avaliar a estabilidade e reduzir a variabilidade inerente aos modelos generativos, cada avaliador executou a classificação de cada instância do corpus em três rodadas independentes. A multiplicação das execuções em número ímpar fornece os insumos necessários para a subsequente consolidação por voto majoritário na fase de Pós-Processamento.

Listagem 1. Prompt resumido para os avaliadores primários. Fonte: Os autores (2025)

```
...
You are a scientific article evaluator. Classify the following
   article as INCLUDED or EXCLUDED
based on the provided inclusion and exclusion criteria. After
   classifying, provide a brief explanation
justifying your decision.
To perform the evaluation, follow this step-by-step reasoning:

1. Check whether the article violates any Exclusion Criteria:
2. Confirm whether the article meets all Inclusion Criteria:
3. If the article does not violate any Exclusion Criteria and
   satisfies the Inclusion Criteria, classify it as "INCLUDED".
   Otherwise, classify it as "EXCLUDED".
4. Provide a brief explanation that references the applied
   criteria justifying your classification.

Title: {0} Abstract: {1}
...
```

A seleção dos modelos avaliadores fundamentou-se na complementaridade de

seus paradigmas de treinamento. A arquitetura *gemma-3-4b-it* emprega a metodologia de destilação de conhecimento (*knowledge distillation*), associada ao *Reinforcement Learning from Human Feedback* para alinhamento de instruções. Esta abordagem visa a otimização do raciocínio lógico e da compreensão de leitura em arquiteturas com menor número de parâmetros [Kamath et al. 2025]. Em contrapartida, o *Qwen3-4b-Instruct* baseia-se em pré-treinamento em corpus massivo seguido de (*Supervised Fine-Tuning*). A arquitetura caracteriza-se pelo processamento de dependências contextuais e pela estabilidade na interpretação de prompts extensos [Yang et al. 2025].

A combinação de arquiteturas distintas, baseadas respectivamente em destilação e generalização supervisionada, é recomendada por [Gu et al. 2025] como uma forma de diversificação analítica. Essa diversificação contribui para a robustez e confiabilidade do sistema, alinhando-se às recomendações do autor para a construção de sistemas *LLM-as-a-Judge*.

5.4. Pós-Processamento

A etapa de pós-processamento unificou saídas geradas nas múltiplas rodadas de inferência, transformando as três execuções independentes de cada avaliador primário em uma decisão consolidada. Para a definição do veredito binário (*INCLUDED* ou *EXCLUDED*), aplicou-se a estratégia de votação majoritária, na qual a classificação final foi determinada pela *label* predominante entre as três rodadas. A seleção da justificativa textual representativa da decisão consolidada, adotou-se uma abordagem baseada em similaridade semântica. As justificativas geradas nas três execuções foram convertidas em vetores de *embeddings* utilizando o modelo *all-MiniLM-L6-v2* [Reimers and Gurevych 2019]. Calculou-se a matriz de similaridade de cosseno entre os vetores, selecionando-se como explicação final aquela que apresentou a maior similaridade média em relação às demais, garantindo que o texto escolhido refletisse o argumento central do modelo.

5.5. Arbitragem

A fase de arbitragem (Figura 3) foi construída como uma camada de validação hierárquica, aplicada exclusivamente às instâncias em que houve divergência entre as decisões consolidadas dos avaliadores primários. Na mediação dos conflitos decisórios, empregou-se um comitê de árbitros composto por três SLMs especializados em avaliação e seguimento de instruções complexas: *M-Prometheus-3B* [Pombal et al. 2025], *Autoj-bilingual-6b* [Wang et al. 2023] e *Phi3-hallucination-judge* [Li et al. 2023].

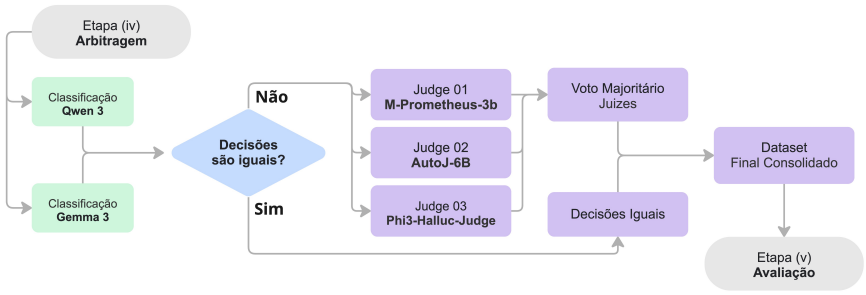


Figura 3. Etapa(iv) Arbitragem - Fonte: Os autores (2025)

A estrutura do comitê de arbitragem integrou três arquiteturas distintas, o que estabelece uma avaliação multidimensional composta por funções técnicas complementares, referenciadas na literatura como avaliação por rubrica, por crítica e por verificação factual.

