Abordagens de Aprendizado de Máquina para Automatização de Etapas do Processo de Meta-Análise no Contexto da Saúde

Marianna D. Verduin[1], **Vinicius K. Lodi**[1], **Julio C. S. Reis**[1]

[1]Universidade Federal de Viçosa (UFV) – Brasil

{marianna.verduin,vinicius.lodi,jreis}@ufv.br

Abstract. *This work evaluates machine learning approaches for automating key stages of healthcare meta-analysis, focusing on PICO extraction and study screening. A BioELECTRA model fine-tuned for NER was compared with Large Language Models – LLMs (i.e., Llama 3 and Gemini) using the EBM-NLP dataset and a derived set of 55 meta-analyses for the task of extracting of PICO entities for article screening. BioELECTRA achieved higher recall and F1-score in entity extraction, while the LLMs outperformed the baseline in article ranking (Precision@1: 1.00 vs. 0.84), suggesting that hybrid NER–LLM pipelines are promising for automating meta-analysis in healthcare.*

Resumo. *Este trabalho avalia abordagens de aprendizado de máquina para automatizar etapas da condução de meta-análises em saúde, com ênfase na extração de entidades PICO e triagem de estudos. Um modelo BioELECTRA ajustado para NER foi comparado a Grandes Modelos de Linguagem – LLMs (i.e., Llama 3 e Gemini) usando a base de dados EBM-NLP e um conjunto derivado de 55 meta-análises na tarefa de extração de entidades PICO para ser utilizada na triagem de artigos. O BioELECTRA obteve maior recall e F1-score na extração de entidades PICO, enquanto as LLMs superaram o baseline no ranqueamento de artigos (Precisão@1: 1,00 vs. 0,84), indicando que pipelines híbridos NER–LLM são promissores para automatizar a condução de meta-análises na área da saúde.*

1. Introdução

A prática da medicina baseada em evidências depende de sínteses confiáveis do conhecimento científico disponível. Nesse cenário, a *revisão sistemática* (RS) organiza, de forma transparente e reprodutível, a identificação, seleção e avaliação crítica de estudos relevantes para uma pergunta bem definida; quando a combinação quantitativa é apropriada, a *meta-análise* aplica modelos estatísticos para agregar estimativas de efeito e produzir uma síntese numérica, levando em conta incertezas e possíveis diferenças entre estudos [Page et al. 2021]. É importante ressaltar que a meta-análise não é o maior nível de evidência em qualquer contexto, a credibilidade do resultado depende diretamente da qualidade metodológica dos estudos incluídos, da consistência entre eles, do risco de vieses e das escolhas analíticas reportadas [Higgins et al. 2024].

Neste contexto, a literatura aponta um crescimento acelerado no volume de revisões sistemáticas e meta-análises indexadas nas últimas décadas, tornando comum a sobreposição de revisões sobre o mesmo tema e elevando o custo (em termos de tempo e mão de obra especializada, por exemplo) para manter sínteses atualizadas [Hoffmann et al. 2021, Jakab 2024]. Embora existam algumas iniciativas no sentido

de automatização deste processo [Tsafnat et al. 2018], este cenário ainda carece da proposição de abordagens computacionais que auxiliem etapas críticas do processo, reduzindo retrabalho e, principalmente, diminuindo erros humanos em tarefas repetitivas.

Em linhas gerais, RS e meta-análise seguem um fluxo estruturado: 1) definição da pergunta e do protocolo, 2) busca em bases relevantes, 3) triagem por critérios de inclusão/exclusão, 4) extração padronizada de variáveis e 5) síntese (qualitativa e/ou quantitativa) [Egger et al. 2008]. A formulação da pergunta costuma ser operacionalizada por critérios como o PICO (*Participant*, *Intervention*, *Comparison* e *Outcome*), que ajuda a tornar explícitos os elementos a serem buscados e extraídos [Richardson et al. 1995, Santos et al. 2007].

Nos últimos anos, técnicas de aprendizado de máquina e processamento de linguagem natural (PLN) vêm sendo investigadas para apoiar diferentes fases desse *pipeline* (por exemplo, triagem de títulos/resumos [Wallace et al. 2010b], extração de informação [O'Mara-Eves et al. 2015] e avaliação de relevância [Guo et al. 2024]), com evidências recentes indicando que Grandes Modelos de Linguagem (LLMs) podem ser úteis em tarefas específicas quando avaliados de forma criteriosa [Ofori-Boateng et al. 2024, Oami et al. 2024]. Ao mesmo tempo, esses modelos introduzem riscos próprios (por exemplo, inconsistência e alucinações), o que reforça a necessidade de validação sistemática e de conjuntos de dados rotulados que permitam comparar abordagens de maneira objetiva [Higgins et al. 2024, Ofori-Boateng et al. 2024].

Diante desse contexto, o objetivo deste trabalho é prover uma avaliação diagnóstica do potencial de abordagens baseadas em aprendizado de máquina e PLN para a realização automatizada de etapa de triagem de artigos para RS e meta-análises no contexto da saúde. Em particular, o trabalho foca em construir uma base de dados rotulada usando LLMs como meio de análise de rotulação e avaliar o desempenho de diferentes modelos no contexto analítico médico, com o propósito de auxiliar (e viabilizar a avaliação) da automatização da etapa de triagem de artigos.

