Small Language Models para a Triagem de Estudos em Revisões Sistemáticas da Literatura:

Abordagem *SLM-as-a-Judge*

Ivan Zichtl Santos

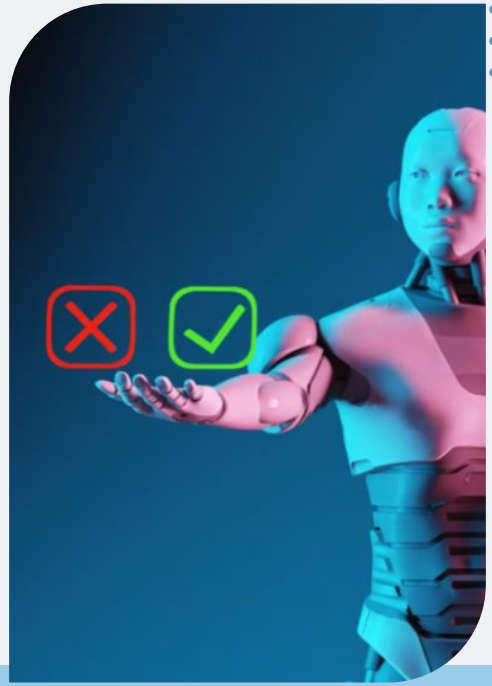

Ilustração. Fonte: Freepik, 2025.

Contexto histórico

A busca por evidências científicas remonta ao desenvolvimento do **método científico**.

Karl Pearson publicou em **1904** as primeiras técnicas formais para combinar resultados de diferentes estudos.

(SHANNON, 2016)

Ilustração. Fonte: Freepik, 2025.

Aumento da produção científica global e brasileira.

62 mil publicações em 2024 no Brasil, um crescimento de **6%**.

(GRABOIS, 2025; HORA DE PUBLICAR, 2025)

Um volume multiplicado 9 vezes entre 1996 e 2022

(AGENCIA BRASIL, 2023)

Ilustração. Fonte: Freepik, 2025.

Revisões Sistemáticas de Literatura (RSL)

Volume de informações disponíveis na literatura científica revelou a necessidade de **métodos eficientes** para a síntese de evidências.

As Revisões sistemáticas emergiram como uma ferramenta crucial nesse cenário.

(MICHELSON; REUTER, 2019; BULLERS et al., 2018).

Problema de Pesquisa

A condução de uma **RSL** é um processo intensivo em **tempo**, **esforço humano** e **recursos financeiros**, apontando custos superiores a **US$ 140 mil** por revisão.

A etapa de **triagem** uma das mais onerosas, tradicionalmente realizada de forma manual e **por mais de um revisor humano.**

(MICHELSON; REUTER, 2019; SHEN et al., 2023; KITCHENHAM, 2007).

Modelos de Linguagem como solução

O uso de LLMs, como o ChatGPT, surge como alternativa para automatizar etapas manuais.

(QURESHI, 2023; WANG et al., 2024; ZUBIAGA et al., 2024).

Qureshi et al. (2023) questiona se LLMs são a resposta para a RSL.

Lieberum et al. (2025) mapeia 37 estudos que abordagem o uso LLMs em RSL.

Objetivo Geral

Avaliar como uma arquitetura baseada em *LLMs-as-a-Judge* multimodelo se compara, em termos de precisão, recall e alinhamento com julgamentos humanos, *Ground Truth.*

Objetivos Específicos

- Medir a estabilidade dos modelos avaliativos A e B através do grau de concordância entre as três execuções, utilizando métricas de similaridade inter-modelo.
-
- Analisar a concordância inter-anotador entre os três modelos árbitros para validar a coerência e consistência das decisões dentro do comitê de arbitragem.
-
- Avaliar e comparar a eficácia das diferentes estratégias de triagem em relação ao *Ground Truth*.

  Modelos Isolados; Pipelines de Árbitro Único; Pipeline de Comitê

Trabalhos Relacionados

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.

Tabela de Trabalhos Relacionados. Fonte: Produzido pelo autor, 2025.

Relevância da pequisa

O estudo é relevante e se diferencia dos apresentados nos seguintes pontos:

- Utiliza de modelos pequenos **open-source** , com custo computacional baixo e com uso livre, **sem custos de API** .

- Avança na lacuna de avaliação da estabilidade e alinhamento de modelos com decisões humanas.

- Apresenta um **pipeline multi-modelo** replicável, que pode avançar para outras etapas em RSL.

- Explora técnicas ***LLM-as-a-Judge*** integrando avaliadores e um comitê decisão, o que pode ajudar a mitigar vieses individuais de modelos.