O *M-Prometheus* atua como o árbitro baseado em Rubrica (*Rubric-Based Judge*). Como apontado por [Pombal et al. 2025], sua função primária é processar a avaliação condicionada estritamente às regras explícitas fornecidas nos critérios de inclusão e exclusão, verificando a aderência das justificativas. O *Auto-J 6B* desempenha o papel de árbitro baseado em Crítica (*Critique-Based Judge*). Conforme definido por [Wang et al. 2023], sua especialização reside na capacidade de articular o raciocínio lógico (*Chain-of-Thought*) para identificar falhas argumentativas ou inconsistências nas explicações geradas pelos avaliadores primários. Por fim, o *Phi3-Hallucination-Judge* opera como o árbitro baseado em Factualidade e Preferência (*Factuality & Preference Judge*). Este modelo foca na detecção de alucinações e na verificação da consistência factual das justificativas em relação ao resumo do artigo [Li et al. 2023].

O processo de arbitragem consistiu na apresentação do título, do resumo, dos critérios da RSL definidos por [Medeiros et al. 2015] e das duas decisões e justificativas conflitantes provenientes dos avaliadores primários para cada um dos árbitros, foi utilizado um prompt estruturado com fluxo instrucional encadeado, *Chain-of-Thought* (Listagem 2). Cada modelo do comitê analisou o contexto e emitiu um voto independente, favorecendo a decisão e a explicação mais coerentes com os critérios de elegibilidade. A decisão final da arbitragem foi obtida pela consolidação dos votos dos três juízes via maioria simples, estabelecendo um veredito definitivo para os casos de incerteza entre os modelos primários. As instâncias na qual os avaliadores primários concordaram em sua decisão foram replicadas, combinadas com as decisões finais do comitê.

Listagem 2. Prompt resumido para os árbitros.

```
...
You are a scientific article evaluation judge. You will receive:
- The title and abstract of a scientific article.
- Two independent evaluations (Decision A and Decision B), each
   containing a classification ("INCLUDED" or "EXCLUDED") and an
    explanation.

Your task is to analyze the article critically and
   determinewhich decision (A or B) better adheres to the
   inclusion and exclusion criteria, and therefore is the most
   accurate.
To perform the evaluation, follow this step-by-step reasoning:

1. Check whether the article violates any Exclusion Criteria:
2. Confirm whether the article meets all Inclusion Criteria:
3. Evaluate Decision A and Decision B:
4. Respond ONLY with a single JSON object in the format:
Title: {0} Abstract: {1} Decision A: {2} Justification A: {3}
   Decision B: {4} Justification B: {5}
...
```

O conjunto de dados produzido é composto pelas *features* originais da RSL de [Medeiros et al. 2015] como, identificador da instância, *status* (*ground truth*), critérios aplicados, titulo, resumo, assim como, todas as decisões e justificativas dos avaliadores primários, todas as decisões e justificativas do comitê, além da decisão final do processo de triagem.

5.6. Avaliação

A etapa avaliativa (Figura 4) adotada, teve com base a comparação das decisões automatizadas contra o *ground truth* estabelecido pela revisão sistemática humana de referência [Medeiros et al. 2015]. O desempenho preditivo foi mensurado através de métricas de classificação binária, incluindo Precisão, Sensibilidade (*Recall*) e *F1-Score*, calculadas tanto para os avaliadores individuais quanto para o sistema consolidado após a arbitragem.

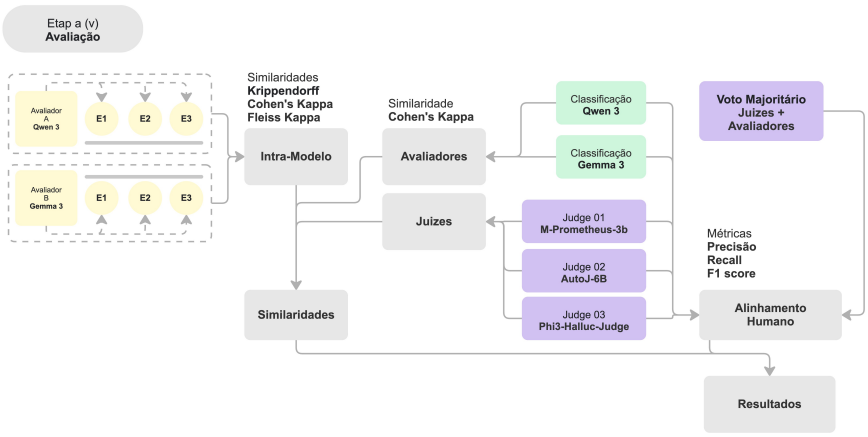


Figura 4. Etapa(v) Avaliação - Fonte: Os autores (2025)

O fluxo iniciou pelo cálculo de métricas de similaridade e concordância entre todos os julgadores automáticos. Foram estimadas as concordâncias intra-modelo, entre as três execuções de cada avaliador primário, em seguidas das concordâncias entre as decisões consolidadas de cada avaliador primário e entre os três árbitros, utilizando o coeficiente Kappa de Cohen para pares, o Kappa de Fleiss e Alfa de Krippendorff para múltiplos classificadores, em linha com as recomendações para meta-avaliação de *LLM-as-a-Judge*. Em seguida, essas mesmas métricas de concordância foram aplicadas para comparar o conjunto de classificadores automáticos em relação ao *Ground Truth* humano, de modo a quantificar o alinhamento estrutural entre SLMs e a triagem de referência.