De forma geral os resultados identificados revelam que modelos especializados em reconhecimento de entidades nomeadas (NER), como o BioELECTRA [Kanakarajan et al. 2021], superam as LLMs na extração estruturada de entidades PICO em termos de métricas como *recall* e *F1-score*, ao passo que as entidades PICO extraídos por LLMs demonstram desempenho superior quando utilizadas na tarefa de ranqueamento e triagem de artigos candidatos, com *Precision@1* de 1,00 *vs.* 0,84 do BioELECTRA. Em suma, essas descobertas sugerem que abordagens híbridas, que combinam modelos NER com LLMs, constituem uma estratégia promissora para a automatização de meta-análises, aproveitando os pontos fortes complementares de cada família de modelos.

O restante do trabalho está organizado da seguinte forma. Na próxima seção, apresentamos trabalhos relacionados. Em seguida, na Seção 3, descrevemos a metodologia experimental proposta para o trabalho. A Seção 4 apresenta os resultados obtidos nas tarefas de extração de entidades PICO e identificação de artigos relevantes. Por fim, a Seção 5 traz as conclusões e aponta direções para trabalhos futuros.

2. Trabalhos Relacionados

Diversos trabalhos investigam a automatização de etapas do processo da revisão sistemática [Marshall and Wallace 2019]. Estudos iniciais que abordam o pro-

cesso de triagem de artigos exploram técnicas de *text-mining* junto com aprendizado supervisionado para ranquear artigos por relevância [Wallace et al. 2010a, Wallace et al. 2010b]. Algumas ferramentas, como Rayyan [Ouzzani et al. 2016], Covidence [Veritas Health Innovation 2024] e Abstrackr [Wallace et al. 2012], utilizam SVM e modelos de regressão logística combinado com aprendizado ativo para ranquear artigos baseado nas decisões feitas por usuários. Porém, essas abordagens apresentam limitações por dependerem de rótulos fornecidos pelo pesquisador no momento de triagem para melhorar a acurácia do ranqueamento, além da quantidade extensa de dados rotulados necessários para treinamento prévio [Marshall and Wallace 2019]. Outras abordagens mais recentes incorporam o uso de LLMs no processo de identificação de relevância de artigos. Um estudo recente propôs templates de *prompts* para formulação de estratégias inteligentes na identificação de artigos relevantes, apresentando alto desempenho de LLMs na triagem de mais de 40.000 resumos [Cao et al. 2025], o que evidencia modelos generativos como uma direção promissora para o processo de triagem.

Ainda neste contexto, um desafio recorrente enfrentado é a falta de rigor que LLMs possuem na busca, triagem e extração de informações relevantes [Wang et al. 2025]. O processo de realização de uma RS é rigoroso, exigindo transparência, rastreabilidade e reprodutibilidade, com guias definitivas (PRISMA [Page et al. 2021], *Cochrane Handbook* [Higgins et al. 2024], etc) que pesquisadores devem usar para assegurar a qualidade da pesquisa [Page et al. 2021]. Para lidar com as limitações de LLMs nesse quesito, trabalhos recentes propõem a automatização do processo de revisão sistemática estilo-PRISMA em etapas modulares, junto com suas 4 fases e *checklist* de 27 partes, com o objetivo de preservar a acurácia e rigidez executada manualmente por humanos [Wang et al. 2025]. O TrialMind-SLR [Wang et al. 2025] permite a realização desta tarefa através de etapas bem definidas e documentadas. O processo de triagem e identificação de relevância tem critérios de elegibilidade claramente definidos e codificados, e cada decisão feita pela LLM é recordada para deixar a possibilidade de avaliação futura. Isso é de acordo com as regras estabelecidas pelo PRISMA, que exige a possibilidade de replicabilidade por pesquisadores futuros [Moher et al. 2015]. De forma complementar, também foi proposto um *framework* para "*living meta-analyses*" que integra LLMs e técnicas de PLN sobre revisões e meta-análises previamente publicadas, mostrando ser possível automatizar a triagem e, em grande parte, a extração mantendo a natureza estrita do processo [Górska and Tacconelli 2024].

Além de propostas baseadas em templates, também há evidências empíricas de que LLMs podem atuar como suporte à triagem de títulos/resumos com sensibilidade aceitável e redução de tempo, desde que avaliadas de forma criteriosa e com salvaguardas contra inconsistências e erros sistemáticos [Oami et al. 2024, Ofori-Boateng et al. 2024]. Esses resultados reforçam que, embora LLMs sejam promissoras em tarefas específicas do *pipeline*, a adoção prática depende de validação sistemática, padronização de decisões e mecanismos de auditoria [Higgins et al. 2024, Ofori-Boateng et al. 2024].