Dataset original

Revisores	ID	CODE	STATUS	Applied Criterio	REASON	CATEG	SUB-CATE	YEAR	COUNTRY	SOURCE	TITLE	AUTHORS	ABSTRACT	URL
Juliana	1	M153	Excluded	ec7				2009	Austria	REC	[vem:xi:] - A Method	Philipp Lie	Service-oriented	http://ieeexplor
Juliana	2	M10	Excluded	ec7				2012	USA	REC	Reconciling Multi-ju	David G. G	Companies that	http://ieeexplor
Juliana	3	M53	Excluded	ec7				2013	Germany	REC	User Feedback in the	Dennis Pag	Application distr	https://wwwbru
Juliana	4	S499	Excluded	ec4						ACM	"Lee's law"	John A. N.		http://dl.acm.or
Juliana	5	M207	Included					2013	Denmark	AGILE	"Scrum Code Camps"	Lene Pries-H	A classic way to	http://dx.doi.or
Juliana	6	S295	Excluded	ec2				2013	Denmark	IEEE	"Scrum Cod	Pries-Heje,	A classic way to	http://ieeexplor
Juliana	7	S1672	Included					2007	EUA	CIENCE_DIR	\{CASE\} 9 - The \{H	Randolph G	Publisher Summ	http://www.scie
Juliana	8	S1927	Excluded	ec7				2004		CIENCE_DIR	\{CHAPTER\} 12 - D	Susan Fowl	er, Victor Stanwic	http://www.scie
Juliana	9	S1596	Excluded	ec4						CIENCE_DIR	\{CHAPTER\} 3 - Ha	Thomas Wi	Publisher Summ	http://www.scie
Juliana	10	S2277	Excluded	ec4						CIENCE_DIR	\{CHAPTER\} 4 - Bu	A. BURGER	, J.-O. NDAP, K	http://www.scie
Juliana	11	S2220	Excluded	ec4						CIENCE_DIR	\{INDEX\}			http://www.scie
Juliana	12	S2106	Excluded	ec3						CIENCE_DIR	\{IT\} support for clin	R. Lenz, R.	Clinical pathway	http://www.scie
Juliana	13	S2282	Excluded	ec2						CIENCE_DIR	\{JOSS\} â€" II: Des	Joseph W S	JOSSâ€¡â€¡JOS	http://www.scie
Juliana	14	S1637	Excluded	ec7						CIENCE_DIR	\{MC\} Sandbox: De	Fredrik Kar	Context Method	http://www.scie
Juliana	15	S1881	Excluded	ec7						CIENCE_DIR	\{MDD\} vs. traditior	Yulkeidi Ma	Context Todayâ€	http://www.scie
Juliana	16	S1909	Excluded	ec7						CIENCE_DIR	\{MDE\} software pr	Julio Ariel H	Software organiz	http://www.scie
Juliana	17	S1753	Excluded	ec7						CIENCE_DIR	\{MNE\} software fo	Alexandre C	Abstract Magnet	http://www.scie
Juliana	18	S1851	Excluded	ec4						CIENCE_DIR	\{REFERENCES\}			http://www.scie

Tabela de Dataset Original. Fonte: Adaptado de (MEDEIROS, 2015).