Na etapa subsequente, foram calculadas métricas clássicas de classificação binária tomando o *Ground Truth* como rótulo de referência. Para cada avaliador primário (*Qwen3-4b-Instruct* e *gemma-3-4b-it*), para cada árbitro (*M-Prometheus*, *Auto-J 6B* e *Phi3-Hallucination-Judge*) e para a decisão final consolidada, estimaram-se Precisão, Sensibilidade (*Recall*) e *F1-Score*. Esses indicadores permitem avaliar, de forma comparável, o desempenho preditivo de cada modelo relação aos critérios de inclusão e exclusão definidos pela revisão sistemática humana.

5.7. Modelos Utilizados

O pipeline foi executado no ambiente Google Colab [Google 2025], com instâncias equipadas com GPU NVIDIA T4, e todos os modelos carregados a partir do Hugging Face Hub [Face 2025a]. Foram utilizadas bibliotecas Python como *transformers* [Face 2025c], *torch* [Paszke et al. 2025] e *huggingface_hub* [Face 2025b]. O processamento total consumiu 68,20 unidades computacionais, o que corresponde a aproximadamente R$ 39,56, considerando o valor de R$ 58,00 para cada 100 unidades, conforme a política do Colab. A Tabela 2 apresenta, em resumo, os modelos utilizados e suas principais características.

Tabela 2. Modelos Utilizados - Fonte: Os autores (2025).

Função no Pipeline	Modelo	Tipo	Referência
Avaliador Primário (A)	Qwen3-4B-Instruct	SLM (Instruct)	Yang et al. (2025). Qwen3 Technical Report. arXiv:2505.09388.
Avaliador Primário (B)	Gemma-3-4b-it	SLM (Instruction-Tuned, Distilled)	Kamath et al. (2025). Gemma 3 Technical Report. arXiv:2503.19786.
Árbitro 1 Rubric-Based Judge	M-Prometheus-6B	LLM Judge	Pombal et al. (2025). M-Prometheus: A Suite of Open Multilingual LLM Judges. arXiv:2504.04953.
Árbitro 2 Critique-Based Judge	Auto-J Bilingual-6B	LLM Judge	Wang et al. (2023). Generative Judge for Evaluating Alignment. arXiv:2310.05470.
Árbitro 3 Factuality & Hallucination Judge	Phi3-Hallucination-Judge	LLM Judge	Li et al. (2023). HaluEval: A Large-Scale Hallucination Evaluation Benchmark. arXiv:2305.11747.
Embeddings para Pós-Processamento	all-MiniLM-L6-v2	Sentence Embedding Model	Reimers & Gurevych (2019). Sentence-BERT. arXiv:1908.10084.

6. Resultados

6.1. Análise de Similaridade e Concordância

A avaliação da confiabilidade da arquitetura proposta revela disparidades significativas entre a consistência interna dos modelos e a sua capacidade de alinhamento com o julgamento humano.

Ao analisar a consistência interna dos SLMs na (Tabela 3), observa-se um comportamento determinístico. As três execuções do modelo (*Qwen3-4b-Instruct* apresentaram uma concordância quase perfeita entre si Kappa de 0.998, assim como as execuções do *gemma-3-4b-it*) Kappa de 0.931. Isso indica que os modelos são estáveis e convictos em suas decisões. No entanto, ao cruzar as decisões entre as diferentes famílias de modelos (*Qwen3-4b-Instruct* × *gemma-3-4b-it*), a concordância colapsa para um Kappa de apenas 0.244 e um índice de Krippendorf de 0.113. Esses dados comprovam que, embora internamente consistentes, os modelos possuem interpretações fundamentalmente distintas

Tabela 3. Concordância Avaliadores - Fonte: Os autores (2025)

Model	Cohen Kappa	Interp CK	Fleiss Kappa	Interp FK	Krippendorf	Interp Krip
Qwen3_01 x Qwen3_02	0.997	Quase Perfeita	–	–	–	–
Qwen3_01 x Qwen3_03	1.0	Quase Perfeita	–	–	–	–
Qwen3_02 x Qwen3_03	0.997	Quase Perfeita	–	–	–	–
Qwen3_01 x Qwen3_02 x Qwen3_03	0.998	Quase Perfeita	0.998	Quase Perfeita	0.998	Excelente
Gemma3_01 x Gemma3_02	0.927	Quase Perfeita	–	–	–	–
Gemma3_01 x Gemma3_03	0.927	Quase Perfeita	–	–	–	–
Gemma3_02 x Gemma3_03	0.939	Quase Perfeita	–	–	–	–
Gemma3_01 x Gemma3_02 x Gemma3_03	0.931	Quase Perfeita	0.931	Quase Perfeita	0.931	Excelente
Qwen3_01 x Gemma3_01	0.244	Razoável	0.113	Leve	0.113	Não confiável

sobre os títulos e resumos dos estudos, gerando um alto volume de conflitos para a etapa de arbitragem.