Um eixo central para tornar decisões e extrações mais verificáveis é a operacionalização por PICO, que explicita o que deve ser buscado e extraído em cada etapa [Richardson et al. 1995, Santos et al. 2007]. Nesse contexto, bases públicas anotadas manualmente (por exemplo, EBM-NLP [Nye et al. 2018]) têm papel importante como referência supervisionada (*gold standard*) para comparar abordagens de extração

estruturada, incluindo *pipelines* baseados em NER (*span extraction*) e extrações geradas por LLMs, que frequentemente exigem pós-processamento e normalização para reduzir variação de formato e permitir avaliação reprodutível [Xia et al. 2024].

Em síntese, a literatura aponta um espectro que vai de *prompting* estruturado para tarefas isoladas (triagem/extração) até *pipelines* modulares fim-a-fim orientados por guias (PRISMA/Cochrane). Ainda assim, persistem lacunas para o cenário de meta-análise automatizada: uma revisão sistemática recente sobre a evolução de *Automated Meta-Analysis (AMA)* sugere que a maior parte dos esforços ainda se concentra na automação de etapas de processamento, enquanto a automação verdadeiramente fim-a-fim permanece rara [Li et al. 2025]; integração consistente entre triagem/relevância, extração PICO e ranqueamento; disponibilidade de conjuntos rotulados comparáveis; e rastreabilidade detalhada para auditoria posterior. Este trabalho se posiciona nesse espaço ao focar explicitamente nessas três etapas e ao adotar referências supervisionadas e estratégias de padronização de saídas para viabilizar comparação objetiva entre abordagens.

3. Metodologia

Nesta seção apresentamos a metodologia experimental proposta para este trabalho.

3.1. Desenho do Estudo e Visão Geral do *Pipeline*

Este trabalho conduz uma avaliação diagnóstica do potencial de abordagens de aprendizado de máquina e PLN para automatizar etapas centrais de uma revisão sistemática com meta-análise, com ênfase em (i) extração estruturada de elementos PICO e (ii) ranqueamento de estudos candidatos para inclusão, conforme ilustrado na Figura 1. Em consonância com a literatura, consideramos o fluxo padrão: 1) pergunta e protocolo, 2) busca, 3) triagem, 4) extração e 5) avaliação/síntese, como referência conceitual para decompor o problema em subtarefas automatizáveis [Page et al. 2021, Higgins et al. 2024].

3.2. Bases de Dados e Estratégias de Rotulação

Base de dados pública rotulada (EBM-NLP) para extração PICO. Como referência supervisionada (*gold standard*) para a tarefa de extração dos critérios PICO em resumos biomédicos, exploramos o conjunto de dados EBM-NLP, composto por aproximadamente 5.000 resumos anotados manualmente com marcações para elementos PICO [Nye et al. 2018]. Essa base é particularmente adequada por refletir o formato clássico de extração em nível de trecho (*span extraction*), permitindo uma comparação controlada entre as abordagens baseadas em aprendizado de máquina e PLN exploradas neste estudo e apresentadas nas seções subsequentes.

Base de dados derivada de meta-análises (rotulagem via LLM) para triagem e ranqueamento. Além do *gold standard* de extração, no contexto deste estudo, construímos

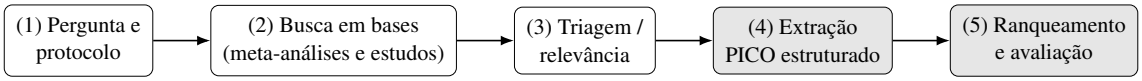

Figura 1. Visão geral do *pipeline* modelado a partir do fluxo de revisão sistemática/meta-análise. O escopo deste trabalho foca especificamente nas etapas (4) e (5) de extração PICO e ranqueamento e avaliação, respectivamente.

Tabela 1. Sumarização das bases de dados utilizadas e/ou geradas. "300 (alvo)" refere-se ao tamanho buscado na etapa automatizada.

Fonte	Unidade	Ordem de grandeza
EBM-NLP	*abstracts* anotados	∼5.000
Meta-análises (base de dados derivada)	meta-análises-âncora	300 (alvo)
Artefatos `.done_ids`	IDs processados	100 / 320

uma base de dados orientada ao cenário de meta-análise automatizada, usando meta-análises publicadas como "âncoras" de consulta. De forma resumida, o processo consiste em: (i) buscar um conjunto de meta-análises com texto completo disponível; (ii) extrair o PICO da meta-análise (PICO-consulta); (iii) coletar a lista de referências do artigo (candidatos); (iv) classificar relevância de cada candidato comparando seu resumo com o PICO-consulta; e, (v) extrair o PICO dos candidatos relevantes para compor pares (meta-análise $\rightarrow$ estudo incluído).

As bases de dados utilizadas para busca e coleta de meta-análises foram extraídas a partir de repositórios amplamente reconhecimentos no contexto da saúde, a saber: PubMed e EuropePMC[1]. Os critérios de busca foram definidos para permitir a recuperação de meta-análises ou revisões sistemáticas publicadas entre os anos 2018 e 2026 com abstrato disponível. Para este propósito, especificamente, palavras-chave como "metanálise", "meta-análise" e *"meta-analysis"*, foram exploradas.