Fluxo Metodológico Geral

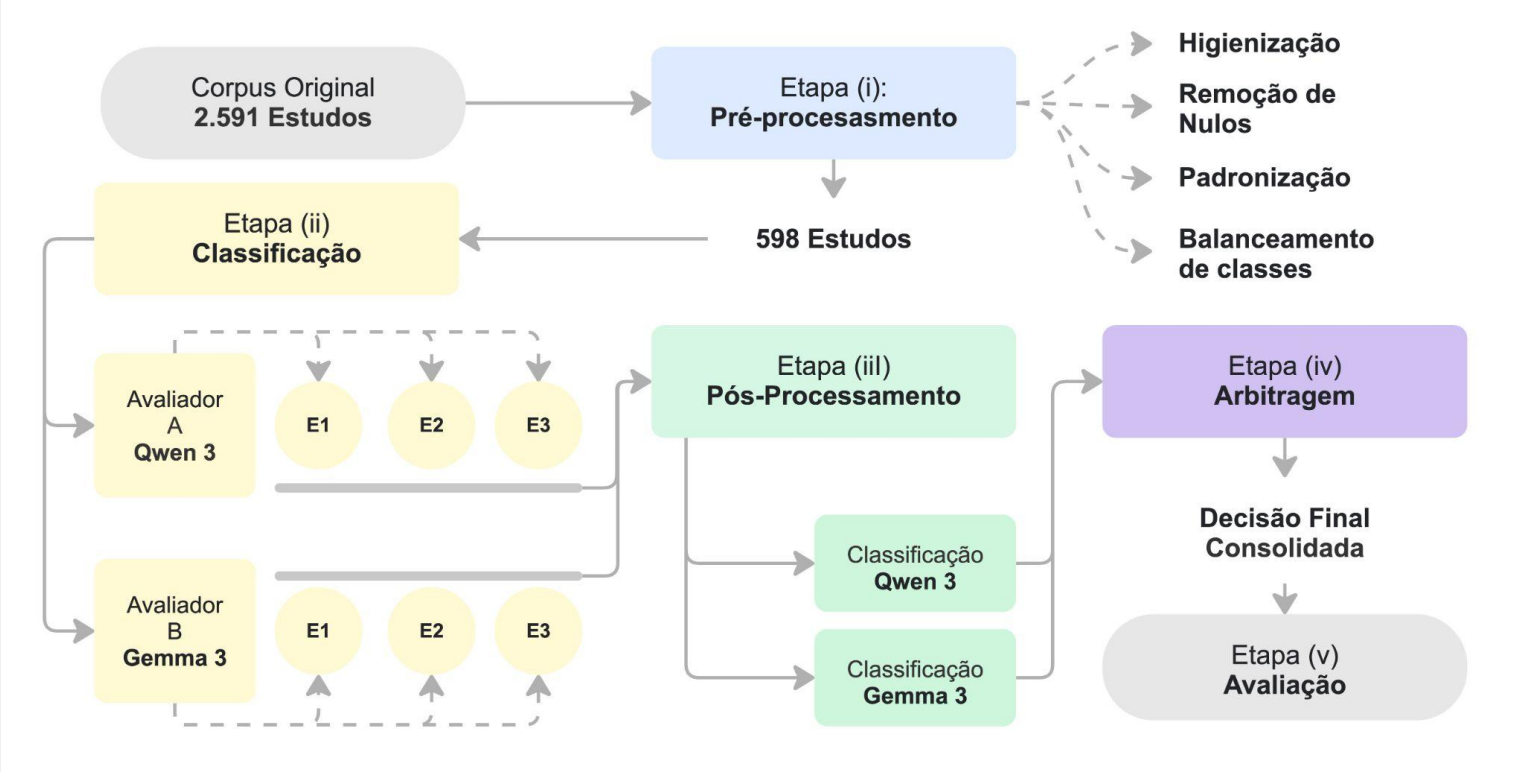

Figura Fluxo Metodológico Geral. Fonte: Produzido pelo autor, 2025.