Na etapa de resolução de conflitos (Tabela 4), destaca-se a alta afinidade entre os modelos *M-Prometheus* e *Auto-J 6B*, que atingiram um Kappa de 0.894. Esse par demonstrou um alinhamento superior a qualquer outra combinação de árbitros. Em contrapartida, o modelo *Phi3-Hallucination-Judge* mostrou-se um *outlier*, apresentando baixa concordância com os demais juízes, Kappa de 0.34 a 0.36, o que sugere que este modelo utiliza critérios de desempate divergentes da maioria.

Tabela 4. Concordância Árbitros - Fonte: Os autores (2025)

Model	Cohen Kappa	Interp CK	Fleiss Kappa	Interp FK	Krippendorf	Interp Krip
M-Prometheus x Autoj-6B	0.894	Quase Perfeita	–	–	–	–
M-Prometheus x Phi3-halluc	0.339	Razoável	–	–	–	–
Autoj-6B x Phi3-halluc	0.36	Razoável	–	–	–	–
Autoj-6B x M-Prometheus x Phi3-halluc	0.531	Moderada	0.442	Moderada	0.442	Não confiável

A comparação final com as decisões humanas, apresentada na (Tabela 5), traz um dos resultados mais críticos do estudo. Individualmente, o modelo *Qwen3-4b-Instruct* obteve o melhor desempenho isolado, com um Kappa de 0.495. Os dados mostram que a complexidade adicionada pelos árbitros e pelo consenso dos avaliadores primários não superou o melhor modelo individual neste cenário específico. O Consenso Final obteve um Kappa de 0.201, alinhando-se mais ao desempenho inferior do *gemma-3-4b-it* 0.204, do que o desempenho superior do *Qwen3-4b-Instruct*. Entre as variações de arbitragem, a configuração avaliadores primários + *Phi3-Hallucination-Judge* foi a única a manter uma concordância moderada 0.411, próxima ao desempenho do *Qwen3-4b-Instruct* isolado, enquanto as configurações com *M-Prometheus* e *Auto-J 6B* caíram para a faixa de concordância 0.20.

Os números indicam que a divergência introduzida pelo modelo *gemma-3-4b-it*,

Tabela 5. Concordância Ground Truth - Fonte: Os autores (2025)

Model	Cohen Kappa	Interp CK	Krippendorf	Interp Krip
Ground Truth x Qwen3 Consolidado	0.495	Moderada	0.493	Não confiável
Ground Truth x Gemma3 Consolidado	0.204	Razoável	0.101	Não confiável
Ground Truth x | A∩B + M-Prometheus |	0.201	Razoável	0.101	Não confiável
Ground Truth x | A∩B + Autoj-6B |	0.227	Razoável	0.138	Não confiável
Ground Truth x | A∩B + Phi3-halluc |	0.411	Moderada	0.411	Não confiável
Ground Truth x Consenso Final	0.201	Razoável	0.101	Não confiável

predominou na formação do consenso final, diluindo a precisão obtida originalmente pelo *Qwen3-4b-Instruct*. Em outras palavras, o mecanismo de votação majoritária entre avaliadores e árbitros direcionou o resultado agregado em direção ao comportamento do modelo com menor alinhamento humano, em vez de potencializar o desempenho do melhor avaliador individual.

Esse efeito evidencia um risco já discutido por [Gu et al. 2025] na literatura de *LLM-as-a-Judge*, sem uma seleção criteriosa de modelos e um esquema de ponderação baseado em desempenho, *ensembles* ingênuos podem degradar a qualidade global da decisão, mesmo quando incluem modelos fortes. O consenso passou a refletir um compromisso entre interpretações incompatíveis dos títulos e resumos, o que reforça a necessidade de (i) excluir ou reponderar avaliadores sistematicamente mais fracos e (ii) validar empiricamente qualquer estratégia de agregação antes de adotá-la como substituto do melhor modelo individual.

6.2. Avaliação de Desempenho e Alinhamento com a Decisão Humana

A análise do desempenho individual e consolidado dos modelos permite distinguir duas estratégias de classificação opostas, evidenciadas pelas métricas de Sensibilidade (*Recall*) e Precisão. Essa dicotomia reflete o trade-off clássico em recuperação da informação entre minimizar a leitura de documentos irrelevantes (Precisão) ou garantir a não exclusão de documentos relevantes (Sensibilidade).