A partir disso, foi construída uma base de dados composta por 55 meta-análises. A Tabela 1 sumariza as bases de dados utilizadas em ambas as etapas de coleta, incluindo o alvo original de 300 meta-análises-âncora. A redução de número de meta-análises na base de dados final é um resultado de falha de recuperação de textos do PubMed ou EuropePMC, ou falha por parte dos modelos na extração de entidades PICO. As referências bibliográficas de cada meta-análise foram então coletadas e processadas por uma LLM utilizada como avaliadora, responsável por determinar se cada referência correspondia a um estudo incluído na meta-análise ou a uma citação complementar.

Em resumo, os estudos identificados como incluídos foram armazenados na base de dados e rotulados como relevantes, juntamente com uma justificativa textual fornecida pela LLM para sua classificação, com o objetivo de garantir rastreabilidade e transparência nas decisões automatizadas. Foi selecionada aleatoriamente uma amostra de 20% das meta-análises (n = 11) e, para cada uma, 20% de suas referências, totalizando 51 artigos. A partir desse conjunto, foi avaliado manualmente se o artigo foi incluído ou não na meta-análise, junto com a justificativa fornecida pela LLM. O objetivo desta etapa foi analisar o comportamento da LLM e rastrear sua linha de raciocínio, identificando possíveis divergências no momento da classificação. Especificamente, foi identificada uma taxa de acerto de 0.76. Entre as amostras avaliadas, observou-se uma tendência de a LLM classificar como relevantes estudos alinhados ao tema de pesquisa, ainda que não contemplem todas as entidades PICO. Como a triagem de abstratos tem a característica de ser mais inclusiva do que a triagem por leitura completa, consideramos essa tendência válida para as experimentações iniciais, porém necessitando uma avaliação mais aprofundada em trabalhos futuros. O objetivo desse desenho é aproximar o comportamento do fluxo manual (busca, triagem e extração) e, simultaneamente, produzir um conjunto rotulado que viabilize avaliação objetiva de modelos e estratégias investigados neste estudo.

[1] `https://pubmed.ncbi.nlm.nih.gov/` e `https://europepmc.org/`

3.3. Modelos e Abordagens Avaliadas

No contexto deste estudo, consideramos duas famílias de abordagens para extração PICO e suporte ao ranqueamento:

Modelo baseado em NER – *Baseline*. O *baseline* segue a formulação de *span extraction* e é diretamente compatível com o EBM-NLP, servindo como referência de desempenho para elementos PICO em resumos biomédicos [Nye et al. 2018].

Extração e triagem assistidas por LLMs. A abordagem por LLMs é empregada em duas frentes: (i) extração do PICO-consulta de meta-análises e (ii) triagem/avaliação de relevância de candidatos e extração de seus respectivos PICOs. Como estratégia prática, exploramos modelos proprietários com maior capacidade para extração (melhor estruturação, como o Gemini[2] e modelos abertos menores para triagem (maior eficiência, como o Llama 3[3]), seguindo a hipótese de que decisões de inclusão/exclusão podem ser obtidas com menor custo computacional do que extrações estruturadas completas.

3.4. Pós-Processamento e Normalização

Saídas de LLMs podem conter texto extra, formatação variável e duplicidades [Geng et al. 2025]. Para tornar as predições comparáveis e consistentes, aplicou-se um pós-processamento com três objetivos: (i) sanitização para manter apenas a estrutura JSON quando requerido, (ii) padronização de chaves e campos, e (iii) deduplicação/normalização superficial de entidades (e.g., remoção de repetições e ajustes simples de caixa/pontuação). Essa etapa é crucial tanto para reduzir falsos erros por formatação quanto para permitir métricas automáticas reprodutíveis.

3.5. Protocolo de Avaliação: Nível Sintático (Léxico) e Nível Semântico

A avaliação é estruturada em dois níveis complementares, contemplando correspondência textual estrita, equivalência semântica e um cenário aplicado de ranqueamento.

Avaliação sintática (Léxica). No nível léxico, comparamos predição e gabarito por sobreposição de conjuntos de termos após normalização simples do texto. Utilizamos: (i) *Jaccard Index* [Jaccard 1912] para verificar se a interseção relativa excede um limiar pré-definido de 0.5[4], (ii) uma métrica de contenção para capturar casos de especificação, verificando se dois trechos extraídos estão contido entre si (e.g., "aspirin" $\subset$ "aspirin 500mg"). Dado um trecho da predição maior que dois tokens, se esse trecho está contido em algum trecho do gabarito, ou vice-versa, é considerado um acerto, e (iii) métricas clássicas no contexto de tarefas de classificação e recuperação de informação: *precision*, *recall* e F1-score por entidade PICO. Esse nível é propositalmente "rígido" e tende a subestimar acertos por sinonímia/paráfrase. Ele avalia a capacidade dos modelos de produzir respostas textualmente idênticas às do gabarito.