Fluxo Metodológico Arbitragem

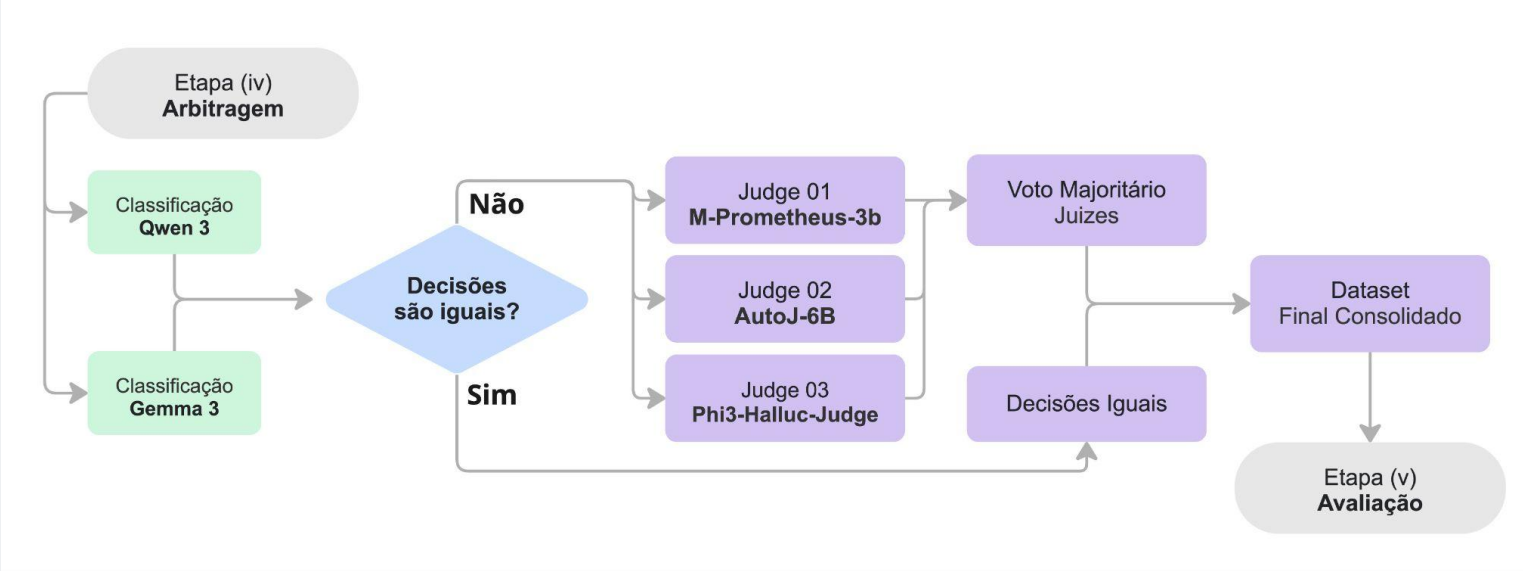

Figura Fluxo Arbitragem. Fonte: Produzido pelo autor, 2025.

Fluxo Metodológico Avaliação

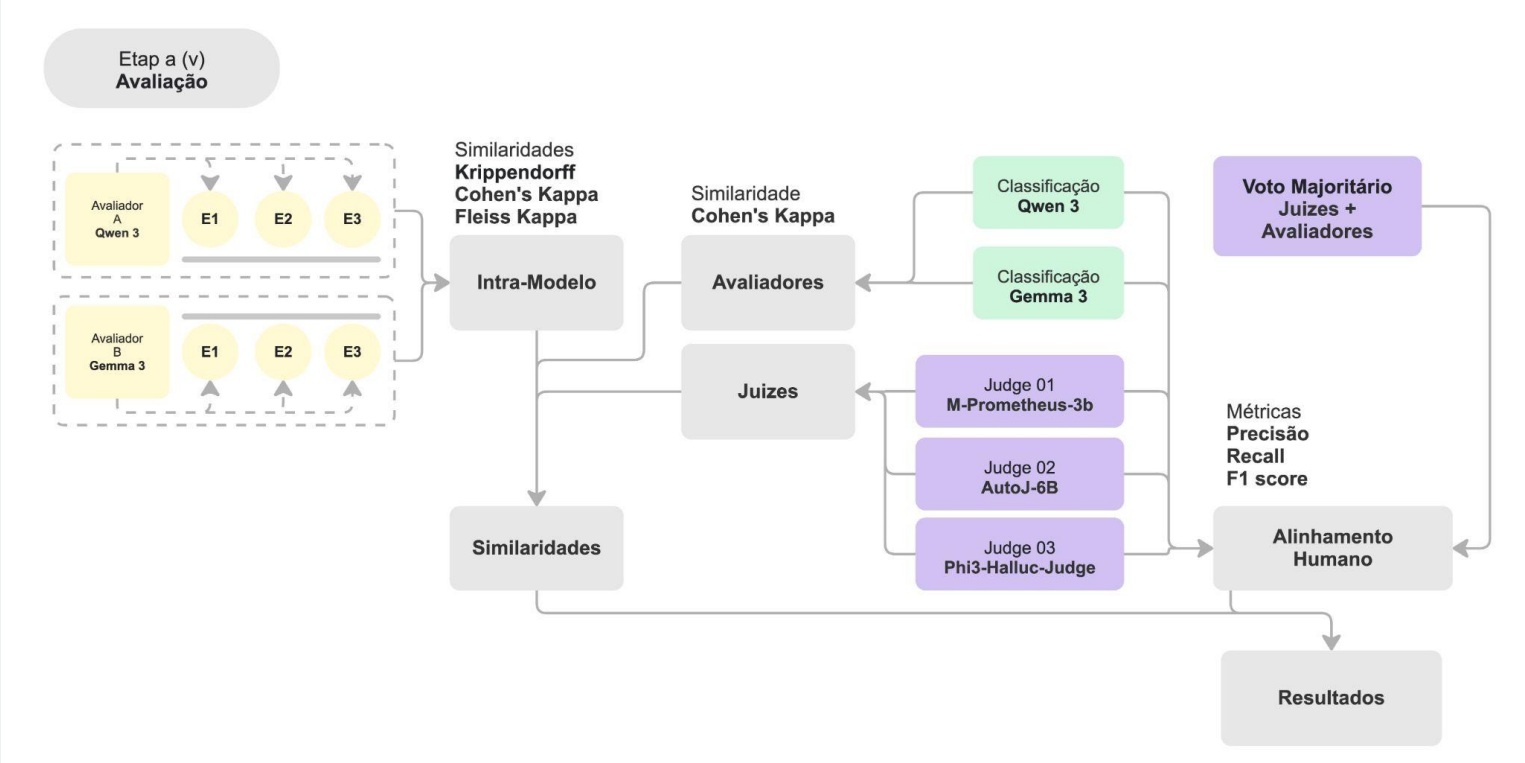

Figura Fluxo Avaliação. Fonte: Produzido pelo autor, 2025.