A avaliação dos classificadores primários revela comportamentos distintos quanto ao rigor da triagem. O modelo *Qwen3-4b-Instruct* demonstrou ser o agente mais equilibrado estatisticamente, porém apresentando riscos para o contexto de uma RSL. Conforme observado na (Tabela 6) e (Tabela 7) este modelo obteve um *F1-Score* estável entre as classes 0.73-0.76. Apesar de sua capacidade de excluir corretamente 242 estudos irrelevantes, o *Qwen3-4b-Instruct* apresentou um risco crítico ao deixar de incluir 94 estudos relevantes, resultando em um *Recall* para a classe *INCLUDED* de apenas 0.686. Ou seja, apesar de preciso, o modelo descarta um volume inaceitável de informação.

Em contrapartida, o modelo *gemma-3-4b-it* operou com uma lógica de prioridade inversa. Os dados indicam uma estratégia de segurança, onde o modelo atingiu um *Recall* expressivo de 0.940 (Tabela 8) e (Tabela 9), perdendo apenas 18 estudos relevantes. O modelo classificou incorretamente 220 estudos irrelevantes como ”Incluídos”(Falsos Positivos). Isso impactou severamente sua Precisão, que caiu para 0.561, indicando que

Tabela 6. Matriz de Confusão *Qwen3* - Fonte: Os autores (2025).

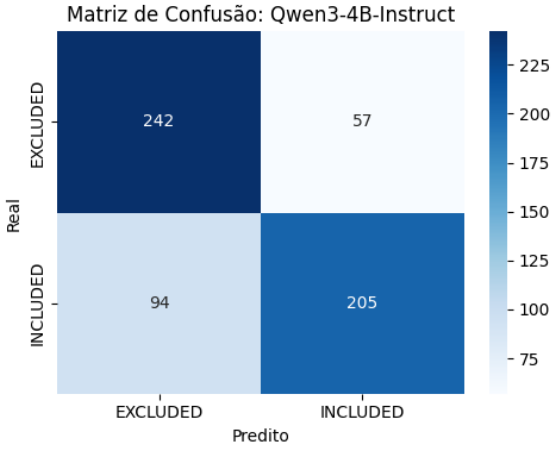


Tabela 7. Métricas *Qwen3* - Fonte: Os autores (2025).

Model	Precisão	Recall	F1-Score	Suporte
EXCLUDED	0.72	0.809	0.762	299
INCLUDED	0.782	0.686	0.731	299
Acurácia Geral	**0.747**	**0.747**	**0.747**	**0**
Média Macro	**0.751**	**0.747**	**0.747**	**598**
Média Ponderada	**0.751**	**0.747**	**0.747**	**598**

quase metade dos estudos selecionados por este modelo não atendem aos critérios de inclusão, transferindo carga de trabalho para a etapa seguinte.

A análise das tabelas referentes à camada de arbitragem revelou que os árbitros tendem a corroborar a lógica permissiva observada no modelo *gemma-3-4b-it*, priorizando a retenção de documentos em caso de dúvida. Os modelos *M-Prometheus* e *Auto-J 6B* replicaram o padrão de alta sensibilidade (Tabela 10) e (Tabela 11). O *M-Prometheus*, por exemplo, obteve um *Recall* de 0.933 para a classe *INCLUDED*, com uma Matriz de Confusão (Tabela 10) quase idêntica à do *gemma-3-4b-it*, 219 Falsos Positivos contra 220. Isso sugere que esses juízes atuam sob a premissa de não descartar evidências, adotando uma posição de segurança.

A exceção a esse comportamento foi o modelo *Phi3-Hallucination-Judge*. Este agente alinhou-se mais ao perfil rígido do *Qwen3-4b-Instruct*, mantendo o *Recall* em um patamar baixo 0.716 para uma triagem primária, descartando indevidamente 85 estudos relevantes.

Ao observar os dados do Consenso Final, nota-se que os resultados são virtualmente idênticos aos obtidos pelo juiz *M-Prometheus*. O sistema final alcançou um *Recall* de 0.933, garantindo uma altíssima recuperação de estudos reais, com uma Precisão de 0.560. A análise da Matriz de Confusão (Tabela 14, 80 exclusões corretas contra 219 falsos positivos, permite concluir que a arquitetura convergiu naturalmente para minimizar o erro do Tipo II - Falso Negativo.

Tabela 8. Matriz de Confusão *gemma-3* - Fonte: Os autores (2025).