Avaliação semântica. Já no nível semântico, mensuramos a equivalência conceitual independentemente das palavras exatas. Para isso, representamos textos e entidades por meio de *embeddings* biomédicos e calculamos a similaridade por cosseno utilizando

[2]`https://ai.google.dev/gemini-api/docs/gemini-3?hl=pt-br`

[3]`https://console.groq.com/docs/model/llama-3.3-70b-versatile`

[4]Limiar definido de forma heurística por representar uma interseção de pelo menos metade dos itens.

o modelo SapBERT[5], treinado com base no UMLS (*Unified Medical Language System*) para identificar sinônimos, abreviações e variações lexicais de conceitos médicos [Liu et al. 2021]. Além da comparação direta (gabarito *vs.* predição), adotamos uma estratégia baseada em matriz de similaridade: comparamos todos os pares (predição $\times$ gabarito) e agregamos os melhores pareamentos para estimar *precision* e *recall* semânticos. Esse desenho tenta capturar casos comuns em PICO, como variações lexicais, sinônimos e reformulações, com o objetivo de avaliar se os modelos conseguem extrair as informações essenciais.

***LLM-as-a-judge* – Avaliação auxiliar qualitativa.** Como avaliação auxiliar, utilizamos uma LLM como "juiz" para atribuir escores discretos de alinhamento semântico entre os itens previstos e o gabarito (escala ordinal), permitindo uma análise qualitativa complementar quando métricas automáticas divergem do julgamento humano. Essa avaliação não substitui métricas quantitativas (especialmente por risco de viés do avaliador), mas serve como ferramenta diagnóstica para explorar a capacidade das LLMs de capturar significado global do conteúdo.

Avaliação em cenário aplicado: Ranqueamento de artigos candidatos. Finalmente, para aproximar o cenário de uso na triagem de estudos, avaliamos o impacto dos PICOs extraídos por cada modelo no desempenho de um modelo ranqueador responsável por ordenar artigos candidatos por similaridade entre o PICO-consulta (derivada da meta-análise) e o PICO do artigo candidato (rotulada como relevante ou irrelevante). Os artigos são ranqueados por similaridade de cosseno entre *embeddings* textuais, utilizando o modelo BioSimCSE-BioLinkBERT[6], que combina a arquitetura BioLinkBERT pré-treinada em literatura médica com o método SimCSE de aprendizagem contrastiva para geração de *embeddings* de alta qualidade [Yasunaga et al. 2022, Gao et al. 2021]. O desempenho é avaliado por métricas em *top-k*, como *Precision@k* e *Recall@k*, utilizando o conjunto previamente coletado de meta-análises e seus estudos incluídos como base de validação. O *Precision@k* é calculado avaliando quantos dos artigos nos top-k posições são realmente relevantes. Já o *Recall@k* é calculado avaliando quantos dos artigos relevantes estão nas top-k posições. Essas métricas refletem a utilidade prática do sistema na redução do esforço manual durante a triagem de estudos [Oami et al. 2024], onde uma *Precision@k* alta indica capacidade do modelo de colocar artigos relevantes nas primeiras posições, e *Recall@k* indica a sensibilidade do modelo em encontrar todos os artigos relevantes.

4. Resultados

Avaliamos o desempenho dos modelos em duas tarefas principais: extração de entidades PICO e identificação de relevância de artigos para inclusão em meta-análises. A extração precisa dessas entidades é fundamental para a compreensão do conteúdo dos estudos e, consequentemente, para a identificação consistente de artigos relevantes para um determinado tópico de pesquisa. Dessa forma, avaliar o desempenho de diferentes modelos nessas tarefas torna-se essencial para estimar seu potencial na automatização do processo de condução de meta-análises no contexto da saúde. Os resultados obtidos são apresentados e discutidos nas seções a seguir.

[5]https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext-mean-token

[6]https://huggingface.co/kamalkraj/BioSimCSE-BioLinkBERT-BASE

4.1. Extração de Entidades PICO

Inicialmente, comparamos o desempenho do modelo baseado em BioELECTRA ajustado para reconhecimento de entidades nomeadas (NER), do Llama 3 e do Gemini na tarefa de extração de entidades PICO a partir de um conjunto de resumos biomédicos. Para realizar essa avaliação, foi utilizada a base de dados EBM-NLP (apresentada na Seção 3.2) como *gold standard* ($\approx$ 5.000 resumos). Foram conduzidos três tipos de avaliação: uma sintática, semântica e baseada na técnica *LLM-as-a-judge*. Observa-se que a base de dados utilizada não contém dados referentes a categoria *Control*, portanto a avaliação realizada foi apenas através de dados *Participants*, *Intervention* e *Outcome* (PIO).

Conforme apresentado na Tabela 2, a avaliação sintática, que é a avaliação mais rigorosa por exigir correspondência textual explícita, indicou que o Gemini demonstrou maior precisão na identificação de *Participants* e *Intervention*, com valores 0.938 e 0.775, respectivamente. No entanto, seu *recall* foi significativamente inferior ao dos demais modelos; um comportamento comum das LLMs em comparação ao modelo BioELECTRA, o que sugere uma tendência a resumir o conteúdo e extrair apenas as informações essenciais. Em contraste, o modelo BioELECTRA apresentou os maiores valores de *recall* e *F1-score* na extração de todas as entidades nesse tipo de avaliação.