Modelos Utilizados

Tabela Modelos Utilizados. Fonte: Produzido pelo autor, 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.

Dataset Resultante

STATUS	RESULT_A_1	EXP	RESULT_A_2	EXP	RESULT_A_3	EXP	FINAL_RESULT_A	FINA	RESULT_B_1	EXP	RESULT_B_2	EXP	RESULT_B_3	FINAL_RESULT_B	FINA	RESULT_JUDGE_A	EXP	RESULT_JUDGE_B	EXP	RESULT_JUDGE_C	EXP	FINAL_RESULT
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	<brie	INCLUDED
INCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED		INCLUDED		INCLUDED		INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	EXCLUDED	Deci	INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	EXCLUDED	The	EXCLUDED		EXCLUDED		EXCLUDED		EXCLUDED
INCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED		INCLUDED		INCLUDED		INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED	Deci	INCLUDED	The	EXCLUDED	Justi	INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED	The	EXCLUDED	As p	EXCLUDED	Justi	INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	EXCLUDED	The	EXCLUDED	EXCLUDED	The	EXCLUDED		EXCLUDED		EXCLUDED		EXCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	EXCLUDED	Justi	INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED	Deci	INCLUDED	The	INCLUDED	<brie	INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	EXCLUDED	The	EXCLUDED		EXCLUDED		EXCLUDED		EXCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	EXCLUDED	Justi	INCLUDED
INCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED		INCLUDED		INCLUDED		INCLUDED
INCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED		INCLUDED		INCLUDED		INCLUDED
INCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED		INCLUDED		INCLUDED		INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED	Deci	INCLUDED		EXCLUDED	Justi	INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED	Deci	INCLUDED	The	EXCLUDED	Justi	INCLUDED
EXCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED		INCLUDED		INCLUDED		INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED	Deci	INCLUDED	Deci	INCLUDED		INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED	Whil	INCLUDED	The	EXCLUDED	Justi	INCLUDED
INCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	Deci	EXCLUDED	Justi	INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	EXCLUDED	The	EXCLUDED		EXCLUDED		EXCLUDED		EXCLUDED
INCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	This	EXCLUDED	This	EXCLUDED	EXCLUDED	This	EXCLUDED		EXCLUDED		EXCLUDED		EXCLUDED
INCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED		INCLUDED		INCLUDED		INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	EXCLUDED	The	EXCLUDED		EXCLUDED		EXCLUDED		EXCLUDED
INCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED		INCLUDED		INCLUDED		INCLUDED
EXCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED		INCLUDED		INCLUDED		INCLUDED
INCLUDED	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED		INCLUDED		INCLUDED		INCLUDED
EXCLUDED	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	EXCLUDED	The	INCLUDED	The	INCLUDED	The	INCLUDED	INCLUDED	The	INCLUDED	Deci	INCLUDED	The	EXCLUDED	Deci	INCLUDED

Tabela de dataset produzido. Fonte: Produzido pelo autor, 2025.

Legenda

🟪 **Consenso**
🟥 **Included**
🟩 **Excluded**
🟦 **Ground Truth**
🟨 **Qwen3**
🟦 **Gemma3**
🟪 **M-Prometheus**
AutoJ
Phi3-Halluc

Resultados - Concordância: Avaliadores Primários

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

Tabela de Concordância entre avaliadores. Fonte: Produzido pelo autor, 2025.

Resultados - Concordância: Árbitros e *Ground Truth*

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

Tabela de Concordância entre juízes. Fonte: Produzido pelo autor, 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

Tabela de Concordância com Groung Truth. Fonte: Produzido pelo autor, 2025.