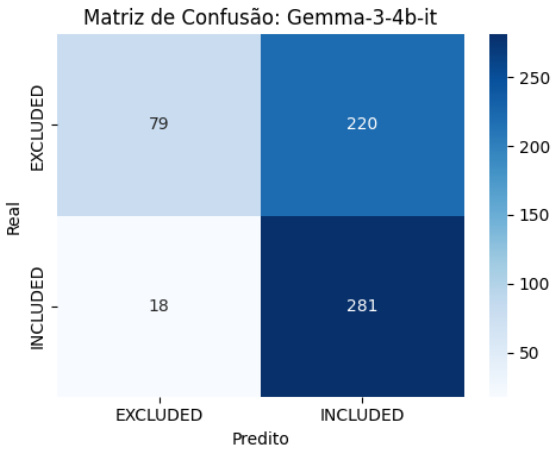
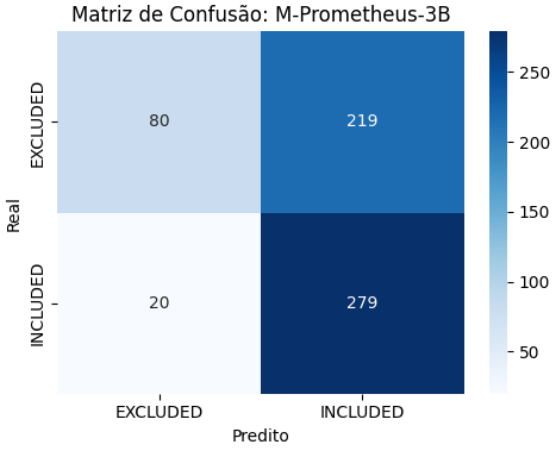


Tabela 9. Métricas *gemma-3* - Fonte: Os autores (2025)

Model	Precisão	Recall	F1-Score	Suporte
EXCLUDED	0.814	0.264	0.399	299
INCLUDED	0.561	0.94	0.702	299
Acurácia Geral	**0.602**	**0.602**	**0.602**	**0**
Média Macro	**0.688**	**0.602**	**0.551**	**598**
Média Ponderada	**0.688**	**0.602**	**0.551**	**598**

Tabela 10. Matriz de Confusão *M-Prometheus* - Fonte: Os autores (2025).



Em um ambiente real, o sistema submeteria o pesquisador à leitura de 219 artigos irrelevantes a arriscar omitir os aproximadamente 20 artigos relevantes que falhou em detectar. Os dados sugerem que a influência dos agentes *gemma-3-4b-it* e *M-Prometheus* predominou na decisão final, sobrepondo-se à rigidez dos modelos *Qwen3-4b-Instruct* e *Phi3-Hallucination-Judge*.

Tabela 11. Matriz de Confusão *Auto-J 6B* - Fonte: Os autores (2025)

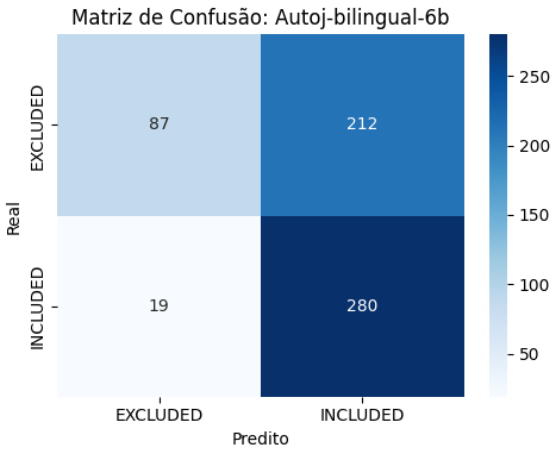


Tabela 12. Matriz de Confusão *Phi3-Halluc* - Fonte: Os autores (2025).

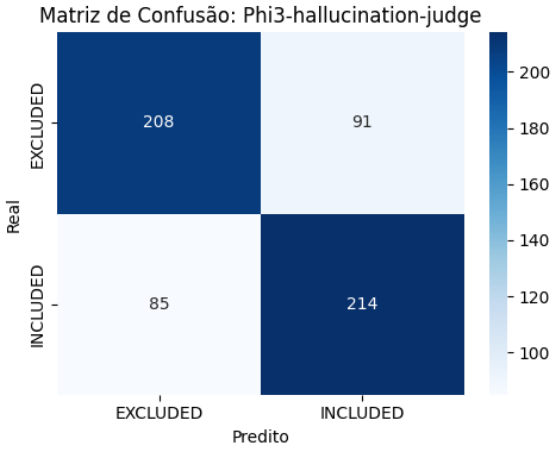


Tabela 13. Métricas *Phi3-Halluc* - Fonte: Os autores (2025).

Model	Precisão	Recall	F1-Score	Suporte
EXCLUDED	0.71	0.696	0.703	299
INCLUDED	0.702	0.716	0.709	299
Acurácia Geral	**0.706**	**0.706**	**0.706**	**0**
Média Macro	**0.706**	**0.706**	**0.706**	**598**
Média Ponderada	**0.706**	**0.706**	**0.706**	**598**

7. Conclusão

Os resultados obtidos apontam que a aplicação de *Small Language Models* em arquiteturas de comitê (*Ensemble*) para triagem de estudos apresenta uma dualidade clara entre segurança e eficiência. Embora o sistema tenha atingido seu objetivo primário de salvaguardar a integridade da RSL, alcançando um *Recall* superior a 93% e minimizando drasticamente a perda de estudos relevantes, isso custou uma baixa Precisão 56%. A predominância de uma postura permissiva no consenso final demonstra que, sem mecanismos

Tabela 14. Matriz de Confusão – Consenso Final - Fonte: Os autores (2025).