Tabela 2. Avaliação sintática na extração de entidades PICO.

Entidade	Modelo	Precisão	*Recall*	*F1-score*
Participants	BioELECTRA	0.730 ± 0.27	$\mathbf{0.839} \pm 0.20$	$\mathbf{0.739} \pm 0.20$
Participants	Llama 3	0.886 ± 0.12	0.656 ± 0.29	0.704 ± 0.20
Participants	Gemini	$\mathbf{0.938} \pm 0.18$	0.507 ± 0.21	0.624 ± 0.20
Intervention	BioELECTRA	0.725 ± 0.22	$\mathbf{0.711} \pm 0.29$	$\mathbf{0.676} \pm 0.17$
Intervention	Llama 3	0.699 ± 0.30	0.617 ± 0.26	0.613 ± 0.15
Intervention	Gemini	$\mathbf{0.775} \pm 0.27$	0.445 ± 0.23	0.520 ± 0.17
Outcome	BioELECTRA	$\mathbf{0.857} \pm 0.22$	$\mathbf{0.607} \pm 0.23$	$\mathbf{0.680} \pm 0.18$
Outcome	Llama 3	0.747 ± 0.27	0.467 ± 0.24	0.538 ± 0.16
Outcome	Gemini	0.818 ± 0.20	0.445 ± 0.16	0.551 ± 0.13

Na Tabela 3, a avaliação semântica, que considera captação de significado e utilização de sinônimos, observamos que o Llama 3 obteve maior precisão na entidade *Participants*, com valor de 0.936. Entretanto, o modelo BioELECTRA novamente apresentou os maiores valores de *recall* e *F1-score* nas demais entidades, sugerindo melhor cobertura na extração de entidades PICO.

Tabela 3. Avaliação semântica na extração de entidades PICO.

Entidade	Modelo	Precisão	*Recall*	*F1-score*
Participants	BioELECTRA	0.846 ± 0.135	$\mathbf{0.914} \pm 0.110$	$\mathbf{0.872} \pm 0.107$
Participants	Llama 3	$\mathbf{0.936} \pm 0.100$	0.790 ± 0.134	0.847 ± 0.090
Participants	Gemini	0.869 ± 0.113	0.731 ± 0.152	0.788 ± 0.122
Intervention	BioELECTRA	$\mathbf{0.877} \pm 0.115$	$\mathbf{0.918} \pm 0.109$	$\mathbf{0.893} \pm 0.099$
Intervention	Llama 3	0.766 ± 0.154	0.708 ± 0.159	0.730 ± 0.143
Intervention	Gemini	0.676 ± 0.141	0.616 ± 0.149	0.638 ± 0.129
Outcome	BioELECTRA	$\mathbf{0.846} \pm 0.111$	$\mathbf{0.752} \pm 0.136$	$\mathbf{0.792} \pm 0.114$
Outcome	Llama 3	0.812 ± 0.107	0.707 ± 0.111	0.752 ± 0.099
Outcome	Gemini	0.782 ± 0.089	0.685 ± 0.084	0.727 ± 0.070

Por fim, na Tabela 4, a avaliação com a abordagem L*LM-as-a-judge* empregou o Llama 3 como avaliador para atribuir uma nota de 1 a 5 as entidades PICO extraídos, com base no conteúdo do abstrato correspondente. Observa-se que o Llama 3 alcançou

Tabela 4. Avaliação *LLM-as-a-judge* na extração de entidades PICO.

Modelo	*Participants*	*Intervention*	*Outcome*
BioELECTRA	3.68	**3.48**	**3.50**
Llama 3	**3.84**	3.33	3.44
Gemini	2.63	3.08	3.30

a maior pontuação na entidade de *Participants*, com valor 3.84, enquanto o modelo BioELECTRA apresentou um desempenho superior nas demais entidades. Esse resultado reforça a tendência observada nas outras avaliações, na qual as LLMs apresentam melhor desempenho na identificação dos *Participants* do estudo, enquanto o modelo BioELECTRA se destaca na identificação da *Intervention* utilizada e seu *Outcome*.

4.2. Identificação de Relevância de Artigos

Finalmente, para analisar o desempenho do modelo ranqueador na identificação de artigos relevantes a partir das entidades PICO extraídas pelos modelos avaliados anteriormente, conduzimos uma avaliação utilizando um conjunto de validação construído e apresentado na Seção 3.2. O conjunto reúne 55 meta-análises, cujos estudos incluídos foram rotulados como positivos. Com base nas entidades PICO extraídas por cada modelo, realizou-se uma avaliação comparativa para identificar a abordagem mais eficaz na detecção de artigos relevantes.

Conforme apresentado na Tabela 5, as entidades PICO extraídos pelas LLMs demonstraram desempenho significativamente superior em comparação com os extraídos pelo modelo BioELECTRA estabelecido como *baseline* no contexto deste estudo. Esse resultado sugere que, embora o BioELECTRA apresente maior acurácia na extração, os PICOs extraídos pelas LLMs capturam melhor o significado dos estudos, resultando em triagem mais eficaz.