Alinhamento com *Ground Truth* - Qwen3-4b-instruct

Matriz de Confusão

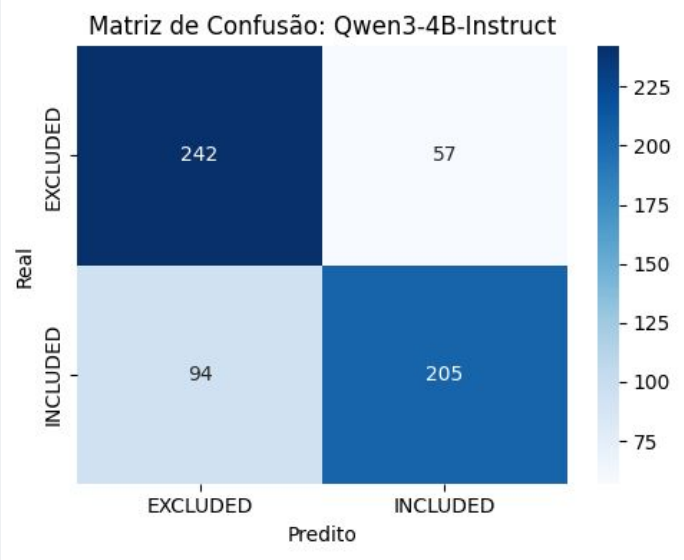

Tabela Matriz de Confusão Qwen3-4b.

Fonte: Produzido pelo autor, 2025.

Métricas

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

Tabela Métricas Qwen3-4b. Fonte: Produzido pelo autor, 2025.

Alinhamento com *Ground Truth* - Gemma3-4b-it

Matriz de Confusão

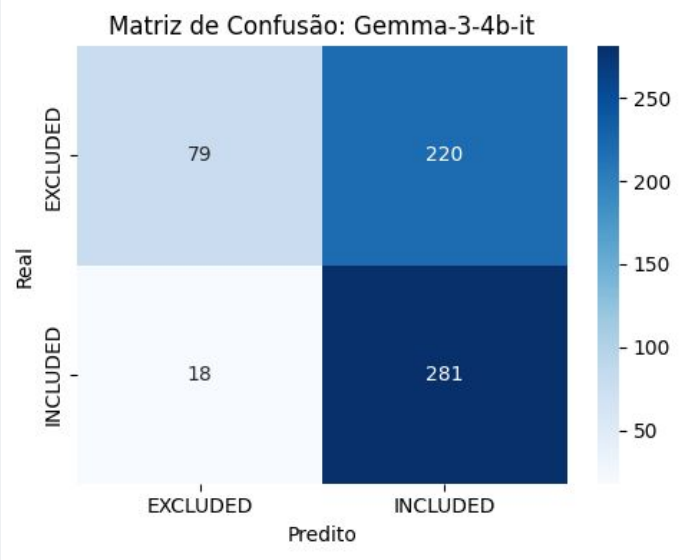

Tabela Matriz de Confusão Gemma3-4b.

Fonte: Produzido pelo autor, 2025.

Métricas

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

Tabela Métricas Gemma3-4b. Fonte: Produzido pelo autor, 2025.

Alinhamento com *Ground ruth* - Árbitros

M-Prometheus

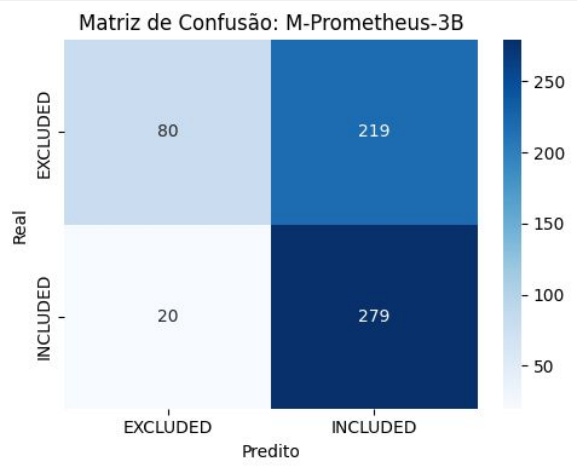

Tabela Matriz de Confusão M-Prometheus.

Fonte: Produzido pelo autor, 2025.

Auto-J

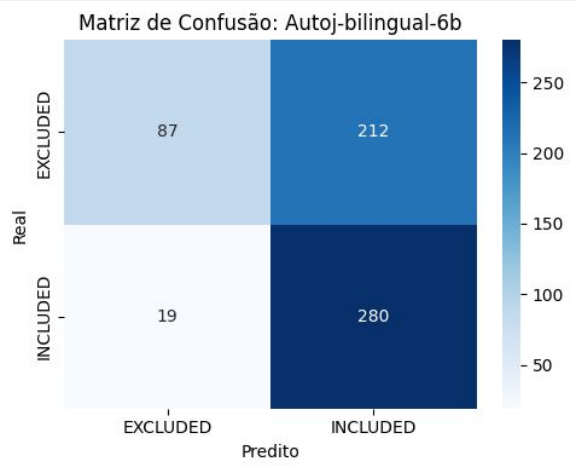

Tabela Matriz de Confusão Auto-J.

Fonte: Produzido pelo autor, 2025.