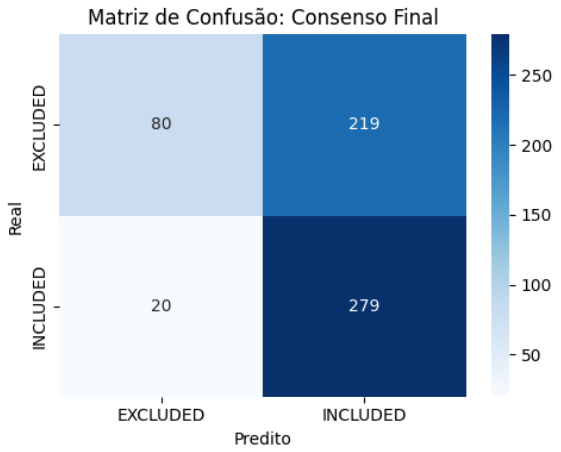


Tabela 15. Métricas – Consenso Final - Fonte: Os autores (2025).

Model	Precisão	Recall	F1-Score	Suporte
EXCLUDED	0.8	0.268	0.401	299
INCLUDED	0.56	0.933	0.7	299
Acurácia Geral	**0.6**	**0.6**	**0.6**	**0**
Média Macro	**0.68**	**0.6**	**0.551**	**598**
Média Ponderada	**0.68**	**0.6**	**0.551**	**598**

de controle, o sistema pode priorizar a redução do erro Tipo II, Falso Negativo.

Adicionalmente, o estudo evidenciou um paradoxo na estratégia de agregação: a combinação de múltiplos modelos não resultou necessariamente em uma inteligência superior à do melhor agente individual. A queda do índice Kappa no Consenso Final 0.201, quando comparado ao desempenho isolado do *Qwen3-4b-Instruct* 0.495, corrobora que a inclusão de modelos com interpretações divergentes ou menos rigorosas, como o *gemma-3-4b-it*, reduziu a qualidade da decisão coletiva.

8. Considerações Finais

8.1. Contribuições da Pesquisa

Como contribuições, este estudo oferece perspectivas sobre a arquitetura de sistemas de triagem automatizada. Primariamente, a pesquisa valida a viabilidade técnica e econômica dos *Small Language Models*, demonstrando que modelos compactos podem oferecer desempenho satisfatório. Esta constatação promove a democratização do acesso a ferramentas de apoio à decisão, permitindo que grupos de pesquisa realizem triagens em larga escala com total privacidade de dados e sem os custos proibitivos associados a modelos proprietários massivos.

Do ponto de vista metodológico, o estudo sugere que a agregação de múltiplos modelos resulta não invariavelmente numa decisão superior, os dados empíricos demonstraram que o consenso final, que apresentou um Kappa de 0.201 foi inferior ao desempenho do melhor avaliador individual *Qwen3-4b-Instruct* com Kappa de 0.495. Este achado

coloca a luz que a inclusão de modelos com perfis de permissivos, como o *gemma-3-4b-it*, pode diluir o rigor lógico de modelos mais precisos, exigindo, portanto, novas abordagens de ponderação em arquiteturas *SLM-as-a-Judge*.

Adicionalmente, o estudo oferece uma caracterização detalhada dos perfis comportamentais de diferentes famílias de modelos, distinguindo entre modelos de mais rigor lógico e outros mais conservadores. Ao comprovar que a arquitetura proposta atinge um *Recall* superior a 93%, a pesquisa estabelece a ferramenta como um assistente seguro para a redução de esforço humano, capaz de filtrar a maioria dos estudos irrelevantes sem comprometer a exaustividade da revisão. Em suma, o trabalho não apenas apresenta uma solução funcional e evidência as limitações da votação majoritária simples e guiando o design de futuros sistemas híbridos de triagem.

8.2. Limitações e Trabalhos Futuros

O fluxo de trabalho experimental não incorporou a abordagem *Human-in-the-loop* durante a execução do processo decisório. A ausência de interação humana para validação intermediária, conforme recomendado por [Gu et al. 2025]. A validação dos modelos foi realizada utilizando conjuntos de dados artificialmente balanceados. Tal distribuição estatística não reflete o severo desbalanceamento de classes inerentes a processos reais de triagem em RSL, onde a proporção de estudos irrelevantes é frequentemente superior. Por fim, a execução dos modelos restringiu-se às configurações de inferência padronizadas, sem a aplicação de estratégias de otimização de hiperparâmetros, como temperatura e top-p. O ajuste fino dessas variáveis poderia ter controlado a permissividade excessiva observada em modelos como o *gemma-3-4b-it*, reduzindo potencialmente a taxa de Falsos Positivos.