Tabela 5. Avaliação da identificação de artigos relevantes.

Modelo	*Precisão@1*	*Precisão@5*	*Recall@10*
BioELECTRA	0.84	0.75	0.89
Llama 3	**1.00**	0.68	**0.93**
Gemini	**1.00**	**0.97**	0.80

5. Conclusão e Trabalhos Futuros

Neste trabalho, avaliamos o desempenho de abordagens de aprendizagem de máquina e PLN distintas para automatização de etapas no processo de condução de meta-análises no contexto da saúde. Na etapa de extração de entidades PICO, os resultados mostraram que o modelo BioELECTRA ajustado para NER apresentou melhor desempenho na avaliação semântica das entidades de *Intervention* e *Outcome*, especialmente em termos de *recall*. O modelo superou o Llama 3 e o Gemini em 0.21 e 0.30, respectivamente, na identificação de intervenção, e em 0.04 e 0.06, respectivamente, na identificação de despecho. Por outro lado, os modelos LLMs, obtiveram melhor desempenho na identificação da entidade de *Participants*. Na avaliação semântica, o Llama 3 obteve maior precisão, superando o Gemini e BioELECTRA por 0.06 e 0.09, respectivamente. Porém, novamente as LLMs apresentaram *recall* menor que o BioELECTRA, reforçando a tendência desses modelos em não identificar todas as palavras pertinentes à entidade. Conjecturamos que parte da discrepância de valores entre o modelo BioELECTRA e as LLMs pode estar relacionada ao formato do gabarito, baseado em *span extraction*. Enquanto o modelo BioELECTRA

produz saídas alinhadas a esse padrão, as LLMs tendem a reformular os trechos e a reduzir expressões específicas a termos essenciais (por exemplo, “aspirin 500mg” para “aspirin”).

Por outro lado, os modelos LLMs, obtiveram melhor desempenho na identificação da entidade de *Participants*. Na avaliação semântica, o Llama 3 obteve maior precisão, superando o Gemini e BioELECTRA por 0.06 e 0.09, respectivamente. Porém, novamente as LLMs apresentaram *recall* menor que o BioELECTRA, reforçando a tendência desses modelos em não identificar todas as palavras pertinentes à entidade. Conjecturamos que parte da discrepância de valores entre o modelo BioELECTRA e as LLMs pode estar relacionada ao formato do gabarito, baseado em *span extraction*. Enquanto o modelo BioELECTRA produz saídas alinhadas a esse padrão, as LLMs tendem a reformular trechos e extrair apenas termos essenciais, resumindo expressões mais específicas (por exemplo, “aspirin 500mg” para “aspirin”). Uma limitação da abordagem *LLM-as-a-Judge* é o uso exclusivo do Llama 3 como avaliador da qualidade dos PICO extraídos, o que pode introduzir viés, especialmente nas entidades geradas pelo próprio modelo.

Em contraste, na tarefa de identificação de artigos relevantes, as entidades PICO extraídas pelas LLMs resultaram em uma triagem com maior precisão. Na métrica *Precisão@1*, o modelo ranqueador colocou um artigo relevante na primeira posição 100% das vezes quando usou as entidades PICO extraídas pelas LLMs, em comparação com um valor de 0.84 para as extraídas pela BioELECTRA. Na métrica *Precisão@5*, o Gemini se destaca com um valor de 0.97, superando o BioELECTRA e Llama 3 por 0.22 e 0.29, respectivamente. Essa maior precisão indica a capacidade das LLMs em gerar representações dos artigos que um modelo ranqueador consegue corretamente atribuir pontuações altas de relevância. Na métrica de *Recall@10*, o Llama 3 apresentou desempenho melhor com um valor de 0.93, superando o BioELECTRA e Gemini por 0.04 e 0.13, respectivamente. Considerando as diferentes métricas ponderadas, o desempenho das LLMs sugere que essa representação mais concisa das entidades PICO contribui para a redução de ruído e melhor captação do significado contextual dos estudos.

Em suma, os resultados indicam que abordagens híbridas, combinando modelos especializados em NER com LLMs, constituem uma estratégia promissora para automatizar etapas do processo de meta-análises em saúde. Como trabalhos futuros, pretende-se ampliar a avaliação para outros modelos e realizar avaliações qualitativas e/ou manuais, a fim de verificar o uso de abordagens baseadas em aprendizado de máquina como suporte ao especialista na área da saúde.

Considerações Éticas e Reprodutibilidade. O uso de IA generativa foi restrito à correção ortográfica, sintática e semântica do texto. Para fins de reprodutibilidade, os códigos estão disponíveis em: `https://github.com/maridverd/PICO-Extraction-and-Ranking-Evaluation`.

Agradecimentos. Este trabalho foi parcialmente financiado pela CAPES, FAPEMIG, UFV (PIBIC-FUNARBIC 2025-2026) e INCT TILD-IAR. Por fim, os autores agradecem à BIP Brasil (antiga Bitka) pelo apoio ao longo das diferentes etapas deste projeto, em especial aos colaboradores Erick Figueiredo, Júlio Resende, Leonardo Porto, Rubens Moraes, Anilton Cardoso Junior e Francisco Fonseca.