Phi-Hallucination-Judge

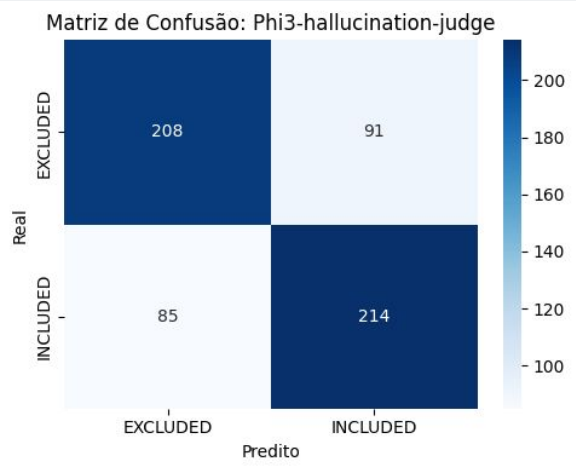

Tabela Matriz de Confusão Phi3-hallucination.

Fonte: Produzido pelo autor, 2025.

Alinhamento com *ground truth* - Consenso Final.

Matriz de Confusão

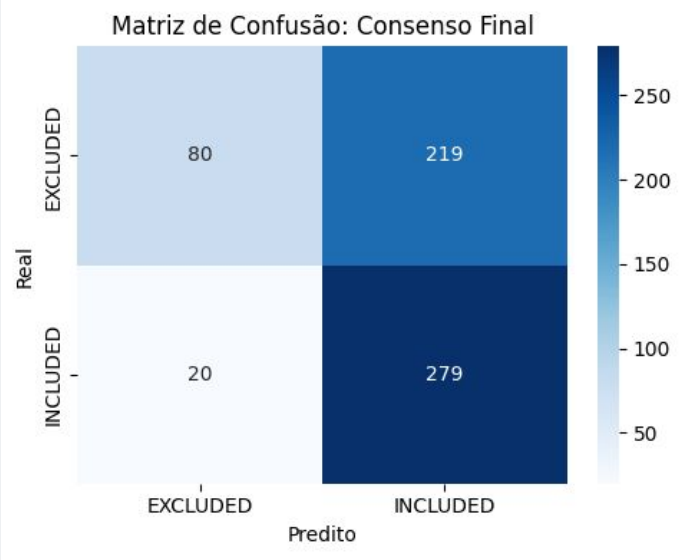

Tabela Matriz de Confusão Consenso Final.

Fonte: Produzido pelo autor, 2025.

Métricas

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

Tabela Métricas Consenso Final. Fonte: Produzido pelo autor, 2025.

Explicabilidade - Exemplo A

Tabela Explicabilidade A. Fonte: Produzido pelo autor, 2025.

Ground Truth	Qwen3	Gemma3	M-prometheus	Auto-J	Phi-Halluc	Consenso
EXCLUDED	EXCLUDED	INCLUDED	INCLUDED	INCLUDED	EXCLUDED	INCLUDED

Justificativas - Respectivamente

Decision B correctly identifies that the article addresses software project requirements using agile methodologies, aligning with Inclusion Criterion 1 (IC1). It also confirms the study's academic nature, meeting Inclusion Criterion 2 (IC2). Additionally, while Decision A incorrectly assumes the article does not engage with agile methodologies due to the broad title, Decision B recognizes the implicit discussion of processes and methodologies in the abstract, suggesting a qualitative or quantitative approach consistent with IC3 and IC4.

Decision B correctly applies the inclusion and exclusion criteria, while Decision A fails to do so. Decision B correctly identifies the article as addressing requirements in software projects using agile methodologies, and as a study from academia. Decision A incorrectly identifies the article as failing to address at least one agile methodology and not specifying any study on software requirements in agile projects.

Justification A provides a clear rationale for exclusion based on the provided EC7 and IC1 criteria, directly referencing the abstract's lack of detail regarding agile methodologies and software requirements in agile projects.

Explicabilidade - Exemplo B

Tabela Explicabilidade B. Fonte: Produzido pelo autor, 2025.

Ground Truth	Qwen3	Gemma3	M-prometheus	Auto-J	Phi-Halluc	Consenso
INCLUDED	EXCLUDED	INCLUDED	INCLUDED	INCLUDED	EXCLUDED	INCLUDED

Justificativas - Respectivamente

Decision B correctly identifies that the article addresses requirements in software projects using agile methodologies, which aligns with Inclusion Criterion IC1. It also confirms that the study is from industry (IC2), is a primary study (IC4), and uses qualitative research (IC3). Moreover, it avoids violating any exclusion criteria, including EC1 through EC9, thus making it the more accurate decision.