Considerando as limitações e os fenômenos observados, sugerem-se as seguintes direções para pesquisas futuras:

- Mecanismos de Votação Ponderada - Substituir o consenso por maioria simples por estratégias, onde o peso do voto de cada modelo seja calibrado dinamicamente com base em métricas de confiança ou desempenho histórico em tarefas de calibração.
- Pipeline Híbrido - Investigar a eficácia de uma abordagem sequencial em vez de paralela, utilizando modelos de alta sensibilidade (como o *gemma-3*) apenas para a fase de recuperação inicial, seguidos por modelos de maior rigor lógico, como o *Qwen3* ou *Phi3*, para uma segunda etapa de filtragem, visando elevar a Precisão sem comprometer o *Recall*.
- Refinamento de Prompts e Raciocínio - Explorar técnicas de *Chain-of-Thought* específicas para resolução de conflitos, forçando os modelos árbitros a explicitar e refletir sobre a lógica de exclusão antes de emitir o veredito, a fim de mitigar a tendência de aceitação passiva de falsos positivos.
- *Human-in-the-loop* - Evoluir o pipeline através da integração com componente humano, estabelecendo uma camada de avaliação humana para a validação de divergências e decisões complexas.

8.3. Artefatos Produzidos

Todos os arquivos utilizados ao longo do desenvolvimento deste trabalho, incluindo os *datasets* produzidos e processados, o *dataset* disponibilizado por [Medeiros et al. 2015],

as apresentações, os códigos-fonte e o executável do pipeline, foram integralmente disponibilizados pelos autores no repositório do projeto no GitHub, disponível em: https://github.com/izichtl/small-language-models-on-systematic-reviews, garantindo transparência, reprodutibilidade e acesso público a todos os materiais.

Referências

Agência Brasil (2023). Impacto acadêmico da ciência brasileira aumentou 21% de 1996 a 2022: Brasil teve nove vezes mais publicações científicas no período. `https://agenciabrasil.ebc.com.br/geral/noticia/2023-11/impacto-academico-da-ciencia-brasileira-aumentou-21-de-1996-2022`. Acesso em: 25 jun. 2025.

Bullers, K. et al. (2018). It takes longer than you think: librarian time spent on systematic review tasks. *Journal of the Medical Library Association*, 106(2):198–207.

Caseli, H. M. and Nunes, M. G. V. (2024). *Processamento de Linguagem Natural: conceitos, técnicas e aplicações em português*. BPLN, 3 edition.

Face, H. (2025a). Hugging face – plataforma de modelos e ferramentas de ia. `https://huggingface.co/`. Acesso em: 08 dez. 2025.

Face, H. (2025b). huggingface_hub – biblioteca para acesso ao hugging face hub. `https://github.com/huggingface/huggingface_hub`. Acesso em: 08 dez. 2025.

Face, H. (2025c). Transformers – state-of-the-art natural language processing. `https://github.com/huggingface/transformers`. Acesso em: 08 dez. 2025.

Gil, A. C. (2002). *Como elaborar projetos de pesquisa*. Atlas, São Paulo.

Google (2025). Google colab – ambiente de computação em nuvem para notebooks python. `https://colab.research.google.com/`. Acesso em: 08 dez. 2025.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Lin, Z., Zhang, B., Ni, L., Gao, W., Wang, Y., and Guo, J. (2025). A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Guo, E. et al. (2024). Automated paper screening for clinical reviews using large language models: data analysis study. *Journal of Medical Internet Research*, 26:e48996.

Kamath, G. T. A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Kitchenham, B. (2007). Guidelines for performing systematic literature reviews in software engineering. EBSE Technical Report EBSE-2007-01, Keele University; University of Durham, Keele; Durham.

Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., and Wen, J.-R. (2023). Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Lieberum, J.-L. et al. (2025). Large language models for conducting systematic reviews: on the rise, but not yet ready for use – a scoping review. *Journal of Clinical Epidemiology*, 181:111746.

Medeiros, F. et al. (2015). Metodologia de revisão sistemática: Guia prático para engenharia de software. *InfoComp*, 14(2):159–180.

Michelson, M. and Reuter, K. (2019). 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*, 16:100443.

National Science Foundation (NSF) (2022). Publication output by country, region, or economy and scientific field. `https://ncses.nsf.gov/pubs/nsb20214/publication-output-by-country-region-or-economy-and-scientific-field`. Science and Engineering Indicators 2022, Arlington. Acesso em: 25 jun. 2025.

Paszke, A., Gross, S., Massa, F., Lerer, A., et al. (2025). Pytorch – an open source machine learning framework. `https://pytorch.org/`. Acesso em: 08 dez. 2025.

Pombal, J., Yoon, D., Fernandes, P., Wu, I., Kim, S., Rei, R., Neubig, G., and Martins, A. F. T. (2025). M-prometheus: A suite of open multilingual llm judges. *arXiv preprint arXiv:2504.04953*.

Qureshi, R. et al. (2023). Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Systematic Reviews*, 12(1):72.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Wang, Y., Wang, Z., Zheng, Y., Zhang, X., Hu, X., Yang, X., Yu, J., et al. (2023). Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z., and Team, Q. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.