Referências

Cao, C., Sang, J., Arora, R., Chen, D., Kloosterman, R., Cecere, M., Gorla, J., Saleh, R., Drennan, I., Teja, B., Fehlings, M., Ronksley, P., Leung, A. A., Weisz, D. E., Ware, H., Whelan, M.,

Emerson, D. B., Arora, R. K., and Bobrovitz, N. (2025). Development of prompt templates for large language model-driven screening in systematic reviews. *Annals of Internal Medicine*, 178(3):389–401.

Egger, M., Smith, G. D., and Altman, D. G., editors (2008). *Systematic Reviews in Health Care: Meta-Analysis in Context*. BMJ Publishing Group, 2 edition.

Gao, T., Yao, X., and Chen, D. (2021). Simcse: Simple contrastive learning of sentence embeddings. In *Proc. of EMNLP*, pages 6894–6910.

Geng, R. et al. (2025). SLOT: Structuring the output of large language models. *arXiv preprint arXiv:2505.04016*.

Górska, A. and Tacconelli, E. (2024). Towards autonomous living meta-analyses: A framework for automation of systematic review and meta-analyses. *Studies in Health Technology and Informatics*, 316:378–382.

Guo, E., Gupta, M., Deng, J., Park, Y.-J., Paget, M., and Naugler, C. (2024). Automated paper screening for clinical reviews using large language models: Data analysis study. *Journal of Medical Internet Research*, 26:e48996.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., and Welch, V. A., editors (2024). *Cochrane Handbook for Systematic Reviews of Interventions*. Version 6.5.

Hoffmann, F., Allers, K., Rombey, T., et al. (2021). Nearly 80 systematic reviews were published each day: Observational study on trends in epidemiology and reporting over the years 2000–2019. *Journal of Clinical Epidemiology*, 138:1–11.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Jakab, M. (2024). How many authors are (too) many? a retrospective, descriptive analysis of authorship in biomedical publications. *Scientometrics*.

Kanakarajan, K. r., Kundumani, B., and Sankarasubbu, M. (2021). BioELECTRA: Pretrained biomedical text encoder using discriminators. In *Proc. of the BioNLP*, pages 143–154.

Li, L., Mathrani, A., and Susnjak, T. (2025). Transforming evidence synthesis: A systematic review of the evolution of automated meta-analysis in the age of ai. arXiv.

Liu, F., Vashishth, S., Uzuner, O., et al. (2021). Self-alignment pretraining for biomedical entity representations. In *Proc. of NAACL*, pages 4228–4238.

Marshall, I. J. and Wallace, B. C. (2019). Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8(1):163.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., and Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. *Systematic Reviews*, 4(1):1.

Nye, B., Li, J. J., Patel, R., Yang, Y., Marshall, I., Nenkova, A., and Wallace, B. (2018). A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proc. of the Annual Meeting of the ACL*, pages 197–207.

Oami, T., Okada, Y., and Nakada, T.-A. (2024). Performance of a large language model in screening citations. *JAMA Network Open*, 7(7):e2420496.

Ofori-Boateng, R., Aceves-Martins, M., Wiratunga, N., and Moreno-Garcia, C. F. (2024). Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review. *Artificial Intelligence Review*, 57:200.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., and Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews*, 4(1):5.

Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1):210.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and Moher, D. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71.

Richardson, W. S., Wilson, M. C., Nishikawa, J., and Hayward, R. S. (1995). The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123(3):A12–A13.

Santos, C. M. d. C., Pimenta, C. A. d. M., and Nobre, M. R. C. (2007). The pico strategy for the research question construction and evidence search. *Revista Latino-Americana de Enfermagem*, 15(3):508–511.

Tsafnat, G., Glasziou, P., Karystianis, G., and Coiera, E. (2018). Automated screening of research studies for systematic reviews using study characteristics. *Systematic Reviews*, 7(1):64.

Veritas Health Innovation (2024). Covidence systematic review software. `https://www.covidence.org/`. Accessed: 2026-04-27.

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., and Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. *Proc. of the ACM SIGHIT*, pages 819–824.

Wallace, B. C., Small, K., Brodley, C. E., and Trikalinos, T. A. (2010a). Active learning for biomedical citation screening. In *Proc. of the ACM SIGKDD*, pages 173–181.

Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., and Schmid, C. H. (2010b). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics*, 11:55.

Wang, Z., Cao, L., Danek, B., Jin, Q., Lu, Z., and Sun, J. (2025). Accelerating clinical evidence synthesis with large language models. *npj Digital Medicine*, 8(1):509.

Xia, C., Xing, C., Du, J., Yang, X., Feng, Y., Xu, R., Yin, W., and Xiong, C. (2024). FOFO: A benchmark to evaluate LLMs' format-following capability. *arXiv preprint arXiv:2402.18667*.

Yasunaga, M., Leskovec, J., and Liang, P. (2022). Linkbert: Pretraining language models with document links. In *Proc. of the Annual Meeting of the ACL*, pages 8003–8016.