The article is more accurately included than excluded. It addresses requirements in software projects using agile methodologies and is a study from industry. It also meets the inclusion criteria IC1, IC2, and IC4. Although it does not explicitly mention agile methodologies, its focus on Feature-Oriented Requirements Engineering (FORE) and its integration with other engineering disciplines suggests that it does address at least one agile methodology. The exclusion justification provided in Decision A is not accurate as the article does not violate Exclusion Criterion EC7.

Decision B incorrectly assumes the article addresses agile methodologies, while the abstract specifically states the use of Feature-Oriented Requirements Engineering (FORE) without mentioning agile methodologies, thus violating EC7.

Principais Contribuições

- Demonstra que SLMs compactos são tecnicamente viáveis e economicamente acessíveis para triagem em larga escala.
- Mostra que o consenso de múltiplos modelos pode ser pior que o melhor modelo individual, exigindo ponderação na agregação.
- Caracteriza perfis de modelos (mais rigorosos vs. mais permissivos) e obtém Recall acima de 93% para a classe *Included*, adequado para apoio seguro à RSL.
- Expõe limitações da votação majoritária simples e aponta direções para sistemas híbridos de triagem mais robustos.

Limitações

- O fluxo de trabalho experimental não incorporou a abordagem ***Human-in-the-loop*** durante a execução do processo decisório.
- A validação dos modelos foi realizada utilizando conjuntos de dados artificialmente **balanceados**.
- Inferência realizada apenas com configurações padrão, sem explorar **otimização de hiperparâmetros**.

Trabalhos futuros

- Aplicar 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.
- Investigar a eficácia de uma abordagem sequencial, utilizando modelos sensíveis apenas para a fase de recuperação inicial, seguidos por modelos de maior rigor lógico para uma segunda etapa de filtragem.
- Evoluir o pipeline através da integração com componente humano, estabelecendo uma camada de avaliação humana para a validação de divergências e decisões complexas.

Até mais, e obrigado pelos peixes

(ADAMS, 1979).

Referências

AGÊNCIA BRASIL. Impacto acadêmico da ciência brasileira aumentou 21% de 1996 a 2022: Brasil teve nove vezes mais publicações científicas no período. 29 nov. 2023.

BULLERS, K. et al. It takes longer than you think: librarian time spent on systematic review tasks. *Journal of the Medical Library Association*, v. 106, n. 2, p. 198–207, 2018. DOI: 10.5195/jmla.2018.323.

FREEPIK. Freepik – Banco de imagens. Disponível em: https://www.freepik.com. Acesso em: 8 dez. 2025.

GRABOIS, A. Brasil retoma sua produção científica com crescimento de 6%. *Grabois*, 26 mar. 2025. Disponível em: https://grabois.org.br/2025/03/26/producao-cientifica-brasil-cresce-6/. Acesso em: 8 dez. 2025.

HORA DE PUBLICAR. Produção científica brasileira em 2025: avanço quantitativo, dilemas estruturais e caminhos possíveis. *Hora de Publicar*, 25 jul. 2025. Disponível em: https://horadepublicar.com.br/producao-cientifica-brasileira-em-2025-avanco-quantitativo-dilemas-estruturais-e-caminhos-possiveis. Acesso em: 8 dez. 2025.

Referências

KITCHENHAM, B. Procedures for performing systematic reviews. *Keele University Technical Report* TR/SE-0401, 2004.
Disponível em: https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf. Acesso em: 8 dez. 2025.

LIEBERUM, J.-L. et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use — a scoping review. *Journal of Clinical Epidemiology*, v. 181, p. 111746, 2025. DOI: 10.1016/j.jclinepi.2025.111746.

QURESHI, R. et al. Are ChatGPT and large language models “the answer” to bringing us closer to systematic review automation? *Systematic Reviews*, v. 12, n. 1, p. 72, 2023.

SHANNON, H. A statistical note on Karl Pearson’s 1904 meta-analysis. *Journal of the Royal Society of Medicine*, v. 109, n. 8, p. 310–311, 2016.

SHEN, Y. et al. ChatGPT and other large language models are double-edged swords. *Radiology*, v. 307, n. 2, e230163, 2023. DOI: 10.1148/radiol.230163.

WANG, Y. et al. Large Language Model Agent: A Survey on Methodology, Applications and Challenges. *arXiv preprint*, arXiv:2404.08536, 2024.

ZUBIAGA, I.; SOROA, A.; AGERRI, R. A LLM-Based Ranking Method for the Evaluation of Automatic Counter-Narrative Generation. *arXiv preprint*, arXiv:2406.15227, 2024.